

# Now or Later: A Simple Policy for Effective Dual Sourcing in Capacitated Systems

Senthil Veeraraghavan

The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104,  
senthilv@wharton.upenn.edu

Alan Scheller-Wolf

Tepper School of Business, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213,  
awolf@andrew.cmu.edu

We examine a possibly capacitated, periodically reviewed, single-stage inventory system where replenishment can be obtained either through a regular fixed lead time channel, or, for a premium, via a channel with a smaller fixed lead time. We consider the case when the unsatisfied demands are backordered over an infinite horizon, introducing the easily implementable, yet informationally rich dual-index policy. We show very general separability results for the optimal parameter values, providing a simulation-based optimization procedure that exploits these separability properties to calculate the optimal inventory parameters within seconds. We explore the performance of the dual-index policy under stationary demands as well as capacitated production environments, demonstrating when the dual-sourcing option is most valuable. We find that the optimal dual-index policy mimics the behavior of the complex, globally optimal state-dependent policy found via dynamic programming: the dual-index policy is nearly optimal (within 1% or 2%) for the majority of cases, and significantly outperforms single sourcing (up to 50% better). Our results on optimal dual-index parameters are generic, extending to a variety of complex and realistic scenarios such as nonstationary demand, random yields, demand spikes, and supply disruptions.

*Subject classifications:* inventory/production; infinite horizon; policies: review/lead times; dual index; dual supply; uncertainty: stochastic.

*Area of review:* Manufacturing, Service, and Supply Chain Operations.

*History:* Received January 2005; revision received January 2006; accepted May 2006.

## 1. Introduction

Many firms are trying to construct supply chains that reduce costs while maintaining customer service, often by incorporating alternatives with respect to sourcing (either by using different suppliers or different modes of delivery with a single supplier). Usually, a supplier who provides material faster has a higher associated cost; thus, to procure materials solely from this premium supplying agent is an expensive and often nonoptimal strategy. On the other hand, due to demand spikes or supply delays, relying exclusively on the slower supplier can likewise prove costly. Thus, companies such as Caterpillar (Rao et al. 2000) often use *dual sourcing*: they get the bulk of their materials from a cheaper *regular* supplier at a lower cost (and longer lead time) but turn to premium *expedited* channels when needed. Along the same lines, in summer 2003 when Amazon promised fast delivery of Harry Potter books, they used FedEx to deliver 400,000 copies on release while also continuing to regularly ship through UPS (Kelleher 2003). Similarly, Nintendo was able to restock shelves in time for the critical pre-Christmas rush using expedited delivery from UPS, selling more than 900,000 games in the United States by the end of the year (Souder 2004). Our problem

is also manifested in several ways, in manufacturing, retail, and service industries. A manufacturer receiving raw materials from suppliers operating with limited capacity may have the option to receive the raw materials faster than the quoted lead time by paying a higher price to the supplier. Likewise, in Internet retail, invariably there exists an option to get materials delivered faster at a premium price.

Firms might use multiple sourcing choices for a variety of strategic reasons, including safeguarding against predatory monopolistic practices and hedging against uncertainties in international markets, such as supply disruptions or exchange rate shifts. Davis (1992) reports that lower price ranks as the most important factor that motivates firms to outsource to external suppliers, but Carter and Vickery (1988) show that under volatile exchange rate movements firms can end up paying substantially more than their contracted price. Gottfredson et al. (2005) argue that a firm's skill in quickly remodeling its sourcing arrangements in response to market conditions may be its strongest competitive advantage. Thus, while low production costs and promising future growth (e.g., the cellular phone market) has led firms such as Nokia to locate production plants in Asia, they have also maintained extant production plants in Finland (see Bellman 2005).

Academically, Li and Kouvelis (1999) study flexible contracts and observe that multisupplier sourcing arrangements can help firms lower sourcing costs when faced with price uncertainty, such as would be caused by international exchange rate fluctuation or inflation in domestic markets. Furthermore, the effect of supply chain disruptions can be quite prominent (see Hendricks and Singhal 2005 and references therein): La-Z-Boy lost 18% of its stock price when its supplier could not deliver normal shipments of polyurethane foam in October 2005 (White 2005). Having dual suppliers in different geographic locations can mitigate this threat of supply chain disruption due to natural disasters or other causes. For example, Chiquita used multiple sourcing to temporarily increase production when under disruption due to Hurricane Mitch when competitors (Dole) suffered loss of revenue due to the lack of alternative supply channels (Tomlin 2006).

Situations like these demonstrate the need for management strategies for supply chains with sourcing options: companies need a simple yet effective way of deciding how much to source, when, and from whom. We focus on the inventory problem when the suppliers differ only by their delivery times and prices because inventory-driven costs can be a significant percentage of a firm's operating margins (see Callioni et al. 2005). Unfortunately, whereas optimal inventory policies are known for quite general single-source models (Tayur et al. 1998), results are much more limited when there are sourcing alternatives, despite the commonality of dual sourcing in practice. Part of the reason for this may be due to the "intractable nature of dual-source models" (Bradley 2004, p. 765). Our study considers this dual-sourcing problem with general lead times, providing an easily implementable, robust, and often near-optimal solution, the *dual-index base-stock* policy. This policy tracks inventory over both regular and expedited lead times, taking advantage of the sourcing flexibility while remaining practically implementable: in every period, if the expedited inventory position is below the expedited order-up-to target level, it is brought back to this level by placing an expedited order. After the expediting order is made, regular orders are placed, restoring the regular inventory position to its regular target level. Despite its simplicity, computational studies show that the optimal dual-index policy is often within 1% or 2% of the globally optimal policy, providing significant savings (up to 50%) compared to single sourcing. Moreover, we find the optimal dual-index parameters in approximately 10 seconds; the globally optimal policy found via dynamic programming takes approximately an hour for even very small instances.

The remainder of this paper is organized as follows: We position our work within the academic literature in §2. In §3, we describe the model in detail. In §4, we describe the dual-index policy, the order of events, and parameter recursions. We also establish separability properties and provide our method for quickly finding the optimal dual-index parameters. We extend this to capacitated systems

in §5. We validate our policies computationally against the optimal policy (found via dynamic programming) and explore issues such as partitioning capacity in §6. We conclude with some directions for future work in §7.

## 2. Our Position in the Literature

The earliest literature on dual sourcing is by Barankin (1961) who studies the one-period problem, work which Daniel (1962) extends to multiple periods. Fukuda (1964) provides optimal policies when the lead times are  $k$  and  $k + 1$ , respectively; he shows that the optimal policies are base-stock, and uses first-order conditions to derive integral parameter expressions. Our dual-index policy reduces to Fukuda's policy for this special case of consecutive lead times. Thus, we are the first to find *globally optimal parameters* for the general  $k, k + 1$  lead time model, although Bulinskaya (1964) derives the optimum inventory policies and parameters for Fukuda's model with  $k = 0$ .

A critical work on dual sourcing is that of Whittmore and Saunders (1977), who consider the problem for multiple periods and lead times of arbitrary fixed lengths  $k$  and  $k + l, l \geq 1$ . They show that for  $l > 1$ , the optimal policy is no longer a simple base-stock; it becomes highly state dependent, requiring multidimensional dynamic programming to find optimal parameters. Moinzadeh and Nahmias (1988) approximate the optimal  $(Q, R)$  policy for a dual-sourcing inventory system with continuous review, assuming there will only be a single outstanding order of each type. They do not benchmark how much the  $(Q, R)$  policy deviates from the complex, globally optimal policy. Moinzadeh and Schmidt (1991) consider an  $(S - 1, S)$  policy where the orders placed are either regular or expedited every period, not both. Alfredsson and Verrijdt (1999) present a similar one-to-one policy with emergency lateral trans-shipments (ELTs) to satisfy demand that cannot be fulfilled using regular ordering. Thus, the emergency shipment option is utilized only after backlogs occur (backlogs do not incur a penalty cost).

Lawson and Porteus (2000) consider a serial multi-echelon system with lead time between each stage equal to one and options to expedite and get materials immediately from the upstream stage, or stop orders in route. They show that a modified base-stock policy is optimal. Tagaras and Vlachos (2001) analyze a system with expedited lead time that can be very different from regular lead time, but restrict the expedited lead time to be smaller than the review period itself. Similarly, Groenevelt and Rudi (2002) analyze a system where the production periods are not smaller than the lead time difference, and Plambeck and Ward (2006) consider a model where emergency lead times are zero, proving a separation principle when 100% service is required. Our model analyzes dual-sourcing systems with no such restrictions on lead times or service levels. Finally, Feng et al. (2004) analyze inventory systems with multiple (consecutive) delivery modes, Tomlin (2006) considers

a manufacturer’s choice of dual sourcing when there are supply chain disruptions, and Scheller-Wolf et al. (2005) define and compare the single-index base-stock policy with a version of the dual-index policy considered here.

Compared to the above, our work analyzes the dual-sourcing problem with arbitrarily differing but constant lead times, under periodic review, where both regular and expedited orders can be made in every period. Demand is stochastic, and, when unsatisfied, is backordered with some penalty. This places us in the framework of Whittmore and Saunders (1977) who show that the optimal policy for this problem is highly complex; optimal ordering decisions are based on the vector of inventory positions covering the entire horizon between the expedited and regular lead times. Because our interest lies in gaining insights for practical implementations, we restrict ourselves to a simpler policy: our dual-index policy provides a simple near-optimal alternative to carrying the entire inventory vector. We also consider the effect of limited capacity for regular or expedited orders (or both). To our knowledge, order-up-to levels for dual-sourcing systems with arbitrary lead times have never been considered in the literature. Further, our work immediately extends to capacitated dual-sourcing systems.

Our separability results for dual-index base-stock policies are new. These results lead to optimal dual-index parameter expressions that can be evaluated through a newsboy fractile, reducing the complex dual-supply problem to a one-dimensional optimization. Our separability results are not constrained by lead times, order crossing, service levels, or demand volumes. Finally, our method is not only computationally simple, but also robust; it is applicable to scenarios including capacities, nonstationary demand, random stoppages, random yields, and certain types of lead time variability. The dual-supply problem under such broad scenarios has never been considered before.

### 3. Our Model

We consider a single-stage, capacitated, manufacturing/service location facing stochastic demand. The manufacturer can order the material through “regular” channels at cost  $c_r$  per unit, or, if the need arises, she can get some or all of the material “expedited” at some premium cost  $c_e$  per unit ( $c_e > c_r$ ), where  $c = c_e - c_r$ . The regular orders arrive after  $l_r$  periods, and the expedited orders arrive after  $l_e$  periods ( $l_e < l_r$ ). The difference in lead times is defined to be  $l = l_r - l_e \geq 1$ . If there is remaining on-hand inventory at the end of period  $n$  after the occurrence of the demand  $d_n$ , these items are carried over to the next period (i.e.,  $I_{n+1} > 0$ ) at a cost of  $h$  per unit. If there is a stock-out due to large demand (i.e.,  $I_{n+1} < 0$ ), there is a penalty cost  $p$  per unit unsatisfied demand. We seek to minimize the infinite-horizon average holding, penalty, and ordering cost.

