

# Capacity Sharing and Cost Allocation among Independent Firms in the Presence of Congestion

Yimin Yu • Saif Benjaafar  
Industrial and Systems Engineering  
University of Minnesota Minneapolis, MN 55455  
yimin@me.umn.edu – saif@umn.edu

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

April 4, 2009

## 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. If firms decide to operate a shared facility they must also decide on a scheme for sharing the costs. We formulate the problem as a cooperative game and identify a cost allocation 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. In settings where unit delay costs are private information, the cooperative capacity sharing game becomes embedded with a non-cooperative information reporting game. We show how a cost allocation rule can be designed to induce all firms to report truthfully this information. Moreover, we show that, under this allocation rule, truth telling is a dominant strategy, with each firm reporting truthfully its private information regardless of the reporting decisions of other firms.

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

# 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 owns all the capacity in the system, has full information, 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, 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. They may also possess private information regarding their costs or service level requirements which they may not report truthfully. 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 *fair* manner that 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 the presence of private information, the cost allocation scheme should also mitigate the possibility of firms not truthfully disclosing their private information.

Capacity sharing among independent firms is increasingly common in the manufacturing, service, and public sectors. In manufacturing, there are numerous instances of independent firms sharing the same production facilities. For example, several car manufacturers, such as Toyota and General Motors, share final assembly plants. Similar sharing arrangements can be found in the electronics industry where firms share printed circuit board assembly facilities or semiconductor

manufacturing plants. In the service sector, the sharing of service facilities is also common. For example, airlines share check-in counters, maintenance facilities, and reservation systems; hospitals share expensive diagnostic equipment, laboratory facilities, and in some cases medical specialists and surgeons. In the public sector, the sharing of resources between independent government entities is also widespread. For instance, local governments in rural communities share fire and police departments, 911 call centers, and other social services. In various sectors, independent organizations are increasingly sharing infrastructure resources such as telecommunication networks, computer services, basic manufacturing resources, and facilities for handling back-office operations. In this paper, we are motivated by such settings. That is, we are motivated by settings where firms have the option of sharing generic infrastructure resources whose capacity can be easily scaled and which can be accessed with equal efficiency by all firms.

Capacity sharing among independent firms<sup>1</sup> raises several important questions. For example, is capacity sharing always beneficial to all firms? Does it always lead to a reduction in total capacity in the system? How should capacity costs be allocated among the different firms? Is capacity sharing among all the firms the best arrangement or would sharing among smaller subsets of the firms be more beneficial to particular firms? Can capacity sharing be beneficial when firms do not report truthfully private information, especially when this information is used in determining capacity levels and cost allocation? Is it possible to induce firms, via cost allocation, to disclose truthfully their private information? If so, would such cost allocation ensure that all firms continue to benefit from capacity sharing?

In this paper, we address these and other related questions for a specific setting. 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 time. 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

---

<sup>1</sup>We use the term *independent firms* broadly to include economic entities who are independently owned and also entities, such as sub-divisions within a single firm, who may have a common owner. What is important to our analysis is that these entities are empowered to make independent decisions that minimize their individual costs.

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 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 reported delay costs and service levels of all the firms), we must also determine how the corresponding cost must be allocated.

The main contributions of our paper are summarized below.

- We provide a framework for modeling capacity sharing in queueing systems with independent firms. We consider systems with and without full information. In systems with full information, parameters of all the firms are common knowledge; in systems with incomplete information, unit delay costs are private information to each firm. To our knowledge, our paper is among the first to model the issue of cooperation and capacity sharing in a queueing context and to do so for systems with and without complete information.
- We formulate capacity sharing as a cooperative game and show that for systems where facilities are modeled as M/M/1 queues (queues with Poisson arrivals and exponential service times) the *core* of the game is non-empty. That is, there always exists a cost allocation rule for which all the firms are better off than under any other alternative sharing arrangement, including being on their own.
- In systems with full information, 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.
- In systems with incomplete information, firms may act strategically and misreport their unit delay costs. This leads to a non-cooperative "information reporting" game embedded in the cooperative "capacity sharing" game. We show that allocation rules that are in the core under full information could lead to significant misreporting of private information in the presence of incomplete information.

- Although firms can act strategically when they possess private information, we show how a cost allocation rule can be designed to induce all firms to truthfully report their private information and to do so regardless of the reporting decisions of other firms. That is, our proposed allocation rule is *incentive-compatible* with truth-telling being a *dominant strategy*. Moreover, we show that our proposed allocation rule is in the core.
- We extend our treatment beyond the M/M/1 queue framework and discuss the extent to which our results continue to hold in more general settings.

Our choice of modeling framework is in part motivated by the fact that types of service and manufacturing facilities can be viewed as queueing systems. There is a rich literature that takes this modeling view (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 services, manufacturing, or elsewhere. We also view it as a contribution, in the form of a potentially rich application domain, to the literature on cooperative games with and without full information.

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 capacity sharing when there is full information. In Section 5, we consider the case of capacity sharing with incomplete information. We extend the analysis to systems with more general arrival and service processes in Section 6. In Section 7, we discuss additional extensions and 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  $\mu$  and demand rate  $\lambda$ , 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. (2004) use approximations to evaluate the benefit of partitioning servers in multiple pools instead of a single large one. Sheikhzadeh et al. (1998), Gurusurthi 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 and Karaesmen (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 four 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, Fourth, we allow for the possibility of private information regarding delay costs or service levels and for the possibility of firms not reporting this information truthfully.

The literature dealing with capacity sharing in the context of independent firms is limited. Gonzalez and Herrero (2004), and also Garcia-Sanz et al. (2007), 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), do not consider delay costs, and assume truthful reporting of all information. 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. We also consider systems where firms might have private information. Anily and Haviv (2008) 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. However, in their case, capacity is exogenously determined so that capacity cost is not included. They also assume full information regarding all parameters.

Dewan and Mendelson (1990) consider the problem faced by the manager of a service facility with multiple users. The manager decides on the capacity of the service facility, which is modeled as a single server queue, and on the price to charge each user. The prices affect the demand rates of the users with higher prices resulting in lower demand rates. This is different from our setting where the demand rates are exogenous. Also, in their case, customers do not have the option of operating independent facilities or forming coalitions.

