MANAGEMENT SCIENCE
Vol. 56, No. 8, August 2010, pp. 1285–1303
ISSN 0025-1909 | EISSN 1526-5501 | 10 | 5608 | 1285
DOI 10.1287/mnsc.1100.1184

Optimal Flexibility Configurations in Newsvendor Networks: Going Beyond Chaining and Pairing

Achal Bassamboo
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208,
a-bassamboo@kellogg.northwestern.edu

Ramandeep S. Randhawa
Marshall School of Business, University of Southern California, Los Angeles, California 90089,
ramandeep.randhawa@marshall.usc.edu

Jan A. Van Mieghem
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208,
vanmieghem@kellogg.northwestern.edu

We study the classical problem of capacity and flexible technology selection with a newsvendor network model of resource portfolio investment. The resources differ by their level of flexibility, where "level-k flexibility" refers to the ability to process k different product types. We present an exact set-theoretic methodology to analyze newsvendor networks with multiple products and parallel resources. This simple approach is sufficiently powerful to prove that (i) flexibility exhibits decreasing returns and (ii) the optimal portfolio will invest in at most two, adjacent levels of flexibility in symmetric systems, and to characterize (iii) the optimal flexibility configuration for asymmetric systems as well. The optimal flexibility configuration can serve as a theoretical performance benchmark for other configurations suggested in the literature. For example, although *chaining* is not optimal in our setting, the gap is small and the inclusion of scale economies quickly favors chaining over *pairing*. We also demonstrate how this methodology can be applied to other settings such as product substitution and queuing systems with parameter uncertainty.

Key words: inventory production; stochastic models; programming; linear; applications; queues; networks; flexibility; newsvendor networks
History: Received January 11, 2009; accepted March 15, 2010, by Yossi Aviv, operations management. Published online in *Articles in Advance* June 9, 2010.

1. Introduction and Summary of Results

When a firm produces several products, should different products share resources or should the firm establish dedicated resources for some of them? The polar extremes of total specialization and full resource sharing or "pooling" are well studied, but intermediate configurations with partial resource sharing are often more appropriate. We study this classic problem of capacity and flexible technology selection with a newsvendor network model, introduced by Van Mieghem (1998) and Van Mieghem and Rudi (2002) as the multidimensional generalization of the familiar two-stage decision problem with recourse in operations research. In stage 1, the firm invests in a portfolio of different resources to produce N different types of products knowing only their demand distribution. In stage 2, product demands are observed and allocated to the resources to maximize profits.

The resources differ by their technology or level of flexibility, which we model as follows. Let "level-k flexibility" refer to the ability to process k different product types. There are $\binom{N}{k} = N!/((N-k)!\,k!)$ different resources with level-k flexibility, including N dedicated or specialized resources with $k=1$ and one fully flexible resource with $k=N$. The firm's capacity or flexibility portfolio (we will use both terms interchangeably) is denoted by the vector K of the capacities of the 2^N-1 different resources.

The products have unit shortage penalty costs denoted by p_i, $i=1,2,\ldots,N$. We first consider a linear cost structure in capacity size and flexibility level: each unit of capacity of a level-k flexible resource costs $c_k = c_1(1+(k-1)\delta)$, where $\delta \in (0,1)$ denotes the flexibility premium. The firm's objective is to select the capacity portfolio $K^\star$ that maximizes its expected profits. This is equivalent to determining the minimal total cost, which is the sum of the expected shortage and capacity costs.

We present an exact set theoretic methodology and characterize the optimal flexibility configuration for parallel newsvendor networks with N products.

Following Van Mieghem (1998), our methodology expresses the marginal value of capacity in terms of demand shortage regions. The analysis in that paper, as in most follow-up work, relies on explicit descriptions of the shortage regions and focuses on only two dimensions. The novelty in our approach is that we express the marginal value of level-k resources using a set theoretic approach that does not require an explicit description of these regions. This abstract yet simple approach extends to N dimensions. This simple approach is sufficiently powerful to prove that (i) flexibility exhibits decreasing returns and (ii) the optimal portfolio will invest in at most two, adjacent levels of flexibility in symmetric systems, and to characterize (iii) the optimal flexibility configuration for asymmetric systems as well.