**Table 1.** Notations.

Description		Description	
$n$	Period index	$d_n$	Demand in period $n$ , ergodic
$l_e$	Expedited lead time	$l_r$	Regular lead time, $l_r > l_e$
$c_e$	Unit expediting cost	$c_r$	Unit regular ordering cost, $c_r \leq c_e$
$c$	$c_e - c_r$	$l$	$l_r - l_e$
$k^e$	Expediting ordering capacity	$k^r$	Regular ordering capacity
$X_n^e$	Period $n$ expedited order ( $X_n^e \leq k^e$ )	$X_n^r$	Period $n$ regular order ( $X_n^r \leq k^r$ )
$IP_n^e$	Period $n$ expedited inventory position	$IP_n^r$	Period $n$ regular inventory position
$z_e$	Expedited order-up-to level	$z_r$	Regular order-up-to level
$I_n$	On-hand inventory at start of period $n$	$\Delta$	$z_r - z_e$

In our dual-index policy, the period  $n$  expediting order,  $X_n^e$ , is based on the on-hand inventory plus the expedited and regular orders that will arrive within  $l_e$  periods; orders that are due to arrive after  $l_e$  periods are not included in the expedited ordering decision. This expedited order,  $X_n^e$ , tries to restore the expedited inventory position,  $IP_n^e$ , to some target parameter level  $z_e$ . The regular order,  $X_n^r$ , is based on the regular inventory position (sum of on-hand inventory and all outstanding orders, including  $X_n^e$ ),  $IP_n^r$ , and tries to restore it to the target parameter  $z_r$ . We define  $\Delta = z_r - z_e$ . Thus, in the dual-index policy we carry two inventory positions, one for expedited ordering and another for regular ordering. There might be capacities on regular and expedited orders, which we denote by  $k^r$  and  $k^e$ , respectively. Notations are summarized in Table 1.

## 4. Analytical Results

In this section, we derive analytical results for the uncapacitated case. We show how they can be modified to admit capacities in §5.

### 4.1. Order of Events

The order of events in a given period  $n$  is as follows: we begin the period with on-hand inventory  $I_n$  and several periods of on-order inventory comprised of expedited orders placed over the past  $l_e$  periods and regular orders placed over the past  $l_r$  periods. Specifically, we have the vector of pipeline regular orders  $\langle X_{n-l_r}^r, \dots, X_{n-1}^r \rangle$  due to arrive in periods in  $n$  through  $n + l_r - 1$ , and pipeline expedited orders  $\langle X_{n-l_e}^e, \dots, X_{n-1}^e \rangle$  due to arrive in periods  $n$  through  $n + l_e - 1$ . The expedited inventory position is comprised of on-hand inventory and all the orders due to arrive in the next  $l_e$  periods:

$$IP_n^e = I_n + (X_{n-l_e}^e + \dots + X_{n-1}^e) + (X_{n-l_r}^r + \dots + X_{n-l-1}^r).$$

The regular inventory position is comprised of on-hand inventory and all the orders that will arrive in the next  $l_r$  periods:

$$IP_n^r = I_n + (X_{n-l_e}^e + \dots + X_{n-1}^e) + (X_{n-l_r}^r + \dots + X_{n-1}^r).$$

At the start of the period  $n$ , orders ( $X_n^e$ , and then  $X_n^r$ ) are placed based on the expedited and regular inventory positions, respectively. The expedited order,  $X_n^e$ , is added to  $IP_n^r$  before  $X_n^r$  is determined. Then, the material due to arrive this period, regular order  $X_{n-l_r}^r$  and expedited order  $X_{n-l_e}^e$ , physically arrive. The demand for the period,  $d_n$ , is revealed and satisfied if enough on-hand inventory is available; any excess demand is backordered. The inventory levels are then updated and holding or penalty costs are incurred.

In marked difference from a standard base-stock policy, in the dual-index policy the expedited inventory position may exceed the target expedited inventory level,  $z_e$ . This is because the order,  $X_{n-l}^r$ , that was made through regular channels  $l = l_r - l_e$  periods in the past enters the information horizon. In some cases, this regular order may push the expedited inventory position above  $z_e$ , causing an *overshoot*:  $O_n \triangleq (IP_n^e + X_{n-l}^r - z_e)^+$ . In this case, no expedited ordering is made. If instead  $IP_n^e$  is lower than  $z_e$ , i.e., there is a *deficit*:  $U_n \triangleq (z_e - IP_n^e - X_{n-l}^r)^+$ , then a positive expedited order of size  $U_n$  is made to restore the inventory position back to  $z_e$ . Thus, it can be observed that  $z_e$  is a lower bound for inventory position after expedited ordering. Note that by definition,  $U_n \cdot O_n = 0$ .

The system recursions are thus

$$IP_{n+1}^e = IP_n^e + X_n^e - d_n + X_{n-l}^r \triangleq z_e + O_n - d_n, \quad (1)$$

$$IP_{n+1}^r = IP_n^r + X_n^e + X_n^r - d_n = z_r - d_n, \quad (2)$$

$$I_{n+1} = I_n + X_{n-l_e}^e + X_{n-l_r}^r - d_n.$$

Holding cost of  $h > 0$  per unit, or penalty cost of  $p > 0$  per unit, is charged on the on-hand inventory excess,  $I_{n+1}^+ = \max(I_{n+1}, 0)$ , or backlog,  $I_{n+1}^- = \max(-I_{n+1}, 0)$ , respectively. The expedited order and regular orders are

$$X_n^e = (z_e - IP_n^e - X_{n-l}^r)^+ \triangleq U_n, \quad (3)$$

$$X_n^r = z_r - (IP_n^r + X_n^e) = d_{n-1} - X_n^e. \quad (4)$$

For all sequences of random variables  $Z_n$ , we define their *long-run time average* as

$$E[Z] = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N Z_n.$$

Similarly, when we refer to a *stationary* (cumulative) distribution of a sequence of random variables  $Z_n$ , we signify  $P(Z \leq x) \triangleq \lim_{N \rightarrow \infty} (1/N) \sum_{n=1}^N \mathbf{I}\{Z_n \leq x\}$ , where  $\mathbf{I}$  denotes the indicator function. (We use time averages because a proper distribution may not exist in our most general settings.) For an integral  $k$  (positive or negative), define  $D_n^k = d_n + d_{n+1} + \dots + d_{n+k}$ . Then, using the above notation, based on the ergodicity of  $d$ , it can be shown (in the case of infinite regular capacity) that if  $P(D^{l-1} \leq \Delta) < 1$ , expedited ordering will take place infinitely often and the entire system will be positive regenerative. If, conversely, this probability is equal to one, then our system reduces

to a single-sourcing system with expedited ordering only. When capacities on regular orders are present, still weaker conditions ensuring positive regeneration may be possible; in all cases, as long as  $k^r + k^e > E[d]$  which we assume, the infinite-horizon average cost converges to

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \pi_n = hE[I^+] + pE[I^-] + (c_e - c_r)E[X^e] + c_rE[d]. \quad (5)$$

If either  $k^r + k^e < E[d]$  or  $E[d] = \infty$ , both sides of (5) diverge.

## 4.2. Recursions for Overshoot and Expedited Deficit

The following properties hold for our system, and are used in our optimization.

LEMMA 4.1. *Overshoot satisfies*  $O_{n+1} = (O_n + X_{n+1-l}^r - d_n)^+$ .

PROOF.

$$\begin{aligned} O_{n+1} &= (IP_{n+1}^e + X_{n+1-l}^r - z_e)^+ \\ &= (z_e + O_n - d_n + X_{n+1-l}^r - z_e)^+ \quad (\text{from Equation (1)}) \\ &= (O_n - d_n + X_{n+1-l}^r)^+. \quad \square \end{aligned}$$

COROLLARY 4.1. *Expedited deficit satisfies*  $U_{n+1} = (O_n + X_{n+1-l}^r - d_n)^-$ .

PROOF. Same as in Lemma 4.1, with  $(\cdot)^-$  replacing  $(\cdot)^+$ .  $\square$

## 4.3. Solution Procedure

Our optimization procedure is based on the following sequence of results in this section:

1. We first show for all  $n$  that the overshoot distribution is a function of  $\Delta$  alone, independent of  $z_e$ . Thus, given a  $\Delta$  we can determine  $O(\Delta)$  independent of  $z_e$ . This is Proposition 4.1. (We will suppress the parameter  $\Delta$  in  $O(\Delta)$  when not required.)

2. Then, we show how given  $O_n, I_n^+$  and  $I_n^-$  can be determined for all  $n$  as a function of  $z_e$  and demand. This is Lemma 4.3.

3. For each  $\Delta$ , we then derive an expression for the optimal  $z_e^*(\Delta)$  as a newsvendor fractile of the lead time demand convoluted with the stationary overshoot. This is Theorem 4.1. Then, for each  $(\Delta, z_e^*(\Delta))$  pair, we find its cost.

4. Using the costs for each  $(\Delta, z_e^*(\Delta))$  pair, we find the lowest-cost pair by one-dimensional search over  $\Delta$ . This yields the optimal dual-index policy within approximately 10 seconds for all the problems we have considered. (Note that nearly all the computational time is spent simulating the convoluted demand and overshoot distribution for each  $\Delta$  because there is no general closed-form distribution for this.)

PROPOSITION 4.1. *The distribution of the overshoot is a function of  $\Delta$ , independent of  $z_e$ .*

PROOF. This is a special case of Lemma 5.1; we defer the proof to there.  $\square$

To prove that the optimal  $z_e$  is a function of the overshoot, we use an alternate expression for overshoot.

LEMMA 4.2.  $O_n = \Delta - (X_n^r + X_{n-1}^r + \dots + X_{n-l_r+1}^r)$ .

