

On the Benefits of Pooling in Production-Inventory Systems

Saif Benjaafar, William L. Cooper

Department of Mechanical Engineering, Graduate Program in Industrial Engineering, University of Minnesota,
Minneapolis, Minnesota 55455 {saif@umn.edu, billcoop@me.umn.edu}

Joon-Seok Kim

Strategy Consulting Team, Samsung SDS, Seoul, 135-918, Korea, jkim@me.umn.edu

We study inventory pooling in systems with symmetric costs where supply lead times are endogenously generated by a finite-capacity production system. We investigate the sensitivity of the cost advantage of inventory pooling to various system parameters, including loading, service levels, demand and production time variability, and structure of the production system. The analysis reveals differences in how various parameters affect the cost reduction from pooling and suggests that these differences stem from the manner in which the parameters influence the induced correlation between lead-time demands of the demand streams. We compare these results with those obtained for pure inventory systems, where lead times are exogenous. We also compare inventory pooling with several forms of capacity pooling.

Key words: production-inventory systems; make-to-stock queues; inventory pooling; demand correlation

History: Accepted by Candace A. Yano, operations and supply chain management; received January 30, 2003.

This paper was with the authors 6 months for 2 revisions.

1. Introduction

Inventory pooling refers to the consolidation of multiple inventory locations into a single one. Inventory locations may be associated with different geographical sites, different products, or different customers. Since the publication of the seminal paper by Eppen (1979), inventory pooling has been an important theme in the operations management literature. An extensive body of work, recent reviews of which can be found in Alfaro and Corbett (2003) and Gerchak and He (2003), has documented the costs and benefits of pooling in a variety of settings. This literature has largely focused on the analysis of single-period problems or problems with multiple periods but with exogenous lead times. 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 correlation between these demands, with the difference reducing to zero when the demands have perfect positive correlation. In a system with identical independent locations, cost increases linearly in the number of demand streams (locations) in a distributed system and proportionally to the square root of the number of demand streams in a pooled one.

In this paper, we study pooling of inventories in a system with symmetric costs in a setup similar to Eppen's. However, we consider a system where

demand arrives dynamically and supply lead times are endogenous and generated by a production system with finite capacity and stochastic production times. Hence, supply lead times are affected by congestion in the production system and are load dependent. The models we describe are *make-to-stock queues*, an overview of which can be found in Buzacott and Shanthikumar (1993). Because they allow the production system to be explicitly modeled, make-to-stock queues are also called *production-inventory systems*. Zipkin (2000) offers a comprehensive discussion of various types of supply systems including those that induce constant, exogenous and sequential, and load-dependent lead times. Treatments of multiclass production-inventory systems include Wein (1992), Zipkin (1995), Veatch and Wein (1996), Ha (1997), de Véricourt et al. (2000a, 2001, 2002), Bertsimas and Paschilidis (2001), and Benjaafar et al. (2004). Despite the large body of work on pooling, the issue has, to our knowledge, not been fully explored in a production-inventory framework, which captures dependencies between production and inventory systems.

Treating production systems and inventory systems as independent units is realistic when they are decoupled through large inventory holding at the production facility or at subsequent stages of the supply chain (for example, when a local retailer is replenished from a large regional warehouse). It may also

be reasonable when the inventory and production systems belong to separate entities, with the owner of the production system guaranteeing a fixed delivery date. Similarly, it may be justified when, e.g., transportation lead times are significantly longer than manufacturing lead times. However, for most integrated systems these assumptions rarely hold. In fact, the increased emphasis on lean manufacturing has resulted in the tight control of finished-goods inventories at factories as well as the reduction of material-handling times between production facilities and finished-goods warehouses. Similar principles are also being applied to entire supply chains, resulting in interdependence among retailers, distributors, and suppliers. Practices such as vendor management of inventory, where the manufacturer is responsible for managing inventory at the retailer, and inventory consignment, where the manufacturer retains ownership of inventory until it is sold, have also strengthened the connection between production and inventory functions. Consequently, distributors and retailers can be immediately affected by congestion and delays on the factory floor. Management decisions regarding finished goods within a factory are, of course, almost always affected by the status of the production facility.

The need to consider pooling does arise in industries where production and inventory are tightly coupled. For example, large contract manufacturers (CMs) in the electronics industry, such as Solectron or Celestica, may own tens of production facilities worldwide and offer manufacturing outsourcing services to several original equipment manufacturers (OEMs). The core manufacturing technology in such CM facilities is typically surface mount technology for printed circuit-board assembly. To meet the needs of a wide range of OEMs, many CMs have invested in flexible technologies that can manufacture a high variety of products in varying quantities with little changeover time or cost (see Kador 2001, Plambeck and Taylor 2005). Terms of the contracts with OEMs often require the CM to fulfill demand on a continuous basis in a just-in-time mode with little or no advance notice (see Barnes et al. 2000). Consequently, many products are manufactured ahead of demand in a make-to-stock fashion. Facilities may produce multiple products for multiple customers, and each product may be produced in multiple facilities, with the finished products held either at the factory, at a customer site, or in regional distribution centers.

For contract manufacturers, the need to assess the benefit of pooling arises in several contexts. For items produced and stocked by multiple factories, there is the possibility of centralizing inventory in a single location, which could then be supplied by the various factories. Alternatively, there is the possibility of

physically keeping inventory at each factory but managing a virtual stock, common to all factories. Orders placed with one factory could then be satisfied by available inventory regardless of its location. Because of strong commonalities between different products (e.g., circuit boards for different mobile phones), there is also the possibility of pooling across products by producing-to-stock common semifinished products that can be quickly differentiated once a specific order is placed. In addition to inventory pooling, there are opportunities to pool capacity, physically or virtually, by either consolidating multiple small facilities, or managing a group of facilities as a shared resource, so that incoming orders can be directed to any of the available facilities. In some cases, there is the opportunity to jointly pool capacity and inventory.

The possibility of pooling in this and other industries raises several questions. How desirable is pooling, and what factors affect the benefits derived from it? In what settings is pooling most or least beneficial? How different are the effects of pooling in production-inventory systems from those observed in pure inventory systems? Is there a difference between inventory pooling and capacity pooling, and are there settings in which one form of pooling is superior to the other? The models we present in this paper are in part a response to these questions and are envisioned to be used by firms to support strategic decisions about when and how to pool. Although the models are stylized, they are useful in generating insights into which factors affect pooling and in developing guidelines as to when pooling is most valuable. The models are also useful as a diagnostic tool for identifying parameters whose improvement would most increase the value of pooling.

We first consider a situation in which multiple inventory locations with identical holding and back-ordering costs, served by a common production facility, are pooled into one inventory location. We show that the cost difference (between pooled and distributed systems) and cost ratio (of pooled to distributed systems) are not monotonic as functions of utilization. The cost difference, although effectively increasing when utilization is high, remains bounded and approaches a finite limit as utilization approaches one. The cost ratio, on the other hand, is largely decreasing when utilization is high and approaches one as utilization approaches one. Although the end demands from individual inventory locations are assumed independent, we show that their lead-time demands are correlated, with lead-time demands becoming nearly perfectly correlated when utilization is near one.

For non-Markovian systems, we find that the effect of increasing variability in demand or production

times parallels that of increasing utilization. In systems where each inventory location is initially supplied by a separate production facility, the effect of utilization on the value of pooling can be quite different. In particular, the cost ratio, when viewed as a function of utilization, is essentially constant, while the cost difference increases without bound in the same parameter. These differences are, in part, due to the lack of correlation in lead-time demands when separate production facilities are used. In production facilities with multiple stages, we show that an increase in the number of stages (although it increases both the mean and variance of lead-time demand) has essentially no effect on the cost ratio. This, as well, can be explained by the fact that an increase in the number of stages leaves correlation unaltered.

A comparison of results for production-inventory systems with those for pure inventory systems reveals the key implicit role played by lead-time demand correlation. The importance of lead-time demand correlation as a determinant of the benefits of pooling has been studied by a number of authors, including Eppen (1979), Lu et al. (2003), and Alfaro and Corbett (2003). These studies focus on lead-time demand correlation induced by correlation in the end demands, while in our case it is induced by the sharing of the supply process. There is also related work that analyzes the influence of demand correlation on the value of flexible capacity; see Netessine et al. (2002) for results and references. One of the distinctive features of the production-inventory setting is that the joint distribution (and consequently, means, variances, correlations, and so forth) of lead-time demand is not assumed, but rather is endogenous and induced by microlevel system characteristics. When viewing the cost benefit of pooling as a function of certain system parameters (e.g., system loading or number of processing stages), we will see different behaviors, apparently caused by subtle differences in the way the parameter in question affects second-order properties of lead-time demand.

In systems where inventory locations are initially supplied by separate production facilities, we also compare inventory pooling to capacity pooling. The analysis reveals that while the relative benefit of inventory pooling tends to diminish with utilization, the relative benefit of capacity pooling tends to increase with utilization. The difference appears to be related to the fact that capacity pooling leads to a larger reduction in lead-time demand variance, with this reduction increasing in utilization. For highly loaded systems, we show that capacity pooling alone achieves nearly the same relative benefit as the joint pooling of capacity and inventory.

The remainder of this article is organized as follows. In §2.1, we present the model and provide preliminary results. In §2.2, we consider the case of Poisson

demand and a cost-based model. In §2.3, we compare the results obtained for the production-inventory system to those obtained from Eppen's model. In §2.4 we examine effects of production-scheduling policy on the benefits of pooling. In §§3.1–3.3, we focus on systems with service-level constraints, non-Markovian demand, and multiple service facilities. In §3.4 we study systems with multiple processing stages. In §4 we present some closing remarks.

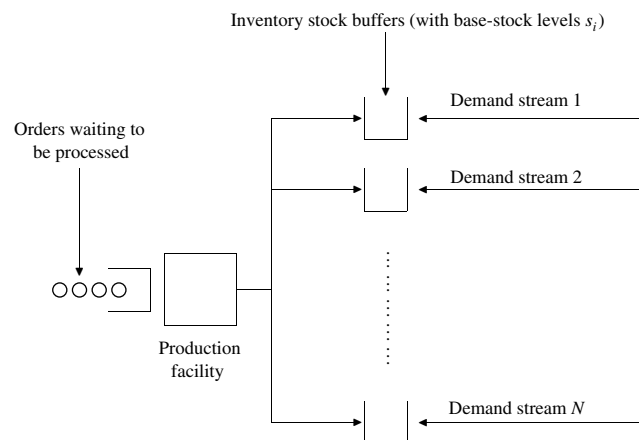
2. The Benefits of Pooling

Consider a production-inventory system with N inventory locations, where for $i = 1, \dots, N$, the location i inventory is managed according to a base-stock policy with base-stock level s_i . See Figure 1. If a demand for type i arrives to find location i empty, the demand is backordered. Supply lead times are endogenous and determined by the production and queuing times at the facility as follows. Whenever a unit of inventory is demanded at location i , an order is triggered at the common (for all inventory locations) production facility. Production and interarrival times are random. Orders are processed one at a time on a first-come-first-served (FCFS) basis, and orders that arrive at a busy production facility must wait in a queue.

2.1. Preliminaries

The term *distributed system* refers to a system as described above with N inventory locations, each dedicated to one demand stream. Likewise, the term *pooled system* refers to the situation where the N demand streams are satisfied from a single inventory location. We assume that the items demanded by each stream in the distributed system are similar enough that any item can satisfy any demand after "pooling."

Figure 1 Production-Inventory System with N Items



Notes. There are N inventory locations, each with its own demand stream and inventory buffer. Each buffer operates under a continuous-review base-stock policy. Replenishments are received from a single production facility, where jobs are processed one-at-a-time on a FCFS basis.

This allows us to treat items and demands as homogeneous in the pooled system, so we need not keep track of what type of item a particular arriving order wants.

