On Service Capacity Pooling and Cost Sharing Among Independent Firms

Yimin Yu  •  Saif Benjaafar
Graduate Program in Industrial Engineering
Department of Mechanical 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

January 27, 2006

Abstract

We analyze the benefit of service capacity pooling for a set of independent firms. Firms have the choice of either operating their own service facilities or investing in a facility that is shared. Firms decide on capacity levels to minimize delay and capacity investment costs. If firms decide to operate a shared facility they must also decide on a scheme for sharing the cost of capacity and on priorities for capacity usage. We describe several settings where it is beneficial for firms to pool capacity and there is always a feasible cost sharing contract. We examine the impact of various factors, including the capacity selection process, service priorities, and cost parameters, on the benefit of capacity pooling.

Key words: Capacity pooling, cost allocation, joint ventures, service operations, economics of queues, cooperative game theory

*Presently visiting the Sauder School of Business of the University of British Columbia
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. This was 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.

Capacity pooling among independent firms is an increasingly common practice in the manufacturing, service, and public sectors. For example, airlines often choose to share aircraft maintenance facilities, airport gates and checkout counters as well as code-share. Hospitals share expensive diagnostic equipment, laboratory facilities, and in some cases medical specialists and surgeons. In the automotive industry, auto manufacturers have invested in shared online auctions and procurement systems. Local governments in rural communities often share fire and police departments, 911 call centers, and other social services. Restaurants in some large cities share booking and reservation systems.

Capacity pooling with independent firms raises several important questions. 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 quality of service guaranteed to be better for all firms? How should priorities for capacity usage be set among 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? Also how are the answers affected by whether capacity can be varied continuously or only in discrete amounts?

In this paper, we address these and other related questions in a setting where each firm provides a service to an independent stream of customers that arrive continuously over time with random inter-arrival times. Each firm can install and operate its own service facility where its customers
are processed. We refer to this scenario as the \textit{distributed system}. If a customer arrives and finds the service facility busy, the customer must wait for service. Hence, each firm can be viewed as a queueing system with finite service capacity. 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. 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 \textit{pooled system}. Since delay costs can vary by firm, it may be system-optimal to assign service priorities to the different firms. In order for a firm to voluntarily join the pooled system, its expected delay or its contribution to the cost of capacity, or both, must be sufficiently reduced.

In this paper, we show that in settings where capacity can be varied continuously (in the form of a service rate), all firms benefit under an appropriate cost sharing scheme. We show that such a cost sharing scheme is always feasible. We find that pooling reduces total capacity in the system but does not guarantee that the expected delay experienced by each firm is also reduced. More specifically, in systems where priorities are assigned to different firms based on their delay costs, some firms may see their delay costs increase. This increase can however be compensated by a sufficient reduction in capacity cost. For systems where capacity can be varied only in discrete amounts (in the form of number of servers), pooling may not be necessarily beneficial to all firms. In those cases, firms may be better off on their own or joining a pool consisting of only a subset of the firms. We do however show that when the firms have identical delay costs, pooling remains beneficial and a single large pool optimal.

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 capacity sharing among the firms in a pooled system. In section 6, 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 7, we analyze systems where firms are assigned priorities based on their delay costs. In section 8, we extend our analysis to two cases: (1) systems with general service time distributions and (2) systems where capacity is discrete. 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. Reiman and Mandelbaum (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 adjacent customer streams and each customer can be routed to two adjacent 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 two 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 firm 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. For pooled systems, we discuss the existence of feasible schemes for sharing capacity cost and for setting priorities for capacity usage among the different firms.

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. An extensive body of work, a review of which can be found for example in Alfaro-Tanco and Corbett (2001), 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. Few papers consider inventory pooling among independent firms. Most focus on how costs should be shared among the firms. Examples include Gerchak and Gupta (1991), Robinson (1993), Hartman and Dror (1996, 2005), Muller et al. (2002), Ben-Zvi and Gerchak (2005) and references therein.

Finally, our work is related to the vast literature in economics of coalition formation and joint ventures. A cooperative game theory framework is typically used to analyze the emergence of coalitions, their stability, and the allocation of gains among coalition members. A review of the main concepts and a discussion of applications in supply chain management can be found in Nagarajan and Sošić (2004).

3 The Distributed System

We consider a system consisting of $N$ independent firms. Each firm makes an independent decision about capacity in the form of a service rate. The objective of each firm is to minimize the sum of expected delay and capacity investment costs. Firm $i, i = 1, \ldots, N$, 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 i.i.d. and exponentially distributed with mean $1/\mu_i$ for firm $i$, where $\mu_i$ denotes the firm’s service rate. Hence, each firm behaves like an M/M/1 queue.
Firm $i$ incurs a delay cost $h_i$ per unit time one of its customers spends in the system (either in the queue or in service), and an amortized linear capacity cost $c$ per unit of service rate per unit time. Let $z_{i,d}(\mu_i)$ denote the expected total cost incurred by firm $i$ given a service rate $\mu_i$ (for stability, we assume that $\lambda_i < \mu_i$). Then,

$$z_{i,d}(\mu_i) = \frac{h_i \lambda_i}{\mu_i - \lambda_i} + c \mu_i. \quad (1)$$

Noting that $z_{i,d}$ is convex in $\mu_i$, the optimal capacity level, $\mu_{i,d}^*$, can be obtained from the first order condition of optimality as

$$\mu_{i,d}^* = \lambda_i + \sqrt{\frac{h_i \lambda_i}{c}}. \quad (2)$$

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 the square root of the ratio $\frac{h_i \lambda_i}{c}$.

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

$$z_{i,d}^* = c \lambda_i + 2 \sqrt{h_i \lambda_i c}. \quad (3)$$

This leads to a total cost in the system, $z_d^*$, given by

$$z_d^* = c \sum_{i=1}^{N} \lambda_i + 2 \sum_{i=1}^{N} \sqrt{h_i \lambda_i c}. \quad (4)$$

In the case of identical firms, with $\lambda_i = \lambda$ and $h_i = h$ for $i = 1, \ldots, N$, the cost reduces to

$$z_d^* = c N \lambda + 2 N \sqrt{h \lambda c}, \quad (5)$$

and the total capacity in the system to

$$\sum_{i=1}^{N} \mu_{i,d}^* = N \left( \lambda + \sqrt{\frac{h \lambda}{c}} \right). \quad (6)$$

As we can see, both the optimal cost and the optimal buffer capacity in the system increase linearly in the number of firms $N$.

### 4 The Pooled System

In the pooled system, the $N$ firms share a single facility with capacity $\mu$ from which the demand of all firms is satisfied. The demand process at the shared facility is Poisson with rate $\sum_{i=1}^{N} \lambda_i$ since the superposition of independent Poisson processes is also a Poisson process. We assume that
customers regardless of their firm affiliation are served in a FCFS fashion (we consider alternative priority schemes in section 6). A central decision maker 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. For a given choice of capacity level $\mu_p$, the cost of the pooled system, $z_p(\mu_p)$, is given by

$$z_p(\mu_p) = \sum_{i=1}^{N} \frac{h_i \lambda_i}{\mu_p - \sum_{i=1}^{N} \lambda_i} + c\mu_p,$$

(7)

where we assume that $\sum_{i=1}^{N} \lambda_i < \mu_p$ for stability. Since $z_p$ is convex in $\mu_p$, the optimal capacity $\mu_p^*$ can be obtained from the first order optimality condition as

$$\mu_p^* = \sum_{i=1}^{N} \lambda_i + \sqrt{\sum_{i=1}^{N} \frac{h_i \lambda_i}{c}},$$

(8)

with the corresponding optimal expected cost given by

$$z_p^* = c \sum_{i=1}^{N} \lambda_i + 2 \sqrt{c \sum_{i=1}^{N} h_i \lambda_i}.$$  

(9)

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 the square root of the sum of the ratios $\frac{h_i \lambda_i}{c}$.