PROOF.

$$\begin{aligned} I_{n+1} &= z_r - d_n - (X_n^e + X_n^r) - \dots - (X_{n-(l_e-1)}^e + X_{n-(l_e-1)}^r) \\ &\quad - X_{n-l_e}^r - \dots - X_{n-l_r+1}^r \\ &= z_r - d_n - d_{n-1} - \dots - d_{n-l_e} - (X_{n-l_e}^r + \dots + X_{n-l_r+1}^r) \\ &\triangleq z_r - D_n^{-l_e} - (X_{n-l_e}^r + \dots + X_{n-l_r+1}^r). \end{aligned} \tag{6}$$

Alternately,

$$\begin{aligned} I_{n+1} &= (z_e + O_n) - d_n - (X_n^e + \dots + X_{n-l_e+1}^e) \\ &\quad - (X_{n-l}^r + \dots + X_{n-l_r+1}^r) \\ &= (z_e + O_n) - d_n - ((d_{n-1} - X_n^r) + \dots + (d_{n-l_e} - X_{n-l_e+1}^r)) \\ &\quad - (X_{n-l}^r + \dots + X_{n-l_r+1}^r) \\ &= (z_e + O_n) - D_n^{-l_e} + (X_n^r + \dots + X_{n-l_e+1}^r) \\ &\quad - (X_{n-l}^r + \dots + X_{n-l_r+1}^r). \end{aligned} \tag{7}$$

Using Equations (6) and (7),

$$\begin{aligned} O_n &= (z_r - z_e) - (X_{n-l_e}^r + \dots + X_{n-l_r+1}^r) \\ &\quad - (X_n^r + \dots + X_{n-l_e+1}^r) + (X_{n-l}^r + \dots + X_{n-l_r+1}^r) \\ &= \Delta - (X_n^r + \dots + X_{n-(l-1)}^r). \end{aligned} \tag{8}$$

LEMMA 4.3.  $I_{n+1} = z_e + O_{n-l_e} - D_n^{-l_e}$ .

PROOF. From Equation (6), we have

$$\begin{aligned} I_{n+1} &= z_r - D_n^{-l_e} - (X_{n-l_e}^r + \dots + X_{n-l_r+1}^r) \\ \Rightarrow I_{n+1} &= z_r - D_n^{-l_e} - (\Delta - O_{n-l_e}) \quad (\text{using Equation 8}) \\ &= z_e + O_{n-l_e} - D_n^{-l_e}. \end{aligned} \tag{9}$$

THEOREM 4.1. *Let  $G_{n,\Delta}(x) = P(D_n^{-l_e} - O_{n-l_e}(\Delta) \leq x)$ , with stationary version  $G_\Delta$ . The optimal level of  $z_e$  given  $\Delta$  is*

$$z_e^*(\Delta) = G_\Delta^{-1}\left(\frac{p}{p+h}\right).$$

PROOF. From Lemma 4.3, the on-hand inventory is equal in distribution to the on-hand inventory in a system with order-up-to level  $z_e$  facing period  $n$  demand  $D_n^{-l_e} - O_{n-l_e}(\Delta)$ , which is a newsvendor problem. Furthermore, the overshoot  $O_n(\Delta)$  is a function of  $\Delta$  independent of  $z_e$

(as is  $D_n^{-l_e}$ ). Because our system is positive regenerative, we have convergence of  $\lim_{N \rightarrow \infty} (1/N) \sum_{n=1}^N \mathbf{I}\{D_n^{-l_e} - O_{n-l_e}(\Delta) \leq x\}$  for all  $x$ ; we denote this as the limiting distribution function of  $G_\Delta$ . Thus,

$$z_e^*(\Delta) = G_\Delta^{-1}\left(\frac{p}{p+h}\right). \tag{10}$$

The expedited order-up-to decision follows a “newsvendor with returns” model: in every period, the demand is reduced by the amount of the overshoot  $l_e$  period past;  $O_{n-l_e}$  items are “returned.” The optimal newsvendor fractile in such a case is as in (10). The returns need not be independent of the demand, i.e., the demand could be dependent on past sales as long as the distribution  $G_\Delta$  exists. For i.i.d. demand this is not an issue:  $G_\Delta$  is comprised of the stationary distribution of next  $l_e + 1$  demands convolved with the negative overshoot before expediting in the current period; they are independent.

### 5. Capacitated Models

Let  $k^r$  and  $k^e$  be the capacity limits on regular and expedited orders, respectively. Let order quantities be as defined in Table 1. We observe that the order quantities  $X^r$ ,  $X^e$  are now constrained by the available regular and expediting capacities  $k^r$  and  $k^e$ , respectively, unlike in the previous sections. If the regular (expedited) order quantity does not bring the regular (expedited) inventory position to the regular (expedited) order-up-to level, regular (expedited) shortfalls occur. Let the regular and expedited shortfalls be  $S^r$  and  $S^e$ , respectively, defined according to

$$S_n^e = (S_{n-1}^e + d_{n-1} - O_{n-1} - X_{n-l}^r - X_n^e)^+, \tag{11}$$

$$S_n^r = S_{n-1}^r + d_{n-1} - (X_n^r + X_n^e). \tag{12}$$

LEMMA 5.1. *Shortfalls  $S_n^r$  and  $S_n^e$  and overshoot  $O_n$  are functions of  $\Delta$  independent of  $z_e$ .*

PROOF. We use induction. Let the inventory process begin in the initial period 1 with expedited orders in  $(z_e/l_e) \wedge k^e$  sizes over the periods  $1, \dots, l_e$  and no regular orders over this horizon. The on-hand inventory at the beginning of the first period is  $I_1 = z_e - l_e((z_e/l_e) \wedge k^e) - d_0$ , where  $d_0$  is the demand at the end of period 0. The regular orders that arrive in periods  $l_e + 1, \dots, l_r$  are all  $(\Delta/l) \wedge k^r$ . Hence,  $IP_1^e = z_e - d_0$  and  $IP_1^r = z_e + l(\Delta/l \wedge k^r) - d_0 = z_e + (\Delta \wedge lk^r) - d_0$ . Using the inventory progression from Equation (1), we have  $O_0 = 0$ . The expedited inventory position before the period 0 demand  $d_0$  occurs is  $z_e$ , hence the expedited shortfall in period 0 is  $S_0^e = 0$ . The regular inventory position before the occurrence of period 0 demand is  $z_e + (\Delta \wedge lk^r)$ . Therefore, the regular shortfall in period 0 is  $S_0^r = z_r - [z_e + (\Delta \wedge lk^r)]$ . Then, for this system we have the relations

$$X_n^e = (S_{n-1}^e + d_{n-1} - O_{n-1} - X_{n-l}^r)^+ \wedge k^e, \tag{13}$$

$$O_n = (O_{n-1} - S_{n-1}^e - d_{n-1} + X_{n-l}^r)^+, \tag{14}$$

$$X_n^r = (S_{n-1}^r + d_{n-1} - X_n^e) \wedge k^r, \quad (15)$$

$$S_n^e = (S_{n-1}^e + d_{n-1} - O_{n-1} - X_{n-1}^r - X_n^e)^+, \quad (16)$$

$$S_n^r = S_{n-1}^r + d_{n-1} - X_n^r - X_n^e. \quad (17)$$

For period 1, using  $\perp$  to denote independence,

$$X_1^e = [d_0 - ((\Delta/l) \wedge k^r)]^+ \wedge k^e \Rightarrow \mathbf{X}_1^e \perp \mathbf{z}_e,$$

$$O_1 = ((\Delta/l) \wedge k^r - d_0)^+ \Rightarrow \mathbf{O}_1 \perp \mathbf{z}_e,$$

$$\begin{aligned} X_1^r &= [z_r - (z_e + l(\Delta/l \wedge k^r)) + d_0 - X_1^e] \wedge k^r \\ &= [z_r - (z_e + (\Delta \wedge lk^r)) + d_0 \\ &\quad - [d_0 - (\Delta/l \wedge k^r)]^+ \wedge k^e] \wedge k^r \\ &= [\Delta - (\Delta \wedge lk^r) + d_0 \\ &\quad - [d_0 - (\Delta/l \wedge k^r)]^+ \wedge k^e] \wedge k^r \Rightarrow \mathbf{X}_1^r \perp \mathbf{z}_e, \end{aligned}$$

$$S_1^e = [d_0 - (\Delta/l \wedge k^r) - [d_0 - (\Delta/l \wedge k^r)]^+ \wedge k^e]^+ \Rightarrow \mathbf{S}_1^e \perp \mathbf{z}_e,$$

$$\begin{aligned} S_1^r &= z_r - [z_e + l(\Delta/l \wedge k^r)] + d_0 - X_1^r - X_1^e \\ &= \Delta - (\Delta \wedge lk^r) + d_0 - X_1^r - X_1^e \Rightarrow \mathbf{S}_1^r \perp \mathbf{z}_e. \end{aligned}$$

Assume that  $X_k^e, X_k^r, O_k, S_k^e, S_k^r$  are independent of  $z_e \forall k = 2, \dots, n-1$ . From the above recursions (13) through (17), it follows that  $X_n^e, X_n^r, O_n, S_n^e, S_n^r$  are independent of  $z_e$ .  $\square$

Note that if either one of the channels is uncapacitated, the above results hold with the appropriate  $k$  set to infinity.

For the capacitated case, reasoning as in Lemma 4.2 can be used to show

$$\begin{aligned} I_{n+1} &= z_r - S_n^r - d_n - (X_n^e + \dots + X_{n-l_e+1}^e) \\ &\quad - (X_n^r + \dots + X_{n-l_r+1}^r), \end{aligned} \quad (18)$$

$$\begin{aligned} I_{n+1} &= z_e + O_n - S_n^e - d_n - (X_n^e + \dots + X_{n-l_e+1}^e) \\ &\quad - (X_{n-l}^r + \dots + X_{n-l_r+1}^r). \end{aligned} \quad (19)$$

Using (18) and (19) provides

$$O_n = \Delta - S_n^r + S_n^e - (X_n^r + \dots + X_{n-l+1}^r). \quad (20)$$

Using (20) in (18), again,

$$\begin{aligned} I_{n+1} &= z_r - S_n^r - d_n - (X_n^e + X_n^r) - \dots - (X_{n-(l_e-1)}^r + X_{n-(l_e-1)}^e) \\ &\quad - X_{n-l_e}^r - \dots - X_{n-l_r+1}^r \\ &= z_r - S_n^r - d_n - (S_{n-1}^r + d_{n-1} - S_n^r) \\ &\quad - \dots - (S_{n-l_e}^r + d_{n-l_e} - S_{n-l_e+1}^r) - X_{n-l_e}^r - \dots - X_{n-l_r+1}^r \\ &= z_r - D_n^{-l_e} - S_{n-l_e}^r - (X_{n-l_e}^r + \dots + X_{n-l_r+1}^r) \\ &= z_r - D_n^{-l_e} - S_{n-l_e}^r - (-O_{n-l_e} + \Delta - S_{n-l_e}^r + S_{n-l_e}^e) \\ &= z_e - S_{n-l_e}^e + O_{n-l_e} - D_n^{-l_e}. \end{aligned}$$

Then, again via the same arguments, the optimal expedited order-up-to level provided  $\Delta$  is

$$z_e^*(\Delta) = F_{(D_n^{-l_e} + S_{n-l_e}^e - O_{n-l_e})(\Delta)}^{-1} \left( \frac{p}{p+h} \right). \quad (21)$$

Note that the regular shortfall is involved only indirectly in determining the overshoot  $O(\Delta)$ . Also note that the relation remains if one of the capacities is infinite, and if both are, we recover (10).

