We analyze the benefit of production/service capacity pooling 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 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 we relax some assumptions in our original problem, we show that capacity pooling may not always be beneficial. We also examine the impact of having one of the firms, instead of a neutral central decision maker, be responsible for making the capacity investment decision for the pooled system.

**Key words:** Capacity pooling, queueing systems, joint ventures, cost allocation, cooperative game theory
1 Introduction

Capacity pooling refers to the consolidation of capacity that resides in multiple facilities into a single one. In a system without pooling, each facility fulfills its own demand relying solely on its capacity. In a pooled system, demand is aggregated and fulfilled from a single shared facility. It has long been known that pooling is 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 pooling is also 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 capacity, service capacity, and inventory.

Capacity pooling has been studied mostly in situations where a single 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 pool 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. If some or all of the firms decide to pool, they must also decide on how to share the cost of the pooled facility. They must do so in a fair manner that prevents any of the firms from defecting and perhaps joining a pooled facility with other firms or staying on their own. Hence, firms that contribute more to the cost of the pooled facility (because of their higher usage of capacity or lower tolerance for delay) are expected to pay a greater share of total cost.

Capacity pooling 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, Isuzu and General Motors share the same plants to produce diesel engines; Fujitsu and AMD jointly produce flash memory; Chrysler and Chery, a Chinese car manufacturer, plan to share assembly plants to build subcompact cars; Acer, IBM, and Compaq shared the same plants in the 1990’s to assemble their own brands of computers. 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, firms are increasingly sharing infrastructure re-
sources such as telecommunication networks, computer services, in-house call centers, and facilities for handling back-office operations.

Capacity pooling among independent firms raises several important questions. For example, is pooling always beneficial to all firms? Does it always lead to a reduction in total capacity in the system? How should capacity costs be shared among the different firms? Is pooling among all the firms the best arrangement or would pooling among smaller subsets of the firms be more beneficial to particular firms? How should priorities for capacity usage be set among the firms and how do these priorities affect cost sharing? How are the answers to the above questions affected by whether an impartial decision maker chooses capacity levels or one of the firms selfishly makes this decision?

In this paper, we address these and other related questions. Our treatment is inspired by the above examples of capacity sharing in practice. However, the setting we study is specific: 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 unit 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. We refer to this scenario as the distributed system. 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. If the firms choose to collectively operate a shared facility, then they must select the capacity of the facility and decide on a scheme to share the corresponding cost. We refer to this scenario as the pooled system.

Our choice of modeling framework is in part motivated by the fact that types of manufacturing and service facilities can be viewed as queueing systems. Such a view is indeed prevalent in the literature. Despite this prevalence, the literature on capacity sharing in queueing systems is limited and few treatments exist on how the associated cooperation among independent firms can be modeled. Although we restrict our analysis in this paper to simple models of queueing systems (for the most part, single server systems where capacity corresponds to a service rate), the models are sufficiently rich to capture many of the important issues that arise when capacity is shared among independent firms. Moreover, the models we consider have been widely studied in settings with a
single firm. Therefore, it is appropriate to enrich that literature with models with multiple firms.

The contributions of this paper are as follows.

- We provide a framework for modeling capacity pooling 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.

- We formulate capacity pooling 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 exists a cost allocation rule for which all the firms are better off than under any other alternative pooling arrangement.

- 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. This allocation rule is perhaps consistent with those observed in practice.

- We extend our treatment beyond the M/M/1 queue framework and identify settings where pooling is not optimal. We also explore the impact of assigning priorities for capacity usage among firms. We use these results to draw various managerial insights.

- We study the effect of who makes the decision regarding capacity (a neutral decision maker or one of the firms) on the desirability of pooling and the role of allocation rules.

More generally, we view our paper as a first 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.

The rest of the paper is organized as follows. In Section 2, we provide a brief review of related literature. In Sections 3 and 4, we analyze the distributed and pooled systems respectively in settings where service facilities are modeled as single server queues with Poisson demand and exponentially distributed service times. In Section 5, we discuss cost allocation among the firms in a pooled system and describe an allocation rule that is in the core. In Section 6, we relax some of the assumptions in the original model and show that this can affect the benefit of pooling. In Section 7, we describe settings where firms are assigned priorities based on their delay costs. In Section 8, we consider the case where one of the firms makes the capacity decision in the pooled
system instead of an impartial decision maker. In Section 9, we provide a summary of the main results 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. He shows, for a wide range of assumptions, that delay cost is minimized by choosing a single server to which all the capacity is assigned. 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. Smith and Whitt (1981) point out that server pooling can be detrimental when customer streams have different service time requirements.

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. He shows that (1) the pooled system reduces expected delay and (2) the benefit of pooling increases as service time variability of the tasks increases. Mandelbaum and Reiman (1998) consider the pooling of general Jackson networks into single server queues with phase-type service time distributions. They show that pooling may or may not be beneficial depending on the values of system parameters.

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), Gurumurthi and Benjaafar (2004) and Jordan et al. (2005) study the chaining of servers, where each server can process customers from two types of 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) and Iravani et al. (2004). These papers belong to the growing literature on queueing systems with server flexibility; see Aksin et al. (2005) and Koole and Pot (2005) for recent reviews.
Our work 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 model 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, either in terms of service rate or number of servers, 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 the pooled system to any other pooling arrangement, including remaining on their own.\footnote{While this paper was already written, we became aware of the papers Gonzales and Herrero (2004) and Garcia-Diaz et al. (2007), who also deal with cooperation in queues. However, these models differ from ours in some key aspects, and thus in some of the issues explored. In particular, they do not optimize capacity (before or after pooling), and the primary object of pooled cost allocation there are fixed capacity costs for each coalition (referred to as the "airport problem").}

The issue of capacity pooling has been treated in other contexts in Operations Management. In particular, pooling has been an important theme in the study of inventory systems. In this case, pooling refers to the consolidation of multiple inventory locations into a single centralized location (physically or virtually). An extensive body of work, a review of which can be found in Alfaro-Tanco and Corbett (2003), has documented the costs and benefits of inventory pooling. Most results are consistent with those of Eppen (1979), who shows that a pooled system yields a lower cost than a distributed system, and that the difference is increasing in the variance of demands and decreasing in their negative correlation; see Gerchak and Mossman (1992) and Gerchak and He (2003) for more recent results and discussion. Benjaafar et al. (2005) study the joint pooling of production capacity and inventory and show that benefits from production capacity pooling can be more significant than those from inventory pooling.

There is a growing number of papers that consider inventory pooling among independent firms. Examples include Gerchak and Gupta (1991), Hartman and Dror (1996, 2005), Muller et al. (2002), Ben-Zvi and Gerchak (2005), Chen and Zhang (2006) and references therein. A review of the main concepts and a discussion of other applications in supply chain management can be found in Nagarajan and Sošić (2007), Kemahlioglu-Ziya (2004), and Chen and Zhang (2007).

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. An important focus of this literature is on determining whether or not there exist cost allocations that are in the core and characterizing various properties of such allocations. 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.

3 The Distributed System

We consider a system consisting of a set $\mathcal{N} = \{1, \ldots, n\}$ of $n$ independent firms. Each firm makes an independent decision about capacity in the form of a service rate. Firm $i$, $i \in \mathcal{N}$, faces an independent demand stream with customers arriving according to a Poisson process with rate $\lambda_i$. Each firm serves its customers 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. We relax this assumption in Section 6.1. 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. Our treatment of the $M/M/1$ queue can be viewed as a complement to that literature in 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), Allon and Federgruen (2007, 2006), Cachon and Zhang (2007), and the vast literature reviewed therein). In Section 6.2, we discuss the case where the service rate is fixed but the number of parallel servers can be varied. In that case, capacity is discrete. 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 (2006, 2007), Dewan and Mendelson (1990), and Stidham (1992), among others), leads to tractable analysis, and provides a useful benchmark for other cost structures. In Section 6.3, we discuss the case where capacity cost
is not linear.

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 of delay costs and service levels are both common in the literature; see for example Dewan and Mendelson (1990, Mendelson and Whang (1990), Allon and Federgruen (2006, 2007) 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 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).