Theorem 4.1 compares the distributed and pooled systems (the proof is straightforward and is omitted for the sake of brevity).

**Theorem 4.1** $z_p^* \leq z_d^*$ and $\mu_p^* \leq \sum_{i=1}^{N} \mu_{i,d}^*$.

Consistent with intuition, 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. In that case, we have

$$\mu_p^* = N\lambda + \sqrt{N h\lambda c}$$  

and  

$$z_p^* = cN\lambda + 2\sqrt{cNh\lambda},$$

from which we can observe that both buffer capacity and the second component of total cost are reduced by a factor of a square root of $N$ (relative to those observed in the distributed system). If we consider the relative reduction in total cost due to pooling, defined by $\delta_z = \frac{z_d^*}{z_p^*}$, then

$$\delta_z = \frac{1 - \frac{1}{\sqrt{N}}}{\frac{\sqrt{h\lambda} + 1}{\sqrt{c}}},$$

(10)

from which we can verify the following two observations:
(1) $\delta_z$ is increasing concave in $N$ and $h$ with lim$_{N \to \infty} \delta_z = \frac{1}{\sqrt{\lambda c/4h+1}}$ and lim$_{h \to \infty} \delta_z = 1 - \frac{1}{\sqrt{N}}$, and

(2) $\delta_z$ is decreasing convex in $\lambda$ and $c$ with lim$_{\lambda \to \infty} \delta_z = \text{lim}_{c \to \infty} \delta_z = 0$.

The fact that the relative benefit from pooling increases with $N$ and $h$ is consistent with intuition. One expects pooling to be relatively more beneficial when delay costs are high or there is a large number of firms participating in the pooled system. Similarly, the fact that the relative benefit of pooling is smaller when $\lambda$ is high is not surprising. Here also, one does expect firms that already have high demand to derive less value from sharing facilities with others. However, observing that higher capacity cost reduces the relative benefit of pooling is perhaps unexpected. This result can be explained by the fact that when capacity cost is high, the optimal capacity level is already low in the distributed system. Therefore, the further reduction in capacity achieved with pooling tends to be modest.

In addition to reducing the total expected cost, pooling reduces the expected delay experienced by each firm’s customers.

**Theorem 4.2** $E[W^*_{i,p}] \leq E[W^*_{i,d}]$ for $i = 1, \ldots, N$, where $E[W^*_{i,p}]$ and $E[W^*_{i,d}]$ are the expected delays for firm $i$ in the pooled and distributed systems respectively under the choice of optimal capacity in each case.

The above result follows by noting that

$$E[W^*_{i,p}] = \frac{1}{\mu^*_p} - \sum_{i=1}^{N} \lambda_i = \sqrt{\frac{c}{\sum_{i=1}^{N} h_i \lambda_i}}$$  \hspace{1cm} (11)$$

and

$$E[W^*_{i,d}] = \frac{1}{\mu^*_i d - \lambda_i} = \sqrt{\frac{c}{h_i \lambda_i}}.$$  \hspace{1cm} (12)

from which it is easy to verify that $E[W^*_{i,p}] \leq E[W^*_{i,d}]$ for $i = 1, \ldots, N$. It is also easy to verify that in a system with identical firms, pooling reduces expected delay by a factor of $\sqrt{N}$. The fact that the expected delay of all firms is reduced is not surprising given the FCFS priority by which customers are processed. In section 6, we show that this ceases to hold when different priority schemes are applied.

Note that the above results are somewhat different from those obtained for other forms of capacity pooling. For example, in the context of inventory pooling, when the delay costs are unequal, a pooled inventory system with FCFS policy can result in higher system cost (Ben-Zvi and Gerchak, 2005).
5 Cost Sharing

Since the firms are independent, they must find a mutually acceptable scheme to share the cost of capacity in the pooled system. Let $\alpha_i$ denote the fraction of capacity cost that firm $i$ pays. Then, expected total cost for firm $i$ is

$$z_{i,p}^*(\alpha_i) = \frac{h_i \lambda_i}{\mu_p - \sum_{i=1}^N \lambda_i} + \alpha_i c \mu_p^*.$$  \hspace{1cm} (13)

A firm is willing to join the pooled system only if it is better off than being on its own. Therefore, the fraction $\alpha_i$, for $i = 1, \ldots, N$, must satisfy the following inequality

$$\frac{h_i \lambda_i}{\mu_p - \sum_{i=1}^N \lambda_i} + \alpha_i c \mu_p^* \leq z_{i,d}^* = c\lambda_i + 2\sqrt{h_i \lambda_i c},$$

or equivalently

$$\alpha_i \leq \alpha_i^{max} \equiv \frac{\lambda_i \sqrt{c \sum_{k=1}^N h_k \lambda_k + 2\sqrt{h_i \lambda_i \sum_{k=1}^N h_k \lambda_k - h_i \lambda_i}}}{\sum_{k=1}^N \lambda_k \sqrt{c \sum_{k=1}^N h_k \lambda_k + \sum_{k=1}^N h_k \lambda_k}},$$ \hspace{1cm} (14)

where $\alpha_i^{max}$ corresponds to the maximum fraction firm $i$ would be willing to pay. The following theorem shows that it is always possible to find a feasible cost-sharing scheme under which all $N$ firms have a preference for the pooled system.

**Theorem 5.1** There exists a feasible cost-sharing contract consisting of a vector $\alpha = (\alpha_1^c, \ldots, \alpha_N^c)$, with $\alpha_i^c \geq 0$ for $i = 1, \ldots, N$ such that $z_{i,p}^*(\alpha_i^c) \leq z_{i,d}^*$ and $\sum_{i=1}^N \alpha_i^c = 1$.

The result follows from the fact that (1) $z_{i,p}^*(\alpha_i^c) \leq z_{i,d}^*$ for $0 \leq \alpha_i^c \leq \alpha_i^{max}$ and (2)

$$\sum_{i=1}^N \alpha_i^{max} = \sum_{i=1}^N \left( \frac{\lambda_i \sqrt{c \sum_{k=1}^N h_k \lambda_k + 2\sqrt{h_i \lambda_i \sum_{k=1}^N h_k \lambda_k - h_i \lambda_i}}}{\sum_{k=1}^N \lambda_k \sqrt{c \sum_{k=1}^N h_k \lambda_k + \sum_{k=1}^N h_k \lambda_k}} \right) \geq 1,$$

where the inequality holds because $\sum_{i=1}^N \sqrt{h_i \lambda_i} \geq \sqrt{\sum_{i=1}^N h_i \lambda_i}$. The difference

$$\Delta = \sum_{i=1}^N \alpha_i^{max} c \mu_p^* - c \mu_p^* = \frac{2\sqrt{\sum_{i=1}^N h_i \lambda_i \left( \sum_{i=1}^N \sqrt{h_i \lambda_i} - \sqrt{\sum_{i=1}^N h_i \lambda_i} \right)}}{\sum_{i=1}^N \lambda_i \sqrt{c \sum_{i=1}^N h_i \lambda_i + \sum_{i=1}^N h_i \lambda_i}}$$

can be viewed as surplus that can be used by the central decision maker to reduce the share of capacity cost that one or more of the firms pays. In a system where the central decision maker is a third party service provider, the surplus would represent profit that the service provider extracts by charging each firm the maximum feasible price $\alpha_i^{max} c \mu_p^*$ for capacity, $i = 1, \ldots, N$.

Finally we note that since delay costs are lower in the pooled system, each firm is willing to spend more on capacity than they are in the distributed system. In other words, $\alpha_i^{max} c \mu_p^* \geq c \mu_{i,d}^*$. 