## 6. Value of Dual Sourcing Under the Dual-Index Policy

We begin by comparing the dual-index policy, single sourcing, and the optimal policies for simple models (for which the optimal policy can be obtained) in §6.1. We then compare the dual index with single sourcing for more general models, investigating the division of costs and effects of lead times in §6.2, and the effects of demand variability in §6.3. Next, we experiment with capacitated systems in §6.4 and the issue of allocating limited capacity between two suppliers in §6.5 before summarizing our results in §6.6. Single-sourcing costs were found using a simple newsvendor solution. Dual-index and optimal policy costs were found using a C++ program on an IBM PC with a Pentium M processor. As mentioned previously, our separation principle is remarkably general: thus, our algorithm also works for cases of correlated demand, capacities, random yields, and disruptions in regular supply (as in Tomlin 2006). We defer such investigations to future work, restricting our experiments here to cases of stationary demand, possibly with capacities on regular and/or expedited orders.

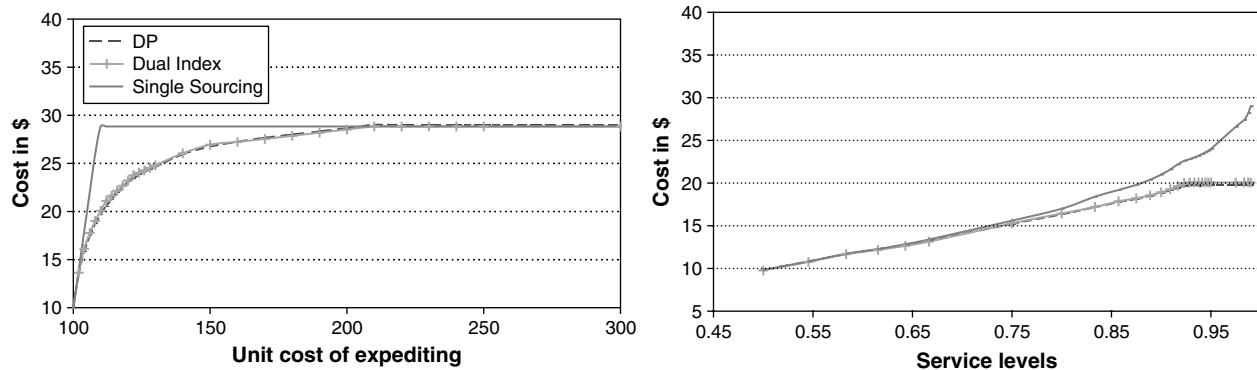
### 6.1. Comparisons Against the Optimal Policy

From the literature reviewed in §2, we know that the optimal policy for the general lead time problem is complex and state dependent, whereas the optimal dual-index solution is simple and easily computed. We now compare the performance of these policies, along with the best single-sourcing option (all materials are always ordered through the regular or expedited supply channel, whichever yields the lowest infinite-horizon average cost). This provides us with two measures: first, how much better the dual-index policy performs than single sourcing, and second, how the dual-index policy compares to the optimal policy.

Because finding optimal policies in general are computationally intensive (involving dynamic programming), our comparisons must be done over a restricted state space (i.e., cases with small  $l$  and a discrete demand distribution with limited support). Within this setting, we vary the desired service level (the newsvendor fractile  $p/(p+h)$ ), the cost of expediting, lead times (both expediting and regular), and demand distribution. Capacity limitations on orders are of course also an important factor, but because these may not be present in general, we defer experimentation with capacities to later sections.

In the following §6.1.1, we analyze the simplest unknown result with lead times  $l_e = 0$  and  $l_r = 2$ . We study the effect of variability in demand distribution in §6.1.2 before we alter regular lead times in §6.1.3, building on these results in §6.1.4, where we address the very significant issue of the width, or granularity of the state space. In §6.1.5 we study the effect of increasing the expedited lead time. Throughout §§6.1.1 to 6.1.5, we hold  $c_r = \$100$  and  $h = \$5$ . When the expediting costs are changed (for all graphs on the left), the penalty cost is held at  $p = \$495$ , and

**Figure 1.** Dual-index, single-sourcing, and optimal DP costs against expediting costs and service levels:  $h = 5$ ,  $c_r = 100$ ,  $d \sim \mathcal{U}[0, 4]$ ,  $l_e = 0$ ,  $l_r = 2$  ( $p = 495$  (left),  $c_e = 110$  (right)).



when required service levels are changed (for all graphs on the right), the expedited cost is held at  $c_e = \$110$ .

**6.1.1. Base Case** ( $l_e = 0$ ,  $l_r = 2$ ,  $d \sim \mathcal{U}[0, 4]$ ). In this section, we analyze the simplest problem for which a static order-up-to policy is not optimal. The expedited lead time is  $l_e = 0$ ; the expedited orders arrive and immediately become on-hand inventory. The regular orders arrive two periods after ordering.

To keep the dynamic program tractable, we restrict the demand  $d \in \{0, 1, \dots, B, \dots, 2B\}$  where  $B = 2$  in this subsection. Because the costs are stationary and orders are uncapacitated, we expect that optimal orders would not exceed the maximum possible demand; we thus limit the regular and expedited orders to  $2B + 1$ . (Computational experiments show that the optimal policies never order  $2B + 1$ , validating this assumption.) As  $p > 0$ , the backlogs are limited to  $l_r + 1$  times the maximum demand, i.e.,  $(l_r + 1)2B$ ; and as  $h > 0$ , the maximum on-hand inventory is restricted to  $(l_r + 1)(2B + 1)$ . We consider a simple discrete uniform demand distribution  $\mathcal{U}[0, 2B]$ . In Figure 1, we illustrate the behavior of single sourcing, dual index (DI), and the optimal DP.

The graph on the left in Figure 1 shows the performance of the policies for various expediting costs, keeping the holding and penalty costs constant at  $h = \$5$  and  $p = \$495$ . The dual-index policy cost is never worse than 3% above the optimal policy for any expediting cost, and brings significant savings compared to single sourcing. The maximum cost for the best single-sourcing option is achieved when the cost of expediting alone matches the cost of using the regular channel alone. The dual-index policy has the highest benefit at this point (typically moderate expediting cost). This result is crucial: *Dual sourcing is of highest utility when the manufacturer is indifferent between two channels of procurement.*

When the marginal unit expediting cost is very low, it is optimal to single source from the expedited supplier; this is exactly what all three policies do at low  $c_e$ . Similarly, as the marginal expediting cost increases, single sourcing through

the regular supplier becomes identical to the optimal policy and the DI policy.

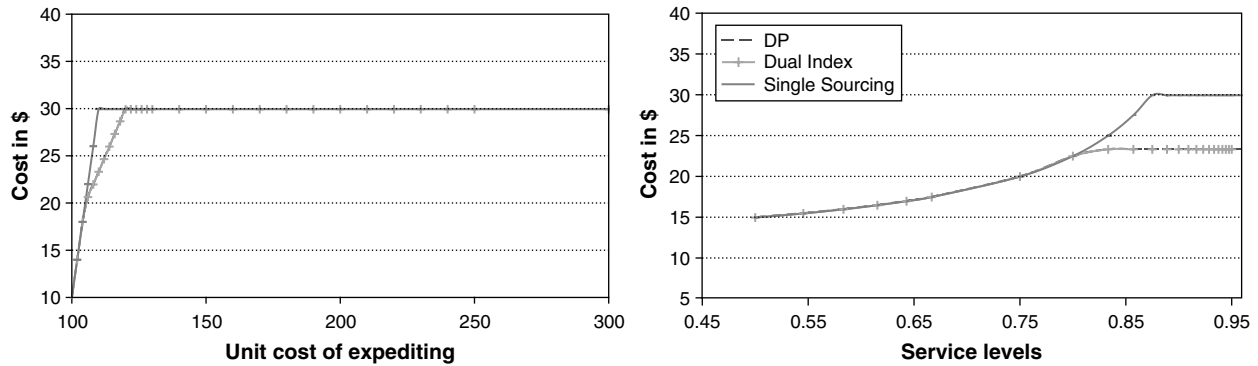
The graph on the right in Figure 1 captures the effect of increasing penalty cost (or increasing the desired service level). The holding cost per item is \$5, the expediting cost per unit item is \$110, and regular ordering cost is \$100. In general, the dual-index policy is within 1% of the optimal cost and its deviation is about 2% at worst. We hypothesize that the benefit from the dynamic program largely comes from the fact that the DP can have different order-up-to levels at different states, achieving a sort of “randomized policy” (see Bertsekas 1995), whereas the dual-index policy follows a static policy. Therefore, one could surmise that any static policy would suffer when the optimal fractile ( $p/(p+h)$ ) is far from any (demand minus overshoot) mass point.

Note that not only the cost, but also the character of the dual-index policy is remarkably similar to the optimal strategy, flattening out at about 92.30% service. At this point, both the DP and DI keep four units on hand at all times, ensuring 100% service. They are able to do this in a cost-efficient manner because they continue to do the bulk of their sourcing using the regular supplier, using the expedited sourcing only when a backlog may occur. Single sourcing lacks this flexibility, and thus is considerably more expensive.

**6.1.2. Effect of Variability in the Demand Distribution.** In this subsection, we increase the coefficient of variation of the demand distribution by shifting the probability mass to its extreme points: keeping  $B = 2$ , we place equal probability mass on zero and four. All other parameters are as in the previous section. From Figure 2, we see that in this case the dual-index policy indeed is optimal. There is still significant savings by implementing the dual index (dual sourcing) rather than single sourcing, especially at moderate expediting costs and high service levels.

**6.1.3. Effect of Varying Regular Lead Times.** We now revert back to uniform demand distribution and increase the regular lead time to three (causing  $l = l_r - l_e$

**Figure 2.** Dual-index, single-sourcing, and optimal DP costs against expediting costs and service levels:  $h = 5$ ,  $c_r = 100$ ,  $d \in \{0, 4\}$ ,  $l_e = 0$ ,  $l_r = 2$  ( $p = 495$  (left),  $c_e = 110$  (right)).



likewise to increase to three), controlling the experiments for the same set of parameters as in §6.1.1. We note that the DI computational times are unaffected, whereas DP is significantly affected due to the multiplication of the state space (see §6.6).