Let $I_i(t)$ and $B_i(t)$ denote inventory and backorder levels at time t for location i in a distributed system. Assume that all locations are fully stocked at time $t = 0$. Define the cumulative total cost accrued in the time interval $(0, t]$ to be

$$\psi_N(t) = \int_0^t \phi_N(u) du,$$

where

$$\phi_N(t) = \sum_{i=1}^N \phi_{i,N}(t), \quad \text{and} \quad \phi_{i,N}(t) = hI_i(t) + bB_i(t),$$

where $h, b > 0$ are the per-unit holding and backorder cost rates. Let $\{T_k\}$ denote the overall sequence of arrival times, so that T_k is the time of the k th arrival, regardless of location. Similarly, let $\{\xi_k\}$ denote the sequence of processing times, so that ξ_k is the time spent in service of the k th job to enter service. Finally, let $L_k \in \{1, \dots, N\}$ indicate to which of the N demand streams the k th arrival belongs. So, the basic exogenous quantities are the sequences $\{T_k\}$, $\{\xi_k\}$, and $\{L_k\}$.

For the distributed system, denote the cumulative demand at location i in the time interval $(0, t]$ by $D_i(t) \equiv |\{k: T_k \leq t, L_k = i\}|$. Let $\{U_k\}$ be the sequence of departure times from the production facility. Note that $\{U_k\}$ is determined by $\{T_k\}$ and $\{\xi_k\}$. Following a setup as in Buzacott and Shanthikumar (1993), let $C_i(t)$ be the total number of type i units completed by the production facility during $(0, t]$, and $R_i(t)$ be the total amount of type i demand satisfied during $(0, t]$. Formally, we have

$$\begin{aligned} C_i(t) &\equiv |\{k: U_k \leq t, L_k = i\}|, \\ R_i(t) &\equiv \min\{s_i + C_i(t), D_i(t)\}, \\ I_i(t) &\equiv s_i + C_i(t) - R_i(t), \\ B_i(t) &\equiv D_i(t) - R_i(t). \end{aligned}$$

Given the distributed system, we define the *corresponding pooled system* to be the one in which the cumulative demand process and base-stock level are given by, respectively, $D(t) \equiv \sum_{i=1}^N D_i(t)$, and $s \equiv \sum_{i=1}^N s_i$, and the sequence of service times is the same as that in the distributed system. Let $\{\bar{U}_k\}$ be the sequence of departure times from the production facility in the pooled system. From the definition of $D(\cdot)$ and the fact that the sequence of service times is the same as in distributed system, it follows that the sequence of departure times from the pooled system is identical to that in the distributed system; that is,

$\{\bar{U}_k\} = \{U_k\}$. For pooled systems, define the processes $C(\cdot), R(\cdot), I(\cdot), B(\cdot), \phi(\cdot)$, and $\psi(\cdot)$ by

$$\begin{aligned} C(t) &\equiv |\{k: \bar{U}_k \leq t\}|, \\ R(t) &\equiv \min\{s + C(t), D(t)\}, \end{aligned}$$

and so forth. Because $\{\bar{U}_k\} = \{U_k\}$, we have $C(t) = \sum_{i=1}^N C_i(t)$. After some algebraic manipulations, we obtain the following result, which states that on a path-by-path basis the cost of the distributed system is at least that of the corresponding pooled system.

PROPOSITION 1. *For all $t \geq 0$, we have $\phi(t) \leq \phi_N(t)$, and $\psi(t) \leq \psi_N(t)$.*

Note that the above does not require distributional assumptions. However, if one is interested in stochastic comparisons, it implies that for any t the cost rate and the cumulative cost of the pooled system are stochastically smaller than those of the distributed system. Proposition 1 also does not need the FCFS assumption and can be extended in a variety of directions. We omit the details. Although proving Proposition 1 is straightforward (the essential step is to state the problem in the right way), it is worth pointing out that it is stronger than analogous results dealing with *statistical* economies of pooling. Specifically, because Proposition 1 holds on sample paths, pooling is beneficial regardless of the statistical properties (i.e., distributions) of the underlying processes. Of course, when evaluating the difference in *expected* costs between distributed and pooled systems, distributional assumptions do matter.

Let $Q_i(t) = D_i(t) - C_i(t)$ and let $\mathbf{Q}(t) = (Q_1(t), \dots, Q_N(t))$, and suppose the individual arrival streams arise from an i.i.d. splitting of a single renewal process, the production times form an i.i.d. sequence with finite variance, the overall arrival rate is strictly less than the service rate, and that all arrival and service processes are independent. By analogy to a $GI/GI/1$ queue, $(\mathbf{Q}(t): t \geq 0)$ and $(\phi_N(t): t \geq 0)$ are regenerative processes; see Wolff (1989). Let $\mathbf{Q} = (Q_1, \dots, Q_N)$ denote a random vector with the time-average limiting distribution of $\mathbf{Q}(t)$. The assumptions ensure that \mathbf{Q} exists, $EQ_i < \infty$, and that the long-run expected total (over all N items) cost per unit-time,

$$TC_N(\mathbf{s}) \equiv \lim_{t \rightarrow \infty} t^{-1} E\psi_N(t; \mathbf{s})$$

exists and equals

$$\sum_{i=1}^N bEQ_i + hs_i - (h + b)E \min\{s_i, Q_i\}.$$

The argument $\mathbf{s} \in \mathbb{Z}_N^+$ of $\psi_N(\cdot)$ indicates the dependence on the base-stock levels. Define

$$TC_N^* \equiv \min_{\mathbf{s} \in \mathbb{Z}_N^+} TC_N(\mathbf{s}),$$

and let $\mathbf{s}^* = (s_1^*, \dots, s_N^*)$ be a vector of base-stock levels that attains the minimum; i.e., $TC_N(\mathbf{s}^*) = TC_N^*$. Consider now a pooled system with corresponding arrival and production processes as described above, but suppose the base-stock level is a decision variable. Let

$$TC(s) \equiv \lim_{t \rightarrow \infty} t^{-1} E\psi(t; s)$$

be the long-run cost for the pooled system with base-stock level s , and define

$$TC^* \equiv \min_{s \in \mathbb{Z}^+} TC(s).$$

Let s^* be a base-stock level that satisfies $TC(s^*) = TC^*$.

PROPOSITION 2. *Under the above assumptions, $TC^* \leq TC_N^*$.*

PROOF. Observe that

$$TC^* \leq TC\left(\sum_{i=1}^n s_i^*\right) \leq TC_N(\mathbf{s}^*) = TC_N^*.$$

The first inequality follows from the definition of TC^* , and the second follows from the definitions of $TC_N(\cdot)$ and $TC(\cdot)$ and Proposition 1. \square

2.2. Markovian Cost-Based Systems

In the previous section, we saw that a pooled system has a lower cost than a distributed one under general conditions. However, this does not tell us about the magnitude of the advantage, or how this advantage is affected by system parameters. In this section, we examine the steady-state average cost per unit-time of both systems under optimal choices of base-stock levels. We first limit our discussion to systems where each demand stream i follows an independent Poisson process with rate λ_i . Production times at the production facility are independent of the arrival process and i.i.d. with exponential distribution with mean μ^{-1} . Let λ denote the overall demand rate, and $p_i = \lambda_i/\lambda$ denote the fraction of overall demand of type i . The superposition of independent Poisson processes is also Poisson, so the arrival process to the production system under the base-stock policy is also Poisson. (Note that this system is distributionally equivalent to one in which there is a single Poisson arrival process with rate λ that is split, in an i.i.d. manner, into N streams according to the probabilities $p_i; i = 1, \dots, N$.) Hence, when viewed in isolation, the production facility is an $M/M/1$ queue. For stability, we assume that $\rho \equiv \lambda/\mu < 1$.

To proceed, we need queue-length distributions and total cost formulas—derivations of the expressions in the next two paragraphs can be found in Buzacott and Shanthikumar (1993, Chapter 4). The joint probability generating function of $\mathbf{Q} = (Q_1, \dots, Q_N)$, define in §2.1, is given by

$$\tilde{p}(\mathbf{z}) = E\left[\prod_{i=1}^N z_i^{Q_i}\right] = \frac{1 - \rho}{1 - \sum_{i=1}^N p_i \rho z_i}. \quad (1)$$

It follows that Q_i is geometrically distributed with parameter r_i (i.e., $P(Q_i = n) = (1 - r_i)r_i^n$) where $r_i \equiv p_i\rho/(1 - \rho + p_i\rho)$. Let

$$(\mathbf{I}(\mathbf{s}), \mathbf{B}(\mathbf{s})) = (I_1(\mathbf{s}), \dots, I_N(\mathbf{s}), B_1(\mathbf{s}), \dots, B_N(\mathbf{s}))$$

denote a random vector with joint distribution equal to that of the steady-state inventory and backorder levels when the vector of base-stock levels is \mathbf{s} . Because $(\mathbf{I}(\mathbf{s}), \mathbf{B}(\mathbf{s}))$ is equal in distribution to $([\mathbf{s} - \mathbf{Q}]^+, [\mathbf{Q} - \mathbf{s}]^+)$, we have

$$E[I_i(\mathbf{s})] = s_i - [r_i(1 - r_i^{s_i})/(1 - r_i)] \quad \text{and}$$

$$E[B_i(\mathbf{s})] = r_i^{s_i+1}/(1 - r_i).$$

To simplify future developments, define

$$\mathcal{C}(p, q, s) \equiv h[s - qp(1 - p^s)/(1 - p)] + bqp^{s+1}/(1 - p).$$

By an ergodic theorem,

$$TC_N(\mathbf{s}) = E \sum_{i=1}^N [hI_i(\mathbf{s}) + bB_i(\mathbf{s})] = \sum_{i=1}^N \mathcal{C}(p_i, 1, s_i).$$

Because the total cost function is strictly convex and separable, the optimal base-stock levels can be obtained by finding the smallest integer s_i that satisfies

$$\mathcal{C}(p_i, 1, s_i + 1) > \mathcal{C}(p_i, 1, s_i).$$

Define $\gamma = h/(h + b)$. Then, optimal base-stock levels are given by

$$\mathbf{s}^* = (s_1^*, \dots, s_N^*) \quad \text{with } s_i^* = \lfloor \tilde{s}_i \rfloor, \quad \text{where } \tilde{s}_i \equiv \ln \gamma / \ln r_i.$$

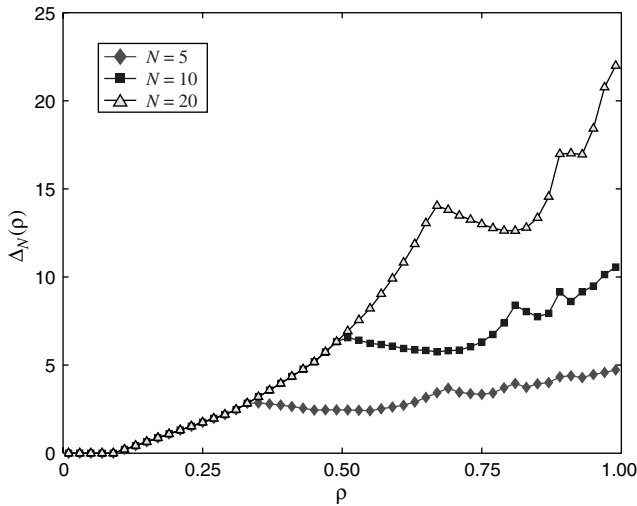
Analysis of the pooled system is similar. In particular, $TC(s) = \mathcal{C}(\rho, 1, s)$, and an optimal base-stock level is given by $s^* = \lfloor \tilde{s} \rfloor$ where $\tilde{s} = \ln \gamma / \ln \rho$; see also Veatch and Wein (1996). For a distributed system, operating under the best base-stock policy is not known to be optimal (among all possible control policies). Nevertheless, we focus exclusively on base-stock policies.

By ignoring the integrality of base-stock levels, we can approximate optimal costs for the distributed and pooled systems with the simpler forms

$$\begin{aligned} TC_N^U &\equiv TC_N(\tilde{\mathbf{s}}) = h \sum_{i=1}^N \tilde{s}_i \\ &= h \sum_{i=1}^N \frac{\ln \gamma}{\ln(p_i\rho) - \ln(1 - \rho + p_i\rho)}, \end{aligned} \quad (2)$$

and $TC^U \equiv TC^U(\tilde{s}) = h\tilde{s}$. In fact, TC_N^U and TC^U are upper bounds that become exact as ρ approaches 1.