8
6 Social versus Individual Optimization

In this section, we consider the case where, in the pooled system, one of the firms chooses selfishly the capacity level instead of an impartial decision maker choosing the socially-optimal level. This arises in practice if, for example, there is a dominant firm that takes on the lead role in installing, managing and pricing the shared facility. We assume in this section that this lead firm sets the capacity level so that its own cost is minimized, subject to the constraint that other firms would find the shared facility still preferable to individual ones in a distributed system.

Let $j$ denote the index of the firm that makes the capacity decision and $\alpha_j$ its fraction of the capacity cost. Given a choice of capacity level $\mu_{j,p}$, the expected cost for firm $j$ is given by

$$z_{j,p}(j, \mu_{j,p}) = \frac{h_j \lambda_j}{\mu_{j,p}} - \sum_{i=1}^{N} \lambda_i + \alpha_j c \mu_{j,p}.$$  \hspace{1cm} (15)

Noting that the above cost function is convex in $\mu_{j,p}$, the optimal capacity level can be obtained from the first order condition of optimality as:

$$\mu_{j,p}^* = \sum_{i=1}^{N} \lambda_i + \sqrt{\frac{h_j \lambda_j}{\alpha_j c}}.$$  \hspace{1cm} (16)

This leads to an optimal cost for firm $j$ given by

$$z_{j,p}^*(j) = \alpha_j c \sum_{i=1}^{N} \lambda_i + 2 \sqrt{h_j \lambda_j \alpha_j c}.$$  \hspace{1cm} (17)

The fraction of capacity cost borne by firm $j$ clearly determines its choice of capacity. In the following theorem, we characterize conditions under which the firm under- or over-invests in capacity relative to the socially-optimal level (the proof is straightforward and therefore we omit it for brevity).

**Theorem 6.1** $\mu_{j,p}^* \geq \mu_p^*$ if and only if $\alpha_j \leq \frac{h_j \lambda_j}{\sum_{i=1}^{N} h_i \lambda_i}$, where if one holds with equality so does the other. Furthermore, let $\mu_{d}^* = \sum_{i=1}^{N} \mu_{i,d}^*$, then $\mu_{j,p}^* \geq \mu_{d}^*$ if and only if $\alpha_j \leq \frac{h_j \lambda_j}{(\sum_{i=1}^{N} h_i \lambda_i)^2}$, where again if one holds with equality so does the other.

The above results are consistent with a result in Balachandran and Radhakrisnan (1996) who also show in a different context that a cost allocation of the form $\alpha_j = \frac{h_j \lambda_j}{\sum_{i=1}^{N} h_i \lambda_i}$ induces multiple users of a shared service facility to choose voluntarily the socially-optimally capacity level; see Hassin and Haviv (Chapter 8, 2003) for related discussion.

A firm $i \neq j$ would be willing to join the pooled system, as designed by firm $j$, if

$$\frac{h_i \lambda_i}{\mu_{j,p}^*} - \sum_{i=1}^{N} \lambda_i + \alpha_i c \mu_{j,p}^* \leq c \lambda_i + 2 \sqrt{h_i \lambda_i c},$$  \hspace{1cm} (9)
or equivalently if

$$\alpha_i \leq \alpha_{i}^{\max}(j) \equiv \frac{c\lambda_i + 2\sqrt{h_i\lambda_i c} - \frac{h_i \lambda_i}{\alpha_j c}}{c\left(\sum_{k=1}^{N} \lambda_k + \sqrt{\frac{h_i \lambda_i c}{\alpha_j c}}\right)}.$$  \hspace{1cm} (18)

In order for the right hand side of the inequality to be non-negative, we must have

$$\alpha_j \leq \alpha_{j}^{\max}(j) \equiv \frac{h_i \lambda_i}{c} \left(\frac{c\lambda_i + 2\sqrt{h_i\lambda_i c}}{h_i \lambda_j c}\right)^2.$$  \hspace{1cm} (19)

Hence, in order for firm $i$ to prefer the pooled system, not only should its fraction of capacity cost not exceed $\alpha_{i}^{\max}(j)$, but also the fraction of capacity cost of firm $j$ should not exceed $\alpha_{j}^{\max}(j)$. The latter requirement follows from the fact that if the fraction paid by firm $j$ is too large then supplier $j$ would not choose a capacity level that is sufficiently high.

If firm $j$ manages to convince every firm $i, i \neq j$ to pay the maximum feasible fraction $\alpha_{i}^{\max}(j)$, its cost function reduces to

$$z_{j,p}(j, \mu_{j,p}) = \frac{\sum_{k=1}^{N} h_k \lambda_k}{\mu} - \sum_{i \neq j}^{N} (c\lambda_i + 2\sqrt{h_i\lambda_i c}).$$  \hspace{1cm} (20)

This leads to an optimal capacity level given by:

$$\mu_{j,p}^{*} = \sum_{k=1}^{N} \lambda_k + \sqrt{\frac{\sum_{k=1}^{N} h_k \lambda_k c}{c}},$$  \hspace{1cm} (21)

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

$$z_{j,p}^{*}(j) = c\lambda_j + 2\sqrt{\sum_{k=1}^{N} h_k \lambda_k c} - \sum_{i \neq j}^{N} 2\sqrt{h_i \lambda_i c}.$$  \hspace{1cm} (22)

Interestingly, $z_{j,p}^{*}(j)$ 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) and, more significantly, extract a profit by having the other firms pay more for capacity than what is justified to acquire it.

### 7 The Pooled System with Service Priorities

We have so far assumed that customers in the pooled system, 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 optimal since it does not minimize expected delay in the system. A policy that does is one where customers with higher delay costs are given higher priority. In this section, we analyze a pooled system under a priority scheme. First, we consider the case where preemption is not allowed, so that the arrival of a higher priority customer does not interrupt the service of a lower priority one. Then, we treat the case with preemption.
7.1 Systems with Nonpreemptive Priorities

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, Cox and Smith (1961), Jaiswal (1968) and Klimov (1974)). Under the \( c\mu \) rule, customers are assigned priorities based on the product of their respective delay costs and service rates. In our setting, this means that higher service priority is given to customers with higher delay costs. More specifically, if we assume (without loss of generality) that \( h_1 \geq h_2 \geq \cdots \geq h_N \), then each time the service facility becomes available, a customer from a firm with the lowest index (highest delay cost), among those waiting, is selected for service; a FCFS policy is used to choose among customers from the same firm. We assume that preemption is not allowed so that the service of a customer, regardless of its class, cannot be interrupted. Under a non-preemptive priority (NPP) scheme of this type, the expected delay

\[
E[W_{i,p}^{NPP}(\mu_p)] = \sum_{k=1}^{N} \frac{\lambda_k}{S_i S_{i-1}} + \frac{1}{\mu_p},
\]

which leads to the following total expected cost

\[
z_p^{NPP}(\mu_p) = \sum_{i=1}^{N} h_i \lambda_i \left( \frac{\sum_{k=1}^{N} \lambda_k}{S_i S_{i-1}} + \frac{1}{\mu_p} \right) + c\mu_p,
\]

where \( S_i = 1 - \sum_{k=1}^{i} \frac{\lambda_k}{\mu_p}, i \geq 1 \) 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 \( z^{NPP}_p \) and the optimal capacity \( \mu^{NPP}_p \) under the NPP policy are lower than those under the FCFS policy.

**Theorem 7.1** \( z^{NPP}_p \leq z^*_p \) and \( \mu^{NPP}_p \leq \mu^*_p \), where \( z^*_p \) and \( \mu^*_p \) refer respectively to the optimal cost and the optimal capacity under the FCFS policy.