The precise statements of these results are provided in later sections of this paper; here we illustrate them for a parallel newsvendor network with $N = 4$ products. A capacity portfolio then can consist of four dedicated resources, six level-2 resources, four level-3 resources, and one fully flexible resource. For expositional simplicity, let us assume here that the demand distribution and unit shortage penalties are product-type independent, or symmetric. The optimal capacity portfolio is then shown to be also product-type independent, and all level-k flexible resources have the same capacity $K_k^\star$. The 15-variable problem thus reduces to one with four variables that we determine as follows.

First, let Ω_i denote the set of demand realizations where there is a shortage of dedicated resource i for a given capacity portfolio K. Then, the expected marginal value of investing additional capacity in dedicated resource i (beyond that in portfolio K) equals the per-unit shortage cost p times the probability that the realized demand lies in the set Ω_i, i.e., $p\mathbb{P}(\Omega_i)$. Now consider a level-2 flexible resource that can process products i and j. Then, one extra unit of its capacity can be used to decrease the contingent shortage of product *i or j*. With a common shortage cost, the marginal value of such a resource is $p\mathbb{P}(\Omega_i \cup \Omega_j)$. Furthermore, in a symmetric system we have that $\mathbb{P}(\Omega_i \cup \Omega_j) = \mathbb{P}(\Omega_1 \cup \Omega_2)$. In general, the marginal value of a level-k flexible resource in a symmetric system equals $p\mathbb{P}(\bigcup_{i=1}^{k} \Omega_i)$. Clearly, as expected, the marginal value increases in the level of flexibility. More importantly, this increase is strictly concave so that there are *decreasing returns to flexibility* in newsvendor networks. Indeed, simple set algebra shows that

$$\begin{aligned} &p\mathbb{P}(\Omega_1 \cup \Omega_2 \cup \Omega_3 \cup \Omega_4) - p\mathbb{P}(\Omega_1 \cup \Omega_2 \cup \Omega_3) \\ &\quad < p\mathbb{P}(\Omega_1 \cup \Omega_2 \cup \Omega_3) - p\mathbb{P}(\Omega_1 \cup \Omega_2) \\ &\quad < p\mathbb{P}(\Omega_1 \cup \Omega_2) - p\mathbb{P}(\Omega_i). \end{aligned}$$