Figure 2 The Effect of Utilization on Absolute Advantage of Pooling
 ($h = 1, b = 10, p_i \equiv 1/N$)

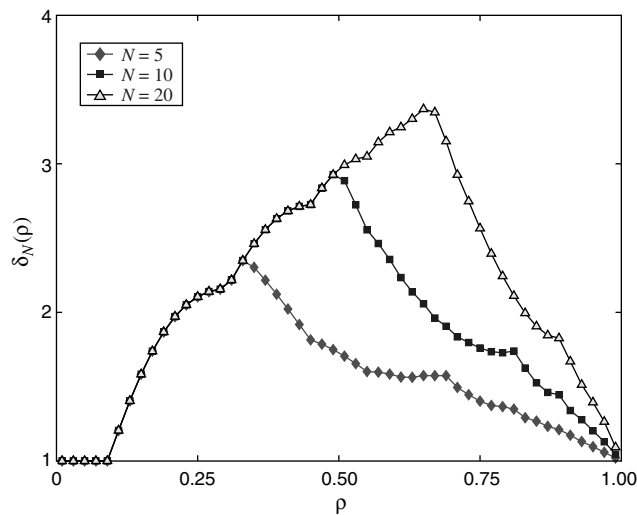


To assess the advantages of pooling, we examine the difference and the ratio of the optimal total cost of the distributed system to that of the pooled system; respectively,

$$\Delta_N \equiv TC_N^* - TC^* \quad \text{and} \quad \delta_N \equiv TC_N^*/TC^*.$$

We will sometimes append a parenthetical ρ to Δ_N and δ_N to indicate dependence on utilization. By virtue of Proposition 2, the ratio is always greater than or equal to one, and the difference is always non-negative. Figures 2 and 3 show the sensitivity of Δ_N and δ_N to ρ . Although both the ratio and the difference are nonmonotonic in ρ , the effect of ρ on the ratio is different from that on the difference. When utilization is relatively low (when $\rho < \gamma$), both pooled and distributed systems operate in a make-to-order

Figure 3 The Effect of Utilization on Relative Advantage of Pooling
 ($h = 1, b = 10, p_i \equiv 1/N$)



fashion, and hence the two systems are equivalent. Initial increases beyond γ tend to increase both the ratio δ_N and the difference Δ_N (in this region the pooled system holds stock while certain locations in the distributed system still produce to order). For the ratio, this initial increase, which is an artifact of the integrality of base-stock levels, persists until base-stocks become relatively large, at which point the ratio begins decreasing.

Characterizing the value of ρ at which δ_N is maximum is difficult in general. However, for a system with symmetric locations ($p_i = 1/N$ for all i), we observed this maximum occurs when $\rho \approx Nh/(Nh + b)$, the point at which the distributed system starts to hold stock. When ρ approaches one, the ratio approaches one also—see Proposition 3 below. On the other hand, the absolute difference exhibits a general upward trend when viewed as a function of ρ ; nevertheless, $\Delta_N(\rho)$ is not an increasing function owing to the integrality of base-stock levels. When utilization is very close to one, $\Delta_N(\rho)$ approaches a finite limit. This means that the absolute benefit of pooling remains bounded, even as utilization approaches one. This limit depends upon N and the cost parameters h and b , but not on the relative size of the demand from each location. As $\rho \uparrow 1$, costs and base-stock levels grow to ∞ (this helps explain the limiting behavior of the ratio in view of that of the difference). Hence, it is (highly) desirable to avoid such situations. Nevertheless, the limit results should be of interest to managers who must work with heavily utilized systems.

PROPOSITION 3. For each $N \in \mathbb{Z}^+$, we have

- (a) $\lim_{\rho \uparrow 1} \Delta_N(\rho) = \frac{1}{2}h(N - 1)\ln(1/\gamma)$,
- (b) $\lim_{\rho \uparrow 1} \delta_N(\rho) = 1$.

PROOF. It can be shown (using a lengthy argument) that

$$\lim_{\rho \uparrow 1} \Delta_N(\rho) = \lim_{\rho \uparrow 1} (TC_N^U - TC^U)$$

and

$$\lim_{\rho \uparrow 1} \delta_N(\rho) = \lim_{\rho \uparrow 1} TC_N^U/TC^U.$$

So, in view of (2) we have

$$\lim_{\rho \uparrow 1} \Delta_N(\rho) = h \ln \gamma \lim_{\rho \uparrow 1} \sum_{i=1}^N \left[\frac{1}{\ln(p_i \rho) - \ln(1 - \rho + p_i \rho)} - \frac{p_i}{\ln \rho} \right].$$

By applying l'Hospital's rule twice, we find that the limit as $\rho \uparrow 1$ of the i th term in brackets is $(p_i - 1)/2$. Therefore, part (a) is proved. For (b), from (2) we have

$$\lim_{\rho \uparrow 1} \delta_N(\rho) = \lim_{\rho \uparrow 1} \sum_{i=1}^N \frac{\ln \rho}{\ln(p_i \rho) - \ln(1 - \rho + p_i \rho)}.$$

Upon application of l'Hospital's rule, the limit as $\rho \uparrow 1$ of the i th term in the above summation can be seen to be p_i , so part (b) is proved. \square

As mentioned earlier, the nonmonotonic behavior of δ_N and Δ_N in ρ is primarily due to the integrality of the base-stock levels. To see this, consider the case where ρ is high enough that $s_i^* \gg 1$ and the integrality requirement may be dropped without significant loss of accuracy. The ratio δ_N can then be approximated by $\delta_N^u = TC_N^u / TC^u$ and the difference Δ_N by $\Delta_N^u = TC_N^u - TC^u$. Examining the effect of ρ , we can show using standard calculus that δ_N^u is indeed strictly decreasing in ρ while Δ_N^u is strictly increasing. In §2.3, we provide additional discussion and an explanation of these effects.

From a managerial perspective, the results of this section indicate that the percentage cost reduction from inventory pooling should be expected to decrease with utilization once utilization reaches a sufficiently high level. In heavily loaded systems, the percentage cost reduction becomes insignificant and a pooled system offers no relative advantage to a distributed one. There is, of course, still value to pooling as indicated by the positive absolute cost savings. A manager would need to take this into account when making a final decision about whether or not to pool. In §2.4, we show that the absolute cost savings from pooling can be smaller and even go to zero in the limit if the distributed system is managed more effectively.

In addition to being sensitive to loading, the performance difference between distributed and pooled systems is sensitive to N , the number of inventory locations. To investigate the effect of N on TC_N^* , we fix ρ , and vary N while assuming that $p_i = 1/N$ for $i = 1, \dots, N$. First note that given ρ , there is a value of N beyond which the optimal cost remains constant. This value can be obtained by noting that

$$\tilde{s}_i = \ln \gamma / \ln(\rho / (N - N\rho + \rho)) < 1 \quad \text{when } N > b\rho / (h - h\rho).$$

This yields $\mathbf{s}^* = (0, \dots, 0)$ and $TC_N^* = b\rho / (1 - \rho)$ whenever

$$N \geq N_{\max} \equiv \lceil 1 + b\rho / (h - h\rho) \rceil.$$

The value N_{\max} marks the number of types beyond which it becomes optimal not to hold any inventory and to produce to order (note that this number is increasing in ρ). Similar observations regarding the effect of N and the existence of N_{\max} have been made by de Véricourt et al. (2000b) in an analysis of delayed product differentiation. In the region where TC_N^* is a nontrivial function of N , (i.e., $N < N_{\max}$) total cost is increasing in N at a rate approximately proportional to $N / \ln N$. Hence, the relative advantage of pooling also increases with the same rate.

We have assumed that items are symmetric in their cost parameters, and under this assumption we have shown that pooling is always desirable. In a system where items have different backorder costs, pooling inventory in a single location does not always

lead to lower costs. In fact, if we continue to fulfill demand on a FCFS basis, serving expensive and cheap classes from a single location can lead to higher total costs (more inventory is kept to meet the requirements of the more expensive class); see Kim (2001) and de Véricourt et al. (2002) for examples and discussion. This effect would, of course, disappear if an optimal inventory rationing policy were used (see de Véricourt et al. 2002). However, even with an optimal rationing policy, one must proceed cautiously. Although total cost is reduced when inventory is pooled, the performance of individual customer classes can deteriorate. In particular, a low demand class is not always guaranteed to experience lower backorder levels in the pooled system.

2.3. Comparisons with Other Inventory Models

In this section we contrast the effect of pooling in a production-inventory system with that in a system with exogenous lead time. A common technique is to approximate lead-time demand (demand that takes place during the supply lead time) with a normal distribution. Using the assumption that for each i , lead-time demand $D_i \sim D$ is normally distributed with identical mean and a standard deviation of σ , Eppen (1979) shows total cost in the distributed and pooled systems can be written as

$$TC_d = KN\sigma \quad \text{and} \quad (3)$$

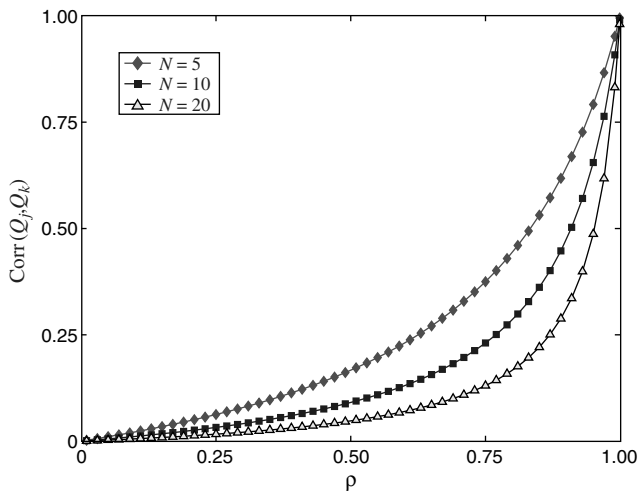
$$TC_p = K\sigma \sqrt{N + \sum_{i \neq j} \nu_{ij}}, \quad (4)$$

respectively, where K is a constant and ν_{ij} is the correlation between lead-time demand for items i and j .

One might argue that with appropriate moment matching, the above model could be used to capture the effect of pooling in a production-inventory system. If the lead-time demands are uncorrelated, then (3)–(4) imply that the cost difference between the distributed and the pooled system is $\Delta = K\sigma(N - \sqrt{N})$ and the cost ratio is $\delta = \sqrt{N}$. Note that δ depends only upon N . Furthermore, as system loading ρ approaches 1 then σ will approach ∞ , so using (3)–(4) with $\nu_{ij} = 0$ will lead us to conclude that the difference Δ grows without bound as ρ increases to 1, thereby suggesting the possibility of arbitrarily large benefits in pooling. As we saw in Proposition 3, this is not the case in the production-inventory setting. Also, the expressions above suggest that the ratio depends only on N . Both results would clearly mislead managers regarding the value (absolute or relative) of inventory pooling when lead times are not exogenous.

The discrepancy arises in part from ignoring the correlation between lead-time demands. In Eppen's model, if lead-time demands are perfectly positively correlated (i.e., $\nu_{ij} = 1 \forall i, j$), then $\Delta = 0$ and

Figure 4 The Effect of Utilization on Lead-Time Demand Correlation ($\rho_j \equiv 1/N$)



$\delta = 1$. In the production-inventory setting of §2.2, although the demand streams are independent, lead-time demands are not. In fact, it follows from (1) that the correlations between the queue lengths (in a Markovian production-inventory system, steady-state queue length is the analog of lead-time demand) in the distributed production-inventory system are given by

$$\text{Corr}(Q_j, Q_k) = \frac{p_j p_k \rho}{\sqrt{p_j p_k (p_j \rho - \rho + 1)(p_k \rho - \rho + 1)}}. \quad (5)$$

The correlation approaches 1 as $\rho \uparrow 1$; see Figure 4. Although the lead-time demands become perfectly correlated in the limit, the cost difference for the production-inventory setting approaches a positive finite limit (not zero). If the correct correlations are used in the Eppen model, one can induce the normal approximation to yield results similar to those of the production-inventory model. However, obtaining the “correct” correlations requires foresight that can be obtained only by solving the production-inventory model.