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 (2006), among others), and news-vendor models (see Muller et al. (2002), Nagarajan and Sošić (2007), Kemahlioglu-Ziya (2004), Chen and Zhang (2007), and Hanany and Gerchak (2008) and the references therein).

Finally, we should note that one could view the decision to invest in a shared facility (instead of dedicated facilities) as a decision by the corresponding firms to outsource. There is a rich literature on outsourcing and procurement, 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, this literature does not deal with cost allocation or coalition formation.

### 3 Systems without Capacity Sharing

Consider a system consisting of a set  $\mathcal{N} = \{1, \dots, n\}$  of  $n$  firms. Firm  $i$ ,  $i \in \mathcal{N}$ , faces an independent demand stream with customers arriving according to a Poisson process with rate  $\lambda_i$  (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  $Y/\mu_i$  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  $\mu_i$ , ( $\mu_i > 0$ ) 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 work content is homogeneous across firms. Given the exponential nature of both customer inter-arrival times and service times, each firm behaves like an M/M/1 queue. There is 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 (e.g., the speed of computing facilities, the bandwidth of communication networks, or the throughput of production lines). It is also 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). It is also consistent with the significant literature on capacity planning, as noted recently by Bassamboo 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 Bassamboo 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), Bassamboo et al. (2008), and Allon and Federgruen (2007, 2008), among many others.<sup>2</sup>

The objective of each firm is to minimize its capacity investment while limiting the amount of delay 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 amount. 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.

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 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). Delay costs may also reflect the cost of work-in-process accumulation when there is a physical product released to the queue with the arrival of each customer, as in many manufacturing applications. The use

---

<sup>2</sup>It is possible to consider systems where the demand rates themselves are random (e.g., when the demand is modulated by another process). However, depending on the assumptions made regarding this modulating process, the analysis could become significantly less tractable and we leave this as a potential area for future research.

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 limit our analysis to the case where service level is expressed in terms of a probability that total delay in the system (time in the queue + time in service) for each customer in steady state does not exceed a specified threshold. We also limit ourselves to 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.

Let  $z_i(\mu_i)$  denote the expected total cost incurred by firm  $i$  given a service rate  $\mu_i$  (for stability, we assume that  $\lambda_i/\mu_i < 1$ ). Let  $W_i$ , a random variable, denote the total time a customer of firm  $i$  spends in the system (customer delay) and  $P(W_i \leq w_0)$  the probability that customer delay does not exceed  $w_0$  where  $w_0 \geq 0$ . The problem faced by firm  $i$  can then be stated as follows

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

subject to

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

and

$$\lambda_i/\mu_i \leq 1. \quad (3)$$

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  $\mu_i$ . 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  $\mu_i^*$  can be obtained from the first order condition of optimality, since  $z_i$  is convex in  $\mu_i$ , as

$$\mu_i^* = \lambda_i + \sqrt{\frac{h_i \lambda_i}{c}}. \quad (4)$$

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

$$\mu_i^* = \lambda_i + \frac{\ln(\frac{1}{1-\alpha_i})}{w_0}. \quad (5)$$

In both cases, the optimal capacity is the sum of two components. The first corresponds to the demand rate,  $\lambda_i$  (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  $\alpha_i$ . The expressions in equations (4) and (5) are not new. Similar expressions have been derived elsewhere; see for example Kleinrock (1976), Allon and Federegrien (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

$$\mu_i^* = \lambda_i + \eta_i, \quad (6)$$

where

$$\eta_i = \max\left\{\frac{\ln(\frac{1}{1-\alpha_i})}{w_0}, \sqrt{\frac{h_i \lambda_i}{c}}\right\}. \quad (7)$$

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

$$z_i^* = c(\lambda_i + \eta_i) + \frac{h_i \lambda_i}{\eta_i}. \quad (8)$$

This leads to a total system cost of  $z_{1,\dots,n}^* = \sum_{i \in \mathcal{N}} z_i^*$ . 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

$$z_i^* = c\lambda_i + 2\sqrt{h_i \lambda_i c}. \quad (9)$$

This leads to a total system cost,  $z_{1,\dots,n}^*$ , given by

$$z_{1,\dots,n}^* = c \sum_{i \in \mathcal{N}} \lambda_i + 2 \sum_{i \in \mathcal{N}} \sqrt{h_i \lambda_i c}. \quad (10)$$

In the case of identical firms, with  $\lambda_i = \lambda$  and  $h_i = h$  for all  $i \in \mathcal{N}$ , the optimal total cost in (10) reduces to

$$z_{1,\dots,n}^* = cn\lambda + 2n\sqrt{h\lambda c}, \quad (11)$$

and the total capacity in the system to

$$\sum_{i \in \mathcal{N}} \mu_i^* = n \left( \lambda + \sqrt{\frac{h\lambda}{c}} \right). \quad (12)$$

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  $\sqrt{\frac{h_i\lambda_i}{c}} \leq \frac{\ln(\frac{1}{1-\alpha_i})}{w_0}$ . 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.

## 4 Capacity Sharing with Full Information

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. In Section 5, we consider the case where firms act strategically and may misreport some of this information. There may of course be firms who are unwilling to share any information, truthfully or not. These firms will not be allowed to participate in the coalition. 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 pooled 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 subsidiaries 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

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 firm 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:

$$\text{Minimize } z_{\mathcal{N}}(\mu_{\mathcal{N}}) = c\mu_{\mathcal{N}} + \frac{\sum_{i \in \mathcal{N}} h_i \lambda_i}{\mu_{\mathcal{N}} - \sum_{i \in \mathcal{N}} \lambda_i} \quad (13)$$

subject to

$$P(W_{\mathcal{N}} \leq w_0) = 1 - e^{(\mu_{\mathcal{N}} - \sum_{i \in \mathcal{N}} \lambda_i)w_0} \geq \alpha_{\mathcal{N}}, \quad (14)$$

and

$$\sum_{i \in \mathcal{N}} \lambda_i / \mu_{\mathcal{N}} \leq 1, \quad (15)$$

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

$$\mu_{\mathcal{N}}^* = \sum_{i \in \mathcal{N}} \lambda_i + \eta_{\mathcal{N}}, \quad (16)$$

where

$$\eta_{\mathcal{N}} = \max\left\{\frac{\ln\left(\frac{1}{1-\alpha_{\mathcal{N}}}\right)}{w_0}, \sqrt{\frac{\sum_{i \in \mathcal{N}} h_i \lambda_i}{c}}\right\}. \quad (17)$$

Similar to the distributed system, 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 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**  $z_{\mathcal{N}}^* \leq z_{1, \dots, n}^*$  and  $\mu_{\mathcal{N}}^* \leq \sum_{i=1}^n \mu_i^*$ , where  $z_{\mathcal{N}}^*$  is the optimal cost in the shared facility.