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

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

subject to

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

and

$\lambda_i/\mu_i \leq 1$. \hspace{1cm} (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 $i \in N$ and (2) the case where $h_i = 0$ for all $i \in N$. 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 Federgruen (2006) 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

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

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 $z_{i_1, \ldots, n}^* = \sum_{i \in N} z_i^*$. In systems where $\sqrt{\frac{h_i \lambda_i}{c}} \geq \frac{\ln(\frac{1}{1-\alpha_i})}{w_0}$ for all $i \in N$, 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, $z_{i_1, \ldots, n}^*$, given by

$$z_{i_1, \ldots, n}^* = c \sum_{i \in N} \lambda_i + 2 \sum_{i \in N} \sqrt{h_i \lambda_i c}. \quad (10)$$
In the case of identical firms, with $\lambda_i = \lambda$ and $h_i = h$ for all $i \in N$, the optimal total cost in (10) reduces to

$$z^*_1,\ldots,n = cn\lambda + 2n\sqrt{h\lambda c},$$

and the total capacity in the system to

$$\sum_{i \in N} \mu^*_i = n \left( \lambda + \sqrt{\frac{h\lambda}{c}} \right).$$

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 $q_i h_i c \ln(1 - q_i) w_0$. 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 The Pooled System

In the pooled system, the $n$ firms in the set $N$ share a single facility with service rate $\mu_N$ from which the demand of all firms is satisfied (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 $\sum_{i \in N} \lambda_i$. Similarly, because the work content for each customer regardless of its firm is exponentially distributed, the processing time at the pooled facility is a random variable $X_N = Y/\mu_N$ with the exponential distribution and mean $1/\mu_N$. We assume that customers regardless of their firm affiliation are served in a FCFS fashion (we consider alternative priority schemes in Section 7). Hence, the pooled system behaves again as an M/M/1 queue.

We assume that a central decision making entity determines the amount of capacity to acquire so as to minimize the sum of expected delay cost experienced by customers of all firms and the cost of capacity subject to the service level constraints associated with each firm. This implies that firms share with this central decision making entity information about delay costs, service level requirements, and the distribution of demand and work content. This is a reasonable assumption in settings where the firms are all subsidiaries of a single large firm. It is also plausible that even independent firms would share information about demand and processing times as they would eventually be observed by the pooled facility. Obviously firms would want to share information about service levels since the pooled facility must guarantee that they are satisfied. In settings where delay costs are directly incurred by the pooled facility (e.g., the pooled facility is responsible
for paying delay penalties directly to the customers), firms would also need to provide the pooled facility with the correct delay costs. However, in settings where delay costs are unobservable by the pooled facility (e.g., delay penalties are handled independently by each firm or the penalties correspond to a private assessment of loss of customer good will), firms may have an incentive not to disclose the correct costs. In such settings, a formulation that includes service level requirements but not delay costs would be more appropriate.

Let $z_N(\mu_N)$ denote total system cost and let $W_N$, a random variable, refer to customer delay. Then the problem faced by the central decision maker for the pooled system can be stated as follows:

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

subject to

$$P(W_N \leq w_0) = 1 - e^{(\mu_N - \sum_{i \in N} \lambda_i)w_0} \geq \alpha_N, \quad (14)$$

and

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

where $\alpha_N = \max(\alpha_1, ..., \alpha_n)$. Then, the optimal capacity is given by

$$\mu^*_N = \sum_{i \in N} \lambda_i + \eta_N, \quad (16)$$

where

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

Similar to the distributed system, the optimal capacity in the pooled system 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 $\alpha_N$.

**Theorem 4.1** $z^*_N \leq z^*_{1,...,n}$ and $\mu^*_N \leq \sum_{i=1}^{n} \mu^*_i$.

The proof of Theorem 4.1 and all subsequent results can be found in the Appendix. Pooling leads to lower total cost for the system and to lower investments in capacity. The potential magnitude of the savings can be more easily seen in a system with identical firms where $\alpha_i = \alpha$, $h_i = h$, and $\lambda_i = \lambda$ for all $i \in N$. First consider the case where $\sqrt{\frac{nh\lambda}{c}} \geq \frac{\ln\left(\frac{1}{1 - \alpha_N}\right)}{w_0}$. This leads to $\mu^*_N = n\lambda + \sqrt{\frac{nh\lambda}{c}}$, $z^*_N = cn\lambda + 2\sqrt{cnh\lambda}$, and $E(W^*_N) = \sqrt{\frac{nh\lambda}{c}}$ 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 distributed system). For the case where $\sqrt{\frac{nh\lambda}{c}} \leq \frac{\ln\left(\frac{1}{1 - \alpha_N}\right)}{w_0}$, we have
\[ \mu_N^* = n\lambda + \frac{\ln(\frac{1}{1-n})}{w_0}, \quad z_N^* = c(n\lambda + \frac{\ln(\frac{1}{1-n})}{w_0} + \frac{nh\lambda w_0}{\ln(\frac{1}{1-n})}), \quad \text{and} \quad E(W_N^*) = \frac{w_0}{\ln(\frac{1}{1-n})}. \]

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 distributed system. Although capacity investments vary in the two cases, it is important to observe that, in both cases, expected delay in the pooled system is no worse than the expected delay experienced in the distributed system.

We conclude this section by noting that, while it is consistent with intuition that pooling would be beneficial when delay costs and service levels of all the firms are identical, it is not at all obvious that it would be so when different firms have different delay costs or different service levels (note that the pooled system imposes the highest of the individual firms’ service levels). In fact, in other settings, such as those where capacity is in the form of inventory, pooling can be detrimental when delay-related costs, such as backorder costs, are sufficiently different and no appropriate priorities are instituted (see for example de Vericourt et al. (2002), Benjaafar et al. (2005) and Ben-Zvi and Gerchak (2005)). In these settings, pooling a firm with a low demand rate but a high delay cost with a firm with a high demand rate but low delay cost can lead to overall higher costs. In Section 6.2, we show that this can also arise in our setting when capacity is in the form of a discrete number of servers instead of a continuous service rate. Finally, let us note that our results show that pooled capacity is always lower than the total capacity in the distributed system. This is not the case with other forms of capacity. For example, it has been observed that inventory pooling, while always beneficial, can increase capacity (Gerchak and Mossman 1992).

5 Cost Sharing

In the previous section, we showed that pooling 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. A firm benefits from pooling if its share of the cost is lower than the cost it would incur from operating its own independent facility. If both capacity costs and delay costs (if any) are absorbed by the pooled facility, each firm is charged a usage fee (per unit time) that corresponds to a fraction of total cost. If, on the other hand, only capacity cost is incurred by the pooled facility with delay costs incurred directly by the individual firms, the usage fee would correspond only to a fraction of capacity cost. The net effect in both cases, is that each firm would incur a fraction of total system cost.

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 pool capacity with only a subset of the firms. This could lead firms to form groupings around multiple smaller shared facilities. A single pooled facility would be preferred by all firms only if there exists a vector of usage fees under which the firms are better off than under any other capacity sharing arrangement, including operating individual facilities. Hence, ideally, the cost allocation in the pooled system 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 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 \( \phi = (\phi_1, \ldots, \phi_n) \) is called an allocation rule if \( \phi_i \) corresponds to the portion of total cost in the grand coalition that is assigned to 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 \( \mathcal{J} \) 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:

\[
\sum_{i \in \mathcal{J}} \phi_i \leq z^*_\mathcal{J}, \quad \forall \mathcal{J} \subseteq \mathcal{N}, \tag{18}
\]

and

\[
\sum_{i \in \mathcal{N}} \phi_i = z^*_\mathcal{N}. \tag{19}
\]

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:

$$\phi_i^* = \frac{h_i \lambda_i}{\eta_N} + c \lambda_i + \gamma_i,$$

(20)

where

$$\gamma_i = \frac{h_i \lambda_i}{\sum_{i \in N} h_i \lambda_i} c \eta_N$$ if \(\frac{\ln(\frac{1}{1 - \alpha_N})}{w_0} \leq \sqrt{\frac{\sum_{i \in N} h_i \lambda_i}{c}}\)

(21)

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

$$\gamma_i = \begin{cases} c \eta_N - c \sqrt{\frac{\sum_{i \in N : i \neq i_{\text{max}}} h_i \lambda_i}{\sum_{i \in N : i \neq i_{\text{max}}}} h_i \lambda_i} & \text{if } i = i_{\text{max}}, \text{ and} \\
\frac{h_i \lambda_i}{c} \frac{\sum_{i \in N : i \neq i_{\text{max}}} h_i \lambda_i}{c} & \text{if } i \neq i_{\text{max}}, \end{cases}$$

(22)

with \(i_{\text{max}} \in \{i : \alpha_i = \max(\alpha_1, \ldots, \alpha_n)\}\). Under the above allocation rule, each firm is charged a fee consisting of two parts: (1) its own delay cost, \(\frac{h_i \lambda_i}{\eta_N}\) 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 \(\frac{\ln(\frac{1}{1 - \alpha_N})}{w_0} \leq \sqrt{\frac{\sum_{i \in N} h_i \lambda_i}{c}}\), this fraction is proportional to the firms’ demand-weighted delay costs. If \(\frac{\ln(\frac{1}{1 - \alpha_N})}{w_0} > \sqrt{\frac{\sum_{i \in N} h_i \lambda_i}{c}}\) (the case where the service level constraint is more restrictive), firm \(i_{\text{max}}\) 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.