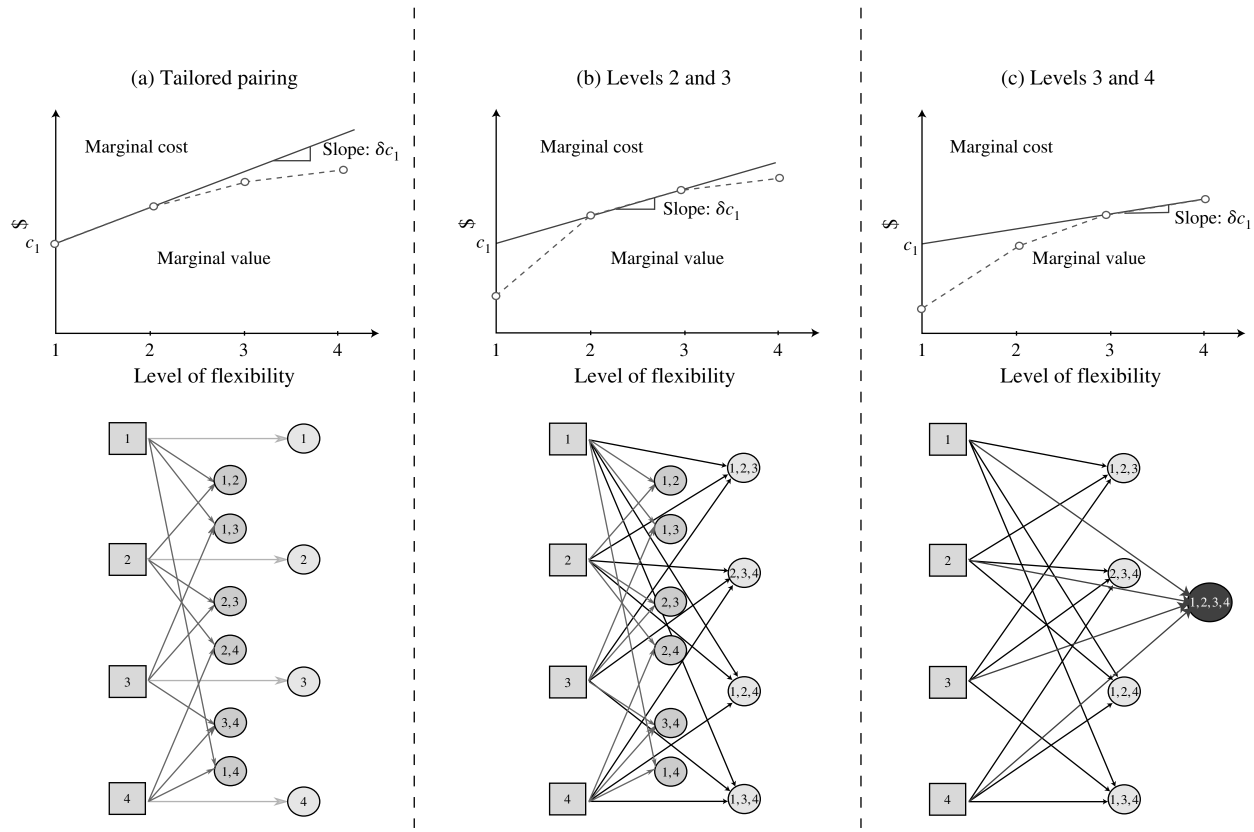
The optimality conditions require that the marginal value of any *positive* investment in a level-k resource must equal its marginal cost. Given that the marginal value is strictly concave in the level of flexibility while the marginal cost is linear, this provides a restriction on potential optimal portfolios. Specifically, for this symmetric system, we obtain that the optimal capacity portfolio invests in at most two adjacent levels of flexibility. Imagine a graph where product types are represented by rectangles and resources by circles. An arc from a rectangle to a circle then represents a possible product–resource assignment; the number of arcs into a circle equals that resource's level of flexibility. Our main result can then be graphically illustrated as in Figure 1. (We use the convention that the numbers in rectangles denote product types, and those in circles represent the products that can be produced by the corresponding resources.) There can only be three optimal flexibility configurations in symmetric newsvendor systems with $N = 4$ products: invest only in resources with flexibility levels (a) one and two, (b) two and three, or (c) three and four. The configuration that invests in levels one and two is referred to as "tailored pairing." In this configuration, each product can be produced by a pair of level-2 resources, and such configurations are referred to as "pairing" in Bassamboo et al. (2010a). The adjective "tailored" refers here to using mostly dedicated resources to serve the average demand and a small amount of only level-2 flexibility to serve variability.

Figure 1 further demonstrates the fact that it will be optimal to invest in higher levels of flexibility only for lower flexibility premiums. We also prove that, as expected, the value of flexibility decreases as any pairwise demand correlation increases.

We investigate the key drivers of the optimal flexibility configurations using numerical studies. When the cost of flexibility δ rises, the investment in higher levels of flexibility is substituted for lower levels. Interestingly, the associated capacity levels are nonmonotone. For example, as δ rises, the investment in level-4 flexibility falls while the level-3 investment initially rises. When level-4 capacity reaches zero, level-3 capacity peaks and then falls (and level-2 flexibility rises as a substitute), as further discussed in §4.

Our main results characterize the optimal flexibility configuration, which can serve as a theoretical performance benchmark for other configurations suggested in the literature. In particular, the seminal paper by Jordan and Graves (1995) showed that "a little flexibility can achieve almost all the benefits of total flexibility" by using *only* level-2 flexible resources in a special configuration called chaining. Chaining represents any flexibility configuration of N level-2 flexible resources that are connected, directly or indirectly,

Figure 1 A Graphical Representation of the Optimality Conditions and Portfolios for an $N = 4$ Product Setting