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

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

where  $i_{max} \in \{i : \alpha_i = \max(\alpha_1, \dots, \alpha_n)\}$ . In order to prove that  $z_{\mathcal{N}}^* \leq z_{1, \dots, n}^*$ , 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

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

(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

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

■

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  $\sqrt{\frac{h\lambda}{c}} \geq \frac{\ln(\frac{1}{1-\alpha})}{w_0}$ . 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  $z_{\mathcal{J}}^* \equiv z_{\mathcal{J}}(\mu_{\mathcal{J}}^*)$ . A vector  $\phi = (\phi_1, \dots, \phi_n)$  is called an allocation rule if  $\phi_i$  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:

$$\sum_{i \in \mathcal{J}} \phi_i \leq z_{\mathcal{J}}^*, \quad \forall \mathcal{J} \subseteq \mathcal{N}, \quad (18)$$

and

$$\sum_{i \in \mathcal{N}} \phi_i = z_{\mathcal{N}}^*. \quad (19)$$

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:

$$\phi_i = \frac{h_i \lambda_i}{\eta_{\mathcal{N}}} + c \lambda_i + \gamma_i, \quad (20)$$

where

$$\gamma_i = \frac{h_i \lambda_i}{\sum_{i \in \mathcal{N}} h_i \lambda_i} c \eta_{\mathcal{N}} \quad \text{if} \quad \frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0} \leq \sqrt{\frac{\sum_{i \in \mathcal{N}} h_i \lambda_i}{c}} \quad (21)$$

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

$$\gamma_i = \begin{cases} c \eta_{\mathcal{N}} - c \sqrt{\frac{\sum_{i \in \mathcal{N}, i \neq i_{max}} h_i \lambda_i}{c}} & \text{if } i = i_{max}, \text{ and} \\ c \frac{h_i \lambda_i}{\sum_{i \in \mathcal{N}, i \neq i_{max}} h_i \lambda_i} \sqrt{\frac{\sum_{i \in \mathcal{N}, i \neq i_{max}} h_i \lambda_i}{c}} & \text{if } i \neq i_{max}, \end{cases} \quad (22)$$

with again  $i_{max} \in \{i : \alpha_i = \max(\alpha_1, \dots, \alpha_n)\}$ . 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  $\phi = (\phi_1, \dots, \phi_n)$  as specified in (20) – (22) 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.*

**Proof.** 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  $z_{\mathcal{J}}^* = \frac{\sum_{i \in \mathcal{J}} h_i \lambda_i}{\frac{\ln(\frac{1}{1-\alpha\mathcal{N}})}{w_0}} + c \sum_{i \in \mathcal{J}} \lambda_i + c \frac{\ln(\frac{1}{1-\alpha\mathcal{N}})}{w_0} \geq \sum_{i \in \mathcal{J}} \phi_i, \forall \mathcal{J} : i_{max} \in \mathcal{J}$ . Consequently, the allocation is in the core.

■

## 5 Capacity Sharing with Incomplete Information

In this section, we consider the case where unit delay costs are private information to each firm. Hence, firms could act strategically and misreport this information if doing so is individually beneficial. In other words, each firm  $i$  makes a decision about what unit delay cost  $\hat{h}_i$  to report, where  $\hat{h}_i$  can be different from the true value  $h_i$ . A firm makes this decision knowing that the reported information will be used to determine the corresponding optimal capacity level. As in Section 4.2, each firm incurs in the long run two costs: (1) a private expected delay cost and (2) a fraction of the total capacity cost, where the latter is determined by the cost allocation rule. For tractability, we restrict our treatment to the case of pure delay costs (i.e., no service level constraints). Treating

systems where both service levels and delay costs appears difficult in the presence of incomplete information. Moreover, it is arguably more important to focus on the case where unit delay costs, and not service level requirements, are private information. The misreporting (at least under-reporting) of service level requirements is less plausible since the only guarantee a firm has that its service level would be fulfilled is to truthfully report it.

We assume the following sequence of events. First, the cost allocation rule, which may depend on the reported information, is announced. Second, the firms report unit delay costs taking into account the announced allocation rule. Third, based on the reported information, the optimal capacity is selected. Finally, firms incur cost based on the realized delay and their share of the capacity cost. Hence, for a given cost allocation rule, the problem faced by the firms can be viewed as a noncooperative information reporting game where the strategy set for each firm consists of the reported values of unit delay costs and service levels. The objective of each firm is to report values that minimize its total expected cost given the reported values by the other firms.<sup>3</sup>

The presence of private information raises several important questions. Would the cost allocation rule described in Section 4.2 lead to misreporting of private information? If so, is it possible to design an alternative cost allocation rule that does lead to truth telling? Would such a cost allocation be in the core and would it preserve desirable fairness properties? In this section, we provide answers to these and other related questions.

We assume that, given reported unit delay costs  $(\hat{h}_1, \dots, \hat{h}_n)$ , total capacity is determined so as to minimize the sum of expected delay costs (based on the reported information) and investment capacity cost. The corresponding capacity selection problem can be stated as follows:

$$\text{Minimize } \hat{z}_{\mathcal{N}}(\mu_{\mathcal{N}}) = c\mu_{\mathcal{N}} + \frac{\sum_{i \in \mathcal{N}} \hat{h}_i \lambda_i}{\mu_{\mathcal{N}} - \sum_{i \in \mathcal{N}} \lambda_i}. \quad (23)$$

The optimal capacity, which we denote by  $\mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n)$ , is then given by

$$\mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n) = \sum_{k=1}^n \lambda_k + \sqrt{\frac{\sum_{k=1}^n \hat{h}_k \lambda_k}{c}}, \quad (24)$$

---

<sup>3</sup>To formulate a game with incomplete information, as for example in a Nash-Bayes game, one would typically require additional assumptions regarding what each firm might know about the private information of other firms. However, as we shall see in Theorem 5.1, we do not need to specify such assumptions, as the described cost allocation rule leads to truth-telling being a dominant strategy.

and the resulting expected delay cost experienced by firm  $i$  is given by  $\frac{h_i \lambda_i}{\mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n) - \sum_{k=1}^n \lambda_k}$ .