Showing that \( z^{NPP}_p \leq z^*_p \) is trivial since the expected delay cost under NPP is always lower than under FCFS for any choice of \( \mu_p \). In order to show that \( \mu^{NPP}_p \leq \mu^*_p \), it is sufficient to show that (1) \( z^{NPP}_p \) is convex in \( \mu_p \) and (2) \( \frac{\partial z^{NPP}(\mu_p)}{\partial \mu_p} \big|_{\mu_p=\mu^*_p} \geq 0 \). A proof that both conditions hold is provided in the Appendix.

Although pooling under NPP is more cost efficient for the system and requires less capacity, it has the appearance of unfairness to firms with lower delay costs and who consequently are assigned lower service priorities. Indeed, as we note in the following observation, it is possible for the expected delay of a lower priority firm to increase when it joins the pooled system.
Observation 7.2 It is possible to have \( E[W_{i,p}^{*}] > E[W_{i,d}^{*}] \), where \( E[W_{i,p}^{*}] \) and \( E[W_{i,d}^{*}] \) refer respectively to the expected delay of firm \( i \) in the pooled (under NPP) and distributed systems, in each case under the corresponding optimal capacity.

The observation can be proven by counter-example as shown in Figure 1. Although pooling may increase the expected delay for certain firms, we show in the following theorem that it is always possible to find a capacity cost-sharing contract under which all firms prefer the pooled system. The proof can be found in the Appendix.

Theorem 7.3 There exists a feasible cost sharing contract consisting of a vector \( \alpha_{NPP} = (\alpha_{1}^{NPP}, \ldots, \alpha_{N}^{NPP}) \), with \( \alpha_{i}^{NPP} \geq 0 \), such that \( z_{i,p}^{*}(\alpha_{i}^{NPP}) \leq z_{i,d}^{*} \) and \( \sum_{i=1}^{N} \alpha_{i}^{NPP} = 1 \), where \( z_{i,p}^{*}(\alpha_{i}^{NPP}) \) is the expected cost of firm \( i \) in the pooled system when it pays a fraction \( \alpha_{i}^{NPP} \) of capacity cost.

The fact that there is a cost sharing scheme under which all firms prefer the pooled system means that firms whose delay costs increase with pooling pay sufficiently less for capacity than in the distributed system. The increase in delay cost is therefore compensated by sufficient reduction in capacity cost.

7.2 Systems with Preemptive/Resume Priorities

If practically feasible, preempting (interrupting) the service of customers with lower delay costs to allow customers with higher delay costs to receive immediate service is desirable since in our setting it reduces expected delay in the system (due to the fact that service times are exponentially distributed and therefore memoryless). In this section, we consider a preemptive resume policy, under which lower priority customers are preempted when a customer with a higher priority arrives. The service on a preempted customer resumes once no customer with a higher priority is present in the system. Under these conditions, it is still optimal to assign priorities based on delay costs with customers with higher delay cost assigned higher priority. Without loss of generality, we again assume that \( h_{1} \geq h_{2} \geq \ldots \geq h_{N} \). Then, under a preemptive/resume priority (PRP) policy, the expected cost \( z_{p}^{PRP}(\mu_{p}) \) for a given choice of capacity \( \mu_{p} \) is given by

\[
z_{p}^{PRP}(\mu_{p}) = \sum_{i=1}^{N} h_{i} \lambda_{i} \left( \frac{R_{i}}{S_{i}S_{i-1}} + \frac{1}{\mu_{p}S_{i-1}} \right) + c\mu_{p},
\]

(25)

where \( R_{i} = \sum_{k=1}^{i} \frac{\lambda_{k}}{\mu_{p}^{k}} \), and \( S_{i} = 1 - \sum_{k=1}^{i} \frac{\lambda_{k}}{\mu_{p}^{k}} \), with \( S_{0} = 1 \).

The following theorem, a proof of which is included in the Appendix, states that both the optimal cost and the optimal capacity are lower under PRP than under NPP.
Theorem 7.4 $z^*_{p,PRP} \leq z^*_{p,NPP}$ and $\mu^*_{p,PRP} \leq \mu^*_{p,NPP}$.

Although PRP is system-optimal, it can lead to even higher expected delay costs than NPP for customers with lower priority. However, as we show in the following theorem (see Appendix for proof), it is still possible to find a feasible cost-sharing contract under which all firms are better off than on their own.

Theorem 7.5 There exists a feasible cost sharing contract consisting of a vector $\alpha^{PRP} = (\alpha_1^{PRP}, \cdots, \alpha_N^{PRP})$, with $\alpha_i^{PRP} \geq 0$, such that $z^*_{i,p}(\alpha_i^{PRP}) \leq z^*_{i,d}$ and $\sum_{i=1}^{N} \alpha_i^{PRP} = 1$, where $z^*_{i,p}(\alpha_i^{PRP})$ is the expected cost of firm $i$ in the pooled system when it pays a fraction $\alpha_i^{PRP}$ of capacity cost.

Since the incentive for low priority customers to participate needs to be larger with PRP than with NPP, their cost allocation needs to be lower still. Put differently, the vector of cost allocations (arranged in order of increasing priority) with PRP will be majorized by the corresponding vector with NPP (Marshall and Olkin, 1979).

Figures 2 and 3 illustrate the impact of using different priority policies on the benefit of pooling relative to the distributed system. As we can see, differences between the three policies, (relative to their differences to the distributed system) either in terms of capacity or cost, do not appear to be significant. This seems to suggest that most of the reduction in cost (relative to the distributed system) is due to capacity pooling, with operational control playing a relatively secondary role. For example, using priority policies further reduces cost relative to FCFS by an amount less than 4 percent in the cases shown.

8 Extensions

In this section, we extend our analysis to two important cases: (1) systems with general service time distributions and (2) systems where capacity is discrete.

8.1 Systems with General Service Time Distributions

We consider a more general version of the single server model discussed in sections 3-5. In particular, we relax the assumption of exponentially distributed service times. Instead we allow service times to be independent and identically distributed random variables denoted by $X$ where $X$ is of the form $Y/\mu$ and $Y$ is a random variable with a general distribution with unit mean and finite variance $\sigma^2$. Hence, expected service time is $E[X] = 1/\mu$, variance of service time is $Var[X] = \sigma^2/\mu^2$, and the coefficient of variation of service is $\sqrt{Var[X]/E[X]} = \sigma$. The parameter $\mu$, ($\mu > 0$) is a scaling.
parameter that corresponds to the service rate or capacity. We retain all other assumptions, except that we restrict ourselves to the case where delay costs are identical for all firms, i.e., \( h_i = h \) for \( i = 1, ..., N \). Given these assumptions, facilities in both the pooled and distributed systems behave like M/G/1 queue systems. A special case is the M/M/1 queue discussed in sections 3-5. Therefore, the approach below provides an alternative treatment to that case as well.

We use a unified notation to describe both the pooled and distributed systems. In particular, we let \( f(\lambda, \mu) \equiv h \lambda E[W(\lambda, \mu)] + c \mu \), where \( E[W(\lambda, \mu)] \) is the expected delay in a M/G/1 queue with arrival rate \( \lambda \) and service rate \( \mu \) (under the assumptions described above). Then, the cost of firm \( i \) in the distributed system, for a choice of service rate \( \mu_{i,d} \), corresponds to \( f(\lambda_i, \mu_{i,d}) \). Similarly, the total cost in the pooled system for a choice of service rate \( \mu_p \) corresponds to \( f \left( \sum_{i=1}^{N} \lambda_i, \mu_p \right) \).

Let \( \mu(\lambda) = \arg \min_{\mu > \lambda} f(\lambda, \mu) \) denote the optimal capacity level for a facility with a demand rate \( \lambda \). Also, let \( \rho(\lambda) = \lambda / \mu(\lambda) \) denote the corresponding average utilization of the facility. The function \( f(\lambda, \mu) \) is convex in \( \mu \) since \( E[W(\lambda, \mu)] \) is convex in \( \mu \) (Webber, 1983). Consequently, \( \mu(\lambda) \) is the unique solution to the first order optimality condition \( \frac{\partial f(\lambda, \mu)}{\partial \lambda} = 0 \).