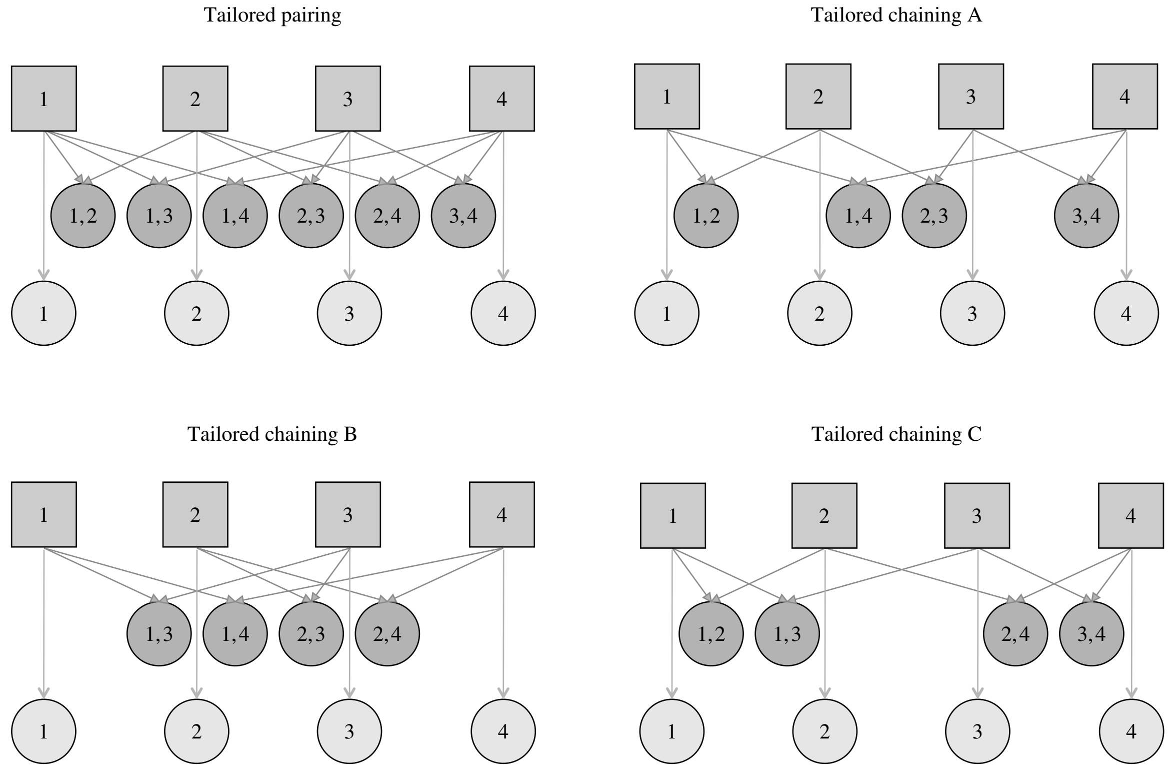
Notes. The optimal configuration invests in flexibility to match the marginal value with the marginal cost. Depending on the cost parameters, only one of three configurations can be optimal: the optimal portfolio invests only in resources with flexibility levels (a) one and two ("tailored pairing"), (b) two and three, or (c) three and four.

to all N product types by product–resource assignments. Chaining allows for shifting capacity from products with lower than expected demand to those with higher than expected demand. Using simulation and providing some analytical justification for the same newsvendor network model as we study, Jordan and Graves (1995) demonstrated that the expected shortfall and capacity utilization of chained level-2 flexible resources is close to the expected shortfall and utilization of fully flexible resources with the same capacity. In other words, a little flexibility goes a long way. Graves and Tomlin (2003) showed that similar chaining benefits extend to multistage systems.

We prove analytically that flexibility exhibits decreasing returns in newsvendor networks. This provides the general, mathematical confirmation that "a little flexibility goes a long way" and corroborates the virtues of chaining. At the same time, we also prove that chaining is *not* optimal in symmetric newsvendor systems with a linear cost structure and more than three products. Indeed, if the optimal configuration in such systems invests in level-2 flexibility, it *must* invest equally in all $N(N-1)/2$ level-2 flexible resources. That is, each product can then be produced by a pair of level-2 resources in this pairing configuration. In contrast, chaining uses only N level-2 flexible resources, or $N(N-3)/2$ less than pairing, and hence is suboptimal for symmetric newsvendor networks. Figure 2 demonstrates the three chaining configurations possible for $N = 4$, and the pairing configuration that invests in all $N(N-1)/2 = 6$ resources. Recall that the adjective "tailoring" refers to using mostly dedicated resources to serve the average demand and a small amount of only level-2 flexibility to serve variability. (Note that with three products, chaining and pairing are identical.)