Let us first consider the cost allocation rule discussed in Section 4.2 where firm  $i$ , for all  $i \in \mathcal{N}$ , incurs privately its delay cost, the cost of capacity for which it is directly responsible, and a fraction of buffer capacity cost that is proportional to its demand-weighted unit delay cost. Consequently, given reported unit delay costs  $\hat{h}_1, \dots, \hat{h}_n$ , the long run expected cost incurred by firm  $i$  is

$$\phi_i(\hat{h}_i, \hat{h}_{-i}) = \frac{h_i \lambda_i}{\sqrt{\frac{\sum_{k=1}^n \hat{h}_k \lambda_k}{c}}} + c \lambda_i + \frac{\hat{h}_i \lambda_i}{\sum_{k=1}^n \hat{h}_k \lambda_k} \sqrt{\frac{\sum_{k=1}^n \hat{h}_k \lambda_k}{c}}, \quad (25)$$

where  $\hat{h}_{-i} = (\hat{h}_1, \dots, \hat{h}_{i-1}, \hat{h}_{i+1}, \dots, \hat{h}_n)$  denotes the set of unit delay costs of firms other than firm  $i$ .

Examining the above cost function, we can see that the expected cost of firm  $i$  is affected by both its true unit delay cost as well as the one it reports. It is also affected by the unit delay cost reported by other firms. For example, by under-reporting (reporting a lower unit delay cost than its true one), a firm could benefit by incurring a smaller fraction of buffer capacity cost. However, it could also incur a higher delay cost because less capacity could be installed. The extent to which a firm benefits from misreporting depends on the reporting decisions of other firms.

Given that firms  $j \neq i$  report unit delay costs  $\hat{h}_{-i}$ , firm  $i$  would choose to report unit delay cost  $\hat{h}_i^*(\hat{h}_{-i})$  that minimizes its total expected cost. Noting that the expected cost function  $\phi_i(\hat{h}_i, \hat{h}_{-i})$  is convex in  $\hat{h}_i$ , the optimal reported costs  $\hat{h}_i^*(\hat{h}_{-i})$  can be obtained from the first order condition of optimality as

$$\hat{h}_i^*(\hat{h}_{-i}) = \max\left\{0, h_i - \frac{2 \sum_{j \neq i} \hat{h}_j \lambda_j}{\lambda_i}\right\}. \quad (26)$$

As we can see, firm  $i$  would always under-report its true delay cost regardless of the reporting decision of other forms. This is the case, which is perhaps surprising, even if other firms are truthful in their reporting. The under-reporting appears due to the proportionality in how buffer capacity cost is allocated among the firms, making it more advantageous for firms to always under-report and reduce their share of buffer capacity than over-report (or report truthfully) and reduce their delay cost. The under-reporting can be significant, leading firm  $i$  in some cases (when its unit delay cost is sufficiently small) to even report a unit delay cost of zero. Thus, the cost allocation rule of Section 4.2 is not incentive compatible. In the remainder of this section, we turn our attention to constructing an allocation rule that is.

Consider the cost allocation rule under which the long run expected cost for firm  $i$  is given by the following

$$\begin{aligned} \phi_i(\hat{h}_1, \dots, \hat{h}_n) = & \frac{h_i \lambda_i}{\mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n) - \sum_{k \in \mathcal{N}} \lambda_k} + \sum_{j \neq i} \frac{\hat{h}_j \lambda_j}{\mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n) - \sum_{k \in \mathcal{N}} \lambda_k} \\ & + c \mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n) - p_i(\hat{h}_{-i}), \end{aligned} \quad (27)$$

where  $p_i(\hat{h}_{-i})$  is a positive function that depends only on the reported unit delay costs of firms  $j \neq i$ . The first term on the right-hand side of (27) corresponds to the private expected delay cost of firm  $i$  while the remaining terms constitute its share of capacity cost. Then, we can show that the following result holds.

**Theorem 5.1** *Under the allocation rule defined in (27),*

$$\phi_i(h_i, \hat{h}_{-i}) \leq \phi_i(\hat{h}_i, \hat{h}_{-i}), \forall \hat{h}_{-i}.$$

*That is, the allocation rule is incentive compatible with truth telling being a dominant strategy for each firm.*

**Proof.** First note that the choice  $p_i(\hat{h}_{-i})$  does not affect firm  $i$ 's choice of  $\hat{h}_i$  since it does not depend on  $\hat{h}_i$ . Given other firms' reported information,  $\hat{h}_{-i}$ , firm  $i$  reports a unit delay cost that minimizes its allocated cost  $\phi_i(\hat{h}_i, \hat{h}_{-i})$ , which can be rewritten as

$$\phi_i(\hat{h}_i, \hat{h}_{-i}) = \frac{h_i \lambda_i + \sum_{j \neq i} \hat{h}_j \lambda_j}{\mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n) - \sum_{k=1}^n \lambda_k} + c \mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n) - p_i(\hat{h}_{-i}) \quad (28)$$

Hence, firm  $i$  would like capacity to be set as  $\sqrt{\frac{h_i \lambda_i + \sum_{j \neq i} \hat{h}_j \lambda_j}{c}} + \sum_{i=1}^n \lambda_i$ , as the above capacity is the unique minimizer of its expected total cost. However, this is possible only if firm  $i$  reports its true delay cost  $h_i$ , regardless of whether other firms report their information truthfully or not. In other words, truth reporting is a dominant-strategy. It is also the unique strategy that minimizes the expected total cost of each firm, by virtue of the fact that the capacity optimization problem in (23) admits the unique minimizer given in (24).  $\blacksquare$