**Theorem 5.1** The cost allocation rule defined by the vector \(\phi^* = (\phi_1^*, \ldots, \phi_n^*)\) is in the core. That is, under this cost allocation, the grand coalition is optimal for all firms in \(N\).

The above allocation rule is valid regardless of whether or not delay costs are absorbed by the pooled facility. If they are, then \(\phi_i^*\) is indeed the fee that is charged to firm \(i\). If, on the other hand, delay costs are incurred directly by the individual firms, the usage fee that is charged to firm \(i\) corresponds only to the portion of the capacity cost given by \(c(\lambda_i + \gamma_i \eta_N)\). In systems without delay costs, firms are charged only the portion of capacity cost. It is perhaps surprising, given the heterogeneous delay costs and service level requirements among the different firms, that the
allocated cost can always be broken down in this fashion, with firms always being charged, directly or indirectly, their own delay costs. 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 Gans and Zhou (2003, 2005) and Aksin et al. (2004).

The allocation rule specified in (20)-(22) is not unique. It is possible to construct other cost allocations that are also in the core. As we show in the following theorem, a well known cost allocation, the Shapley value, is in the core for two important special cases.

**Theorem 5.2** In systems with either only delay costs or only service level constraints, the capacity sharing game is a concave game, i.e.,

\[ z^*_S + z^*_T \geq z^*_{S \cup T} + z^*_{S \cap T}, \quad \forall S, T \subseteq N. \]

Therefore, the Shapley value is in the core.

The Shapley value is an allocation scheme that has been shown to possess several desirable fairness properties and has been widely studied in other contexts (see for example Moulin (1995)). However, the Shapley value does not necessarily separate the allocated cost the way our proposed allocation does.

### 6 Is Capacity Pooling Always Beneficial?

We have so far studied a setting where pooling is always beneficial to each firm and there is a fair cost allocation scheme under which all firms prefer to participate in a single pooled facility. In this section, we highlight situations where capacity pooling may not be desirable and firms may prefer to either operate independent facilities or to form smaller coalitions. In particular, we illustrate the effect of three factors that impact the desirability of pooling: (1) work content heterogeneity among the firms, (2) capacity discreteness, and (3) convexity of capacity cost. For the sake of brevity, we focus on systems with only delay costs (i.e., without service level constraints). The treatment can be easily extended to the case where delay costs and service level constraints are both in effect.

#### 6.1 The Effect of Heterogeneous Work Content

In the model described in Sections 3 and 4, we assumed that work content is homogenous 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 content associated with customers vary from firm to firm. Customer processing time is still a random variable of the form \( Y_i / \mu \), for firm
i, but \( Y_i \) is now exponentially distributed with mean \( \omega_i \) (recall that, in the original model, \( \omega_i = 1 \) for all \( i \in \mathcal{N} \)). Customers are still served on a FCFS basis. In a pooled system, with \( n \) firms sharing a single facility, service times are no longer exponentially and identically distributed. Instead the distribution of service time is a mixture of exponential distributions (i.e., the distribution is hyper-exponential) with mean \( \mathbb{E}(S) = \sum_{i \in \mathcal{N}} p_i \omega_i / \mu_N \) and second moment \( \mathbb{E}(S^2) = 2 \sum_{i \in \mathcal{N}} p_i \omega_i^2 / \mu_N^2 \), where \( p_i = \lambda_i / \sum_{i \in \mathcal{N}} \lambda_i \). The resulting squared coefficient of variation \( C_s^2 = \frac{\mathbb{E}(S^2) - \mathbb{E}(S)^2}{\mathbb{E}(S)^2} \) 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.

In a distributed system, the problem faced by firm \( i \) can be stated as

\[
\text{minimize } z_i(\mu_i) = c \mu_i + \frac{h_i \lambda_i}{\mu_i / \omega_i - \lambda_i}
\]

subject to

\[
\lambda_i \omega_i / \mu_i \leq 1,
\]

and in the pooled system as

\[
\text{minimize } z_N(\mu_N) = \sum_{i \in \mathcal{N}} h_i \lambda_i \left\{ \frac{(1 + C_s^2) \mathbb{E}(S)}{2(1 - \rho)} + \frac{\omega_i}{\mu_N} \right\} + c \mu_N
\]

subject to

\[
\rho = \frac{\sum_{i \in \mathcal{N}} \lambda_i \omega_i}{\mu_N} \leq 1.
\]

The optimization problem in both cases is convex and therefore can be easily solved numerically. However, a closed form analytical solution is difficult to obtain in the case of the pooled system because the pooled facility no longer behaves like an M/M/1 queue and, instead, behaves like a 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 \( C_s \).

It is difficult to characterize analytically the optimal capacity in a shared facility. Therefore, it is also difficult to characterize analytically conditions under which the core may or may 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 \( i \in \mathcal{N} \). In this case, the expression for \( C_s^2 \) simplifies to \( C_s^2 = 2 C_\omega^2 + 1 \) where \( C_\omega \) is the coefficient of variation of the means of the individual work contents. Total cost in the pooled system can now be expressed as

\[
z_N(\mu_N) = c \mu_N + h \left\{ \frac{(1 + C_\omega^2) \rho^2}{(1 - \rho)} + \rho \right\}.
\]

It is easy to see that, everything else remaining the same, total cost for a given service rate \( \mu_N \) is increasing in \( C_\omega \) and, consequently, the optimal capacity level is also increasing in \( C_\omega \). This
suggests that for sufficiently high $C_\omega$, the pooled system could be less desirable than a system with independent facilities.

This observation is confirmed by the numerical results shown in Figures 1 where the percentage cost difference between the pooled and distributed system, $(z^*_N - z^*_1,\ldots,n)/z^*_1,\ldots,n$, is shown for different values of $C_\omega$. There is a threshold on $C_\omega$ above which the distributed system becomes preferable. The effect of increasing $C_\omega$ 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).

Note that the fact that total pooling may not be superior to multiple independent facilities does not mean that all forms of pooling are not desirable. In Figure 2, we compare the performance of three facility pooling scenarios in a system with 4 firms. Scenario 1 corresponds to the grand coalition, scenario 2 to four independent facilities, and scenario 3 to two pooled facilities: one shared by firms 1 and 2 and the other by firms 3 and 4. The firms are identical with respect to all parameters, except for their mean work contents which are given by $\omega_1 = 2.5, \omega_2 = 2.5, \omega_3 = 100$, and $\omega_4 = 450$. As we can see, scenario 3 is superior to both scenarios with the difference increasing as delay costs increase. That is, firms with similar work content find it beneficial to form smaller shared facilities.

6.2 The Effect of Discrete Capacity

We have so far assumed that capacity can be varied continuously in the form of a service rate. This allowed us to model the service facility, regardless of the number of participating firms and their corresponding demand rates, as a single server queueing system. In this section, we consider settings where capacity can be acquired only in discrete quantities with each unit of capacity corresponding to a separate server (e.g., servers are human operators, computer servers, or telecommunication lines). The service rate of each server is fixed but the number of servers is variable. Hence, facilities can now be viewed as multi-server queues with a fixed service rate per server.