Once we obtain the correct moments using the production-inventory model, the normal approximation can be used to “reproduce” limit results such as Proposition 3. For example, if we approximate the lead-time demand in both the distributed and pooled systems by normal distributions with matching moments, then the optimal cost in the pooled system can be approximated as $TC^* \approx \zeta \equiv K\sigma_p$, where $\sigma_p^2 = \rho/(1 - \rho)^2$ is the exact variance of lead-time demand (which is geometric with parameter ρ) for the pooled production-inventory system. Similarly, the approximate cost for the distributed system is $TC_d^* \approx \zeta_N \equiv NK\sigma_d$, where $\sigma_d^2 = r/(1 - r)^2$ is the exact variance of a lead-time demand (which is geometric with

parameter $r = \rho/(N - N\rho + \rho)$) in a symmetric distributed production-inventory system. Observe that, as expected,

$$\sigma_p = \sigma_d \sqrt{N + \sum_{i \neq j} \text{Corr}(Q_i, Q_j)},$$

because $\text{Corr}(Q_i, Q_j) = r$ for symmetric systems by (5).

In addition, we have

$$\zeta = NK\sigma_d / \sqrt{N - N\rho + \rho} = \zeta_N / \sqrt{N - N\rho + \rho}.$$

Hence, the approximations suggest that

$$TC_N^*/TC^* \approx \zeta_N/\zeta = \sqrt{N - N\rho + \rho},$$

which is decreasing in ρ with $\lim_{\rho \uparrow 1} \zeta_N/\zeta = 1$. Furthermore,

$$\lim_{\rho \uparrow 1} \zeta_N - \zeta = K(N - 1)/2.$$

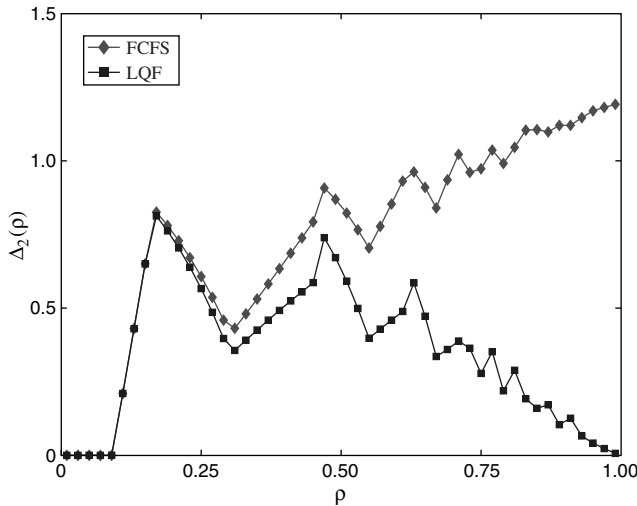
So, using the normal approximation with correct moments allows us to recover the essential asymptotic behavior described in Proposition 3. The results from the normal approximation confirm those we obtained using exact analysis, and also clearly highlight the roles of variance and correlation as implicit determinants of the benefits of pooling (when viewed as a function of system loading) for production-inventory systems.

2.4. The Longest-Queue-First Policy

The results in §2.2 rely on the assumption of FCFS processing at the production facility. Our focus on FCFS is motivated by its widespread use in practice, its ease of implementation, its perceived fairness, and its analytical tractability. However, it is of interest to know the extent to which results of §2.2 apply if optimal policies are used. For a pooled system, the base-stock policy with FCFS is optimal; see de Véricourt et al. (2002a) and references therein. Therefore, it follows that the limit expressions in Propositions 3(a) and 3(b) are, respectively, an upper bound and an equality for the benefit of pooling in relation to a distributed system that operates under an optimal production policy. Hence, qualitatively, the limit results remain valid in the sense that ratio approaches one and the difference remains bounded.

Completely characterizing an optimal policy for a distributed system is a difficult problem that to date remains unresolved for the general case; see de Véricourt et al. (2000a). In principle, the problem can be formulated as a Markov decision process and solved numerically. However, such an approach is practically feasible only for small systems (e.g., with two items) and for relatively low utilization levels.

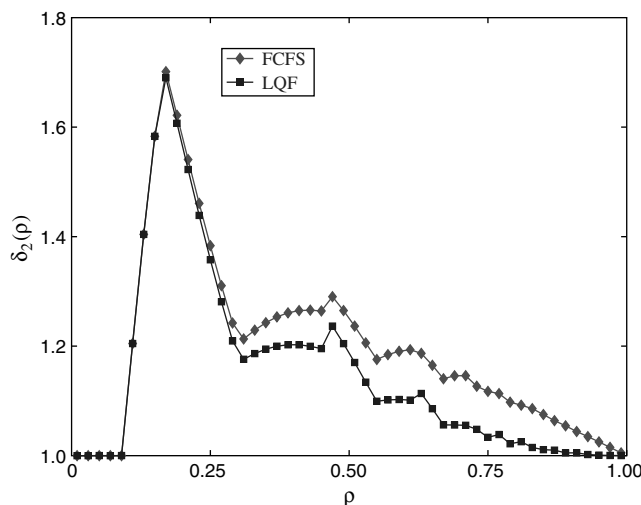
Figure 5 The Absolute Advantage of Pooling Under FCFS and LQF
 ($h = 1, b = 10, N = 2, \rho_i \equiv 1/2$)



For symmetric systems with $N = 2$, Zheng and Zipkin (1990) show that the “serve the longest queue first” (LQF) policy at the production facility is better than FCFS. They also provide an algorithm for computing the queue-length distributions under LQF for $N = 2$. Using their algorithm, we computed the steady-state costs and then chose the best base-stock policy for the LQF scheduling policy. We then used the obtained optimal costs for LQF to evaluate the cost difference and the cost ratio between the distributed system under LQF and the pooled system. Figures 5 and 6 show these differences and ratios, and also show analogous quantities when the distributed system operates under FCFS.

As expected, there is less benefit, absolute or relative, to pooling when the distributed system is operating under the better scheduling policy. For low

Figure 6 The Relative Advantage of Pooling Under FCFS and LQF
 ($h = 1, b = 10, N = 2, \rho_i \equiv 1/2$)



and moderate values of ρ , the effect of the scheduling policy does not appear to be significant. However, for high utilization, the cost difference between the two policies becomes larger. Interestingly, for this example, there is virtually no cost difference between pooled systems and distributed systems under the LQF policy when utilization is very high. With the better policy, the benefit of pooling becomes even smaller with higher utilization and is eventually eliminated.

As an aside, it is interesting to point out that the choice of scheduling policy for the distributed system appears to play little role in determining the best base-stock level—among the 50 data points plotted in the figures, the base-stock levels for FCFS and LQF were identical 46 times. As shown in Table 2 of Van Houtum et al. (1997), base-stock levels for the two types of scheduling rules are also quite similar for systems with service-level objectives.

The results of this section suggest that managers may trade off the benefit of physically pooling inventory with that of improving production scheduling control. In particular, for high utilization levels it appears that a well-managed system with multiple locations will perform nearly as well as a system where inventory is pooled in a single location.

3. Extensions

In this section, we describe various extensions to the basic model.

3.1. Systems with Service-Level Constraints

Instead of using backordering costs, it is more meaningful in some settings to require that the probability of stocking out remains bounded below a certain desirable level, which we refer to as the service level. This would be the case, for example, when it is difficult to quantify backordering costs or when a service level is mandated by the customer. Let $B_i(s_i)$ denote a random variable equal in distribution to the steady-state backorder level for item i when item i operates with a base-stock level of s_i . (Recall from §2.2 that the distribution of $B_i(s_i)$ does not depend upon $s_j, j \neq i$.) Keeping in mind the PASTA (Poisson Arrivals see Time Averages) property, our objective is to minimize inventory holding cost subject to constraints

$$P(B_i(s_i) > 0) \leq \alpha_i \quad i = 1, \dots, N, \quad (6)$$

where $\alpha_i \in (0, 1]$ is a specified service level. Under the above Markovian assumptions the probability of a backorder can be obtained as

$$P(B_i(s_i) > 0) = 1 - P(B_i(s_i) = 0) = 1 - \sum_{n=0}^{s_i-1} (1 - r_i)r_i^n = r_i^{s_i}.$$

Expected inventory is increasing in s_i , thus, constraints (6) are always binding. Therefore, the optimal base-stock level for type i is obtained by finding the smallest integer that satisfies $r_i^{s_i} \leq \alpha_i$. This yields

$$\mathbf{s}^{(\alpha)} = (s_1^{(\alpha_1)}, \dots, s_N^{(\alpha_N)}), \quad \text{where } s_i^{(\alpha_i)} = \lceil \ln \alpha_i / \ln r_i \rceil.$$

Substituting into $TC_N(\mathbf{s})$ gives

$$TC_N^{(\alpha)} \equiv TC_N(\mathbf{s}^{(\alpha)}) \\ = \sum_{i=1}^N h \left[\lceil \ln \alpha_i / \ln r_i \rceil - \frac{r_i}{1-r_i} (1 - r_i^{\lceil \ln \alpha_i / \ln r_i \rceil}) \right]. \quad (7)$$

Expression (7) shares several of the qualitative characteristics of the optimal cost function of §2.2. There are key differences, however, which can best be seen when $\alpha_i = \alpha$ and $\lambda_i = \lambda/N$ for all i , in which case $r_i = r \equiv \rho/(N - N\rho + \rho)$ for all i . We make these assumptions for the remainder of this section.

Total cost $TC_N^{(\alpha)}$ is not, in general, increasing in N . Although this is largely caused by the integrality of the stock, which forces us to hold more inventory than is necessary to meet exactly the service-level constraint, the decrease in total cost due to an increase in N can sometimes be large. In contrast to the cost-based model of §2.2, there is no value of N at which we switch from a make-to-stock to a make-to-order mode of production. There is, however, a value

$$N_{\max}^{(\alpha)} = \rho(1 - \alpha)/(\alpha(1 - \rho)),$$

beyond which the optimal cost reduces to $TC_N^{(\alpha)} = Nh(1 - r)$, which is almost linearly increasing in N for large N . Hence pooling becomes highly desirable when N is large. These effects are driven by the fact that for any nontrivial value of α , the base-stock levels must be at least one to satisfy the service-level requirement (a zero base-stock level leads to a zero service level). In general, the integrality of inventory levels can result in a choice of base stock that offers a higher service level than what is strictly required. This effect is compounded when N is large.

To measure the benefits of pooling, we compare the distributed system to the pooled system, whose optimal base-stock level is $s^{(\alpha)} = \lceil \ln \alpha / \ln \rho \rceil$ and whose total cost is $TC^{(\alpha)} \equiv TC(s^{(\alpha)})$. We consider the cost difference $\Delta_N^{(\alpha)} \equiv TC_N^{(\alpha)} - TC^{(\alpha)}$ and the cost ratio $\delta_N^{(\alpha)} \equiv TC_N^{(\alpha)} / TC^{(\alpha)}$. Our first result of the section states that pooling is indeed beneficial in the service-level context.

PROPOSITION 4. Suppose $\alpha_i = \alpha$ and $\lambda_i = \lambda/N$ for $i = 1, \dots, N$. Then, $\Delta_N^{(\alpha)} \geq 0$.