In practice, capacity investment often enjoys scale economies that induce a firm to invest in fewer but bigger resources. Although general scale economies pose mathematical challenges, we are able to extend our main analytic result to a setting where a positive capacity investment incurs a fixed cost. Clearly, with setup costs, the optimal capacity portfolio need no longer be symmetric (not even in a symmetric system). The contribution of this result is to significantly reduce computational time in finding the optimal strategy. As expected, our numerical study confirms that scale economies diminish the practical value of

Figure 2 With $N = 4$ Product Types There Exists Only One Tailored Pairing Configuration But Three Tailored Chaining Configurations



pairing and favor chaining, and increasingly so when the number of products rises. We also extend our results to a setting where capacity costs are concave in the level of flexibility. The optimal levels of flexibility increase with increasing concavity or "scope economies."

Our work continues the line of literature on flexible technology, started by Fine and Freund (1990) and followed by Gupta et al. (1992), Jordan and Graves (1995), and Van Mieghem (1998), among others. Bish and Wang (2004) and Chod and Rudi (2005) added pricing to the flexibility problem. Newsvendor network models have also been used to study sourcing or input flexibility (e.g., Tomlin and Wang 2005, Tomlin 2006), transshipment (e.g., Dong and Rudi 2004 and references therein), and part substitution and commonality (e.g., Gerchak and Henig 1989, Van Mieghem 2004). Goyal and Netessine (2007) studied flexibility strategies in competitive newsvendor networks. We briefly consider the case of product substitution by the firm (e.g., Bassok et al. 1999, Netessine et al. 2002). We show how the insights and methodology derived in this paper can be extended to include cases where the firm can satisfy demand by substituting products.

We also relate our findings to recent studies of flexibility in queuing systems, e.g., Sheikhzadeh et al. (1998), Gurumurthi and Benjaafar (2004), Iravani et al. (2005), Wallace and Whitt (2005), and Gurvich and Whitt (2010). In a recent work, Bassamboo et al. (2010a) proved that "a little flexibility is all you need" in symmetric queuing systems. To be precise, they show that tailored pairing is asymptotically optimal in queuing systems with large arrival rates. Queuing systems with independent arrival and service times enjoy statistical pooling that ultimately make variability a second-order effect compared to the mean demand. The resulting asymptotic optimality of tailored pairing agrees with our optimality results here. Indeed, we show that if the optimal capacity portfolio invests in dedicated resources, tailored pairing is also optimal in newsvendor networks. Yet variability can be a first-order effect in newsvendor networks, and we show that the optimal flexibility configurations exhibit more richness and can "go beyond" tailored pairing and invest in higher levels of flexibility.

Realizing that traditional asymptotic queuing analysis relegates uncertainty to a second-order effect, recent studies have investigated queuing systems with arrival rate uncertainty, e.g., Harrison and Zeevi (2005), Whitt (2006), and Bassamboo et al. (2006). In such queuing systems, capacity decisions reduce to a newsvendor network problem of the type studied here. This shows that our results also apply to flexibility configurations in queuing systems.

The remainder of this paper starts with a model description. Section 3 illustrates our approach for symmetric systems and characterizes the optimal flexibility configuration. The key drivers of that configuration are investigated with a numerical study in §4. Section 5 extends our analytic results to newsvendor networks with (a) asymmetric demand distributions, (b) scale economies, and (c) scope economies. We discuss the case of product substitution and the connection of our results to queuing systems in §6 and present concluding remarks in §7. For pedagogical reasons, the proofs of results in §4 are provided in the main text, whereas those of results in §§5 and 6 (which have the same underlying principles as those in §4) and supporting results are relegated to the appendices.

2. A Newsvendor Network Model of Flexibility

We consider a multidimensional two-stage decision problem with recourse. In stage 1, the firm invests in a portfolio of different resources to produce N different types of products knowing only their demand distribution. In stage 2, product demands are observed and allocated to the resources to maximize profits.