Consistent with our treatment in our original model of Sections 3 and 4, let us assume that firm $i$, for all $i \in \mathcal{N}$, faces a Poisson demand with rate $\lambda_i$, the service rate of each standard server is $\mu$, service times are exponentially distributed (with mean $1/\mu$), customers are served on a FCFS basis, and the amortized cost per server is $k$.

Consider first the distributed system. Let $m_i$ denote the number of servers in firm $i$’s facility. Then, each facility can be modeled as an $M/M/m_i$ queue. The problem faced by firm $i$ in
determining the optimal number of servers can be stated as follows:

\[ \text{minimize } z_i(m_i) = h_i \lambda_i \left( \frac{1}{\mu} + \frac{P(m_i, \lambda_i)}{m_i \mu - \lambda_i} \right) + km_i, \]  

subject to

\[ \rho_i = \frac{\lambda_i}{m_i \mu} < 1, \]  

and

\[ m_i \geq 1, \quad m_i : \text{integer}, \]  

where

\[ P(m_i, \lambda_i) = \frac{(m_i \rho_i)^{m_i}}{m_i!(1-\rho_i)} + \frac{(m_i \rho_i)^{m_i}}{m_i!(1-\rho_i)}. \]  

Using the fact that \( P(m_i) \) is convex in \( m_i \) (see for example Dyer and Proll (1977)), we can show that \( z_i(m_i) \) is also convex. Therefore, the optimal number of servers is given by the smallest integer \( m_i \) that satisfies the inequality:

\[ z_i(m_i + 1) - z_i(m_i) > 0. \]

We denote the optimal capacity level by \( m_i^* \). Although it is difficult to obtain a closed form expression for \( m_i^* \) and the corresponding optimal cost \( z_i^* \equiv z_i(m_i^*) \), both are easy to compute numerically.

For the pooled system, the analysis is similar. Expected cost, given a choice of number servers \( m_N \), is given by:

\[ z_N(m_N) = \sum_{i \in N} h_i \lambda_i \left( \frac{1}{\mu} + \frac{P(m_N, \sum_{i \in N} \lambda_i)}{m_N \mu - \lambda_i} \right) + km_N, \]  

where

\[ P(m_N, \sum_{i \in N} \lambda_i) = \frac{(m_N \rho_N)^{m_N} m_N! (1-\rho_N)}{\sum_{n=0}^{m_N-1} \frac{(m_N \rho_N)^n}{n!} + \frac{(m_N \rho_N)^m}{m_N!(1-\rho_N)}}. \]  

and \( \rho_N = \frac{\sum_{i \in N} \lambda_i}{m_N \mu} \). The optimal capacity \( m_N^* \) and the optimal cost \( z_N^* \) can be obtained using a similar approach as in the case of the distributed system.

In contrast to the case where capacity is in the form of a service rate that can be varied continuously, pooling is not always beneficial here. We show this by counter example. In Figure 3, we provide numerical results for a system with two firms, 1 and 2, with the following characteristics: \( \mu = 1, \lambda_1 = \epsilon, \epsilon > 0, h_1 = \frac{1000}{\epsilon} \), \( \lambda_2 = 1 \) and \( h_2 = 10 \). The figure shows the impact of varying \( \epsilon \) on the optimal cost for the pooled and distributed systems. As we can see, for sufficiently small values of \( \epsilon \), the total cost in the distributed system is lower than the cost in the pooled system. As \( \epsilon \) gets small, it is possible to invest in fewer servers for firm 1 in the distributed system, although
the delay cost is large. This is possible because customers of firm 1 either experience no wait or wait very little for service. This is not the case in the pooled system. The wait for customers of firm 1 can be significant even though $\epsilon$ is small (e.g., when servers are busy with customers from firm 2). This leads to higher delay cost since $h$ is increasing in $\epsilon$. In turn, this could prevent the reduction in the optimal number of servers seen in the distributed system. In the example shown in Figure 3, reducing $\epsilon$ from 2.5 to 0.5 reduces the number of servers of firm 1 from 10 to 6 (the number of servers of firm 2 is unaffected). However, in the pooled system the number of server remains unchanged at 23.

In general, we have observed that pooling becomes less beneficial as variability in delay costs among customers of different firms increases. Interestingly, this effect is absent in the model of sections 2-4 where capacity is in the form of a service rate. In that case, the benefit of pooling are much more significant. In particular, the reduction in service times that accompanies pooling is always sufficient to offset the negative impact of any increased variability in delay costs.

While it is possible for total or partial capacity pooling to be beneficial (e.g., when variability in delay cost among firms sharing the same facility is relatively moderate), it is difficult to characterize conditions under which this occurs. Therefore, it is also difficult to identify analytically conditions under which the grand coalition is optimal. An exception is when the firms have identical delay costs, with $h_i = h$ for all $i \in \mathcal{N}$. In this case, we can show that pooling is system-optimal, with $z^*_N = z_N(m^*_N) \leq z^*_1, \ldots, n = \sum_{i \in \mathcal{N}} z_i(m^*_i)$; see the Appendix for a proof. Unfortunately, even in this case, it is difficult to show that the core is not empty in general. However, it is possible to do so if we restrict ourselves to the limit regime where $\lambda_N \to \infty$ and rely on approximations that are asymptotically exact. This regime has gained a lot of attention in the literature recently; see Halin and Whitt (1981), Reiman (1984), and Borst et al. (2004). It is important in practice because it is useful in modeling large scale service systems such as large call centers, say with hundreds of operators. We briefly sketch this approach and show how our allocation rule is in fact in the core in this case.

Following the approach in Borst et al. (2004), the optimal number of servers can be approximated by the continuous parameter $\hat{m}^*_N$, where

$$\hat{m}^*_N = \lambda_N/\mu + \sqrt{\lambda_N/\mu}, \quad x^* = \arg \min_{x > 0} \{ kx + \frac{hP(x)}{x} \},$$

$$P(x) := \frac{1}{1 + e^{-x}},$$

and $H(\cdot)$ is the hazard rate function of the standard normal distribution. The value of $\hat{m}^*_N$ is asymptotically optimal with $\lim_{\lambda_N \to \infty} \frac{\hat{m}^*_N}{m^*_N} = 1$. Expected delay can now be approx-
imated (the approximation is also asymptotically exact) as
\[ E(\hat{W}_N^*) = \frac{P(x)}{m_N^* \mu - \lambda_N^*} + \frac{1}{\mu}, \]
and expected cost as
\[ \hat{z}_N^*(m_N^*) = k\hat{m}_N^* + h\lambda_N^*E(\hat{W}_N^*). \]
Consider now the allocation rule specified by
\[ \hat{\phi}_i^* = h\lambda E(\hat{W}_N^*) + k\lambda_i/\mu + (\lambda_i/\lambda_N^*)kx^*\sqrt{\lambda_N^*/\mu}, \]
for \( i \in \mathcal{N} \). This is a similar allocation rule to the one specified in equations (20)-(22). Each firm is charged a fee consisting of three parts: (1) its own delay cost, (2) the cost of capacity for which it is directly responsible, \( k\lambda_i/\mu \), and (3) a portion of the cost of buffer capacity \( (\lambda_i/\lambda_N^*)kx^*\sqrt{\lambda_N^*/\mu} \) which is proportional to its demand rate (recall that \( h_i = h \) for all \( i \in \mathcal{N} \)).

**Theorem 6.1** The allocation rule specified by \( \hat{\phi}_i^* \) for \( i \in \mathcal{N} \) is in the core under the cost function \( \hat{z}_N^* \).

The above result suggests that the effect of discrete capacity becomes less significant when demand is sufficiently large and consequently the number of servers is also large. The result also provides evidence of the robustness of the allocation function originally specified for systems with single servers.

### 6.3 The Effect of Diseconomies of Scale

In some settings, capacity cost may not be linear (in service rate or number of facilities) but instead may feature either economies or diseconomies of scale. The first case may correspond to situations where operating a facility involves fixed costs that are largely independent of the size of the facility. In this case, the benefits of pooling, if any, would not be diminished and we expect the grand coalition to continue to be optimal. The second case may correspond to situations where increasing capacity gets increasingly expensive (e.g., due to limits on technology or to increasing managerial complexity). Here, pooling may not always be beneficial. It is easy to construct examples where either the distributed system or a system consisting of multiple small coalitions is optimal.