Although the above cost allocation rule is incentive-compatible, it must also be efficient so that

the sum of the allocated costs equals the total actual cost incurred by the system. This means that we must have  $\sum_{i \in \mathcal{N}} \phi_i(\hat{h}_i, \hat{h}_{-i}) = z_{\mathcal{N}}^*(\mu_{\mathcal{N}}(\hat{h}_1, \dots, \hat{h}_n))$ , where

$$z_{\mathcal{N}}^*(\mu_{\mathcal{N}}(\hat{h}_1, \dots, \hat{h}_n)) = \frac{\sum_{k=1}^n \hat{h}_k \lambda_k}{\mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n) - \sum_{k=1}^n \lambda_k} + c \mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n), \quad (29)$$

which corresponds to the actual total cost in the system. Noting that

$$\sum_{i \in \mathcal{N}} \phi_i(\hat{h}_i, \hat{h}_{-i}) = z_{\mathcal{N}}^*(\mu_{\mathcal{N}}(\hat{h}_1, \dots, \hat{h}_n)) + (n-1) \hat{z}(\mu_{\mathcal{N}}(\hat{h}_1, \dots, \hat{h}_n)) - \sum_{i \in \mathcal{N}} p_i(\hat{h}_{-i}),$$

where

$$\hat{z}(\mu_{\mathcal{N}}(\hat{h}_1, \dots, \hat{h}_n)) = \frac{\sum_{k=1}^n \hat{h}_k \lambda_k}{\mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_{-i}) - \sum_{k=1}^n \lambda_k} + c \mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n),$$

we can see that the functions  $p_i(\hat{h}_{-i})$  must satisfy the equality

$$\sum_{i \in \mathcal{N}} p_i(\hat{h}_{-i}) = (n-1) \hat{z}^*(\mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n)). \quad (30)$$

In what follows, we describe how the functions  $p_i$  can be constructed to satisfy the above condition. First, note that  $\hat{z}_{\mathcal{N}}^*(\mu_{\mathcal{N}}(\hat{h}_1, \dots, \hat{h}_n)) = 2\sqrt{c \sum_{k=1}^n \hat{h}_k \lambda_k} + c \sum_{k=1}^n \lambda_k$ . Next, suppose that  $h_i \geq a$  and  $\lambda_i \geq b$  for some  $i$  and some positive  $a$  and  $b$  (we rule out the trivial cases where  $h_i = 0$  and  $\lambda_i = 0$  for all  $i \in \mathcal{N}$ ). This would then ensure the existence of a Taylor expansion for  $\sqrt{c \sum_{k=1}^n \hat{h}_k \lambda_k}$ . In particular, we have

$$\sqrt{c \sum_{k=1}^n \hat{h}_k \lambda_k} = \sqrt{cab} + \sum_{m=1}^{\infty} \frac{f^{(m)}(cab)}{m!} (c \sum_{k=1}^n \hat{h}_k \lambda_k - cab)^m,$$

where  $f(x) = \sqrt{x}$  and  $f^{(m)}(x)$  is the  $m$ -th derivative of  $f$  evaluated at  $x$ . Define now the function  $p_i(\hat{h}_{-i})$  as

$$\begin{aligned} p_i(\hat{h}_{-i}) &= 2 \sum_{m=1}^{\infty} \frac{f^{(m)}(cab)}{m!} (c^m \sum_{j, k \neq i} \sum_{l=0}^m \binom{m}{l} (\hat{h}_k \lambda_k)^l (\hat{h}_j \lambda_j)^{m-l} \\ &+ c^m \sum_{k \neq i} \sum_{l=1}^{m-1} \binom{m}{l} (\hat{h}_k \lambda_k)^l (-ab)^{m-l} + (-cab)^m + 2\sqrt{cab} + \sum_{j \neq i} c \lambda_j. \end{aligned} \quad (31)$$

Then it is easy to verify that we indeed have  $\sum_{i=1}^n p_i(\hat{h}_{-i}) = (n-1)z_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n)$ . Hence, the cost allocation function specified by (27) and (31) is efficient. Note that under this cost allocation, each firm continues to incur its own delay cost, the cost of capacity for which it is directly responsible, and a fraction of buffer capacity cost that is decreasing in its own unit delay cost and demand rate.

What now remains to show is that if the above cost allocation rule is always used to allocate cost in any coalition, then firms would prefer the grand coalition to any other coalition. First note that if the cost allocation rule, as specified in (27) and (31), is applied to every coalition  $\mathcal{J}$ , then a firm  $i \in \mathcal{J}$  would report truthfully its unit delay cost and incurs expected cost

$$\phi_i(h_i, h_{-i}|\mathcal{J}) = \frac{\sum_{j \in \mathcal{J}} h_j \lambda_j}{\mu_{\mathcal{J}}^* - \sum_{k=1}^n \lambda_k} + c\mu_{\mathcal{J}}^* - p_i(h_{-i}|\mathcal{J}), \quad (32)$$

where  $p_i(h_{-i}|\mathcal{J})$  is defined similarly as  $p_i(h_{-i})$  for the set of firms in  $\mathcal{J}$ .

**Theorem 5.2** *The cost allocation rule specified in (32) is in the core. That is,  $\phi_i(h_i, h_{-i}|\mathcal{N}) < \phi_i(h_i, h_{-i}|\mathcal{J})$  for any subset  $\mathcal{J}$  of  $\mathcal{N}$ .*

**Proof.** Noting that

$$\begin{aligned} \phi_i(h_i, h_{-i}|\mathcal{N}) = & 2 \sum_{m=1}^{\infty} \frac{f^{(m)}(cab)}{m!} (c^m \sum_{j=1, j \neq i}^n \sum_{l=1}^{m-1} \binom{m}{l} (h_j \lambda_j)^{m-l} (h_i \lambda_i)^l + c^m \sum_{l=1}^{m-1} \binom{m}{l} (-ab)^{m-l} (h_i \lambda_i)^l \\ & + c^m (h_i \lambda_i)^m) + c\lambda_i, \end{aligned}$$

and

$$\begin{aligned} \phi_i(h_i, h_{-i}|\mathcal{J}) = & 2 \sum_{m=1}^{\infty} \frac{f^{(m)}(cab)}{m!} (c^m \sum_{j \in \mathcal{J}, j \neq i} \sum_{l=1}^{m-1} \binom{m}{l} (h_j \lambda_j)^{m-l} (h_i \lambda_i)^l + c^m \sum_{l=1}^{m-1} \binom{m}{l} (-ab)^{m-l} (h_i \lambda_i)^l \\ & + c^m (h_i \lambda_i)^m) + c\lambda_i, \end{aligned}$$

we have

$$\begin{aligned} \phi_i(h_i, h_{-i}|\mathcal{N}) - \phi_i(h_i, h_{-i}|\mathcal{J}) &= 2 \sum_{m=1}^{\infty} \frac{f^{(m)}(cab)}{m!} \left( c^m \sum_{j \notin \mathcal{J}} \sum_{l=1}^{m-1} \binom{m}{l} (h_j \lambda_j)^{m-l} (h_i \lambda_i)^l \right) \\ &= 2 \sum_{j \notin \mathcal{J}} \left( \sqrt{c(h_i \lambda_i + h_j \lambda_j)} - \sqrt{c h_i \lambda_i} - \sqrt{c h_j \lambda_j} \right) < 0. \end{aligned}$$