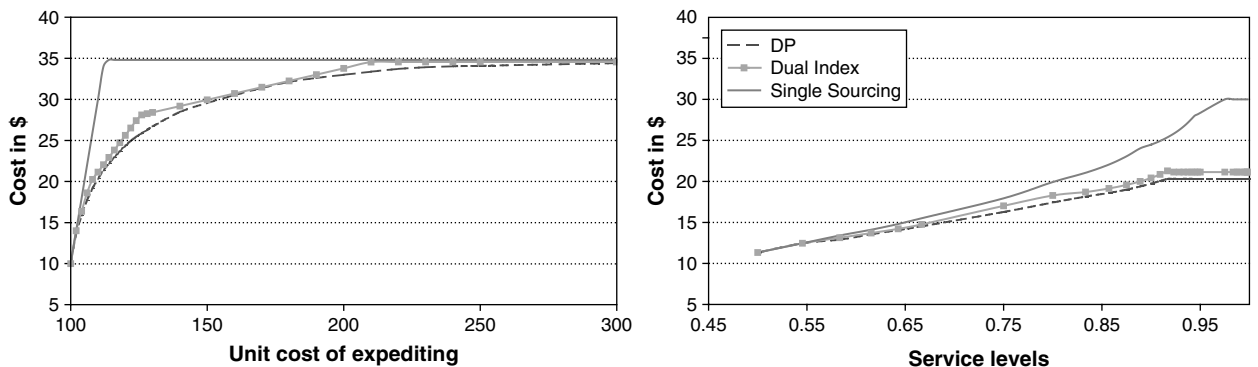
Increasing  $l$  causes both the dual-index policy and single sourcing to perform significantly worse as compared to DP—at times the dual index costs increase linearly (see the left-hand part of Figure 3), approximating the DP less accurately. Note specifically the peaks at the end points of the linear cost segment when  $c_e = 126$  and  $c_e = 210$ . Once again, this is due to the effect of limited demand support which leads to optimal fractiles being far from chosen discrete demand points. This is more acute now that the DP has more information, and can make dynamic decisions more finely. (In experiments with  $l_e = 0$ ,  $l_r = 3$  and exponential demand, these peaks disappeared; see Veeraraghavan 2004.) Even with these peaks, the worst-case performance of the dual-index policy is still within 8% of optimal for any unit expediting cost and within 5% for all service levels, significantly outperforming single sourcing.

**6.1.4. Effect of Increased Demand Mean and Support.** We argued in §§6.1.1 and 6.1.3 that limited state space and discretization might work against the dual-index

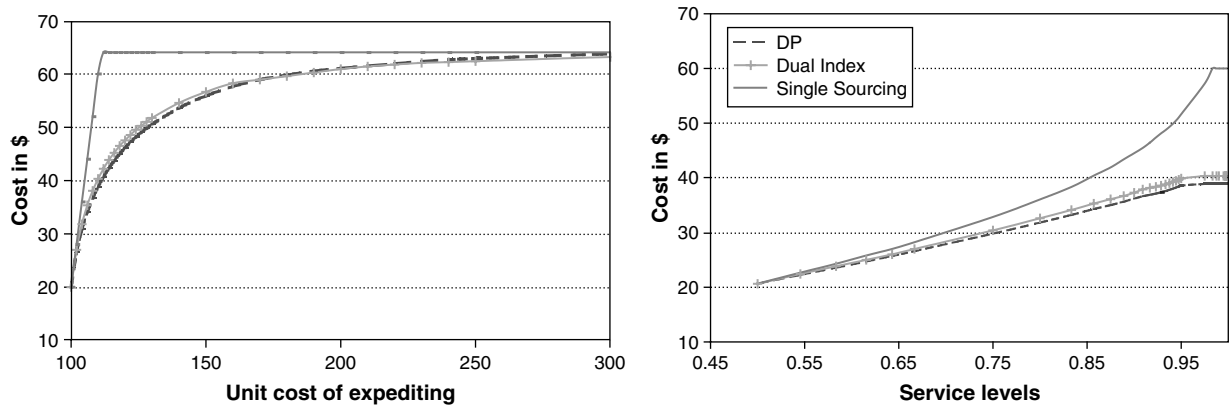
policy because the fractiles could only be used to cover  $(2B + 1)$  demand outcomes. In this section, we study the effect of making the demand “finer” and simultaneously study the effect of higher mean demand: we consider the case of §6.1.3 with discrete uniform demand  $d \sim \mathcal{U}[0, 8]$ . All other parameters remain the same.

Looking at Figure 4, and comparing it to Figure 3, we see that the dual-index policy once again performs much better than single sourcing, as expected. More significantly, the dual-index performance against the DP has improved, with maximum deviation reducing from 8% to 4%, and 5% to 3.7%, for varying expediting costs and service levels, respectively. In addition, the DI policy now mimics the DP behavior more accurately; the linearities have been reduced because we have finer information about the convolved overshoot minus demand, enabling the DI policy to meet the optimal fractile more closely. Even with this finer information though, DP still more closely approaches the target fractile by “randomizing” actions. The dual-index policy (or any static policy) simply cannot do this, but this difference becomes less and less important as discretization becomes dense or grids get finer. As the demand distribution becomes continuous (and dynamic programming becomes prohibitive), we expect the dual-index policy to approach the DP cost even more closely.

**Figure 3.** Dual-index, single-sourcing, and optimal DP costs against expediting costs and service levels:  $h = 5$ ,  $c_r = 100$ ,  $d \sim \mathcal{U}[0, 4]$ ,  $l_e = 0$ ,  $l_r = 3$  ( $p = 495$  (left),  $c_e = 110$  (right)).



**Figure 4.** Dual-index, single-sourcing, and Optimal DP costs against expediting costs and service levels:  $h = 5$ ,  $c_r = 100$ ,  $d \sim \mathcal{U}[0, 8]$ ,  $l_e = 0$ ,  $l_r = 3$  ( $p = 495$  (left),  $c_e = 110$  (right)).



**6.1.5. Increasing Expediting Times.** Keeping  $l = 3$  (as in §6.1.3), we increase the expedited lead time  $l_e$  to one period; expedited orders arrive in one period instead of immediately, and regular orders arrive in four.

Comparing the corresponding charts in Figures 3 and 5, we observe that the relative performance of the dual-index policy is remarkably improved—the deviation from the optimal policy is now within 2.5% everywhere. Because expediting does not immediately convert the orders into on-hand inventory, all information on expedited orders is less valuable, having to filter through a second demand. Therefore, the advantage gained by increased DP information over the DI policy, or increased DI information over single sourcing is reduced—note that both the DP and DI cost curves flatten out at a much higher service level in Figure 5 than in Figure 3, and the worst-case performance of the single-sourcing policy is about 30% from optimal, and only then at very high service levels.

**6.2. Performance of the Dual-Index Policy Under Continuous Distribution of Demand**

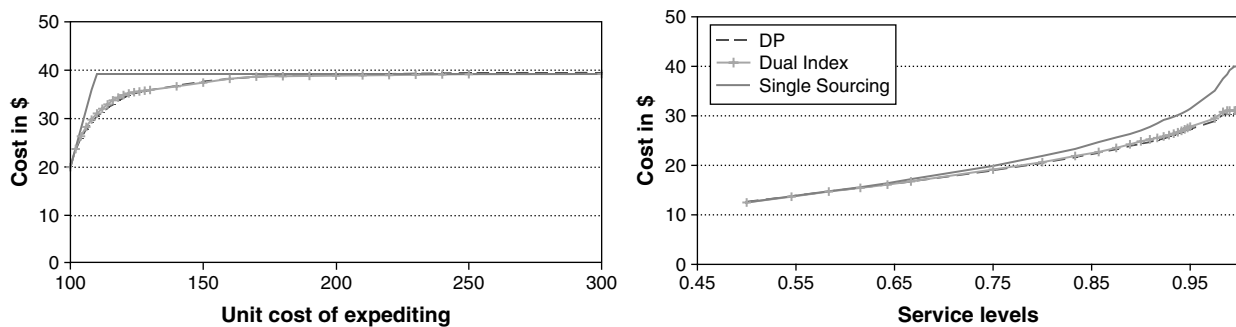
In all the experiments in the previous section, the demand was limited so that the optimal policy could be evaluated by dynamic programming. In this section, we are interested in understanding how well the dual-index policy

performs under general demand and lead times when dynamic programming becomes computationally impractical; specifically, we examine cases with unbounded continuous demand and larger  $l$ .

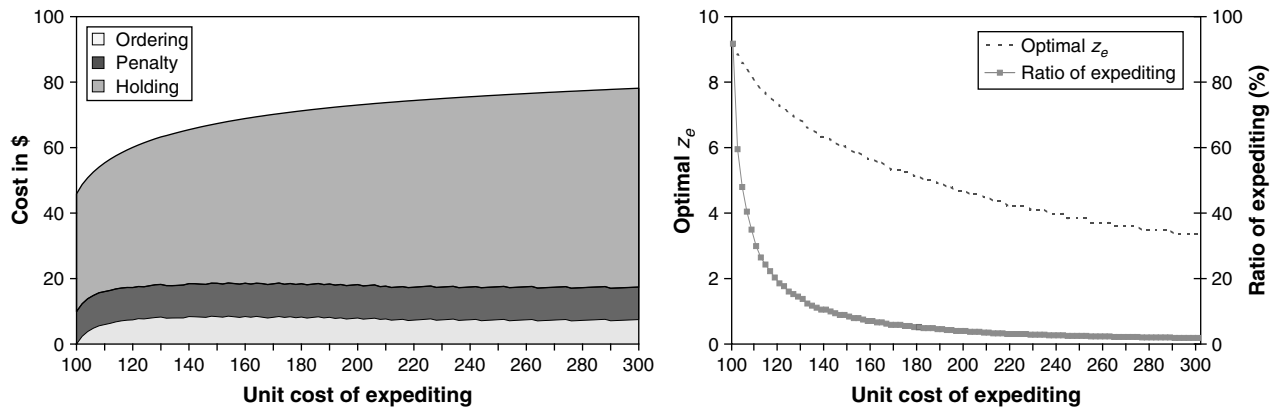
Because we have continuous demand, finding the optimal parameter pair becomes more delicate. Given a  $\Delta$ , the distribution of  $D_n^{-l_e} + S_{n-l_e}^e - O_{n-l_e}(\Delta)$  is found via simulation. Using this distribution, the critical newsvendor fractile is established. Again using this distribution, the costs for each  $(\Delta, z_e^*(\Delta))$  are calculated. Our experiments indicate that while the cost curve is not convex, it appears to be unimodal in  $\Delta$ . Therefore, any simple one-dimensional search method (e.g., golden search) can be effected to reduce the optimization process.

To understand the behavior of the dual-index policy better, we now illustrate how the costs are split between holding, penalty, and ordering costs as the expediting costs, and therefore the level of expedited sourcing, change. As we would expect, the chart on the right in Figure 6 shows that both the fraction of demand received through expedited channels and the optimal expediting parameter decrease as unit expediting cost increases. The chart on the left displays the different optimal cost breakdowns. Note that as expediting unit cost increases, less expediting is done and holding

**Figure 5.** Dual-index, single-sourcing, and optimal DP costs against expediting costs and service levels:  $h = 5.0$ ,  $c_r = 100$ ,  $d \sim \mathcal{U}[0, 4]$ ,  $l_e = 1$ ,  $l_r = 4$  ( $p = 495$  (left),  $c_e = 110$  (right)).