PROOF. The key idea of the proof is to show that the base-stock level

$$s^+ \equiv \sum_{i=1}^N s_i^{(\alpha)} = Ns_1^{(\alpha)}$$

(by symmetry $s_1^{(\alpha)} = \dots = s_N^{(\alpha)}$) satisfies the service-level constraint for the pooled system; i.e., that $\rho^{s^+} \leq \alpha$. The optimal base-stock level for each location in the distributed system satisfies $r^{s_i^{(\alpha)}} \leq \alpha$. Hence, it suffices to prove that

$$\rho^N \leq \rho/(N - N\rho + \rho).$$

After some work, it follows that this is equivalent to $N\rho^{N-1} \leq \sum_{k=0}^{N-1} \rho^k$, which is true, since $\rho < 1$. \square

For

$$\rho \leq \rho^{(\alpha)} \equiv N\alpha/(\alpha N - \alpha + 1),$$

we have $s^{(\alpha)} = 1$, and consequently, $\delta_N^{(\alpha)}(\rho) = N$. This is different from the cost-based model, where $\delta_N = 1$ for sufficiently small ρ . However, for large ρ the two models behave similarly; $\lim_{\rho \uparrow 1} \delta_N^{(\alpha)} = 1$, and $\Delta_N^{(\alpha)}$ remains bounded in a left-neighborhood of 1, although interestingly $\lim_{\rho \uparrow 1} \Delta_N^{(\alpha)}$ does not, in general, exist. In addition to being affected by ρ and N , the benefits of pooling also depend upon α , as described in the following result.

PROPOSITION 5. Suppose $N \geq 2$, and $\alpha_i = \alpha$ and $\lambda_i = \lambda/N$ for $i = 1, \dots, N$. Then,

- (a) $\lim_{\alpha \downarrow 0} \Delta_N^{(\alpha)} = \infty$,
- (b) $\lim_{\alpha \downarrow 0} \delta_N^{(\alpha)} = N \ln \rho / \ln r$.

PROOF. For part (b), successive applications of l'Hospital's rule give

$$\lim_{\alpha \downarrow 0} \delta_N^{(\alpha)} = \lim_{\alpha \downarrow 0} N \frac{1/\alpha \ln r + r/(1-r)}{1/\alpha \ln \rho - \rho/(1-\rho)} = \lim_{\alpha \downarrow 0} N \frac{1/\alpha^2 \ln r}{1/\alpha^2 \ln \rho}.$$

Evaluating the final expression yields (b). Part (a) follows from part (b), because $N \ln \rho / \ln r > 1$ for $N \geq 2$ and both $TC_N^{(\alpha)}$ and $TC^{(\alpha)}$ grow to ∞ as $\alpha \downarrow 0$. \square

It is interesting to note that the effect of α is different from that of ρ described in §2.2. Letting α approach 0 causes costs in both the distributed and pooled systems to increase without bound. In §2.2 we saw a similar phenomenon when we let ρ approach 1. However, when service levels become progressively tighter, Proposition 5 shows that the absolute cost benefit of pooling grows without bound, whereas Proposition 3 shows that as utilization becomes progressively higher, the cost benefit of pooling does remain bounded. For the cost-based systems in Proposition 3, larger values of ρ induce higher pairwise correlations for lead-time demands, thereby dampening the beneficial effects of pooling, even as costs increase to ∞ in both the pooled and distributed systems. In Proposition 5, however, changes in α do not affect lead-time demand distributions, and hence there is no such correlation-induced dampening of pooling effects.

The results of this section suggest that the benefits from inventory pooling can be more significant in systems with service-level requirements than in cost-based systems, particularly when either N is

large or ρ is small. The savings are also significant when service-level requirements are high.

3.2. Systems with Non-Markovian Demand and General Processing Times

In this section, we relax the assumptions of Poisson demand and exponential service times to examine the impact of variability. Initially, one might expect that as either demand or process variability increases, the value of pooling would also. In fact, we show that the relative advantage to pooling is largely decreasing in variability and in the limiting case of “very high variability,” the relative advantage to pooling disappears while the absolute benefit approaches a finite bound.

3.2.1. The Effect of Demand Variability. To isolate the effect of demand variability, we consider the following setup. Demand occurs one unit at a time according to a renewal process with mean interarrival time λ^{-1} . Independent of everything else, each arrival demands exactly one of the N items. The probability that a demand wants item i is p_i ; $i = 1, \dots, N$. When viewed in isolation the production system behaves like a $GI/M/1$ queue. The distribution of the total numbers of customers in a $GI/M/1$ queue in steady state is given by

$$P(Q = n) = \rho(1 - \beta)\beta^{n-1} \quad \text{for } n \geq 1$$

and

$$P(Q = 0) = 1 - \rho,$$

where β is the unique solution in $(0, 1)$ of the equation $\beta = \tilde{G}(\mu - \mu\beta)$, and \tilde{G} is the Laplace-Stieltjes transform of the interarrival distribution. Letting

$$r_i(\beta) = p_i\beta/(1 - \beta + p_i\beta)$$

and arguing as in Buzacott and Shanthikumar (1993, p. 133), it follows that

$$P(Q_i = 0) = 1 - \frac{\rho r_i(\beta)}{\beta} \quad \text{and}$$

$$P(Q_i = n) = \frac{\rho}{\beta}(1 - r_i(\beta))(r_i(\beta))^n \quad n \geq 1. \quad (8)$$

An optimal individual base-stock level is

$$s'_i = \max\{0, \lfloor \ln(\gamma\beta/\rho) / \ln r_i(\beta) \rfloor\},$$

and the optimal expected cost is

$$TC_N^V = \sum_{i=1}^N \mathcal{C}(r_i(\beta), \rho/\beta, s'_i).$$

Let TC^V be the optimal total cost in the pooled system.

One method for comparing the variability of random variables is the *convex order*—see Shaked and Shanthikumar (1994). A random variable A is smaller than random variable B in the convex order (written $A \leq_{cx} B$) if $Ef(A) \leq Ef(B)$ for all convex f for which the expectations exist. Observe that $A \leq_{cx} B$ implies that $EA = EB$ and $\text{Var}(A) \leq \text{Var}(B)$. If we compare two separate $GI/M/1$ systems with common service rates

in which the interarrival distribution of system 1 is smaller than that of system 2 according to the convex order, then (see, e.g., §11-5 of Wolff 1989) the respective solutions β_1, β_2 of $\beta = \tilde{G}(\mu - \mu\beta)$ for the two systems satisfy $\beta_1 \leq \beta_2$. Benjaafar and Kim (2004) study effects of convexly ordered interarrival distributions on lead-time demand and safety stock for production-inventory systems.

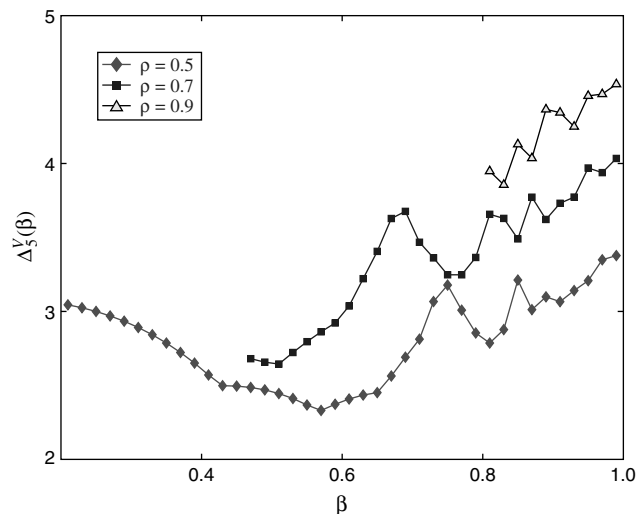
We will examine the effect of demand variability in the production-inventory setting by keeping ρ fixed, but varying β in the expressions above. In this scheme, a larger β can be interpreted as corresponding to *more-variable* interarrival times. (For a specific example, the Gamma distribution with shape parameter a_1/t and scale parameter a_2/t has an expected value of a_1/a_2 and variance of a_1t/a_2 . Hence, varying t allows us to fix the mean and change the variance. Furthermore, a larger t does correspond to being larger in the convex order, and hence to a larger β . In fact, $\beta \rightarrow 1$ as $t \rightarrow \infty$ and $\beta \rightarrow \beta_{\min}$ as $t \rightarrow 0$ where β_{\min} is the solution to $\beta = \tilde{G}(\mu - \mu\beta)$ for deterministic interarrival times of a_1/a_2 .)

Returning to the issues at hand, fix ρ and let

$$\Delta_N^V \equiv TC_N^V - TC^V \quad \text{and} \quad \delta_N^V \equiv TC_N^V/TC^V.$$

The variability parameter β plays a similar role to ρ . In particular, δ_N^V is largely decreasing in β , although there is no monotonicity because of the integrality of base-stock levels. In the limit, as β approaches 1, the ratio δ_N^V also goes to 1. Similarly, the difference Δ_N^V is largely increasing in β and approaches a finite bound as β approaches 1. The results are illustrated numerically in Figure 7. All values of $\beta \in (0, 1)$ are not shown, because when parameterizing according to \leq_{cx} , the smallest possible value of β (for a fixed $\rho = \lambda/\mu$) is β_{\min} for deterministic interarrival times of $1/\lambda$. The limit expressions are summarized

Figure 7 The Effect of Demand Variability on Absolute Advantage of Pooling ($h = 1, b = 10, N = 5, p_i \equiv 1/N$)



below. We omit the proof, which is similar to that of Proposition 3.

PROPOSITION 6. For each $N \in \mathbb{Z}^+$ and $\rho \in (0, 1)$, we have

- (a) $\lim_{\beta \uparrow 1} \Delta_N^V(\beta) = \frac{1}{2}h(N-1)\ln(\rho/\gamma)$,
- (b) $\lim_{\beta \uparrow 1} \delta_N^V(\beta) = 1$.

The interpretation of Proposition 6 is similar to that given for the effect of utilization. Higher demand variability increases both the mean and variance of queue size for each demand stream, but it also increases the correlation between queue sizes. Hence, the effect of variability parallels that of utilization in §2.2. As we will see shortly, however, under different assumptions on the structure of the supply system, we will obtain quite different behavior when viewing the benefit of pooling as a parameterized function in which an increase in the parameter causes an increase in the marginal queue-size distributions. In any case, managers need to be aware that higher variability in demand interarrival times induces greater correlation in the lead-time demands of individual items, which in turn tends to reduce the relative value of pooling.

3.2.2. The Effect of Service Time Variability. We now turn to the case where there are Poisson arrivals and processing times form an independent i.i.d. sequence. Here, the number of units on order of each type is equal in distribution to the number in system in a multiclass $M/G/1$ queue. From the stationary distribution of number in system in an $M/G/1$ queue, it is possible to obtain various performance measures of interest for both pooled and distributed systems in a manner analogous to the one described in §2.2. Although numerical analysis is straightforward, an explicit expression for the mass function of the number in system in an $M/G/1$ queue is not available.

In what follows, to obtain some analytical insights, we approximate the queue size in an $M/G/1$ queue by a geometric distribution of the form

$$P(Q = n) \approx \rho(1 - \sigma)\sigma^{n-1} \quad \text{for } n \geq 1$$

and

$$P(Q = 0) = 1 - \rho, \quad \text{where } \sigma \equiv (EQ - \rho)/EQ$$

and EQ is the expected number customers in a steady-state $M/G/1$ queue; that is

$$EQ = \rho + (\lambda ES^2)/(2(1 - \rho)).$$

Such approximations are described in Buzacott and Shanthikumar (1993) and Tijms (1986). The approximation is exact when the processing times are exponentially distributed. Furthermore, the geometric approximation has been shown to be accurate in estimating the p th percentile of the distribution of queue

size when p is large (Tijms 1986). Because for our purposes only the tail probability is needed to obtain the optimal base-stock level (the optimal base-stock level is the solution of the critical fractile equation $P(Q \leq s) = b/(b + h)$) and because in most applications b/h is large, the corresponding approximation error is generally small.

As in (8) we now have

$$P(Q_i = 0) = 1 - r_i(\sigma) \quad \text{and}$$

$$P(Q_i = n) = \frac{\rho}{\sigma}(1 - r_i(\sigma))(r_i(\sigma))^n \quad n \geq 1$$

where $r_i(\sigma) = p_i\sigma/(1 - \sigma + p_i\sigma)$. Here, the parameter σ plays the same role that β plays in the model with non-Markovian demand; so, performance measures of interest can be obtained in a similar fashion. The effect of processing-time variability can be studied via the parameter σ by noting that for ρ fixed, σ is smaller when the processing-time distribution is smaller in the convex ordering (because EQ is larger; see Wolff 1989). Hence, the effect of σ on the cost difference and cost ratio of distributed to pooled systems is similar to the effect, described earlier, of β . For brevity, the details are omitted.