### 7 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 since it does not minimize the overall expected delay. A policy that does is one where customers with higher delay costs are given higher priority. For an M/M/1 queue with multiple customer classes, where each class may have a different delay cost, the so-called $c\mu$ rule minimizes expected delay in the system (see for example, Jaiswal (1968), Klimov (1974), Harrison (1975)). 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 higher service priority is given to customers with higher delay costs. In this section, we study the impact of using such a policy on the desirability of pooling and whether or not the grand coalition continues to be optimal.

Other than the priority policy, the assumptions of the model we consider are the same as those of our original model described in Sections 3 and 4. Because the analysis of systems with priority is considerably more difficult than the FCFS case, we limit our analysis to systems with no service level constraints.

Without loss of generality, we let $h_1 \geq h_2 \geq \cdots \geq h_n$. Under the priority policy, each time the service facility becomes available, the customer with the highest delay cost, among those waiting, is selected for service; a FCFS policy is used to choose among customers with the same delay cost. We assume that preemption is allowed so that the service of a customer, regardless of its class, can be interrupted for the service of a higher priority customer (the analysis can be easily extended to systems where preemption is not allowed and the results are qualitatively the same). Under a priority scheme of this type, total expected cost in a pooled facility with $N$ firms and service rate $\mu_N^P$ is given by

$$z_N^P(\mu_N^P) = \sum_{i=1}^{n} h_i \lambda_i \left( \sum_{k=1}^{n} \frac{\lambda_i}{S_i S_{i-1}} + \frac{1}{\mu_N^P S_i} \right) + c\mu_N^P,$$

(34)

where $S_i = 1 - \sum_{k=1}^{i} \frac{\lambda_k}{\mu_N^P}$, $i \geq 1$ and $S_0 = 1$.

Although it is not possible to provide an explicit expression for the optimal cost and the optimal capacity level, it is possible to show that both the optimal cost $z_N^{P*}$ and the optimal capacity $\mu_N^{P*}$ under the priority policy are lower than those under the FCFS policy.

**Theorem 7.1** $z_N^{P*} \leq z_N^*$ and $\mu_N^{P*} \leq \mu_N^*$, where $z_N^*$ and $\mu_N^*$ refer respectively to the optimal cost and the optimal capacity under the FCFS policy.

Showing that $z_N^{P*} \leq z_N^*$ is trivial since the expected delay cost under the priority policy is always lower than under the FCFS policy for any choice of $\mu_N$. In order to show that $\mu_N^{P*} \leq \mu_N^*$, it is sufficient to show that (1) $z_N^P$ is convex in $\mu_N^P$ and (2) $\frac{\partial z_N^P(\mu_N^P)}{\partial \mu_N^P}|_{\mu_N^P=\mu_N^{P*}} \geq 0$. A proof that both conditions hold is provided in the Appendix.
Although pooling under the priority policy is more cost efficient for the system and requires less capacity, it does not guarantee that all firms would experience a lower expected delay.

**Observation 7.2** It is possible to have \( E[W^*_i; N] > E[W^*_i] \), where \( E[W^*_i; N] \) and \( E[W^*_i] \) refer respectively to the expected delay of firm \( i \) in the pooled and distributed systems, in each case under the corresponding optimal capacity.

The observation can be proven by counter-example. Consider the case where \( h_i = h \) for \( i = 1, \ldots, n \), \( \lambda_n = \frac{h}{c} > 1 \), and \( \lambda_i > 0 \) for \( i \neq n \). Customers are processed according to the priority policy \( 1, \ldots, n \) with firm \( n \) receiving the lowest priority. Then,

\[
E[W^*_n; N] = \frac{\sum_{i=1}^n \lambda_i}{\sqrt{b \sum_{i=1}^n \Lambda_i + \lambda_n}} + \frac{1}{\sqrt{b \sum_{i=1}^n \Lambda_i + \lambda_n}} \\
\geq \sqrt{\frac{c}{h\lambda_n}} = E[W^*_n].
\]

The above result illustrates the fact that a firm with a low priority may experience higher expected delay (this can be the case for example when the capacity cost is high and the demand rate for a low priority firm is low, or vice-versa). The fact that expected delay increases for some firms could make the priority policy difficult to implement in practice. This could of course be addressed by specifying service level constraints in terms of upper bounds on expected delay for each firm. A priority policy with constraints on expected delay can be shown to be still superior to a FCFS policy. Therefore, the pooled system is still a superior configuration to the distributed one. In the absence of service level constraints, a firm who sees its expected delay increase may still prefer the pooled system if its overall share of total cost is lower than its cost in the distributed system. In the following Theorem, we show that it is possible to design a cost allocation scheme under which all firms are better off in the pooled system.

**Theorem 7.3** For a pooled system under a priority policy, there always exists a feasible cost allocation consisting of a vector \( \phi^P = (\phi^P_1, \ldots, \phi^P_n) \) where \( \phi^P_i = h_i \lambda_i E[W^*_i; N] + \beta_i \), and \( \beta_i \geq 0 \), such that \( \phi^P_i \leq z^*_i \) and \( \sum_{i \in N} \phi^P_i = z^*_N \).

Under the above cost allocation scheme, each firm in the pooled system is responsible for its own delay cost and a portion of total capacity cost. The allocation is individually rational for each firm (each firm is better off in the pooled system than being on its own) and is efficient (the total allocated cost is equal to the total cost in the system). However, we have not been able to show that the allocation is stable for each coalition \( J \subseteq N \) (i.e., \( \sum_{i \in J} \phi^P_i \leq z^*_J \), \( \forall J \subseteq N \)). Therefore, it is not clear whether or not the core is empty. Extensive numerical experiments we carried out
seem to support the fact that the core is non-empty. Pooling any set of coalitions into a single larger coalition always reduces cost. Showing that this is true in general has proven challenging because of lack of closed form expressions for the optimal cost.

8 Social versus Individual Optimization

In some settings, the decision regarding how much capacity is selected is not carried out by a neutral central authority as we have assumed so far, but by one of the individual firms instead. For example, in addition to fulfilling its own demand, one of the firms may scale up its facility to fulfill demand from other firms (for example in the 1990’s, Acer expanded its production facilities to produce computers for IBM and Compaq in addition to producing its own). There are clearly many arrangements under which one firm could offer capacity to outside firms. In this section, we focus on one particular scenario where one firm determines the amount of total capacity in which to invest. The firm makes this decision so as to minimize the sum of its own delay cost and its share of capacity cost subject to constraints on service level requirements. Consistent with the cost allocation discussed in section 5, we restrict our treatment to allocation schemes where firms incur their own delay cost and pay a portion of total capacity cost.

The problem faced by the lead firm, indexed by $j$, can then be stated as follows:

\[
\text{Minimize } z^j_N(\mu^j_N) = \theta_j c \mu^j_N + \frac{h_j \lambda_j}{\mu^j_N - \sum_{i \in N} \lambda_i} \\
\text{subject to } P(W^j_N \leq w_0) = 1 - e^{(\mu^j_N - \sum_{i \in N} \lambda_i)w_0} \geq \alpha_N, \\
\text{and } \sum_{i \in N} \lambda_i / \mu^j_N \leq 1,
\]

where $\mu^j_N$ is the service rate firm $j$ chooses, $\theta_j$ is the fraction of capacity cost it incurs, and $W^j_N$ is the expected delay in the system given capacity $\mu^j_N$. Firm $j$ chooses capacity $\mu^j_N$ to minimize its own cost. In doing so, it does not account for the delay costs of other firms, but it does take into account their service level requirements.

Noting that the objective function is convex in $\mu^j_N$, the optimal capacity level can be obtained from the first order condition of optimality as:

\[
\mu^j_N = \sum_{i \in N} \lambda_i + \max\{\frac{\ln(\frac{1}{1-\alpha_N})}{w_0}, \sqrt{\frac{h_i \lambda_i}{\theta_j c}}\}.
\]
If \( \ln \left( \frac{1}{w_0} \right) \geq \sqrt{\frac{h_j \lambda_j}{\theta_j c}} \), then firm \( j \) is forced to acquire the same amount of capacity that a neutral central authority would choose (i.e., \( \mu_j^* = \mu_N^* \)). However, if \( \ln \left( \frac{1}{w_0} \right) < \sqrt{\frac{h_j \lambda_j}{\theta_j c}} \), then the amount of capacity invested would depend on the fraction of capacity cost \( \theta_j \) borne by firm \( j \). In particular,

- if \( \theta_j < \frac{h_j \lambda_j}{\sum_{i \in \mathcal{N}} h_i \lambda_i} \), \( \mu_j^* > \mu_N^* \),
- if \( \theta_j = \frac{h_j \lambda_j}{\sum_{i \in \mathcal{N}} h_i \lambda_i} \), \( \mu_j^* = \mu_N^* \), and
- if \( \theta_j > \frac{h_j \lambda_j}{\sum_{i \in \mathcal{N}} h_i \lambda_i} \), \( \mu_j^* < \mu_N^* \).