The following result is due to Stidham (1992) which we restate using our notation.

**Lemma 8.1** \( f(\lambda, \mu(\lambda)) \) is concave in \( \lambda \) and \( \rho(\lambda) \) is non-decreasing in \( \lambda \).

We are now ready to state our main result.

**Theorem 8.2** \( z_p^* \leq z_d^* \) and \( \mu_p^* \leq \mu_d^* \) where \( z_p^* \) and \( z_d^* \) refer to the optimal total cost for the pooled and distributed systems (described in this section) respectively and \( \mu_p^* \) and \( \mu_d^* \) to the corresponding optimal capacity levels.

The result regarding the optimal cost follows from the fact that \( f(\lambda, \mu(\lambda)) \) is concave. Therefore, \( f \left( \sum_{i=1}^{N} \lambda_i, \mu(\sum_{i=1}^{N} \lambda_i) \right) \leq \sum_{i=1}^{N} f \left( \lambda_i, \mu(\lambda_i) \right) \). The result regarding optimal capacity follows from the fact that

\[
\sum_{i=1}^{N} \mu(\lambda_i) = \sum_{i=1}^{N} \frac{1}{\rho(\lambda_i)} \lambda_i \geq \sum_{i=1}^{N} \frac{1}{\rho \left( \sum_{i=1}^{N} \lambda_i \right)} \lambda_i = \frac{1}{\rho \left( \sum_{i=1}^{N} \lambda_i \right)} \sum_{i=1}^{N} \lambda_i = \mu \left( \sum_{i=1}^{N} \lambda_i \right),
\]

where the inequality is due to \( \rho(\lambda) \) being nondecreasing in \( \lambda \).

**Theorem 8.3** There is always a feasible cost sharing contract under which all firms prefer the pooled system.

The following simple contract can be shown to be always feasible. Let \( x_i = f \left( \sum_{j=1}^{i} \lambda_j, \mu(\sum_{j=1}^{i} \lambda_j) \right) - f \left( \sum_{j=1}^{i-1} \lambda_j, \mu(\sum_{j=1}^{i-1} \lambda_j) \right) \) correspond to the share of total cost assigned to firm \( i, i = 1, ..., N, \) where
\[ \sum_{j=1}^{0} \lambda_j = 0. \] The above allocation guarantees that \[ \sum_{i=1}^{N} x_i = z^*_i \] and \[ x_i \leq z^*_{i,d}, \] where the inequality follows from the concavity of \( f(\lambda, \mu(\lambda)) \). This cost sharing contract implies that firms \( i \) pays a fraction \( \alpha_i = \frac{x_i - h_i E[W^*_p]}{\sigma^2_{p_i}} \) if \( x_i \geq h_i E[W^*_p] \). Otherwise, \( \alpha_i = 0 \) (firm \( i \) pays no capacity cost) and instead receives a compensation in the amount of \( h_i E[W^*_p] - x_i \) to offset its higher delay cost. However since \( h_i E[W^*_p] \leq z_{i,d} \) (see Appendix for proof), we have \( h_i E[W^*_p] - x_i \leq 0 \) and, consequently, \( \alpha_i \geq 0 \).

The above cost sharing scheme is of course not unique. In fact, there are at least \( N! \) such schemes that can be generated by simply relabeling the different firms. Note also that a cost sharing scheme in which one firm does not pay any fraction of total cost is unlikely to occur in practice. Considerable research on desirable properties of cost sharing contracts and how to generate them can be found in the cooperative game theory literature; see for example (Nagarajan and Sošić, 2004) for a discussion and references.

Although we have restricted ourselves in this section to systems where the firms have identical delay costs, we expect the results to continue to hold when the costs are non-identical. This conjecture is supported by numerical results and the results for the M/M/1 case; but an analytical proof appears difficult and we do not attempt it here. Also, although we have limited our discussion to settings where the demand process is Poisson, our results have the potential of being useful for other demand processes as well. The only requirement for the results to be true is for Lemma 8.1 to hold.

We conclude this section by examining the effect of service time variability on the benefit of pooling. In Figure 4, we show the impact of increasing \( \sigma \), the coefficient of variation of service time. As we can see, the relative benefit of pooling is increasing (at a decreasing rate) in service time variability. Pooling reduces the variance in service time from \( \sigma^2 / (\mu^*_{i,d})^2 \) to \( \sigma^2 / (\mu^*_p)^2 \) for firm \( i = 1, ..., N \). Hence, it is perhaps not surprising that pooling is more beneficial when \( \sigma \) is large.

### 8.2 Pooling with Discrete Capacity

In this section we consider systems where capacity can be acquired only in discrete amounts. For example, capacity could be determined by the number of human operators, computer servers, or telecommunication lines. In these settings, a more appropriate modeling framework is one where a server’s capacity is fixed but the number of servers is variable. This means that facilities are modeled as multi-server queues with a fixed service rate per server. Consistent with our treatment in our basic model, we assume that firm \( i, i = 1, ..., N \), 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.

First we consider the distributed system. Let \( m_{i,d} \) denote the number of servers in firm \( i \)'s facility. Then, each facility can be modeled as an \( M/M/m_{i,d} \) queue and the expected cost of firm \( i \) is given by:

\[
\begin{align*}
  z_{i,d}(m_{i,d}) &= h_i \lambda_i \left( \frac{1}{\mu} + \frac{P(m_{i,d}, \lambda_i)}{m_{i,d} \mu - \lambda_i} \right) + k m_{i,d}, \\
  \end{align*}
\]

where

\[
\begin{align*}
  P(m_{i,d}, \lambda_i) &= \frac{(m_{i,d} \rho_i)^{m_{i,d}}}{m_{i,d}! (1 - \rho_i)} + \frac{(m_{i,d} \rho_i)^{m_{i,d}}}{m_{i,d}! (1 - \rho_i)}, \\
  \end{align*}
\]

and \( \rho_i = \frac{\lambda_i}{m_{i,d} \mu} \); for stability, we assume \( \rho_i < 1 \).

Using the fact that \( \frac{P(m_{i,d}, \lambda_i)}{m_{i,d} \mu - \lambda_i} \) is convex in \( m_{i,d} \) (see for example Dyer and Proll (1977)), we can show that \( z_{i,d}(m_{i,d}) \) is also convex. Therefore, the optimal capacity (i.e., optimal number of servers) is given by the smallest integer \( m_{i,d} \) that satisfies the inequality:

\[
\begin{align*}
  z_{i,d}(m_{i,d} + 1) - z_{i,d}(m_{i,d}) > 0. \\
  \end{align*}
\]

We denote the optimal capacity level by \( m^*_{i,d} \). Although it is difficult to obtain a closed form expression for \( m^*_{i,d} \) and the corresponding optimal cost \( z_{i,d}(m^*_{i,d}) \), both are easy to compute numerically.

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

\[
\begin{align*}
  z_p(m_p) &= \sum_{i=1}^{N} h_i \lambda_i \left( \frac{1}{\mu} + \frac{P(m_p, \sum_{i=1}^{N} \lambda_i)}{m_p \mu - \sum_{i=1}^{N} \lambda_i} \right) + k m_p, \\
  \end{align*}
\]

where

\[
\begin{align*}
  P(m_p, \sum_{i=1}^{N} \lambda_i) &= \frac{(m_p \rho_p)^{m_p}}{m_p! (1 - \rho_p)} + \frac{(m_p \rho_p)^{m_p}}{m_p! (1 - \rho_p)}, \\
  \end{align*}
\]

and \( \rho_p = \frac{\sum_{i=1}^{N} \lambda_i}{m_p \mu} \). The optimal capacity \( m^*_p \) and the optimal cost \( z^*_p \) can be obtained using a similar approach as in the case of the distributed system.