3.3. Systems with Multiple Production Facilities

In certain applications, each inventory location in the distributed system is associated with its own production facility. In these settings, inventory pooling may represent the replacement of factory warehouses with a single shared warehouse. To study this form of pooling, we consider a system with N production facilities, each with capacity μ_i . In the distributed case, each production facility is dedicated to one of the items/inventory locations. In the pooled case, we consider three possible scenarios: (a) the N inventory locations are replaced by a single one but replenishment orders for each demand stream are still placed with the original facility (i.e., inventory is pooled, but not capacity); (b) the N inventory locations are replaced by a single one and replenishment orders, regardless of their origin, are processed by the first available facility and all facilities pull jobs from a common queue (i.e., both inventory and capacity are pooled)—see Figure 8; and (c) the N production facilities are consolidated as in (b), but the N inventory locations are kept distinct (i.e., capacity is pooled, but not inventory).

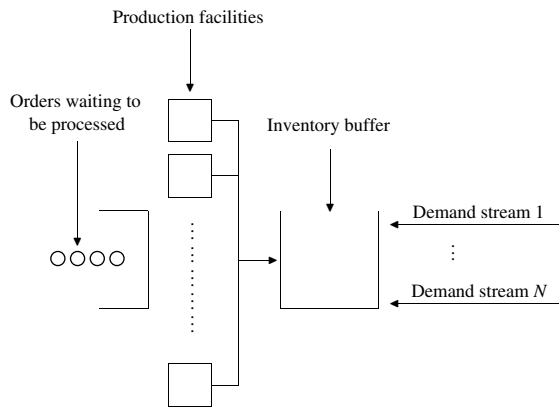
Under the Markovian assumptions on demand and production times as in §2.2, the optimal total cost for the distributed system is

$$TC_N^F = \sum_{i=1}^N \mathcal{C}(\rho_i, 1, s_i^+)$$

where

$$s_i^+ = \lceil \ln \gamma / \ln \rho_i \rceil \quad \text{and} \quad \rho_i = \lambda_i / \mu_i.$$

Figure 8 Pooled Scenario (b)



Notes. There are N demand streams that all draw inventory from a single buffer. Replenishment orders are sent to a common queue, from which the N production facilities pull jobs.

The numerical examples throughout the remainder of this section assume that the systems in question are symmetric; that is, $\lambda_i = \lambda/N$ and $\mu_i = \mu_0 = \mu/N$ for all i , so

$$\rho_i = \lambda_i/\mu_i = \lambda/\mu = \rho \quad \text{for all } i.$$

3.3.1. Inventory Pooling Without Capacity Pooling. For pooled scenario (a), we have a single inventory location supplied by N production facilities. If a demand originates from stream i , then the replenishment order is placed with production facility i . Hence, in the pooled system each facility i still sees the same Poisson arrival process with the same rate λ_i as in the unpooled system, and the individual facilities can still be viewed as independent $M/M/1$ queues. In practice, this may correspond to plant warehouses being replaced by a centralized warehouse, while customer demand is still handled by regional sales offices that simultaneously place a shipping order with the warehouse and a production order with the plant. Alternatively, this may correspond to systems where inventory is virtually pooled, although inventory is still physically stored in multiple plants where initial customer orders are processed and production orders placed. Note that the model is also applicable to systems where incoming production orders are assigned probabilistically to one of the plants.

Let $\tilde{Q} = (\tilde{Q}_1, \dots, \tilde{Q}_N)$ be a random vector whose joint distribution is that of the steady-state number of jobs on order at the facilities. Since the entries of \tilde{Q} are independent geometric random variables, the probability mass function of the total number of jobs on order, $\tilde{Q}_{(a)} \equiv \sum_{i=1}^N \tilde{Q}_i$, is given by

$$P(\tilde{Q}_{(a)} = n) = \sum_{i=1}^N \frac{\rho_i^{n+N-1} \prod_{l=1}^N (1 - \rho_l)}{\prod_{k:k \neq i} (\rho_i - \rho_k)}, \quad (9)$$

provided $\rho_i \neq \rho_k$ for all $i \neq k$ (for a derivation, see, Rubio and Wein 1996, p. 263). In a balanced system

with $\rho_i = \rho$ for $i = 1, \dots, N$, the total number of jobs on order has the negative binomial distribution.

Let $I_{(a)}$ and $B_{(a)}$ denote, respectively, the amount of inventory and the number of backorders in pooled scenario (a). The expected inventory is given by

$$EI_{(a)} = \sum_{n=0}^s (s - n) P(\tilde{Q}_{(a)} = n),$$

and the expected number of backorders is given by

$$EB_{(a)} = E\tilde{Q}_{(a)} + EI_{(a)} - s, \quad \text{where } E\tilde{Q}_{(a)} = \sum_{i=1}^N \rho_i / (1 - \rho_i).$$

An optimal base-stock level is given by the smallest integer s that satisfies

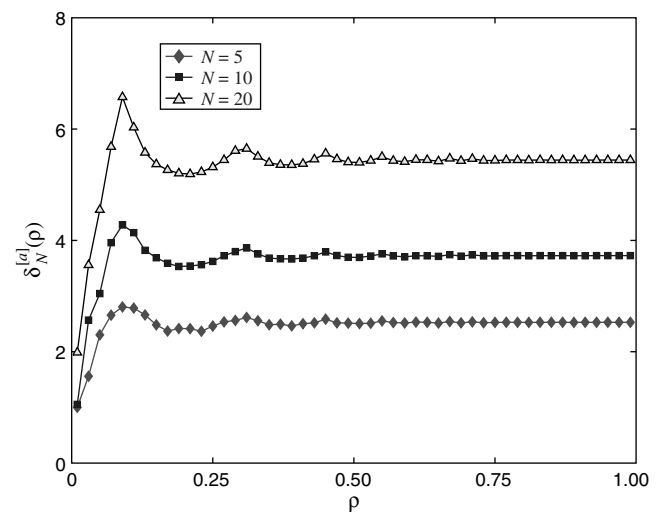
$$TC_{(a)}^F(s + 1) - TC_{(a)}^F(s) \geq 0,$$

where

$$TC_{(a)}^F(s) = hEI_{(a)} + bEB_{(a)}.$$

This is equivalent to choosing the smallest integer s , which we denote $s_{(a)}^*$, that satisfies $P(\tilde{Q}_{(a)} \leq s) \geq 1 - \gamma$. Although we do not have an explicit expression for $s_{(a)}^*$ and the corresponding optimal cost, it is straightforward to compute them numerically from (9) or the negative binomial mass function. Figure 9 shows the cost ratio of distributed to pooled systems for different levels of utilization and different numbers of items (for ease of illustration, the results are for balanced systems). As depicted, the ratio is essentially independent of ρ for large ρ . Note also that because the costs of the pooled and distributed systems both grow to infinity as ρ approaches 1, the fact that the ratio in Figure 9 is approaching a limit greater than 1 implies that the difference is unbounded in ρ . These results are similar to those observed in Eppen's model with independent demands.

Figure 9 The Effect of Utilization on Relative Advantage of Pooling in Scenario (a) ($h = 1, b = 10, \rho_i \equiv 1/N$)



The difference between these effects and those observed in earlier sections appears to be related to the degree to which lead-time demand correlation is present. In the original model with a single shared facility, there is positive correlation in the lead-time demands in the distributed system. This correlation increases with increases in utilization. In contrast, in systems where inventory locations are associated with independent facilities, lead-time demands are uncorrelated. These results suggest that how capacity is shared among the individual demand streams can have a significant impact on both the relative and absolute advantage of inventory pooling. Managers need to be aware of these subtle but important differences when estimating the impact of a pooling strategy.

3.3.2. Inventory Pooling with Capacity Pooling.

For pooled scenario (b), both inventory and capacity are pooled so that a replenishment order can be placed with any facility, regardless of its origin. In this case, it is best to postpone the assignment of orders to facilities until at least one of the facilities is available. In this section we assume that production facilities are identical (i.e., all servers have rate μ_0), so it is optimal under a base-stock policy to assign orders to the first available facility. In light of the Markovian assumptions on the demand process and production times, it follows that the production system is an $M/M/N$ queue. See Figure 8. Therefore, the distribution of jobs on order is given by

$$P(\tilde{Q}_{(b)} = n) = \begin{cases} \frac{p_0(N\rho)^n}{n!} & \text{if } n < N, \\ \frac{p_0\rho^n N^N}{N!} & \text{otherwise,} \end{cases} \quad (10)$$

where

$$p_0 = P(\tilde{Q}_{(b)} = 0) = \left[\frac{(N\rho)^N}{(1-\rho)N!} + \sum_{j=0}^{N-1} \frac{(N\rho)^j}{j!} \right]^{-1},$$

and $\rho = \lambda/(N\mu_0)$; see, e.g., Wolff (1989, p. 256).

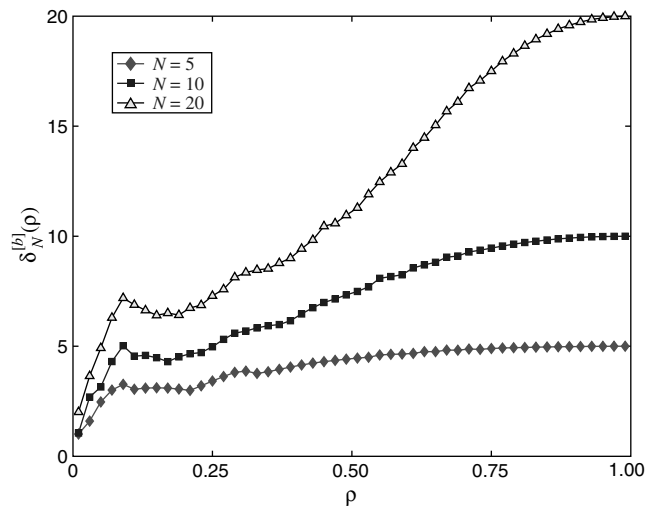
We are again interested in minimizing the long-run average total cost, $TC_{(b)}^F(s) = hEI_{(b)} + bEB_{(b)}$ over base-stock levels. For $s \geq N$, we obtain from (10) the following explicit expression for the total cost as a function of base-stock level s :

$$TC_{(b)}^F(s) = h \left[s - \frac{\rho^{s+1} - \rho^{N+1}}{\theta(1-\rho)} - N\rho \right] + b \left[\frac{\rho^{s+1}}{\theta(1-\rho)} \right] \quad s \geq N, \quad (11)$$

where $\theta = [N!(1-\rho)]/[N^N p_0]$. When $s < N$, it is also straightforward to compute $TC_{(b)}^F(s)$ exactly using (10), because

$$EI_{(b)} = \sum_{n=0}^s (s-n)P(\tilde{Q}_{(b)} = n), \quad EB_{(b)} = E\tilde{Q}_{(b)} + EI_{(b)} - s,$$

Figure 10 The Effect of Utilization on Relative Advantage of Pooling in Scenario (b) ($h=1, b=10, \rho_i \equiv 1/N$)



and the mean number in system for the $M/M/N$ queue is

$$E\tilde{Q}_{(b)} = \rho^{N+1}/[\theta(1-\rho)] + N\rho.$$

For the following, let $TC_{(b)}^F = TC_{(b)}^F(s_{(b)})$, where $s_{(b)}$ is a minimizer of $TC_{(b)}^F(\cdot)$. As before, we consider the cost ratio $\delta_N^{[b]} = TC_N^F/TC_{(b)}^F$ and the cost difference $\Delta_N^{[b]} = TC_N^F - TC_{(b)}^F$ between unpooled and pooled systems. Figure 10 shows the cost ratio of distributed to pooled systems for different levels of utilization and different numbers of items. Both the cost ratio and the cost difference are generally increasing in ρ . The ensuing proposition describes the behavior of the ratio and difference for heavily loaded systems.

PROPOSITION 7. For each $N \geq 2$ we have

- (a) $\lim_{\rho \uparrow 1} \Delta_N^{[b]}(\rho) = \infty$,
- (b) $\lim_{\rho \uparrow 1} \delta_N^{[b]}(\rho) = N$.