The proof is straightforward and is omitted for brevity. As we can see, the lead firm may under- or over-invest in capacity relative to what is socially-optimal. Interestingly, an allocation rule with \( \theta_j = \frac{h_j \lambda_j}{\sum_{i \in \mathcal{N}} h_i \lambda_i} \), always leads a firm to voluntarily choose what is socially optimal so that \( \mu_j^* = \mu_N^* \) (see Balachandran and Radhakrishnan (1996) for a similar result in a different context). In practice, how much of the total capacity cost is incurred by firm \( j \) would depend on the market power of the different firms and on whether or not firms have the option of operating their own independent facilities or joining other firms to form smaller coalitions.

If firms have the option of operating their own facilities, then they would join a facility designed by firm \( j \) only if the fraction of capacity cost for which they are responsible (assuming they must absorb their own delay cost) is sufficiently small. If we let \( \theta_i \) denote the fraction of capacity cost allocated to firm \( i \), then the following condition must be satisfied:

\[
\frac{h_i \lambda_i}{\mu_N^* - \sum_{i \in \mathcal{N}} \lambda_i} + \theta_i c \mu_N^* \leq z_j^* = c \lambda_i + 2 \sqrt{h_i \lambda_i c},
\]

or equivalently if

\[
\theta_i \leq \theta_i^{\max}(j) \equiv \frac{c \lambda_i + 2 \sqrt{h_i \lambda_i c} - \frac{h_i \lambda_i}{\sqrt{\theta_j c}}}{c(\sum_{i \in \mathcal{N}} \lambda_i + \sqrt{\frac{h_i \lambda_i}{\theta_j c}})}.
\] (39)

For the right hand side of the inequality to be non-negative, we must have

\[
\theta_j \leq \alpha_j^{\max}(j) \equiv \frac{h_i \lambda_i}{c} \left( \frac{c \lambda_i + 2 \sqrt{h_i \lambda_i c}}{h_i \lambda_j} \right)^2.
\] (40)

Hence, in order for firm \( i \) to prefer the pooled system, not only should its fraction of capacity cost not exceed \( \theta_i^{\max}(j) \) (this would make its cost in the pooled system the same as the cost it would incur operating an independent facility), but also the fraction of capacity cost of firm \( j \) should not exceed \( \theta_j^{\max}(j) \). The latter requirement follows from the fact that if the fraction paid by firm \( j \) is too large then firm \( j \) would not choose a capacity level that is sufficiently high. Hence, perhaps surprisingly, it is in the best interest of all firms to limit the capacity cost incurred by the lead firm.
We conclude this section by noting that in the extreme scenario where each firm is charged by the lead firm the maximum feasible cost (i.e., the cost associated with each firm operating its own independent facility), the cost function for the lead firm $j$ reduces to

$$z^j_N(\mu^*_N) = \frac{\sum_{i \in N} h_i \lambda_i}{\mu - \sum_{i \in N} \lambda_i} + c\mu_N - \frac{\sum_{i \neq j} (c\lambda_i + 2\sqrt{h_i \lambda_i c})}{\mu - \sum_{i \in N} \lambda_i}. \quad (41)$$

This scenario can be viewed as one where the lead firm absorbs all the costs in the system (total capacity cost + delay costs of all firms) and in exchange receives payments equal to the costs firms would have incurred being on their own. Since these payments are independent of the choice of capacity, the total cost that firm $j$ attempts to minimize is the same as that of a neutral central authority. Consequently, the optimal capacity level is given by:

$$\mu^*_N = \sum_{i \in N} \lambda_i + \sqrt{\frac{\sum_{i \in N} h_i \lambda_i}{c}}, \quad (42)$$

which is equal to the socially-optimal capacity level. The optimal cost for firm $j$ becomes

$$z^*_N = c\lambda_j + 2\sqrt{\sum_{i \in N} h_i \lambda_i c} - \sum_{i \neq j} 2\sqrt{h_i \lambda_i c}. \quad (43)$$

The optimal cost $z^*_N$ can be negative, in which case firm $j$ realizes a profit. In other words, the lead firm may be able to free-ride (i.e., pay no cost for capacity) or even extract a profit by having the other firms pay more for capacity usage than what it costs to acquire it. This is perhaps not surprising, as the prospects of such profit is what may lead firms in practice to take the leadership role in capacity investment.

9 Summary and Concluding Comments

We presented models to study the benefit of capacity pooling among independent firms. We formulated the capacity pooling problem as a cooperative game and showed that for systems where facilities are modeled as M/M/1 queues the core of the game is non-empty. We introduced 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. This allocation rule is perhaps consistent with those observed in practice where combinations of volume based and capacity/service level based fees are common.

We also investigated settings where capacity pooling among all the firms may not be optimal and that some firms may be better off participating in other capacity sharing arrangements or
remaining on their own. We showed that this could be the case when (1) the work contents of customers from different firms are sufficiently heterogenous, (2) facilities consist of multiple servers with the number of servers determining capacity, and (3) there are diseconomies of scale to the cost of capacity. We also explored the impact of assigning priorities for capacity usage among firms and showed that this may lead some firms to experience longer delays in a pooled system than the delays they experience operating their own facilities. However, we showed that these firms can always be compensated for their higher delay costs.

Finally, we studied the effect of having one of the firms (instead of an impartial decision maker) be in charge of making capacity decisions. We showed that this can lead to either under- or over-investments in capacity relative to what is socially optimal. We characterized conditions under which the firm that makes the capacity decision voluntarily chooses the socially optimal capacity level.

This paper is a first step toward a more comprehensive study of cooperation in queueing systems. Therefore, there are numerous possible avenues for future research. It is of course of interest to study systems with alternative assumptions about customer arrival processes, customer processing times, and customer service priorities. It is also of interest to study more complex queueing systems (e.g., systems with parallel servers, servers in series, or servers in a general network configuration) and systems with alternative cost structures. Most such systems have no closed form expressions for performance measures of interest, but perhaps it is still possible to obtain ordinal results leading to non-emptiness of the core in some cases. However, we suspect that as more of our original assumptions are relaxed, pooling may not always be beneficial. For those cases, it is important to characterize conditions under which pooling is (or is not) beneficial and to identify corresponding allocation rules.

In this paper, we assumed that cooperation among independent firms, in the form of capacity sharing, affects only their costs. In practice, cooperation often takes place between competing firms (e.g., cooperation among airlines, hospitals, or original equipment manufacturers in the same industry). In that case, cooperation, by reducing the firms’ costs, could affect the intensity of the competition. In turn, this could affect the desirability of cooperation. Clearly, there is a need for models that capture both competition and cooperation and incorporate the effect of competition in the design of allocation rules.

In this paper, we briefly discussed settings in which one firm is responsible for choosing the capacity level and examined the impact of the fraction of capacity cost incurred by this lead firm on overall capacity investment. In practice, the cost portion paid by the lead firm is likely to be
affected by the bargaining power of that firm. Therefore, it would be useful to construct models that explicitly capture the relative bargaining power of the different firms and the role bargaining power plays in cost allocation. This might also be useful for studying settings where all firms outsource to a common (third party) supplier. The supplier still decides on how much capacity to invest and on the fees to charge each firm based on its volume, service level requirement, and delay penalties.
Figure 1 – The effect of work content variability
$(c = 30, \lambda_1 = \lambda_2 = 20, \omega_1 + \omega_2 = 20)$

Figure 2 – The benefit of partial pooling
$(c = 50, \lambda_1 = \lambda_2 = \lambda_3 = 20, \omega_1 = \omega_2 = 2.5, \omega_3 = 100, \omega_4 = 450)$
Figure 3 – The effect of discrete capacity

\( h_1 = \frac{1000}{\epsilon}, \lambda_1 = \epsilon, h_2 = 10, \lambda_2 = 1, c = 2, \mu = 1 \)
Appendix\textsuperscript{2}

Proof of Theorem 4.1

To prove that $\sum_{i \in N} \mu_i^* = \mu_N^*$, note that

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

In order to prove that $z_N^* \leq z_1^*, \ldots, z_n^*$, we distinguish two cases.