**Observation 8.4** In a system where capacity is discrete, pooling may not always be socially-optimal.

The above observation can be illustrated using the following example. Consider a system with two firms, 1 and 2, with the following characteristics: \( \mu = 1 \), \( \lambda_1 = \epsilon \), \( \epsilon > 0 \), \( h_1 = \frac{1}{\epsilon} \), \( \lambda_2 = 1 \) and \( h_1 = 1 \). In Figure 5, we show 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. This effect appears to be due to the fact that capacity can be varied only in discrete amounts.
Although pooling may not be preferable in general, it is so when the firms have identical delay costs, i.e., \( h_i = h \) for \( i = 1, ..., N \).

**Theorem 8.5** In a system where the firms have identical delay costs, \( z_p^* = z_p(m_p^*) \leq z_d^* = \sum_{i=1}^{N} z_{i,d}(m_{i,d}^*) \). Furthermore, it is always possible to find a cost sharing contract under which all firms are better off in the pooled system.

**Proof:** 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, \( g(m, \sum_{i=1}^{N} \lambda_i) \) corresponds to expected delay in a pooled system with \( m_p \) servers and \( g(m_{i,d}, \lambda_i) \) to expected delay of firm \( i \) with \( m_{i,d} \) servers in the distributed system. Smith and Whitt (1981) (see also Benjaafar (1996)) show that \( g \left( \sum_{i=1}^{N} m_i, \sum_{i=1}^{N} \lambda_i \right) \leq \sum_{i=1}^{N} g(m_i, \lambda_i) \) for any integer \( m_i > 0 \) and \( \lambda_i \geq 0 \), with \( \lambda_i/m_i \mu < 1 \). This leads to

\[
 z_p(m_p^*) = h \sum_{i=1}^{N} \lambda_i g \left( m_p^*, \sum_{i=1}^{N} \lambda_i \right) + km_p^* \leq h \sum_{i=1}^{N} \lambda_i g \left( \sum_{i=1}^{N} m_{i,d}^*, \sum_{i=1}^{N} \lambda_i \right) + k \sum_{i=1}^{N} m_{i,d}^* 
\]

and since

\[
 h \sum_{i=1}^{N} \lambda_i g \left( \sum_{i=1}^{N} m_{i,d}^*, \sum_{i=1}^{N} \lambda_i \right) + k \sum_{i=1}^{N} m_{i,d}^* \leq \sum_{i=1}^{N} [h \lambda_i g(m_{i,d}^*, \lambda_i) + m_{i,d}^*] = \sum_{i=1}^{N} z_{i,d}(m_{i,d}^*),
\]

we have

\[
 z_p(m_p^*) \leq \sum_{i=1}^{N} z_{i,d}(m_{i,d}^*).
\]

A proof that there is feasible cost sharing contract is similar to the one of Theorem 8.3 and we omit it for brevity. Here we are not able to show that the expected delay cost for each firm is no greater than its total cost in the distributed system. Therefore, a feasible contract may include a positive financial transfer to some firms to compensate for the increase in their delay costs.

\[\blacksquare\]

**9 Summary and Concluding Comments**

We presented models to study the benefit of capacity pooling among independent service firms. We showed that in settings where facilities are modeled as M/M/1 queues and capacity is determined by the service rate, capacity pooling reduces both total cost and total capacity in the system. We also showed that all firms can benefit under an appropriate cost sharing scheme and that such a scheme is always feasible. However, we found that pooling does not guarantee that the expected delay experienced by each firm would be reduced. In systems where priorities are assigned to
different firms based on their delay costs, some firms may in fact see their delay costs increase. We showed that in systems where there is a lead firm that makes the capacity decision for the pooled system, the firm may under- or over-invest in capacity relative to what is socially-optimal. We characterize conditions under which the lead firm voluntarily chooses the socially optimal capacity level. We extended our results to the case where service times have general distributions and provided a general condition for pooling to continue to being preferable for all firms. We also considered systems where facilities are modeled as M/M/m queues and capacity is determined by the number of servers. In this case, we found that pooling may not be necessarily beneficial to all firms, although it is so when firms have identical delay costs.

There are several possible avenues for future research. It is of interest to consider systems where the inter-arrival times of customers have general distributions that vary from firm to firm. We expect the analysis to become considerably more difficult since there are no general results that characterize the distribution of inter-arrival times when heterogenous arrival processes are superposed. However, it may be possible to construct approximations which could be useful in generating insights into the impact of demand variability (see Gerchak and Gupta 1991 for a discussion of this issue in the context of inventory models with multiple customer classes). We expect pooling to be detrimental to some firms in some cases here. For example, firms with low demand variability may be better off on their own than joining a pooled facility that caters mostly to firms with high demand variability. Similarly, it is of interest to consider firms with heterogenous service time requirements. Here too, we suspect that there may be cases where firms with short service times and/or low service time variability could be better off on their own than sharing a facility with firms with long service times and/or high variability. However, it might be possible to characterize priority schemes that take into account service time distributions, for which pooling is beneficial to all firms.

Finally, it would be of interest to investigate whether or not partial forms of pooling are more desirable than total pooling. Specifically, are there settings under which a subset of the firms are better off breaking away from the pooled system with $N$ firms and forming an independent smaller pooled system? In cooperative game theory, this issue relates to the notion of core and whether or not the core of the game is empty. For the M/M/1 system with FCFS priority and homogenous service times, it is not difficult to show that there always exist a cost sharing scheme under which all subsets of firms are better off in a system with total pooling (the so-called grand coalition). Therefore, the core is not empty. However, it is not clear if this is also true for systems with service priorities or for systems with heterogenous demand and service time distributions.
Appendix

Proof of Theorem 7.1: Let $E[Q_i^{NPP}(\mu)]$ refer to the expected waiting time in the queue for customers of firm $i$ under the NPP policy. Then

$$z_p^{NPP}(\mu) = \frac{\sum_{i=1}^N h_i \lambda_i}{\mu} + \frac{\sum_{i=1}^N h_i \lambda_i E[Q_i^{NPP}(\mu)]}{\mu} + c\mu.$$  

Furthermore, since

$$\sum_{i=1}^N \lambda_i E[Q_i^{NPP}(\mu)] = \frac{\lambda}{\mu - \lambda} - \frac{\lambda}{\mu},$$

where $\lambda = \sum_{i=1}^N \lambda_i$, we also have

$$z_p(\mu) = \sum_{i=1}^N \frac{h_i \lambda_i}{\mu - \lambda} = \sum_{i=1}^N \frac{h_i \lambda_i}{\mu} + \frac{\sum_{k=1}^N h_i \lambda_k}{\lambda} \lambda_i E[Q_i^{NPP}(\mu)] + c\mu,$$

where $z_p(\mu)$ is expected total cost under the FCFS policy given capacity level $\mu$. Noting that

$$\frac{\partial E[Q_i^{NPP}(\mu)]}{\partial \mu} = -\left( \frac{\sum_{k=1}^N \lambda_k - \sum_{i=1}^N \lambda_i}{(\mu - \sum_{k=1}^N \lambda_k)^2} \right) + \frac{\sum_{k=1}^N \lambda_k}{(\mu - \sum_{k=1}^N \lambda_k)^2},$$

we have $\frac{\partial E[Q_i^{NPP}(\mu)]}{\partial \mu} \geq \cdots \geq \frac{\partial E[Q_i^{NPP}(\mu)]}{\partial \mu}$. Since $h_1 \geq h_2 \cdots \geq h_N$, and $\sum_{i=1}^N \left( \frac{\sum_{k=1}^N h_k \lambda_k}{\lambda} \right) \lambda_i = \sum_{i=1}^N h_i \lambda_i$, it is easy to verify that $\frac{\partial z_p(\mu)}{\partial \mu} \geq \frac{\partial z_p(\mu)}{\partial \mu}$. Using the fact that $z^{NPP}(\mu)$ is a convex function of $\mu$ (a proof is straightforward and is omitted for brevity) leads to $\mu^*_{NPP} \leq \mu^*_{p}$. \hfill \blacksquare