**Figure 6.** Optimal split costs,  $z_e$ , and expediting ratio for various expediting costs:  $c_r = 100$ ,  $h = 5$ ,  $p = 495$ ,  $d \sim \exp(2)$ ,  $l_r = 6$ ,  $l_e = 0$ .



cost increases; note that penalty costs remain relatively constant. Thus, expediting is not used to lower penalty costs—rather it serves to lower inventory levels.

This illustrates a unique aspect of our problem: the *three-way interaction* between holding, penalty, and ordering costs; in traditional newsvendor problems, the trade-offs are between penalty and holding only. Our separation result decouples these three costs, making the dual-sourcing problem tractable. Once  $\Delta$  and thus expediting costs are fixed, our expression for optimal  $z_e(\Delta)$  accounts for the interaction between  $p$  and  $h$  through the critical fractile.

**6.2.1. Effect of Increasing Lead Times.** In this subsection, we compare models with different lead time combinations: expediting lead time is held at zero, but the regular lead time is increased from three (in Figure 7) to six (in Figure 8). Comparing these figures, we see that, not surprisingly, the savings over single sourcing is significantly greater in the  $l_r = 6$  case than when  $l_r = 3$ . In-depth comparisons of the savings between single sourcing and the dual-index policy are more delicate. When single sourcing is using the regular supplier (at lower service levels or higher expediting cost), dual-index savings are greater in the  $l_r = 6$  case. But when expedited single sourcing is

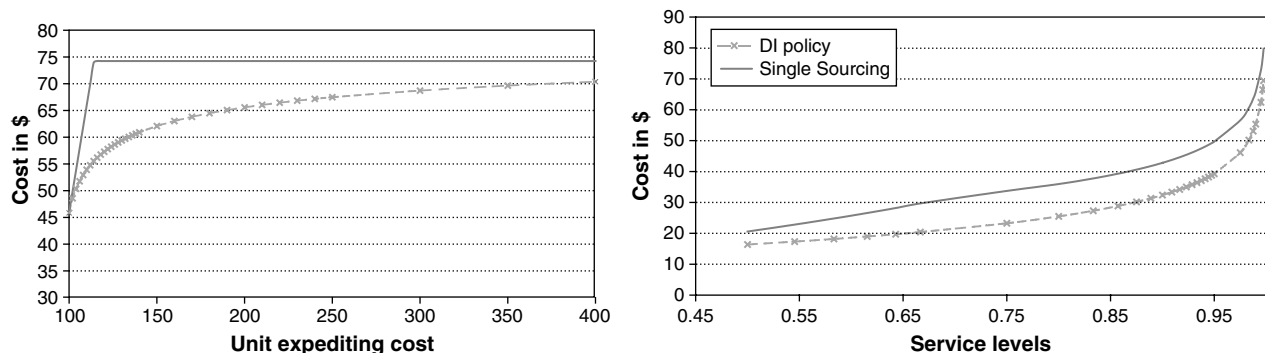
better (at higher service levels or lower expediting costs), dual-index savings are greater in the opposite case, when  $l_r = 3$ . What is important is the lead time of the additional channel the dual-index model is using—this is the advantage it has over single sourcing. Dual sourcing behaves best when this additional channel has as short a lead time as possible.

Experiments with a regular lead time of six periods and an expedited lead time of three (seen in Figure 9) show similar behavior to the figures above: the dual-index policy performs very well at low expediting costs, and the savings available due to the dual-index policy are increasing through almost the entire range of service levels. In addition, as in §6.1.5, the savings possible by using dual sourcing are reduced by a longer expedited lead time, although they are typically still significant. Overall, this section shows that the performance of the dual-index policy brings significant savings when the sourcing options differ significantly in lead times, as often is the case.

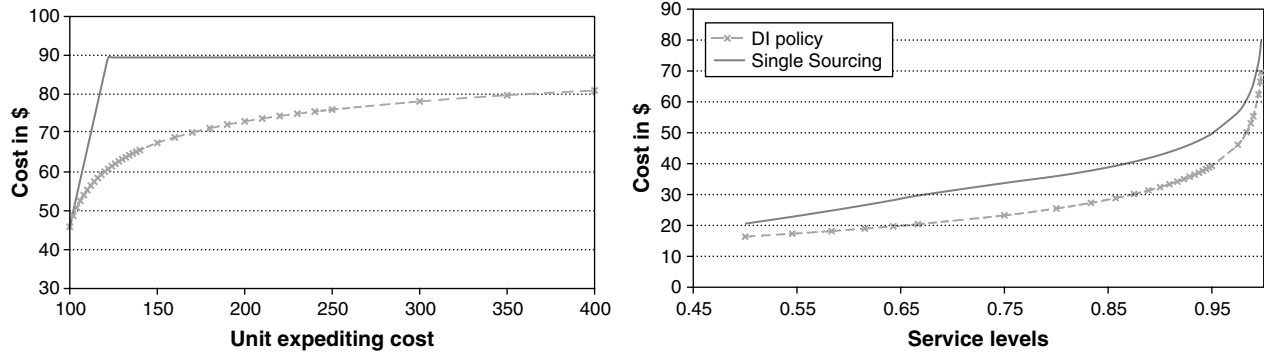
### 6.3. Effect of Increasing Variability and Sudden Demand Surges

When there is a sudden surge in demand either due to seasonal sales (e.g., the holiday season) or due to unscheduled

**Figure 7.** Optimal dual-index and single-sourcing costs and savings due to the dual-index policy:  $d \sim \exp(2)$ ,  $l_e = 0$ ,  $l_r = 3$ ,  $c_r = 100$ ,  $h = 5$  ( $p = 495$  (left),  $c_e = 110$  (right)).



**Figure 8.** Optimal dual-index and single-sourcing costs:  $d \sim \exp(2)$ ,  $l_e = 0$ ,  $l_r = 6$ ,  $c_r = 100$ ,  $h = 5$  ( $p = 495$  (left),  $c_e = 110$  (right)).



or natural events (e.g., sales of wooden boards before hurricanes), it may be crucial to dual source to maintain stable service. There is rich literature with respect to forecasting demand spikes, but when such spikes cannot be effectively forecast, a *robust reactive measure*, that can sustain and serve the demand is necessary. In this section, we examine the performance of the dual-index policy under such demand conditions.

To achieve this dual purpose of higher variability and infrequent large demands, we model the demand as a mixture of Erlang distributions. The mean demand per period is still two units as in previous sections. However, the standard deviation is higher,  $\sigma = 6$ . This corresponds to a demand which is exponential with  $\mu = 1$  with probability  $p = 0.971428$  and a mixture of 18 exponential distributions with  $\mu = 2$  otherwise. Thus, we expect small demands for most periods, with occasional very large demands.

Figure 10 shows the scenario when the expedited materials arrive immediately (i.e.,  $l_e = 0$ ) and regular lead time is three periods (the behavior of the other lead time combinations are similar). Increasing variability drives up overall costs, but the savings from the dual-index policy remain significant except at low expediting costs or very high service, in which cases very little is ordered by the regular channel. One interesting point seen in all our experiments varying service level is that there is a point at high service

level where there is an abrupt jump in percentage savings over single sourcing (see Veeraraghavan 2004). This point illustrates the reaction of the dual-index and single-sourcing policies to the necessity of maintaining high service levels when facing highly variable demands. When this is the case, huge demand spikes need to be planned for; this jump corresponds to the point where this becomes crucial, and inventory levels in both models rise precipitously, although less so for the dual-index model.

Figure 11 shows the effect of both increased variability of the demand and bigger lead time differences by increasing the regular lead time from three to six. In this case, the effect of the burstiness of demand becomes more pronounced and hence dual sourcing becomes more valuable because it takes longer for the regular source to recover from shocks.

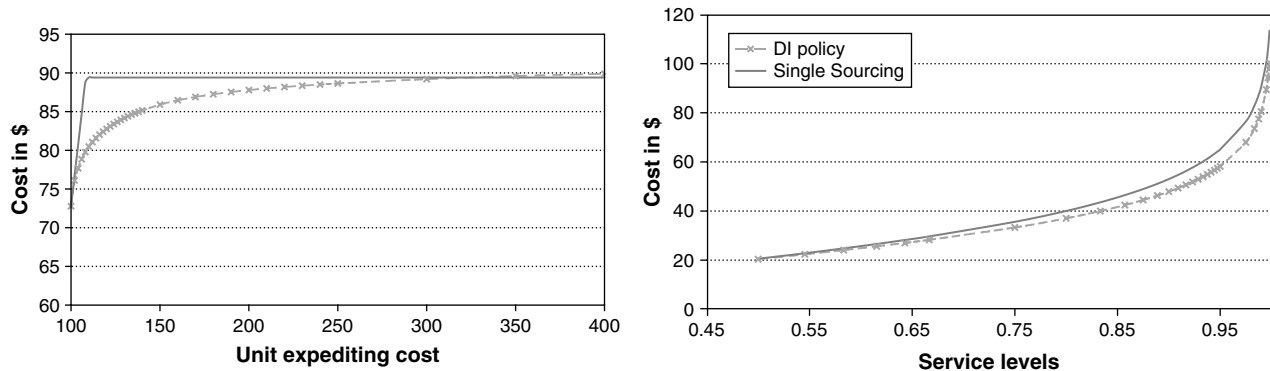
**6.4. Capacitated Ordering Systems**

In this section, we consider the effect of regular or expedited capacity limitations for three different lead time scenarios:

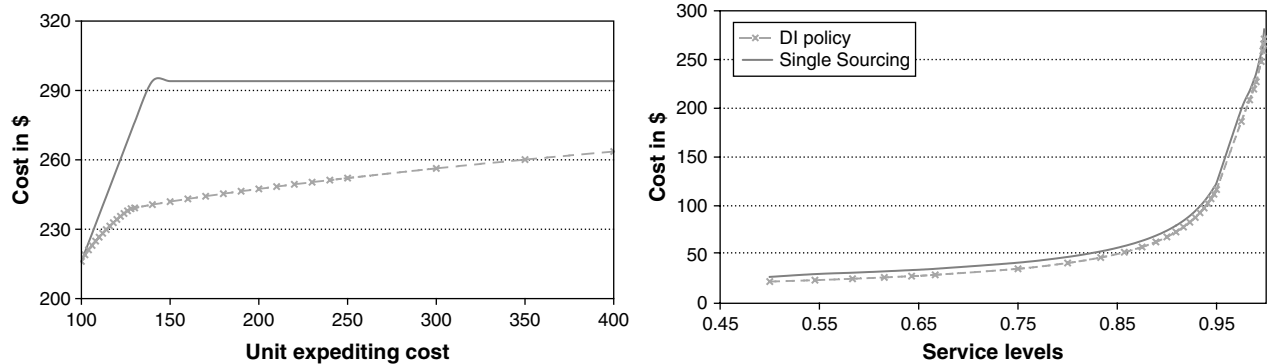
- Case 1: ( $l_e = 3$ ,  $l_r = 6$ ).
- Case 2: ( $l_e = 0$ ,  $l_r = 6$ ).
- Case 3: ( $l_e = 0$ ,  $l_r = 3$ ).