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

$$
z_N^* = \frac{\sum_{i \in N} h_i \lambda_i}{\sqrt{\sum_{i \in N} h_i \lambda_i}} + c \sum_{i \in N} \lambda_i + c \sqrt{\sum_{i \in N} h_i \lambda_i} \frac{\sum_{i \in N} h_i \lambda_i}{c}
\leq \sum_{i \in N} \left( \frac{h_i \lambda_i}{\sqrt{h_i \lambda_i}} + c \lambda_i + c \frac{h_i \lambda_i}{c} \right)
\leq \sum_{i \in N} \left( c(\lambda_i + \eta_i) + \frac{h_i \lambda_i}{\eta_i} \right) = z_1^*, \ldots, z_n^*.
$$

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

$$
z_N^* = \frac{\sum_{i \in N} h_i \lambda_i}{\ln(\frac{1}{1-\alpha_N})} + c \sum_{i \in N} \lambda_i + c \frac{\ln(\frac{1}{1-\alpha_N})}{w_0}
\leq \frac{h_{i_{max}} \lambda_{i_{max}}}{\ln(\frac{1}{1-\alpha_{i_{max}}})} + c \lambda_{i_{max}} + c \frac{\ln(\frac{1}{1-\alpha_{i_{max}}})}{w_0} + \sum_{i \in N, i \neq i_{max}} \left( \frac{h_i \lambda_i}{\sqrt{h_i \lambda_i}} + c \lambda_i + c \frac{h_i \lambda_i}{c} \right)
\leq \sum_{i \in N} \left( c(\lambda_i + \eta_i) + \frac{h_i \lambda_i}{\eta_i} \right) = z_1^*, \ldots, z_n^*.
$$

\[\blacksquare\]

Proof of Theorem 5.1

We distinguish two cases here.

\textsuperscript{2}Should this paper be accepted for publication, this appendix may be placed in an "Online Companion" to the paper.
(1) \( \frac{\ln \left( \frac{1}{1 - \alpha_N} \right)}{w_0} \leq \sqrt{\frac{\sum_{i \in N} h_i \lambda_i}{c}} \): First note that

\[
z_J^* \geq c \sum_{i \in J} \lambda_i + 2 \sqrt{c \sum_{i \in J} h_i \lambda_i}.
\]

Since

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

we have

\[
z_J^* \geq \sum_{i \in J} \phi_i^*, \ \forall J \subseteq N.
\]

It follows that the allocation rule is in the core.

(2) \( \frac{\ln \left( \frac{1}{1 - \alpha_N} \right)}{w_0} > \sqrt{\frac{\sum_{i \in N} h_i \lambda_i}{c}} \).

For coalition \( J \subseteq N \setminus \{i_{\max}\} \), we have

\[
z_J^* \geq c \sum_{i \in J} \lambda_i + 2 \sqrt{c \sum_{i \in J} h_i \lambda_i}.
\]

Since

\[
c \sum_{i \in J} \lambda_i + 2 \sqrt{c \sum_{i \in J} h_i \lambda_i} \\
\geq \sum_{i \in J} \left[ \frac{h_i \lambda_i}{\sqrt{\sum_{i \in N \setminus \{i_{\max}\}} h_i \lambda_i}} + c \lambda_i + c \frac{h_i \lambda_i}{\sum_{i \in N \setminus \{i_{\max}\}} h_i \lambda_i} \sqrt{\frac{\sum_{i \in N \setminus \{i_{\max}\}} h_i \lambda_i}{c}} \right] \geq \sum_{i \in J} \phi_i^*,
\]

we have

\[
z_J^* \geq \sum_{i \in J} \phi_i^*, \ \forall J \subseteq N \setminus \{i_{\max}\}.
\]

If \( i_{\max} \in J \), then

\[
z_J^* = \frac{\sum_{i \in J} h_i \lambda_i}{\ln \left( \frac{1}{1 - \alpha_N} \right) \frac{w_0}{w_0}} + \frac{c \sum_{i \in J} \lambda_i \ln \left( \frac{1}{1 - \alpha_N} \right)}{w_0} \geq \sum_{i \in J} \phi_i^*, \forall J : i_{\max} \in J.
\]

Consequently, the allocation is in the core.
Proof of Theorem 5.2
Consider first the case of pure delay costs. Since
\[ z_N^* = c \sum_{i \in N} \lambda_i + 2 \sqrt{c \sum_{i \in N} h_i \lambda_i} \]
is a jointly concave function of \( \lambda_i, i = 1, \ldots, n \) \( g(x) = f(ax + b) \) is a concave function of \( x \) if \( ax + b \) is an affine mapping and if \( f(\cdot) \) is a concave function), the corresponding game is a concave game by Topkis (1998). For the case with pure service level constraints, let \( S = \max \{\alpha_S, \alpha_T\} \) and \( \alpha_S, \alpha_T \leq \min \{\alpha_S, \alpha_T\} \) \( S, T \subseteq N \), we have \( \ln(\frac{1}{1-\alpha_S}) + \ln(\frac{1}{1-\alpha_T}) \geq \ln(\frac{1}{1-S}) + \ln(\frac{1}{1-T}) \). Consequently, the game is a concave.

Proof of Theorem 6.1
For any coalition \( J \subseteq N \), we have
\[
\sum_{i \in J} \hat{\phi}_i = h \sum_{i \in J} \lambda_i / \mu + h \sum_{i \in J} \lambda_i / \mu P(x^*) / x^* \sqrt{\lambda_N / \mu} + k \sum_{i \in J} \lambda_i / \mu + \sum_{i \in J} \lambda_i / \lambda_N x^* \sqrt{\lambda_i / \mu} \\
\leq h \sum_{i \in J} \lambda_i / \mu + h \sum_{i \in J} \lambda_i / \mu P(x^*) / x^* \sqrt{\sum_{i \in J} \lambda_i / \mu} + k \sum_{i \in J} \lambda_i / \mu + k x^* \sqrt{\sum_{i \in J} \lambda_i / \mu} \\
= \hat{z}_J(m^*_J).
\]
Consequently, the allocation rule is in the core.

Proof of Result in Section 6: \( z_N^* = z_N(m_N^*) \leq \sum_{i \in N} z_i(m_i^*) \).
We define the following unified notation to describe both the distributed and pooled systems. Let \( g(m, \lambda) \) refer to the expected delay in a facility with \( m \) servers and demand rate \( \lambda \). Then, \( g(m_N, \lambda_N) \) corresponds to expected delay in a pooled system with \( m_N \) servers and \( g(m_i, \lambda_i) \) to expected delay of firm \( i \) with \( m_i \) servers in the distributed system. Smith and Whitt (1981) (see also Benjaafar (1996)) show that \( g(\sum_{i=1}^n m_i, \lambda_N) \leq \sum_{i=1}^n g(m_i, \lambda_i) \) for any integer \( m_i > 0 \) and \( \lambda_i \geq 0 \), with \( \lambda_i / m_i \mu < 1 \). This leads to
\[
z_N(m_N^*) = h \lambda_N g(m_N^*, \lambda_N) + km_N^* \leq h \lambda_N g \left( \sum_{i=1}^n m_i^*, \lambda_N \right) + k \sum_{i=1}^n m_i^*
\]
and since
\[
h \lambda_N g \left( \sum_{i=1}^n m_i^*, \lambda_N \right) + k \sum_{i=1}^n m_i^* \leq \sum_{i=1}^n \left[ h \lambda_i g(m_i^*, \lambda_i) + m_i^* \right] = \sum_{i=1}^n z_i(m_i^*),
\]
we have
\[ z_N(m^*_N) \leq \sum_{i=1}^{n} z_i(m^*_i). \]