Proof of Theorem 7.3: We make use of the results and proofs of Theorems 7.4 and 7.5 (see below). First, we note that

$$E[L_i^{NPP}(\mu)] - E[L_i^{PRP}(\mu)] = \lambda_i \left( \frac{R_i}{S_i S_{i-1}} + \frac{1}{\mu} \right) - \sum_{i=1}^N h_i \lambda_i \left( \frac{R_i}{S_i S_{i-1}} + \frac{1}{\mu S_{i-1}} \right)$$

$$= \lambda_i \left( \frac{\sum_{k=1}^N \lambda_k - \mu}{(\mu - \sum_{k=1}^N \lambda_k)^2} + \frac{1}{\mu} \right).$$

Since for stability we must have $\mu > \sum_{i=1}^N \lambda_i$, we also have

$$h_i E[L_i^{NPP}(\mu)] < h_i E[L_i^{PRP}(\mu)] + h_i \lambda_i \frac{1}{\mu}.$$  

By Theorem 7.4, $\mu^*_{p} \geq \mu^*_{p^{PRP}} \geq \mu^*_{p^{B_i}}$, where $\mu^*_{p^{B_i}}$ refers to the optimal capacity in a system consisting of only firms 1, 2, ..., $i$ as defined in the proof of Theorem 7.5, for $i = 1, \ldots, N$. Therefore,
it follows that
\[
\begin{align*}
\hat{h}_i L^{NPP}_{i,p}(\mu^*_NPP) &\leq \hat{h}_i L^{PRP}_{i,p}(\mu^*_PRP) + \hat{h}_i \lambda_i \frac{1}{\mu_p} \\
&\leq c\lambda_i + \sqrt{\hat{h}_i \lambda_i c} + \hat{h}_i \lambda_i \frac{1}{\sum_{k=1}^{i} \lambda_k + \sqrt{\hat{h}_i \sum_{k=1}^{i} \lambda_k}} \\
&\leq c\lambda_i + \sqrt{\hat{h}_i \lambda_i c} + \frac{1}{\hat{h}_i \lambda_i} \\
&= c\lambda_i + 2\sqrt{\hat{h}_i \lambda_i c} = z^*\iota_{i,\delta}.
\end{align*}
\]

Similar arguments to those used in the proof of Theorem 7.5 complete the proof.

**Proof of Theorem 7.4:** First note that \(z^{NPP}_p(\mu^*_NPP) \geq z^{PRP}_p(\mu^*_NPP) \geq z^{PRP}_p(\mu^*_PRP)\). Next, observe that
\[
\begin{align*}
z^{NPP}_p(\mu) - z^{PRP}_p(\mu) &= \sum_{i=1}^{N} \hat{h}_i \lambda_i \left( \frac{R^N}{S_i S_{i-1}} + \frac{1}{\mu} \right) - \sum_{i=1}^{N} \hat{h}_i \lambda_i \left( \frac{R_i}{S_i S_{i-1}} + \frac{1}{\mu S_{i-1}} \right) \\
&= \sum_{i=1}^{N} \hat{h}_i \lambda_i \left( \frac{\sum_{k=1}^{N} \lambda_k - \mu}{(\mu - \sum_{k=1}^{i} \lambda_k)(\mu - \sum_{k=1}^{i-1} \lambda_k)} + \frac{1}{\mu} \right). \\
&= \sum_{i=1}^{N} \hat{h}_i \lambda_i \left( \frac{\sum_{k=1}^{N} \lambda_k - \mu}{(\mu - \sum_{k=1}^{i} \lambda_k)(\mu - \sum_{k=1}^{i-1} \lambda_k)} + \frac{1}{\mu} \right). \\
&= \sum_{i=1}^{N} \hat{h}_i \lambda_i \left( \frac{\sum_{k=1}^{N} \lambda_k - \mu}{(\mu - \sum_{k=1}^{i} \lambda_k)(\mu - \sum_{k=1}^{i-1} \lambda_k)} + \frac{1}{\mu} \right).
\end{align*}
\]

Define \(T_j, j = 1, \cdots, N - 1\) as follows
\[
T_j = \sum_{i=1}^{j} \lambda_i \left( \frac{\sum_{k=1}^{N} \lambda_k - \mu}{(\mu - \sum_{k=1}^{i} \lambda_k)(\mu - \sum_{k=1}^{i-1} \lambda_k)} + \frac{1}{\mu} \right).
\]

Using the fact that the expected total number of customers in the system is unaffected by the priority policy, we have
\[
\sum_{i=1}^{N} \lambda_i \left( \frac{R^N}{S_i S_{i-1}} + \frac{1}{\mu} \right) = \sum_{i=1}^{N} \lambda_i \left( \frac{R_i}{S_i S_{i-1}} + \frac{1}{\mu S_{i-1}} \right) = \frac{\lambda}{\mu - \lambda},
\]
we can then express \(T_j\) as
\[
T_j = \frac{\sum_{k=j+1}^{N} \lambda_k}{\mu - \sum_{k=1}^{j} \lambda_j} - \frac{\sum_{k=j+1}^{N} \lambda_k}{\mu}, \quad j = 1, \cdots, N - 1.
\]

It is easy to verify that (1) \(\frac{\partial T_j}{\partial \mu} < 0, j = 1, \cdots, N - 1\) for \(\mu > \sum_{i=1}^{N} \lambda_i\) and (2)
\[
z^{NPP}_p(\mu) - z^{PRP}_p(\mu) = \sum_{i=1}^{N-1} (h_i - h_{i+1})T_i.
\]
Since \( h_1 \geq h_2 \geq \cdots \geq h_N \), we have
\[
\frac{\partial (z_{pp}^{N}(\mu) - z_{pp}^{PRP}(\mu))}{\partial \mu} \leq 0,
\]
or equivalently,
\[
\frac{\partial z_{pp}^{N}(\mu)}{\partial \mu} \leq \frac{\partial z_{pp}^{PRP}(\mu)}{\partial \mu}.
\]
Noting that \( z_{pp}^{PRP}(\mu) \) is a convex function of \( \mu \) leads to
\[
\mu_{pp}^* \leq \mu_{pp}^{NPP}.
\]