The resources have different levels of flexibility, which we model as follows. Let "level-k flexibility" refer to the ability to process $k \in \{1, 2, \ldots, N\}$ different product types. To specify which k product types a given resource can produce, we refer to that resource by the set of product types $F \subseteq \{1, 2, \ldots, N\}$ it can produce. The cardinality of F thus equals the resource's level of flexibility. We assume that each unit of production consumes one unit of capacity, irrespective of the resource and product, and we denote the maximal number of units that resource F can produce by its capacity K_F. There are $\binom{N}{k} = N!/((N-k)!\,k!)$ different resources with level-k flexibility, including N dedicated or specialized resources with $k=1$ and one fully flexible resource with $k=N$. The firm's capacity or flexibility portfolio (we will use both terms interchangeably) is denoted by the vector $K := \{K_F \colon F \subseteq \{1, \ldots, N\}\}$ and can comprise up to $\sum_{k=1}^{N}\binom{N}{k} = 2^N - 1$ different resources.

We denote the demand for product i by the random variable D_i. The product vector D has a general distribution with probability measure $\mathbb{P}$, and $\mathbb{E}_D$ will denote the associated expectation operator. For analytic simplicity, we will assume that the probability density function exists and is positive over $\mathbb{R}_+^N$.

The products have unit shortage penalty costs denoted by $p_i > 0$. We first consider a linear cost structure in capacity size and flexibility level: each unit of capacity of a level-k flexible resource costs $c_k = c_1(1 + (k-1)\delta)$, where $\delta \in (0,1)$ denotes the flexibility premium. Clearly, a fully flexible resource would dominate all resources with lower levels of flexibility if $\delta = 0$. Similarly, k dedicated resources would dominate one level-k flexible resource if $\delta \geq 1$.

The firm's objective is to select the capacity portfolio $K^\star$ that maximizes its expected profits. This profit maximization problem is equivalent to determining the minimal total cost $\Pi^\star$, which is the sum of the expected shortage and capacity costs, as follows:

$$\Pi^\star = \min_K \left\{ \mathbb{E}_D \pi(K, D) + \sum_{F \subseteq \{1, \ldots, N\}} c_{|F|} K_F \right\}, \tag{1}$$

where $\pi(K, D)$ is the optimal contingent operating profit:

$$\pi(K, D) = \min_{x \geq 0} \sum_{i=1}^{N} p_i \left(D_i - \sum_{\{F \colon i \in F\}} x_{i,F} \right), \tag{2}$$

$$\sum_{\{F \colon i \in F\}} x_{i,F} \leq D_i \quad \text{for all } i = 1, \ldots, N, \tag{3}$$

$$\sum_{i \in F} x_{i,F} \leq K_F \quad \text{for all } F \subset \{1, \ldots, N\}. \tag{4}$$

In the second-stage problem (2)–(4), the firm must allocate the chosen capacity portfolio K to the observed demand vector D to minimize shortage costs. In this allocation problem, $x_{i,F}$ denotes the amount of product i produced by resource F. Thus, $\sum_{\{F \colon i \in F\}} x_{i,F}$ denotes the total production of product i and $D_i - \sum_{\{F \colon i \in F\}} x_{i,F}$ represents its shortage. Summing the shortages of all products yields the total shortfall. This optimization is subject to the usual demand and capacity constraints. The demand constraints (3) ensure that supply (production) does not exceed the demand for any product; the capacity constraints (4) reflect that supply cannot exceed capacity.

Linear programming theory shows that the optimal contingent operating profit $\pi(K, D)$ is jointly convex in K and D. Furthermore, it also is supermodular in D as we show in Appendix A (cf. Corollary 3). As a linear continuous superposition, the expected operating profit $\mathbb{E}_D \pi(K, D)$ is *strictly* convex in the capacity portfolio K. Hence, the optimal portfolio $K^\star$ is unique and solves the necessary and sufficient first-order or Karush–Kuhn–Tucker (KKT) conditions:

$$K^\star \cdot (c - \nabla_K \mathbb{E}_D \pi(K^\star, D)) = 0, \tag{5}$$

$$c \geq \nabla_K \mathbb{E}_D \pi(K^\star, D). \tag{6}$$

Differentiation and expectation can be interchanged and the optimality conditions simplify to

$$\nabla_K \mathbb{E}_D \pi(K, D) = \mathbb{E}_D \nabla_K \pi(K, D) = \sum_j \tilde{\lambda}_j \mathbb{P}(\tilde{\Omega}_j(K)), \tag{7}$$

where $\{\tilde{\Omega}_j\}$ represents a partition of the demand space such that $\bigcup_j \tilde{\Omega}_j = \mathbb{R}_+^N$ and $\tilde{\lambda}_j$ are the constant Lagrange multipliers of the capacity constraints for $D \in \tilde{\Omega}_j$. Equation (7) can be made rigorous by using arguments similar to those of Harrison and Van Mieghem (1999, Proposition 2).