Proof of Theorem 7.1
Let \( E[W^P_i(\mu)] \) refer to the expected waiting time in the queue for customers of firm \( i \) under the priority policy. Then
\[ z_N^P(\mu) = \sum_{i=1}^{n} h_i \lambda_i E[W^P_i(\mu)] + c \mu. \]
Since
\[ \sum_{i=1}^{n} \lambda_i E[W^P_i(\mu)] = \frac{\lambda_N}{\mu - \lambda_N}, \]
we also have
\[ z_N^P(\mu) = \sum_{i=1}^{n} \frac{h_i \lambda_i}{\mu - \lambda_N} + c \mu = \sum_{i=1}^{n} \left( \frac{\sum_{k=1}^{i} h_k \lambda_k}{\lambda_N} \right) \lambda_i E[W^P_i(\mu)] + c \mu, \]
where \( z_N^P(\mu) \) is expected total cost under the FCFS policy given capacity level \( \mu \). Noting that
\[ \frac{\partial E[W^P_i(\mu)]}{\partial \mu} \geq \cdots \geq \frac{\partial E[W^P_i(\mu)]}{\partial \mu}, \]
we have \( \frac{\partial E[W^P_i(\mu)]}{\partial \mu} \geq \cdots \geq \frac{\partial E[W^P_i(\mu)]}{\partial \mu} \). Since \( h_1 \geq h_2 \cdots \geq h_n \), and \( \sum_{i=1}^{n} \left( \frac{\sum_{k=1}^{i} h_k \lambda_k}{\lambda_N} \right) \lambda_i = \sum_{i=1}^{n} h_i \lambda_i \), it is easy to verify that \( \frac{\partial z^P_N(\mu)}{\partial \mu} \geq \cdots \geq \frac{\partial z^P_N(\mu)}{\partial \mu} \). Using the fact that \( z^P_N(\mu) \) is a convex function of \( \mu \) (a proof is straightforward and is omitted for brevity) leads to \( \mu^*_N \leq \mu' \).

Proof of Theorem 7.3
In order to show that there exists an individual rational capacity cost sharing contract, it suffices to show that the expected delay cost for each firm in the pooled system is no greater than the firm’s total expected cost in the distributed system. To do so, we first consider a system where \( h_i = h_n \) for \( i = 1, \ldots, n - 1 \) and the order of priority of the firms is \( 1, 2, \ldots, n \), the same as in the original one. We refer to this system as system \( B_n \). The total cost for this system \( z^P_N(\mu) \) given capacity level \( \mu \) is
\[ z^P_N(\mu) = \frac{h_n \sum_{i=1}^{n} \lambda_i}{\mu - \sum_{i=1}^{n} \lambda_i} + c \mu, \]
and the corresponding optimal capacity is

$$
\mu^*_N = \sum_{i=1}^{n} \lambda_i + \sqrt{\frac{h_n \sum_{i=1}^{n} \lambda_i}{c}}.
$$

Let $E[L^*_N]$ refer to the expected number of customers of firm $n$ in system $B_n$. Then,

$$
h_n E[L^*_N] = h_n \lambda_n \left( \frac{\sum_{i=1}^{n} \lambda_i}{(\mu^*_N - \sum_{i=1}^{n} \lambda_i)(\mu^*_B - \sum_{i=1}^{n} \lambda_i)} + \frac{1}{\mu^*_N - \sum_{i=1}^{n} \lambda_i} \right)
$$

$$
\leq h_n \lambda_n \left( \frac{\sum_{i=1}^{n} \lambda_i}{\sqrt{h_n \sum_{c \lambda_i}} + \lambda_n} + \frac{1}{\sqrt{h_n \lambda_n}} \right)
$$

$$
= c \lambda_n + \sqrt{h_n \lambda_n c}.
$$

Noting that the optimal expected cost for firm $n$ in the distributed system is $z_n^* = c \lambda_n + 2\sqrt{h_n \lambda_n c}$, we have $h_n E[L^*_N] \leq z_n^*$.

We now consider our original case of $h_1 \geq h_2 \geq \cdots \geq h_n$ and refer to this system as system $A_n$. Let $E[L^A_{i,N}(\mu)]$ denote the expected number of customers of firm $i$ in the system given capacity level $\mu$. Then

$$
\sum_{i=1}^{n} h_i E[L^A_{i,N}(\mu)] = \sum_{i=1}^{n} (h_i - h_n) E[L^A_{i,N}(\mu)] + h_n \sum_{i=1}^{n} E[L^A_{i,N}(\mu)]
$$

$$
= \sum_{i=1}^{n} (h_i - h_n) E[L^A_{i,N}(\mu)] + h_n \sum_{i=1}^{n} E[L^B_{i,N}(\mu)],
$$

where the last equality follows from the fact that the expected number of customers of firm $i$ in the system is the same in systems $A_n$ and $B_n$. Since (1) $h_i - h_n \geq 0$ for all $i$ and (2) $\frac{\partial E[L^A_{i,N}(\mu)]}{\partial \mu} \leq 0$, we have $\mu^*_A \geq \mu^*_B$. Consequently, we also have

$$
h_n E[L^A_{n,N}(\mu^*_A)] = h_n E[L^B_{n,N}(\mu^*_B)] \leq h_n E[L^B_{n,N}(\mu^*_B)] \leq c \lambda_n + \sqrt{h_n \lambda_n c}.
$$

Now suppose we only pool firms $N - 1 = \{1, \cdots, n - 1\}$ with delay costs $h_1, \cdots, h_{n-1}$. We refer to this system as $A_{n-1}$ and to the corresponding optimal capacity as $\mu^*_N$. Consider also a system where the same $n - 1$ firms are pooled but the delay cost is the same for all firms and equal to $h_{n-1}$. We refer to this system as system $B_{n-1}$ and the corresponding optimal capacity as $\mu^*_N$. In both systems the firm priority order is $1, \cdots, n - 1$. It is easy to verify that $\mu^*_N = \mu^*_N$. We now consider our original case of $h_1 \geq h_2 \geq \cdots \geq h_n$ and refer to
\[ \sum_{i=1}^{n-1} \lambda_i + \sqrt{\frac{h_{n-1} \sum_{i=1}^{n-1} \lambda_i}{c}}. \] Using the fact that the fact \( \frac{\partial^2 z_N^P}{\partial N_i \partial \mu} \leq 0 \) for all \( i \), i.e., \( z_N^P \) is submodular in \((\lambda_i, \mu)\) for \( i = 1, \cdots, n \), it follows (see Theorem 6.1 in Topkis (1978)) that \( \mu_{N-1}^{A_n} \leq \mu_{N-1}^{A_{n-1}} \leq \mu_{N}^{A_n} \), which leads to

\[
h_{n-1} E[L_{n-1,N}^{A_n}(\mu_{N}^{A_n})] \leq h_{n-1} E[L_{n-1,N-1}^{A_{n-1}}(\mu_{N-1}^{A_{n-1}})] \leq h_{n-1} E[L_{n-1,N-1}^{B_{n-1}}(\mu_{N-1}^{B_{n-1}})] \leq c\lambda_n + \sqrt{h_n \lambda_n c} \leq z_{n-1}^*
\]

where we have again used the fact that the expected number of customers in the system for each firm is the same for systems \( A_n, A_{n-1} \) and \( B_{n-1} \).

Similarly, by constructing systems \( A_{n-2}, \cdots, A_1 \) and \( B_{n-2}, \cdots, B_1 \) we can show for \( i = 1, \cdots, n-2 \) and \( I = \{1, \cdots, i\} \) that

\[
\mu_{I}^{*A_i} = \sum_{k=1}^{i} \lambda_k + \sqrt{h_i \sum_{k=1}^{i} \lambda_k} \leq \mu_{I}^{*A_i} \leq \mu_{I}^{*A_n},
\]

and

\[
h_i L_{i,N}^{A_n}(\mu_{N}^{A_n}) \leq h_i L_{i,I}^{A_i}(\mu_{I}^{*A_i}) \leq h_i L_{i,I}^{B_i}(\mu_{I}^{*B_i}) \leq c\lambda_i + \sqrt{h_i \lambda_i c} \leq z_i^*.
\]

Hence, there always exists a feasible capacity cost sharing contract under which all firms are no worse off in the pooled system than in the distributed one.
References


M. E. Dyer and L. G. Proll, “On the Validity of Marginal Analysis for Allocating Servers in M/M/c