The second equality can be obtained by letting  $cab \rightarrow 0$ . Consequently, the cost allocation scheme defined by  $\phi_i$  is in the core. ■

The above results are rather remarkable. Not only is it possible to design a cost allocation rule to induce all firms to reveal their true unit delay costs, and for this to be a dominant strategy, but this allocation scheme also ensures that all firms prefer the grand coalition. Furthermore, the cost allocation ensures that each firm incurs its own delay cost and the cost of capacity for which it is directly responsible, with the allocated cost being increasing in each firm's unit delay cost and demand rate. That is, the desirable fairness properties observed in the allocation of Section 4.2 continue to hold.

Intuitively, the above cost allocation scheme is truth revealing because each firm, in addition to incurring its own delay cost, incurs a fraction of the delay costs of all other firms. This, coupled with the fact that capacity investment decisions are optimized based on reported information, leads firms to prefer truth telling. In this sense, the cost allocation rule can be viewed as an example of a Groves mechanism from the theory of mechanism design; see for example, Groves (1973, 1976) and Groves and Loeb (1979). In fact, just like a Groves mechanism, the cost allocation rule has broad applicability for any queueing system where the expected delay is well defined and for which the  $p_i$  functions can be specified; see the Appendix for details.

## 6 Extensions to Systems with General Demand and Processing Times

In this section, we briefly discuss systems where the customer inter-arrival times and processing times are not necessarily exponentially distributed. Our objective here is not to provide a compre-

hensive analysis, which is outside the scope of this paper, but rather to offer preliminary insights into the impact of relaxing assumptions made so far and the extent to which results we obtained under these assumptions would continue to hold. Exact analysis for general systems is difficult. Therefore, to obtain these preliminary insights, we rely throughout on approximations that have been extensively used in the literature. For simplicity, we also restrict our treatment to the case of pure delay costs, although the analysis can be extended to systems with service level constraints.

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  $Y/\mu$ , where  $Y$  has a mean equal to one and coefficient of variation  $c_s$ . The parameter  $\mu$  is again a scaling factor that corresponds to the service rate. 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  $\lambda_i \rightarrow \infty$ ). In particular, we approximate the expected number of customers at firm  $i$ , given capacity level  $\mu_i$ , as follows

$$E[Q_i(\mu_i)] \approx \sigma_i^2 \frac{\lambda_i}{\mu - \lambda_i},$$

where  $\sigma_i = \sqrt{\frac{c_{a_i}^2 + c_s^2}{2}}$ . Motivation and supporting arguments for this approximation can be found in Harrison (1985) and more recently in Bassombo et al. (2008) and the references therein. The problem faced by each firm can then be restated as

$$\text{Minimize } z_i(\mu_i) \approx \frac{h_i \lambda_i}{\mu_i - \lambda_i} \frac{c_{a_i}^2 + c_s^2}{2} + c \mu_i.$$

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

$$z_i^* = c \lambda_i + 2 \sigma_i \sqrt{c h_i \lambda_i}. \quad (33)$$

Bassombo et al. (2008) show that this capacity is asymptotically optimal when the demand rate is high (i.e.,  $\lambda_i \rightarrow \infty$ ). 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  $c_{a_N}^2 = \sum_{i \in \mathcal{N}} \frac{\lambda_i c_{a_i}^2}{\sum_{i=1}^n \lambda_i}$ . For the case of full information, the capacity optimization problem can be stated as follows:

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

where  $\sigma_{\mathcal{N}} = \sqrt{\frac{c_{a_N}^2 + c_s^2}{2}}$ . 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 by

$$z_{\mathcal{N}}^* = c \sum_{i \in \mathcal{N}} \lambda_i + 2\sigma_{\mathcal{N}} \sqrt{c \sum_{i \in \mathcal{N}} h_i \lambda_i}. \quad (34)$$

**Observation 6.1** *Capacity sharing can lead to higher total cost in the system. That is, it is possible to have  $z_{\mathcal{N}}^* > \sum_{i \in \mathcal{N}} z_i^*$ .*

**Proof.** Comparing the optimal costs in (33) and (34), we can see that  $z_{\mathcal{N}}^* \leq \sum_{i \in \mathcal{N}} z_i^*$  does not always hold. To see that, let  $\lambda = \lambda_i$ . Then, in order to have  $z_{\mathcal{N}}^* \leq \sum_{i \in \mathcal{N}} z_i^*$  we must have

$$\frac{1}{n} \sum_{i=1}^n \sigma_i^2 \sum_{i=1}^n h_i \leq \left( \sum_{i=1}^n \sqrt{h_i \sigma_i^2} \right)^2.$$

But if  $n = 2$ ,  $\lambda_1 = \lambda_2$ ,  $\sigma_1 = 1$ ,  $\sigma_2 = 0$ ,  $h_1 = 0$ ,  $h_2 = 1$ , we have

$$\frac{1}{n} \sum_{i=1}^n \sigma_i^2 \sum_{i=1}^n h_i = \frac{1}{2} > 0 = \left( \sum_{i=1}^n \sqrt{h_i \sigma_i^2} \right)^2,$$

which is a counterexample. Note that in this counterexample, firms, when they are on their own,

do not need any buffer capacity as either the unit delay cost is zero or variability is zero. When the firms share the same facility, the overall variability is positive and, therefore, there is congestion which leads to delay costs being incurred by customers of firm 2. ■

Although capacity sharing is not always beneficial, it is still possible to identify plausible ranges of parameter values for which capacity sharing is beneficial. The following result describes such a setting.