In the interest of brevity, we do not compare with single-sourcing costs, concentrating instead on the sensitivity of

**Figure 9.** Optimal dual-index and single-sourcing costs:  $d \sim \exp(2)$ ,  $l_e = 3$ ,  $l_r = 6$ ,  $c_r = 100$ ,  $h = 5$ ,  $p = 495$ .



**Figure 10.** Optimal dual-index and single-sourcing costs and savings due to the dual-index policy:  $d \sim$  mixed Erlang( $\mu = 2$ ,  $\sigma = 6$ ),  $l_e = 0$ ,  $l_r = 3$ ,  $c_r = 100$ ,  $h = 5$  ( $p = 495$  (left),  $c_e = 110$  (right)).



dual-index performance under highly and moderately variable demand in Figures 12 and 13, respectively.

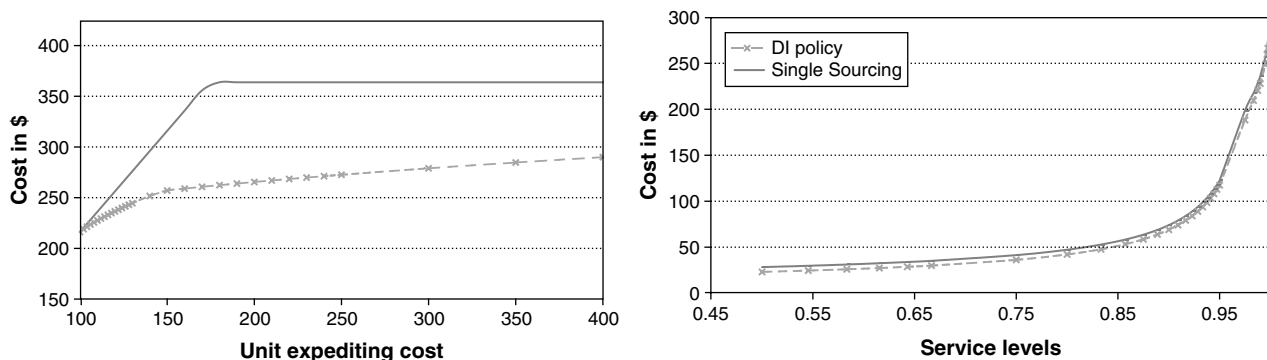
We first consider highly variable (again mixed Erlang) demand in Figure 12, reporting the change in costs when the regular orders are constrained (on the left) and the expedited orders are constrained (on the right). We observe in the left-hand chart that when the regular capacity is low, the cost for Case 1 is significantly higher, and Cases 2 and 3 are almost identical. This is because items arriving via the regular channel are used up immediately in Cases 2 and 3, while expediting is used as a reactive measure, producing goods immediately to satisfy the rest of the demand. In contrast to this, Case 1 ( $l_e = 3$ ) cannot react immediately, and inventory must be held. We also observe that all three curves decrease with capacity only until the regular capacity is greater than or equal to one (recall that the mean demand is two); after the regular capacity reaches one, additional capacity is not useful. Why? The optimal dual-index policies continue to expedite to meet demand spikes, and for smaller orders regular capacity of one unit is sufficient. This underlines a significant advantage of dual sourcing: it may facilitate significant disinvestment in regular capacity.

When expedited capacity is constrained (see the chart on the right of Figure 12), the cost behavior is different.

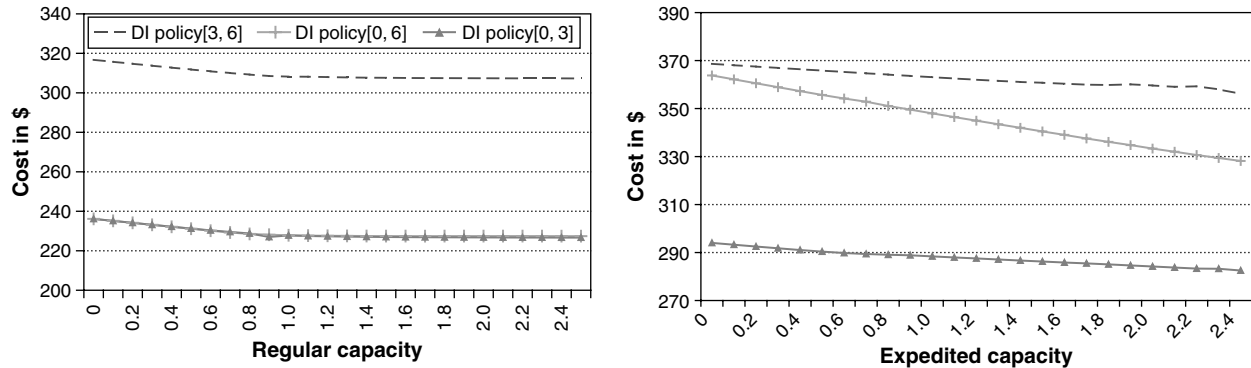
When expediting is extremely limited ( $k^e \sim 0$ ), most of the demand is satisfied through the regular channel; Case 1 and Case 2 must carry more inventory to accommodate their longer regular lead times. As the expedited capacity increases, costs fall, indicating the *value of expediting*. This marginal value is greatest for Case 2; the alternate channel is the most beneficial in this case because it reduces lead times by six. We also notice that for all three cases, the benefit of additional expedited capacity lasts much longer when compared to additional regular capacity. This is because the expedited capacity is used to recover from large demand shocks, and not the common smaller demands. Thus, in dual-sourcing systems the appropriate level of each type of capacity is determined not by the overall mean demand, but rather by the magnitude of the demands each type of capacity is satisfying.

In the left-hand chart of the exponential demand experiment in Figure 13, Cases 2 and 3 have the same cost behavior as Figure 12 for  $k^r < 1$ ; they are again holding very little inventory and using the expedited channel as a reactive measure. Once the regular capacity is greater than 1.5, the cost curves again flatten out (demand is less than 1.5 with probability 0.53), but they do diverge slightly, indicating that the differing regular lead time comes into play. As expected, the costs are again highest for Case 3 because it

**Figure 11.** Optimal dual-index and single-sourcing costs and savings due to the dual-index policy:  $d \sim$  mixed Erlang( $\mu = 2$ ,  $\sigma = 6$ ),  $l_e = 0$ ,  $l_r = 6$ ,  $c_r = 100$ ,  $h = 5$  ( $p = 495$  (left),  $c_e = 110$  (right)).



**Figure 12.** Optimal costs for various  $k^r$  (left) and  $k^e$  (right):  $c_r = 100$ ,  $c_e = 110$ ,  $h = 5$ ,  $p = 495$ ,  $d \sim \text{mixed Erlang}(\mu = 2, \sigma = 6)$ .



lacks the immediate reactive capacity. The right-hand chart of Figure 13 exhibits behaviors similar to that previously discussed in Figure 12.

Comparing both charts of Figure 13 to their counterparts in Figure 12, we arrive at a surprising conclusion—the decrease in cost with additional capacity for the two demand models are approximately equal in absolute magnitude, and significantly greater proportionally for the case of less variable demand. We hypothesize that this is because in the high variability mixed-Erlang case, most demands are small, and lower capacities are often sufficient. Thus, in the complex dual-sourcing environment, the variance of the demand in itself cannot predict the marginal value of capacity—the modality of the demand distribution is also crucial. In general, one must understand the type of variability exhibited by the demand (in this case regular demands with occasional large shocks) as well as how the optimal dual-index policy copes with the variability.

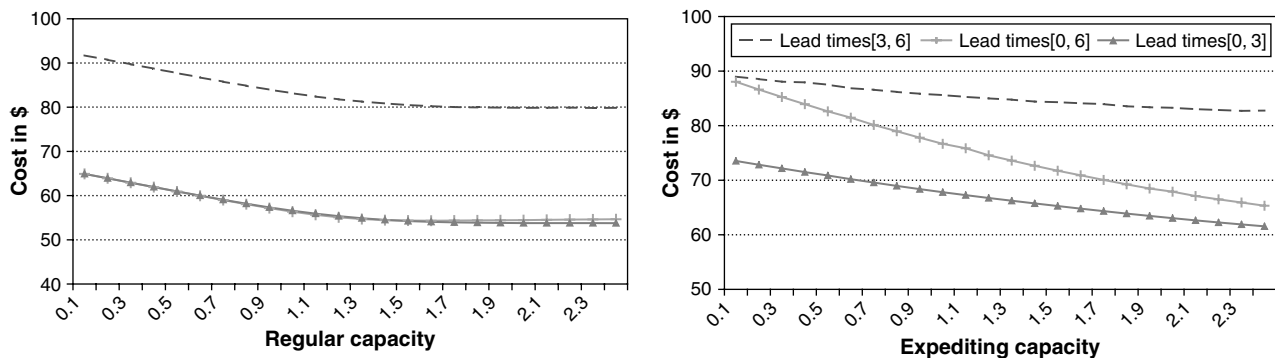
**6.5. Partitioning Capacity**

Finally, we consider the question of how to partition limited capacity. Once the capacity is partitioned, firms can reserve some portion of the capacity for each supplier. This issue might be crucial to firms that have to decide on allocating capacity (investing capital) between different supply modes, for example, when contracting for rail versus

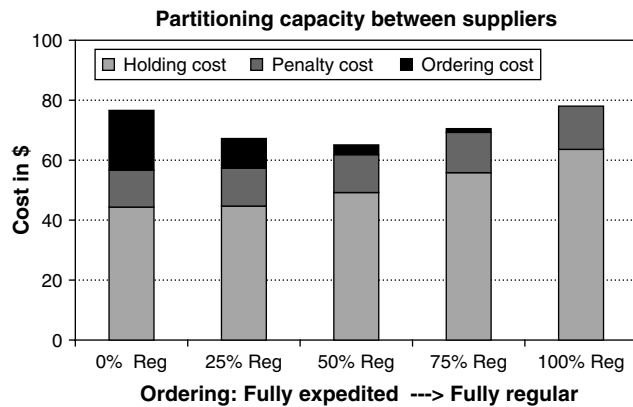
truck shipments. For instance, a firm might want to allocate 75% of its capacity to supplier A (rail) and 25% to faster but more expensive supplier B (truck), but before doing so would like to explore if a better capacity (and capital) allocation is possible. In general, solving for such a partitioning under dual sourcing is a complicated issue. In contrast, under the dual-index policy this question can be answered very quickly. We consider such an instance below.