PROOF. To prove part (b), observe that for ρ sufficiently large (with N fixed), we will have $s_{(b)} \geq N$. Proof of this assertion follows from the facts that

$$TC_{(b)}^F = bE\tilde{Q}_{(b)} + hs - (h+b)E \min\{s, \tilde{Q}_{(b)}\}$$

and $E \min\{s, \tilde{Q}_{(b)}\} \rightarrow s$ pointwise in s as $\rho \uparrow 1$. Therefore, a base-stock level that minimizes $TC_{(b)}^F(s)$ over all s is given by a base-stock level that minimizes (11) over $s \geq N$. Such an s can be found as the smallest integer for which $TC_{(b)}^F(s+1) > TC_{(b)}^F(s)$, which yields $s_{(b)} = \lfloor \ln(\theta\gamma)/\ln\rho \rfloor$.

As before, when computing $\lim_{\rho \uparrow 1} \delta_N^{[b]}$ we can, without loss of generality, ignore the integrality of base-stock levels in the evaluation of $TC_{(b)}^F$ and TC_N^F . So,

$$\lim_{\rho \uparrow 1} \delta_N^{[b]} = \lim_{\rho \uparrow 1} \frac{Nh \ln \gamma / \ln \rho}{h[\ln(\theta\gamma)/\ln\rho - N\rho + \rho(\theta - \rho^N)/(\theta(1-\rho))]}.$$

Working with the reciprocal of the expression on the right-hand side above, and applying L'Hopital's rule yields $\lim_{\rho \uparrow 1} 1/\delta_N^{[b]} = 1/N$, thereby completing the proof of (b). Both TC_N^F and $TC_{(b)}^F$ grow to infinity as $\rho \uparrow 1$; consequently, part (b) implies part (a). \square

An alternative form of capacity pooling is one where upon consolidation, the N production facilities are replaced by a single facility with one server working at rate $N\mu_0 = \mu$. Here, the optimal long-run average cost for the pooled system is the same as that of the pooled system in §2.2, and therefore the ratio of the cost of the distributed system to that of the pooled system is N , regardless of the utilization level. This result is consistent with the limiting behavior seen in scenario (b) as $\rho \uparrow 1$, since in heavy traffic an $M/M/N$ queue with rate μ/N per server behaves essentially the same as a $M/M/1$ queue with server rate μ .

The analysis of this section further confirms that the manner in which capacity is shared among demand streams affects the relative and absolute advantage of inventory pooling. In contrast to the previous situations, when inventory pooling is accompanied by capacity pooling as in scenario (b), the cost difference and the cost ratio are increasing in ρ , with limits different from those observed in earlier sections. In the next section, we provide an interpretation of these results and contrast the effects of inventory pooling and capacity pooling.

3.3.3. Capacity Pooling Without Inventory Pooling. In scenario (c), the production facilities are consolidated so that they draw from a common queue as in scenario (b). However, inventory locations are kept distinct. For a system with N identical facilities, this means replacing the N separate $M/M/1$ production systems by a single $M/M/N$ multiclass FIFO production system. As we did for the previous scenarios, we define $TC_{(c)}^F = TC_{(c)}^F(s_{(c)})$, where $s_{(c)}$ is a minimizer of the scenario (c) cost function $TC_{(c)}^F(\cdot)$. Without loss of optimality, we assume that the base-stock level for each inventory location is the same. Let

$$\delta_N^{[c]} = TC_N^F/TC_{(c)}^F, \quad \text{and} \quad \Delta_N^{[c]} = TC_N^F - TC_{(c)}^F.$$

To develop the scenario (c) cost function, observe that the steady-state total number of jobs in the production facility $\tilde{Q}_{(c)}$ has the same distribution (10) as $\tilde{Q}_{(b)}$ from scenario (b). Let \hat{Q} have the steady state distribution of the number of type i jobs in the production system. By symmetry, the distribution of \hat{Q} does not depend upon i . Moreover, conditional upon $\tilde{Q}_{(c)} = n$, the number of type i jobs \hat{Q} has the binomial distribution with parameters n and $p = 1/N$. For $k \geq N$, it follows that

$$P(\hat{Q} = k) = r^k(1-r)/\theta \quad \text{where} \quad r = \rho/(N - N\rho + \rho).$$

For $k < N$, we can evaluate the mass function of \hat{Q} numerically by

$$P(\hat{Q} = k) = \sum_{n \geq k} b_{n,k} P(\tilde{Q}_{(c)} = n), \quad \text{where} \quad b_{n,k} = P(Z = k),$$

and Z is binomial with parameters n and $p = 1/N$. If type i inventory is managed according to base-stock level $s \geq N$, we can obtain the following explicit expression for the total cost

$$TC_{(c)}^F(s) = N \left[h \left(s - \frac{\rho^{N+1}}{\theta N(1-\rho)} - \rho + \frac{r^{s+1}}{\theta(1-r)} \right) + b \left(\frac{r^{s+1}}{\theta(1-r)} \right) \right]. \quad (12)$$

If $s < N$, it is also straightforward to compute $TC_{(c)}^F(s)$ exactly as in scenario (b) using the probability mass function of \hat{Q} , because

$$E\hat{I} = \sum_{k=0}^s (s-k)P(\hat{Q} = k), \quad E\hat{B} = E\hat{Q} + E\hat{I} - s,$$

and

$$E\hat{Q} = E\tilde{Q}_{(c)}/N = E\tilde{Q}_{(b)}/N.$$

Figure 11 shows the cost ratio of distributed to pooled systems for different levels of utilization and different numbers of items. The cost difference, not shown, increases without bound as ρ approaches one. These results are formally stated below. We omit the proof, which is similar to that of Proposition 7, and rests on ignoring the integrality of $\lceil \ln(\theta\gamma)/\ln(r) \rceil$, which minimizes (12).

PROPOSITION 8. For each $N \geq 2$, we have

- (a) $\lim_{\rho \uparrow 1} \Delta_N^{[c]}(\rho) = \infty$,
- (b) $\lim_{\rho \uparrow 1} \delta_N^{[c]}(\rho) = N$.

Figure 11 The Effect of Utilization on Relative Advantage of Pooling in Scenario (c) ($h = 1, b = 10, p_i = 1/N$)

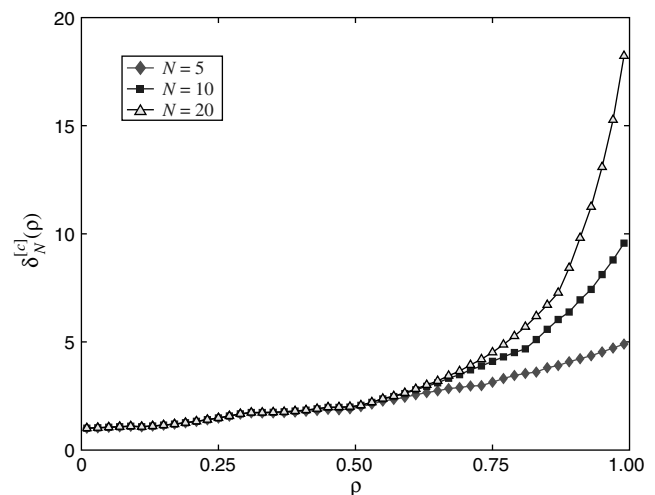
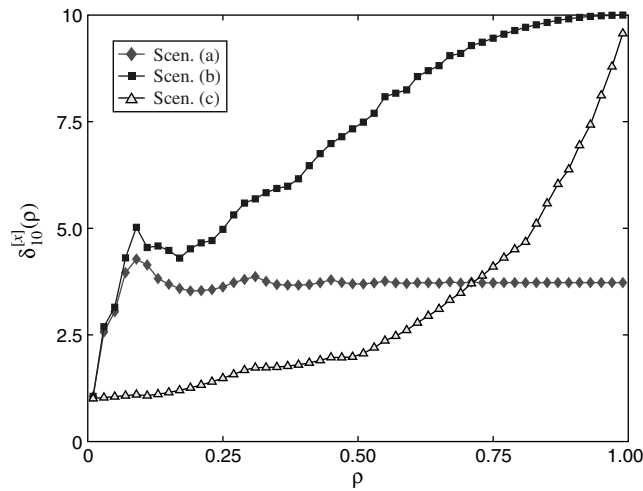


Figure 12 The Relative Advantage of Pooling in Scenarios (a), (b), and (c) ($h = 1, b = 10, N = 10, \rho_i \equiv 1/N$)



Scenarios (a), (b), and (c) are compared in Figure 12. For very high utilization, the relative advantage of capacity pooling is nearly the same as that of the joint capacity and inventory pooling in scenarios (b). Note, however, that utilization must indeed be quite high for this to be the case; at $\rho = 0.93$, scenario (b) offers a roughly 25% larger relative advantage of pooling than does scenario (c), and at $\rho = 0.99$ (the highest value of ρ shown) scenario (b) gives about 4% larger relative benefit than (c). In addition, modest differences in relative advantage can translate into large differences in absolute advantage for high values of ρ , because absolute costs are very high.

For an alternative form of capacity pooling, consider scenario (c'), where inventories remain distinct, but the N production facilities are replaced by a single facility with one server that is N times faster. This leads to results that are qualitatively the same as those for scenario (c) in the sense that the limits for the absolute and relative benefit of pooling as $\rho \uparrow 1$ are the same as in Proposition 8.

Our results suggest that (i) the effect of capacity pooling is more significant than that of inventory pooling and (ii) the relative benefit of capacity pooling increases with utilization. The latter is due to the fact that more efficient use of capacity matters more when system loading is high. In particular, variance of lead-time demand (which tends to drive inventory-related costs) is observed to be decreasing in N . For example, in scenario (c') the standard deviation of lead-time demand for each inventory location is given by

$$\sqrt{r/(1-r)^2} = \sqrt{\rho(N - N\rho + \rho)/(N - N\rho)^2}.$$

For the distributed system, the standard deviation of lead-time demand for each location is $\sqrt{\rho/(1-\rho)^2}$.

Hence, capacity pooling reduces standard deviation by a factor of

$$N/\sqrt{N - N\rho + \rho}.$$

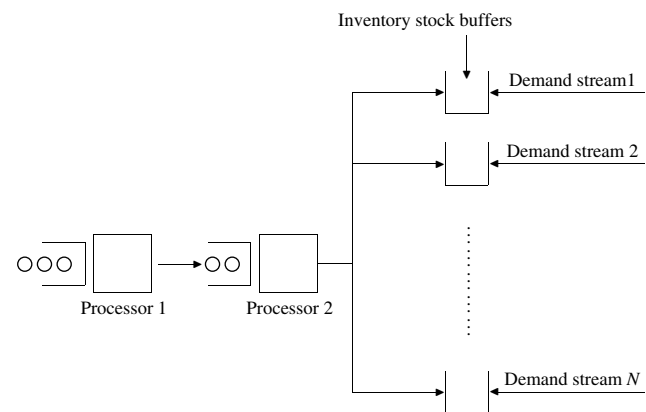
This factor is increasing in ρ with limit equal to N as ρ approaches 1.

In summary, the effect of ρ differs depending on whether we pool inventory, capacity, or both. The relative advantage of capacity pooling increases with utilization, whereas the relative advantage of inventory pooling, depending on the structure of the supply process (i.e., is capacity shared as in §2.2 or dedicated as in §3.3.1), either decreases or remains essentially invariant with utilization. In systems with very high utilization, pooling capacity alone achieves most of the relative benefit of pooling both capacity and inventory.