**Theorem 6.2** *Capacity sharing is beneficial if, for each pair of firms  $i$  and  $j$ ,  $h_i \geq h_j$  if and only if  $\sigma_i \geq \sigma_j$ . In other words, capacity sharing is beneficial if firms with higher delay costs also have higher demand variability.*

**Proof.** To show that  $z_{\mathcal{N}}^* \leq \sum_{i=1}^n z_i^*$ , it suffices to show that

$$\sum_{i=1}^n \sqrt{c h_i \lambda_i \sigma_i^2} \geq \sqrt{\frac{\sum_{i=1}^n \lambda_i \sigma_i^2}{\sum_{i=1}^n \lambda_i}} \sqrt{c \sum_{i=1}^n h_i \lambda_i}.$$

Taking the square of both sides of the inequality, it is enough to show that

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

But to prove  $\sum_{i=1}^n \lambda_i \left( \sum_{i=1}^n h_i \lambda_i \sigma_i^2 + 2 \sum_{j \neq i} \sqrt{h_i \lambda_i \sigma_i^2 h_j \lambda_j \sigma_j^2} \right) \geq \sum_{i=1}^n \lambda_i \sigma_i^2 \sum_{i=1}^n h_i \lambda_i$ , it is sufficient to show that  $\sum_{i=1}^n \lambda_i \sum_{i=1}^n h_i \lambda_i \sigma_i^2 \geq \sum_{i=1}^n \lambda_i \sigma_i^2 \sum_{i=1}^n h_i \lambda_i$ . 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  $a(i, j) = h_i \lambda_i \lambda_j \sigma_i^2$  and  $b(i, j) = h_i \lambda_i \lambda_j \sigma_j^2$ . 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  $(h_i - h_j)(\sigma_i^2 - \sigma_j^2) \geq 0$ , we must also have

$$\sum_{j \neq i} h_i \lambda_i \lambda_j \sigma_i^2 \geq \sum_{j \neq i} h_i \lambda_i \lambda_j \sigma_j^2 \quad \text{and} \quad \sum_{i=1}^n \lambda_i \sum_{i=1}^n h_i \lambda_i \sigma_i^2 \geq \sum_{i=1}^n \lambda_i \sigma_i^2 \sum_{i=1}^n h_i \lambda_i,$$

which completes the proof. ■

Interestingly, the condition in Theorem 6.2 is independent of the firms' demand rates. In particular, if  $\sigma_i^2 = \sigma_j^2$  for all  $i$  and  $j$ , then capacity sharing is always beneficial. The case of exponential inter-arrival times and processing of course satisfies this condition, but it is also satisfied by a broader class of problems. The condition in Theorem 6.2 is also satisfied if firms have the same unit delay cost with  $h_i = h_j$  for all  $i$  and  $j$  or if the ratio  $h_i/\sigma_i$  is the same for all  $i \in \mathcal{N}$ . These results make intuitive sense, firms that have high demand variability but low delay costs (or high delay cost but low demand variability) could get away with little buffer capacity. However, this ceases to be the case if firms with high demand variability but low delay costs share the same facility with firms with high delay cost but low demand variability. The condition in Theorem 6.2 can be viewed as a requirement that firms be relatively "alike" in the sense that the magnitude of their unit delay costs are consistent with the magnitude of their variability parameters.

In systems where unit delay costs are private information, it is always possible to design a cost allocation rule under which all firms that decide to share a single facility would truthfully report their private information. In particular, we can show that an allocation rule of the same form as the one we considered in Section 5.2 is incentive compatible with truth telling being a dominant strategy for each firm.

**Theorem 6.3** *Consider the cost allocation rule under which the expected cost of firm  $i$  is given by*

$$\begin{aligned} \phi_i(\hat{h}_1, \dots, \hat{h}_n) = & \frac{\sigma_{\mathcal{N}} h_i \lambda_i}{\mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n) - \sum_{k \in \mathcal{N}} \lambda_k} + \sum_{j \neq i} \frac{\sigma_{\mathcal{N}} \hat{h}_j \lambda_j}{\mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n) - \sum_{k \in \mathcal{N}} \lambda_k} + c \mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n) \\ & - p_i(\hat{h}_{-i}). \end{aligned} \tag{35}$$

where  $p_i(\hat{h}_{-i})$  is. Then, under this cost allocation, all firms report truthfully their unit delay costs. Moreover, truthful reporting is a dominant strategy.

A proof for the above result is similar to the proof of Theorem 5.1 and is therefore omitted for brevity. The functions  $p_i$  can be constructed in a similar fashion to the one described in Section 5 to obtain an efficient allocation. Furthermore, we can show that under some conditions (e.g., when the ratio  $h_i/\sigma_i^2 = h_j/\sigma_j^2$  for all  $i, j \in \mathcal{N}$ ) the resulting cost allocation rule is in the core with all firms preferring the grand coalition to any other capacity sharing arrangements.

We conclude this section by noting that the results of this section highlight the fact that, although capacity sharing among all firms may not be always beneficial, it may be so among subsets of the firms that satisfy certain conditions. For these firms, the results we obtained in this paper can be used as a basis for determining capacity and for allocating cost.

## 7 Summary and Concluding Comments

In this paper, we presented models to study the benefit of capacity sharing among independent firms. We formulated the capacity sharing problem as a cooperative game. We showed that capacity sharing can lead to significant savings in total system cost. However, we found that whether or not firms choose to share capacity, and with whom, depends on how the associated cost are allocated. It also depends on whether or not firms are truthful about their private information. We showed that it is possible to design a cost allocation rule that induces all firms to report truthfully their private information and for this allocation to be in the core. We showed that there exists settings for which capacity sharing among all the firms may not be beneficial. In such settings, capacity sharing among subsets of the firms with similar characteristics may still be beneficial.

We view the results of this paper as a step toward a better understanding of the issue of cooperation among independent firms via capacity sharing in the presence of congestion. The results identify some of the important mechanisms that might be needed to make capacity sharing desirable and to mitigate the effect of incomplete information. Much more work obviously remains to be done. The analysis could be extended to more general and complex systems. For example, it would be of interest to study systems with service priorities, systems with multiple servers, and systems where firms have heterogenous work content. Our preliminary analysis of such systems (see Benjaafar et al. 2008) suggest that, although these systems are significantly less tractable, many

of the first order effects observed in the simpler models continue to hold. We also observe that, although capacity sharing is not always beneficial, it continues to be the case among subsets of firms that share sufficiently similar characteristics. The analysis could also be extended to systems with more general cost structures. For example, it would be of interest to study systems where capacity costs are either concave or convex or exhibit discontinuities and how such cost structures affect the desirability of capacity sharing and the design of cost allocation rules.