**Proof of Theorem 7.5:** In order to show that there exists a feasible 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 \( i = 1, ..., N - 1 \) and the order of priority of the firms (under PRP) is 1, 2, \cdots, N, the same as in the original one. We refer to this system as system \( B_N \). The total cost for this system \( z_{pp}^{B_N}(\mu) \) given capacity level \( \mu \) is
\[
z_{pp}^{B_N}(\mu) = \frac{h_N \sum_{i=1}^{N} \lambda_i}{\mu - \sum_{i=1}^{N} \lambda_i} + c \mu,
\]
and the corresponding optimal capacity is
\[
\mu_{pp}^{B_N} = \sum_{i=1}^{N} \lambda_i + \sqrt{\frac{h_N \sum_{i=1}^{N} \lambda_i}{c}}.
\]
Let \( E[L_{N,p}^{B_N}] \) refer to the expected number of customers of firm \( N \) in system \( B_N \). Then,
\[
h_N E[L_{N,p}^{B_N}] = h_N \lambda_N \left( \frac{\sum_{i=1}^{N} \lambda_i}{(\mu_{pp}^{B_N} - \sum_{i=1}^{N} \lambda_i)(\mu_{pp}^{B_N} - \sum_{i=1}^{N-1} \lambda_i)} + \frac{1}{\mu_{pp}^{B_N} - \sum_{i=1}^{N-1} \lambda_i} \right)
\]
\[
= h_N \lambda_N \left( \frac{\sum_{i=1}^{N} \lambda_i}{\sqrt{h_N \sum_{i=1}^{N} \lambda_i} \left( \frac{h_N \sum_{i=1}^{N} \lambda_i}{c} + \lambda_N \right)} + \frac{1}{\sqrt{h_N \sum_{i=1}^{N} \lambda_i} \left( \frac{h_N \sum_{i=1}^{N} \lambda_i}{c} + \lambda_N \right)} \right)
\]
\[
\leq h_N \lambda_N \left( \frac{\sum_{i=1}^{N} \lambda_i}{\sqrt{h_N \sum_{i=1}^{N} \lambda_i} \frac{h_N \sum_{i=1}^{N} \lambda_i}{c} + \lambda_N} + \frac{1}{\sqrt{h_N \lambda_N} \frac{h_N \lambda_N}{c}} \right)
\]
\[
= c \lambda_N + \sqrt{h_N \lambda_N c}.
\]
Noting that the optimal expected cost for firm \( N \) in the distributed system is \( z_{N,d}^* = c \lambda_N + 2 \sqrt{h_N \lambda_N c} \), we have \( h_N E[L_{N,p}^{B_N}] \leq z_{N,d}^* \).

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 \( E[L_{i,p}^{A_N}(\mu)] \) denote the expected number of customers of firm \( i \) in the system given capacity level \( \mu \). Then
\[
\sum_{i=1}^{N} h_i E[L_{i,p}^{A_N}(\mu)] = \sum_{i=1}^{N} (h_i - h_N) E[L_{i,p}^{A_N}(\mu)] + h_N \sum_{i=1}^{N} E[L_{i,p}^{A_N}(\mu)]
\]
\[
= \sum_{i=1}^{N} (h_i - h_N) E[L_{i,p}^{A_N}(\mu)] + h_N \sum_{i=1}^{N} E[L_{i,p}^{B_N}(\mu)],
\]
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) $\frac{\partial E[L_{i,p}(|\mu|)]}{\partial \mu} \leq 0$, we have $\mu_p^{*A_N} \geq \mu_p^{*B_N}$. Consequently, we also have

$$h_N E[L_{N,p}^{A_N}(\mu_p^{*A_N})] = h_N E[L_{N,p}^{B_N}(\mu_p^{*A_N})] \leq h_N E[L_{N,p}^{B_N}(\mu_p^{*B_N})] \leq c\lambda_N + \sqrt{h_N \lambda_N c}.$$

Now suppose we only pool firms $\{1, \cdots, N-1\}$ with delay costs $h_1, \ldots, h_{N-1}$. We refer to this system as $A_{N-1}$ and to the corresponding optimal capacity as $\mu_p^{*A_{N-1}}$. 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 $\mu_p^{*B_{N-1}}$. In both systems the firm priority order is $1, \ldots, N-1$ and the policy is PRP. It is easy to verify that $\mu_p^{*B_{N-1}} = \sum_{i=1}^{N-1} \lambda_i + \sqrt{\frac{h_{N-1} \sum_{i=1}^{N-1} \lambda_i}{c}}$. Using the fact the fact $\frac{\partial^2 z_{PRP}}{\partial \lambda_i \partial \mu} \leq 0$ for all $i$, i.e., $z_{PRP}$ is submodular in $(\lambda_i, \mu)$ for $i = 1, \cdots, N$, it follows (see Theorem 6.1 in Topkis (1978)) that $\mu_p^{*B_{N-1}} \leq \mu_p^{*A_{N-1}} \leq \mu_p^{*A_N}$, which leads to

$$h_{N-1} E[L_{N-1,p}^{A_N}(\mu_p^{*A_N})] \leq h_{N-1} E[L_{N-1,p}^{A_{N-1}}(\mu_p^{*A_{N-1}})] \leq h_{N-1} E[L_{N-1,p}^{B_{N-1}}(\mu_p^{*B_{N-1}})] \leq c\lambda_N + \sqrt{h_N \lambda_N c} \leq z_{N-1,d},$$

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 $i = 1, \ldots, N-2$ that

$$\mu_p^{*B_i} = \sum_{k=1}^{i} \lambda_k + \sqrt{\frac{h_i \sum_{k=1}^{i} \lambda_k}{c}} \leq \mu_p^{*A_i} \leq \mu_p^{*A_N},$$

and

$$h_i L_{i,p}^{A_N}(\mu_p^{*A_N}) \leq h_i L_{i,p}^{A_i}(\mu_p^{*A_i}) \leq h_i L_{i,p}^{B_i}(\mu_p^{*B_i}) \leq c\lambda_i + \sqrt{h_i \lambda_i c} \leq z_{i,d}.$$

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.

**Proof of Theorem 8.3:** In order to complete the proof, we need to show that $h\lambda_j E[W_p^*] \leq z_{j,d}$ for $j = 1, \ldots, N$. Suppose the opposite is true. Then there exists some $j$ such that

$$h\lambda_j E[W_p^*] \geq z_{j,d} = f(\lambda_j, \mu(\lambda_j)).$$

Multiplying both sides of the above inequality by $\sum_{i=1}^{N} \lambda_i / \lambda_j$ we obtain

$$h \left( \sum_{i=1}^{N} \lambda_i \right) E[W_p^*] > \left( \sum_{i=1}^{N} \frac{\lambda_i}{\lambda_j} \right) f(\lambda_j, \mu(\lambda_j)).$$

22
Since \( f \left( \sum_{i=1}^{N} \lambda_i, \mu(\sum_{i=1}^{N} \lambda_i) \right) / \sum_{i=1}^{N} \lambda_i < f \left( \lambda_j, \mu(\lambda_j) \right) / \lambda_j \) due to the strict concavity of the function \( f \), we have

\[
h \left( \sum_{i=1}^{N} \lambda_i \right) E[W^*_p] \geq \left( \sum_{i=1}^{N} \lambda_i / \lambda_j \right) f \left( \lambda_j, \mu(\lambda_j) \right) > f \left( \sum_{i=1}^{N} \lambda_i, \mu(\sum_{i=1}^{N} \lambda_i) \right).
\]

However, \( f \left( \sum_{i=1}^{N} \lambda_i, \mu(\sum_{i=1}^{N} \lambda_i) \right) \geq h \left( \sum_{i=1}^{N} \lambda_i \right) E[W^*_p] \), which leads to a contradiction. ■
References


Figure 1 – An example illustrating how pooling under NPP can lead to higher expected delay to a firm with lower priority ($\lambda_1 = 200$, $\lambda_2 = 2$, $h_1 = 310$, $h_2 = 300$)
Figure 2 – The effect of service policies on the relative reduction in optimal cost
\( (h_1 = 500, h_2 = 350, h_3 = 250, \lambda_1 = 200, \lambda_2 = 300, \lambda_3 = 400) \)

Figure 3 – The effect of service policies on the relative reduction in optimal capacity
\( (\delta = (\mu - \mu') / \mu', h_1 = 500, h_2 = 350, h_3 = 250, \lambda_1 = 200, \lambda_2 = 300, \lambda_3 = 400) \)
Figure 4 – The effect of service time variability and capacity cost on the relative benefit of pooling ($N = 2$, $h = 200$, $\lambda_1 = \lambda_2 = 350$)

Figure 5 – An example illustrating how pooling can lead to higher total cost ($h_1 = 1/\varepsilon$, $h_2 = 1$, $\lambda_1 = \varepsilon$, $\lambda_2 = 1$, $\mu = 1$)