Let the demand be exponential with mean two and costs as described in Figure 14. The firm can ship from a regular supplier over three periods but can also expedite and receive it in one period. The firm would like to consider a few different options of splitting the available ordering capacity: for example, whether to allocate 0%, 25%, 50%, 75%, or 100% of the total of four units of capacity (twice the mean demand) to the faster supply channel. The best dual-supply allocation choice among the choices considered can be found by conducting a one-dimensional search over  $\Delta$  as many times as there are allocation choices (five in this case). Even with more options, this is a computationally inexpensive method (50 seconds in this case; see Table 2). Figure 14 summarizes the cost of the optimal dual-index policy for each of these scenarios. Shipping everything through the expedited supplier reduces holding plus penalty costs slightly but greatly increases the

**Figure 13.** Optimal costs as a function of  $k^r$  and  $k^e$ :  $c_r = 100$ ,  $c_e = 110$ ,  $h = 5$ ,  $p = 495$ ,  $d \sim \text{exp}(\mu = 2)$ .



**Figure 14.** Partitioning capacity over different options:  
 $h = 5, p = 495, c_e - c_r = 10, d \sim \exp(2)$ .



expediting costs. This scenario is shown in the extreme left choice of Figure 14. At the extreme right of the figure, the choice of allocating the entire capacity to the regular supplier (100% regular ordering) is represented; the holding costs are much higher but the ordering costs are reduced. The optimal policy in this case allocates one half of the capacity to expediting. This dominates the earlier allocation that was considered by the firm (25% of the capacity to the expedited supplier). Interestingly, only a relatively small amount of expediting is done, roughly 10% of the demand, but this is able to reduce the holding costs significantly as compared to 100% regular ordering. The dual-index policy provides a quick method of calculating and choosing the best capacity allocation between the channels.

### 6.6. Summary of Computational Study

In §6.1, we see that the dual-index policy always performed within 5% of the optimal solution for all service levels and significantly better than single sourcing, especially at higher service levels (percentage savings greater than 30%). The worst dual-index performance for any unit expediting cost is within 8% for the considered cases but is often within 2%, and at moderate expediting costs may outperform single sourcing by 50%. As the demand distribution grows finer, either due to larger support or increased  $l_e$ , the dual-index policy performs extremely well in all cases; for cases where  $l_e > 0$ , i.e., when expediting goods are not

**Table 2.** Representative computational times.

$(l_e, l_r)$	Demand (min, max)	DP computational time (mins.)	Dual index time (secs.)
(0, 2)	(0, 4)	12	10
(0, 2)	(0, 8)	20	10
(0, 3)	(0, 4)	30	10
(0, 3)	(0, 8)	55	10
(1, 3)	(0, 4)	15	10
(0, 4)	(0, 4)	90	10

delivered immediately, the dual-index policy appears to be nearly optimal.

Any benefit in cost brought by using the optimal policy is tempered by the following disadvantages. The optimal policy is state dependent, and therefore complicated to implement. Further, finding the optimal actions requires computational effort; Table 2 shows representative computational times. The dual-index policy, being insensitive to problem size, is computationally far more efficient (up to 50 times as the state space grows).

Our experiments support the following observations:

- There are almost always significant savings in using the dual-index policy as compared to single sourcing, (up to 50% in some cases).
- When single sourcing is done through the regular supplier, faster expediting lead times yield greater savings in the dual-index policy. When single sourcing uses the expedited channel, the value of the dual-index policies is greatest with shorter regular lead times. Thus, the critical parameter is not the difference in the lead times,  $l$ , but rather the speed of the mode of delivery the single sourcing solution is not using, because this is the degree of additional flexibility the dual-index policy adds.
- The savings of the dual-index policy are highest when the operational costs of getting material solely through one or the other channel are equivalent: dual sourcing is of greatest value when the manufacturer is indifferent between two sourcing channels.
- Increased regular capacities are crucial when they are lower than or comparable to the mean “typical” demand. The benefit of additional expedited capacity lasts much longer compared to additional regular capacity. Thus dual sourcing may facilitate disinvestment in regular capacity.
- Surprisingly, our experiments show that extra capacity may not be more valuable when demand is highly bursty than when it is more regular, particularly if the bulk of the demands are small, and can be served with the lower capacities. Thus, in dual sourcing the role of capacities is very complex.
- Partitioning capacities to expedited orders can lead to significant savings in holding costs with only a small amount of expedited ordering.
- Expediting primarily drives down holding, rather than penalty costs; it allows high service with less inventory.

## 7. Conclusions and Future Directions

The dual-sourcing decision under general lead times has been a challenging problem for over 40 years, despite its frequency in practice. We have devised an easy-to-implement order-up-to policy that uses the regular and expedited inventory positions in making ordering decisions. We have likewise shown how to efficiently find optimal parameters for our policy. Our policy is globally optimal when lead times differ by one period, and thus we provide a method of efficiently finding the globally optimal policy parameters for this case. Finally, our method easily

extends to a variety of related but more general models including capacities, random yield, nonstationary demand, returns, supply disruptions, and some cases of random lead times. Further work in these settings is anticipated in the future.

Computational experiments show that our dual-index policy mimics the behavior of the optimal policy remarkably well, and that dual sourcing is especially beneficial when service levels are high (high loss of goodwill costs), when expediting costs are moderate, or when single sourcing via the expedited or regular channels have similar costs. Thus, we have provided a simple, practical policy which allows the industry to take nearly full advantage of dual-sourcing flexibility in a variety of very general environments.

## Acknowledgments

The authors thank Fangruo Chen, an anonymous associate editor, and three anonymous referees for their suggestions, which improved the exposition and focus of the paper.

## References

- Alfredsson, P., J. Verrijdt. 1999. Modeling emergency supply flexibility in a two-echelon inventory system. *Management Sci.* **45** 1416–1431.
- Barankin, E. W. 1961. A delivery-lag inventory model with an emergency provision. *Naval Res. Logist. Quart.* **8** 285–311.
- Bellman, E. 2005. Nokia to build cellphone plant in India to meet rising demand. *Wall Street Journal* (April 7) B4.
- Bertsekas, D. P. 1995. *Dynamic Programming and Optimal Control*, Vol. 2. Athena Scientific, Belmont, MA.
- Bradley, J. R. 2004. A Brownian motion approximation of a production-inventory system with a manufacturer that subcontracts. *Oper. Res.* **52**(5) 765–784.
- Bulinskaya, E. 1964. Some results concerning optimal inventory policies. *Theory Probab. Appl.* **9** 502–507.
- Callioni, G., X. De Montgros, R. Slagmulder, L. Van Wassenhove, L. Wright. 2005. Inventory-driven costs. *Harvard Bus. Rev.* **83**(3) 135–141.
- Carter, J. R., S. K. Vickery. 1988. Managing volatile exchange rates in international purchasing. *J. Purchasing and Materials Management* **24**(4) 13–20.
- Daniel, K. H. 1962. A delivery-lag inventory model with emergency order. H. Scarf, D. Gilford, M. Shelly, eds. *Multistage Inventory Models and Techniques*, Chapter 2. Stanford University Press, Stanford, CA.
- Davis, E. W. 1992. Global outsourcing: Have U.S. managers thrown the baby out with the bath water? *Bus. Horizons* **35**(4) 58–65.
- Feng, Q., G. Gallego, S. Sethi, Y. Houmin, H. Zhang. 2004. Optimality and non-optimality of base-stock policies in inventory systems with multiple delivery modes. *J. Indust. Management Optim.* **2**(1) 19–42.
- Fukuda, Y. 1964. Optimal policies for the inventory problem with negotiable lead time. *Management Sci.* **10** 690–708.
- Gottfredson, M., R. Puryear, S. Phillips. 2005. Strategic sourcing: From periphery to the core. *Harvard Bus. Rev.* **83**(2) 132–139.
- Groenevelt, H., N. Rudi. 2002. A base stock inventory model with possibility of rushing part of order. Working paper, Simon School of Business, University of Rochester, Rochester, NY.
- Hendricks, K. B., V. R. Singhal. 2005. Association between supply chain glitches and operating performances. *Management Sci.* **51**(5) 695–711.
- Kelleher, K. 2003. Why FedEx is gaining ground. *Business 2.0* (October) 56–57.
- Lawson, D. G., E. L. Porteus. 2000. Multistage inventory management with expediting. *Oper. Res.* **48**(6) 878–893.
- Li, C., P. Kouvelis. 1999. Flexible and risk sharing supply contracts under price uncertainty. *Management Sci.* **45**(10) 1378–1398.
- Moinzadeh, K., S. Nahmias. 1988. A continuous review model for an inventory system with two supply modes. *Management Sci.* **26** 483–494.
- Moinzadeh, K., C. P. Schmidt. 1991. An  $(S-1, S)$  inventory system with emergency orders. *Oper. Res.* **39**(3) 308–321.
- Plambeck, E., A. Ward. 2006. Note: A separation principle for a class of assemble-to-order systems with expediting. *Oper. Res.* **55**(3) 603–609.
- Rao, U., A. Scheller-Wolf, S. Tayur. 2000. Development of a rapid-response supply chain at Caterpillar. *Oper. Res.* **48**(2) 189–204.
- Scheller-Wolf, A., S. Veeraraghavan, G. J. Van Houtum. 2005. Inventory models with expedited ordering: Single index policies. Working paper, Tepper School of Business, Carnegie Mellon University, Pittsburgh.
- Souder, E. 2004. Retailers rely more on fast deliveries. *Wall Street Journal* (January 14).
- Tagaras, G., D. Vlachos. 2001. A periodic review inventory system with emergency replenishments. *Management Sci.* **47**(3) 415–429.
- Tayur, S. R., M. J. Magazine, R. Ganeshan. 1998. *Quantitative Models for Supply Chain Management*. Kluwer Academic Publishers, Norwell, MA.
- Tomlin, B. T. 2006. On the value of mitigation and contingency strategies for managing supply-chain disruption risks. *Management Sci.* **52**(5) 639–657.
- Veeraraghavan, S. 2004. Supply choice and capacity decisions under uncertainty. Doctoral dissertation, Tepper School of Business, Carnegie Mellon University, Pittsburgh.
- White, C. 2005. Supply chain “best practices” raise risks of disruption. *Dow Jones News Letter* (November 2).
- Whittmore, A. S., S. C. Saunders. 1977. Optimal inventory under stochastic demand with two supply options. *SIAM J. Appl. Math.* **32** 293–305.