**Acknowledgments:** We thank seminar participants at Northwestern University, University of Wisconsin, and Lehigh University for useful feedback and comments. We particularly thank Albert Ha, Eran Hanany, Donald Hausch, Moshe Haviv, the Associate Editor, and two anonymous referees for insightful comments and suggestions on earlier versions of the paper.

## 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.
- Aksin O.Z., 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," working paper, Tel Aviv University, 2008.
- 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: Optimality of Tailored Chaining and Pairing," working paper, Northwestern University, 2008.
- 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.
- 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," working paper, New York University, 2007.
- S. Dewan and H. Mendelson, "User Delay Costs and Internal Pricing for a Service Facility," *Management Science*, **36**, 1502-1517, 1990.
- 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," to appear in *European Journal of Operational Research*, 2007.
- P. Gonzalez and C. Herrero, "Optimal Sharing of Surgical Costs in the Presence of Queues," *Mathematical Methods of Operations Research*, **59**, 435-446, 2004.
- T. Groves, "Incentives in Teams," *Econometrica*, **41**, 617-633, 1973.
- T. Groves, "Incentive Compatible Control of Decentralized Organizations," in *Directions in large-scale systems many-person optimization and decentralized control*, ed. by Y. C. Ho and S. K. Mitter, New York, 149-185, 1976.
- T. Groves and M. Loeb, "Incentives in a Divisionalized Firm," *Management Science*, **25**, 221-230, 1979.
- S. Gurusurthi 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.
- A. Y. Ha, "Incentive-compatible pricing for a service facility with joint production and congestion externalities," *Management Science*, **44**, 1623-1636, 1998.

- E. Hanany and Y. Gerchak, "Nash Bargaining over Inventory and Pooling Contracts," forthcoming in *Naval Research Logistics*, 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.
- 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. Koole and A. Pot, "An Overview of Routing and Staffing in Multi-Skill Customer Contact Centers," working paper, Vrije Universiteit, 2005.
- 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.
- 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 Queueing 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," to appear in *European Journal of Operational Research*, 2007.
- 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," working paper, Northwestern University, 2004.

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

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

## Appendix

In this appendix, we show how the incentive compatible allocation rule described in Section 4.2 can be extended to settings much more general than the M/M/1 setting of Section 5. To illustrate, we consider the case where facilities are modeled as GI/G/1 queues as in Section 6 (the applicability of the allocation rule is however significantly broader). Let  $E[W_{\mathcal{N}}(\mu_{\mathcal{N}})]$  refer to expected delay when firms in coalition  $\mathcal{N}$  share a single facility with capacity level  $\mu_{\mathcal{N}}$ . Given reported unit delay costs  $(\hat{h}_1, \dots, \hat{h}_n)$ , let  $\mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n)$  denote the capacity level that minimizes the expected cost

$$\hat{z}_{\mathcal{N}}(\mu_{\mathcal{N}}) = \sum_{i \in \mathcal{N}} \hat{h}_i \lambda_i E[W_{\mathcal{N}}(\mu_{\mathcal{N}})] + c\mu_{\mathcal{N}}.$$

Consider now the cost allocation rule under which the long run expected cost of firm  $i$  is given by

$$\begin{aligned} \phi_i(\hat{h}_1, \dots, \hat{h}_n) = & h_i \lambda_i E[W_{\mathcal{N}}(\mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n))] + \\ & \sum_{j \neq i} \hat{h}_j \lambda_j E[W_{\mathcal{N}}(\mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n))] + c\mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_n) - p_i(\hat{h}_{-i}), \end{aligned} \quad (36)$$

where  $p_i(\hat{h}_{-i})$  is a positive function that depends only on the reported unit delay costs of firms  $j \neq i$ . It is easy to verify that the cost of firm  $i$  would be minimized if the capacity level for the shared facility is set equal to  $\mu_{\mathcal{N}}^*(\hat{h}_1, \dots, \hat{h}_{i-1}, h_i, \hat{h}_{i+1}, \dots, \hat{h}_n)$ . However, this is possible only if firm  $i$  reports truthfully its unit delay cost  $h_i$ . Hence, the allocation rule specified in (36) is incentive-compatible with truth telling being a dominant strategy.

The above arguments are similar to those used to construct a Groves mechanism; see for example Groves (1973, 1976) and Groves and Loeb (1979). We briefly summarize the key steps below. In a typical application of a Groves mechanism, the capacity of a public good is determined based on the reported valuation of this public good by the users. Each user's valuation depends on her type, which is private information. Let  $\theta_i$  denote the true type of user  $i$  and  $\hat{\theta}_i$  be the reported type which may be different from the true one (type plays the role of unit delay cost in our application). User  $i$  derives private utility  $v(x, \theta_i)$  from the public good if the capacity level of the public good is  $x$  (private utility plays the same role as expected delay cost in our application). Under a Groves mechanism, the amount of capacity of the public good is chosen based on the reported types so as to maximize  $\sum_{i \in \mathcal{N}} v(x, \hat{\theta}_i) - g(x)$ , where  $v(x, \hat{\theta}_i)$  is the estimated utility user  $i \in \mathcal{N}$  would derive from capacity level  $x$  based on her reported type and  $g(x)$  is the cost of investing in capacity level  $x$ . Let  $x^*(\hat{\theta}_1, \dots, \hat{\theta}_n)$  denote this capacity level. Consider a mechanism under which each user is charged a fee such that user  $i$ 's overall utility is given by

$$v(x^*(\hat{\theta}_1, \dots, \hat{\theta}_n), \theta_i) + \sum_{j \neq i} v(x^*(\hat{\theta}_1, \dots, \hat{\theta}_n), \hat{\theta}_j) - g(x^*(\hat{\theta}_1, \dots, \hat{\theta}_n)) - p_i(\hat{\theta}_{-i}),$$

where  $p_i(\hat{\theta}_{-i})$  is a function that depends on the reported information of firms other than firm  $i$ . Then, using arguments similar to the ones used for our application, it is easy to verify that such a mechanism would induce users to report truthfully their type and to do so regardless of the reporting decisions of other firms.