3.4. Systems with Multiple Production Stages

In this section, we investigate production systems with multiple stages to determine what effect the structure of the supply system has on the value of inventory pooling. For this, we return to a setup similar to that in §2, but now rather than having a single processor at the production facility, we have a series of M sequential processors—see Figure 13. As in §2 the production system does not change with pooling. Each order first goes to processor 1, then to processor 2, and so on through processor M , after which the order is sent back to the appropriate inventory buffer. Each processor works on a FCFS basis, and there are intermediate queues between processors. The external arrival process to the inventory buffers is identical to that described in §2. Processing times at processor j are assumed to be i.i.d. exponential random variables (independent of the arrivals and other processing time sequences) with mean μ_j^{-1} . Hence, the production system is a Jackson network

Figure 13 N -Item Production-Inventory System with $M = 2$ Stages



Notes. There are N inventory locations, each with its own demand stream and its own inventory buffer. Replenishment orders go first to processor 1 and then to processor 2, before being sent to the appropriate buffer.

with trivial routing probabilities. (Using techniques from Rubio and Wein 1996, similar analysis can be done when the supply system is a general open Jackson network.)

Assume that $\rho_j \equiv \lambda/\mu_j < 1$, let $\mathbf{Q} = (Q_1, \dots, Q_N)$ be the steady-state vector of type i jobs in the production facility, and define $Q = \sum_{i=1}^N Q_i$. Then, restricting ourselves to base-stock policies, the optimal base-stock levels for the distributed system are again given by the smallest integer that satisfies

$$P(Q_i \leq s_i) \geq 1 - \gamma.$$

Similarly, for the pooled system, the optimal base-stock level is given by the smallest integer so that

$$P(Q \leq s) \geq 1 - \gamma.$$

The total number of jobs at processor k is geometric with parameter ρ_k , the number of jobs of type i at processor k is geometric with parameter

$$r_{ik} = p_i \rho_k / (1 - \rho_k + p_i \rho_k),$$

and the stationary distribution has the well-known Jackson product form; see Buzacott and Shanthikumar (1993, §7.5) or Rubio and Wein (1996). As in (9), it follows that

$$P(Q = n) = \sum_{j=1}^M \frac{\rho_j^{n+M-1} \prod_{l=1}^M (1 - \rho_l)}{\prod_{k:k \neq j} (\rho_i - \rho_k)}, \quad \text{and}$$

$$P(Q_i = n) = \sum_{j=1}^M \frac{r_{ij}^{n+M-1} \prod_{l=1}^M (1 - r_{il})}{\prod_{k:k \neq j} (r_{ij} - r_{ik})}.$$

It is now simple to obtain performance measures of interest;

$$EI_i = \sum_{n=0}^{s_i} (s_i - n) P(Q_i = n),$$

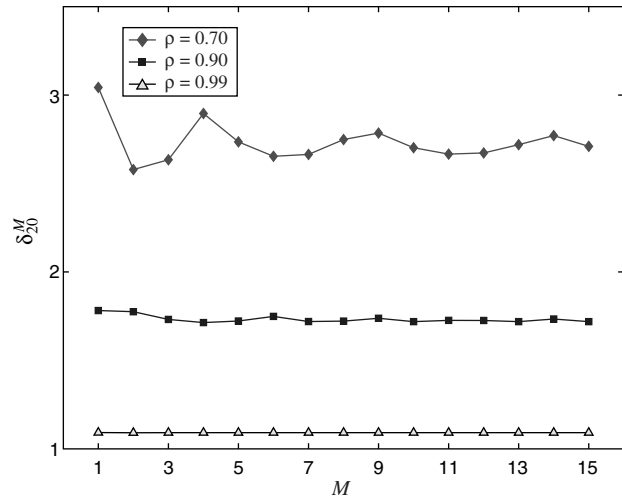
and

$$EB_i = EQ_i + EI_i - s_i, \quad \text{where } EQ_i = \sum_{k=1}^M r_{ik} / (1 - r_{ik}).$$

For balanced and symmetric systems (those with $p_i = 1/N$ for $i = 1, \dots, N$ and $\mu_j = \mu$ for $j = 1, \dots, M$) Q_i is the sum of i.i.d. geometric random variables, and hence has the negative binomial distribution. As described in Rubio and Wein (1996), the above expressions can be used to numerically compute optimal base-stock levels.

Using this approach, we are able to examine how various system parameters affect the value of pooling. For example, we have observed that the effect of utilization is consistent with our earlier results (i.e., the cost ratio is largely decreasing in the utilization of individual processors and approaches one in the limit,

Figure 14 The Effect of M on Relative Advantage of Pooling for Series Systems ($h = 1, b = 10, N = 20, p_i \equiv 1/N$)



while the ratio is largely increasing and approaches a finite bound). We can also examine the effect of the number of processors M on both the cost ratio and the cost difference. Note that an increase in M leads to both longer replenishment lead times and (stochastically) larger lead-time demands. In fact, increasing M increases both the mean and the variance of lead-time demand. It is tempting then to argue that increasing M will have a similar effect to increasing utilization. However, numerical results show this not to be true. As illustrated in Figure 14, the cost ratio δ_N^M remains effectively constant when viewed as a function of M , once M is large enough, and the cost difference increases without bound (not shown).

The difference between the effect of M and that of ρ appears to be related to the fact that an increase in M affects the correlation in lead-time demands differently than does an increase in ρ as in §2.2. In fact, an increase in M leaves correlation in lead-time demands unchanged. This is, of course, quite different than the effect seen when ρ grows in (5). To make this more precise, let $\text{Corr}_M(Q_i, Q_k)$ and $\text{Cov}_M(Q_i, Q_k)$ denote the correlation and covariance of Q_i and Q_k for $i \neq k$ when there are M processors in a balanced system (with $\mu_j = \mu$ for all j). Let Q_{ij} be the number of orders of type i at processor j . For $i \neq k$, define $\chi_{ik} = \text{Cov}(Q_{ij}, Q_{kj})$. This quantity does not depend on j , because the system is balanced. Similarly, let $\sigma_i^2 = \text{Var}(Q_{ij})$. We now have

$$\begin{aligned} \text{Cov}_M(Q_i, Q_k) &= \text{Cov}\left(\sum_{j=1}^M Q_{ij}, \sum_{j=1}^M Q_{kj}\right) \\ &= \sum_{j=1}^M \sum_{l=1}^M \text{Cov}(Q_{ij}, Q_{kl}) = M\chi_{ik}, \end{aligned}$$

where the final equality follows from the independence of Q_{ij} and Q_{kl} when $j \neq l$. Furthermore, from

the product-form of the stationary distribution, we have

$$\text{Var}(Q_i) = \sum_{j=1}^M \text{Var}(Q_{ij}) = M\sigma_i^2.$$

Hence,

$$\text{Corr}_M(Q_i, Q_k) = \chi_{ij}/(\sigma_i\sigma_j),$$

which is independent of M . The expression for the $\text{Corr}_M(Q_i, Q_k)$ highlights important differences between an increase in lead-time demand due to an increase in congestion (i.e., tight capacity or high variability) versus an increase in lead-time demand due to an increase in the number of processing steps. In the former, lead-time demands become perfectly correlated (see (5)), whereas in the latter correlation is unchanged.

4. Conclusion

We examined inventory pooling in production-inventory systems and showed that utilization, demand and process variability, control policy, service levels, and the structure of the production process all play a role in determining precisely how valuable pooling might be. We described how correlation in the lead-time demands explains differences between the various models. In particular, we showed that in a system where the supply process is shared, there can be significant correlation in the lead-time demands of the different items, even when the individual demand streams are independent. We showed that the amount of correlation is affected by utilization, demand and service time variability, and supply structure. Consequently, the value derived from pooling is also affected by these factors. We also compared inventory pooling and capacity pooling and showed that the effects of the two differ. While the relative benefit of inventory pooling tends to diminish with utilization, the relative benefit of capacity pooling tends to increase with utilization. For highly loaded systems, we showed that capacity pooling alone achieves nearly the same relative benefit as the joint pooling of capacity and inventory.

Acknowledgments

The authors acknowledge support from the National Science Foundation under grant DMI 9988721 and Honeywell Laboratories.

References

Alfaro, J. A., C. J. Corbett. 2003. The value of SKU rationalization in practice (the pooling effect under suboptimal inventory policies and nonnormal demand). *Production Oper. Management* 12 12–29.

Barnes, E., J. Dai, S. Deng, D. Down, M. Goh, H. C. Lau, M. Sharafali. 2000. Electronics manufacturing service industry. TLI-AP white paper, The Logistics Institute-Asia Pacific, National University of Singapore, Singapore.

Benjaafar, S., J.-S. Kim. 2004. On the effect of demand variability in production-inventory systems. Working paper, Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN.

Benjaafar, S., M. ElHafsi, F. de Véricourt. 2004. Demand allocation in multiproduct, multifacility make-to-stock systems. *Management Sci.* 50 1431–1448.

Bertsimas, D., I. Paschalidis. 2001. Probabilistic service level guarantees in make-to-stock manufacturing systems. *Oper. Res.* 49 119–133.

Buzacott, J. A., J. G. Shanthikumar. 1993. *Stochastic Models of Manufacturing Systems*. Prentice-Hall, Englewood Cliffs, NJ.

Eppen, G. D. 1979. Effects of centralization on expected costs in a multi-location newsboy problem. *Management Sci.* 25 498–501.

Gerchak, Y., Q.-M. He. 2003. On the relation between the benefits of risk pooling and the variability of demand. *IIE Trans.* 35 1027–1031.

Ha, A. Y. 1997. Optimal dynamic scheduling policy for a make-to-stock production system. *Oper. Res.* 45 42–53.

Kador, J. 2001. Contract manufacturing grows up. *Electronic Business Magazine* (September).

Kim, J.-S. 2001. Modeling and analysis of production-inventory systems. Ph.D. thesis, University of Minnesota, Minneapolis, MN.

Lu, Y., J.-S. Song, D. D. Yao. 2003. Order fill rate, leadtime variability, and advance demand information in an assemble-to-order system. *Oper. Res.* 51 292–308.

Netessine, S., G. Dobson, R. A. Shumsky. 2002. Flexible service capacity: Optimal investment and the impact of demand correlation. *Oper. Res.* 50 375–388.

Plambeck, E. L., T. A. Taylor. 2005. Sell the plant? The impact of contract manufacturing on innovation, capacity and profitability. *Management Sci.* 51(1) 133–150.

Rubio, R., L. M. Wein. 1996. Setting base stock levels using product-form queueing networks. *Management Sci.* 42 259–268.

Shaked, M., J. G. Shanthikumar. 1994. *Stochastic Orders and Their Applications*. Academic Press, New York.

Tijms, H. C. 1986. *Stochastic Modelling and Analysis: A Computational Approach*. John Wiley & Sons, Chichester, U.K.

Van Houtum, G.-J., I. Adan, J. Van Der Wal. 1997. The symmetric longest queue system. *Stochastic Models* 13 105–120.

Veatch, M. H., L. M. Wein. 1996. Scheduling a make-to-stock queue: Index policies and hedging points. *Oper. Res.* 44 634–647.

Véricourt, F. de, F. Karaesmen, Y. Dallery. 2000a. Dynamic scheduling in a make-to-stock system: A partial characterization of optimal policies. *Oper. Res.* 48 811–819.

Véricourt, F. de, F. Karaesmen, Y. Dallery. 2000b. Dynamic stock rationing in a make-to-stock queue: Optimal policies and some implications on delayed product differentiation. *INFORMS Manufacturing Service Oper. Management Conf. Proc.*, Ann Arbor, MI.

Véricourt, F. de, F. Karaesmen, Y. Dallery. 2001. Assessing the benefits of different stock-allocation policies for a make-to-stock production system. *Manufacturing Service Oper. Management* 3 105–121.

Véricourt, F. de, F. Karaesmen, Y. Dallery. 2002. Optimal stock allocation for a capacitated supply system. *Management Sci.* 48 1486–1501.

Wein, L. M. 1992. Dynamic scheduling of a multiclass make-to-stock queue. *Oper. Res.* 40 724–735.

Wolff, R. W. 1989. *Stochastic Modeling and the Theory of Queues*. Prentice-Hall, Englewood Cliffs, NJ.

Zheng, Y.-S., P. Zipkin. 1990. A queueing model to analyze the value of centralized inventory information. *Oper. Res.* 38 296–307.

Zipkin, P. H. 1995. Performance analysis of a multi-item production-inventory system under alternative policies. *Management Sci.* 41 690–703.

Zipkin, P. H. 2000. *Foundations of Inventory Management*. McGraw-Hill, New York.