Intuitively, the (polyhedral) demand sets $\tilde{\Omega}_j$ correspond to a set of demand samples for which some resources are capacity constrained. We will refer to these sets as shortage regions and to the corresponding resources as conditional bottlenecks. Their associated component in the Lagrange vector $\tilde{\lambda}_j$, that corresponds to the optimal dual vector, is the conditional marginal value of their capacity. In general, the Lagrange vectors and the domains depend on the capacity portfolio K and on the relative profitability of products. We observe that any positive component of λ_j must equal the shortage penalty of some product p_i. Indeed, an $\epsilon > 0$ increase in the capacity of a conditional bottleneck resource can be used to reduce the shortage of some product i by $p_i\epsilon$ (see Lemma 1 in Appendix A for a formal proof).

In this paper, we use set theory to express the marginal value of a flexible resource in terms of the shortage regions of the dedicated resources. This allows us to characterize the value of flexible resources and develop insights into optimal flexibility portfolios.

To illustrate the mode of analysis, we begin by analyzing a symmetric system, where the shortage penalty is the same for all products and denoted by p, and the demand distribution $G(x) = \mathbb{P}(D < x)$ is symmetric in its components, meaning that $G(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_N) = G(x_1, \ldots, x_j, \ldots, x_i, \ldots, x_N)$ for any $1 \le i < j \le N$. Henceforth, we shall simply call such distributions "symmetric," and we shall generalize to asymmetric demand distributions in §5.1.

The symmetry assumption allows further simplifications. First, any positive components in any Lagrange multiplier vector $\tilde{\lambda}_j$ equal the common shortage penalty p. Furthermore, we will be able to specify the optimality equations in terms of N shortage regions Ω_i, where Ω_i is the shortage region of dedicated resource i. (Formally, it is the set of demand realizations where the dual variable corresponding to resource $\{i\}$ equals p. See Definition 1 in Appendix A for further details.) We will see that these shortage regions Ω_i typically are not disjoint, and hence differ from the regions $\tilde{\Omega}_j$ in (7).

3. Optimal Flexibility Portfolios for Symmetric Systems

System symmetry together with the uniqueness of the optimal portfolio $K^\star$ implies that the optimal portfolio must also be symmetric and product-type independent. This greatly simplifies the analysis because we can restrict attention to portfolios of the form $(K_1, K_2, \ldots, K_N)$, where K_k is the capacity of *each* of the $\binom{N}{k}$ different resources with level-k flexibility. Thus, instead of having $2^N - 1$ variables, the optimal solution is characterized by only N variables.

3.1. Shortage Regions and Marginal Value of Dedicated Resources

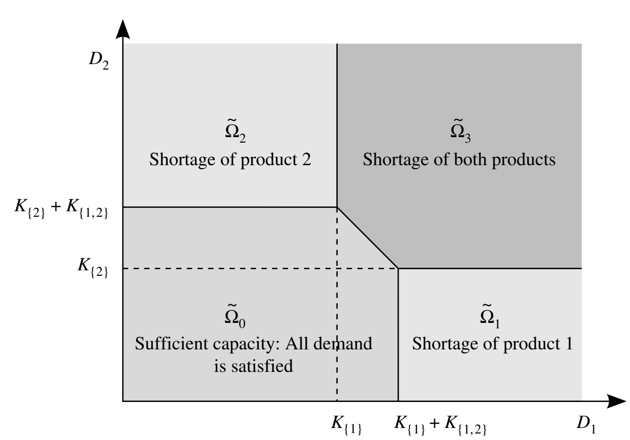
Let $\Omega_i(K)$ denote the shortage region for product i under the capacity portfolio K. The marginal value of investing in dedicated resource i then is $p\mathbb{P}(\Omega_i)$. To better understand the shortage regions, we will illustrate them for the one- and two-product cases.

The single-product case ($N = 1$) is the standard newsvendor problem and the capacity portfolio consists of only a dedicated resource with capacity $K_{\{1\}}$. The resource is a contingent bottleneck when $D_1 > K_{\{1\}}$ so that its shortage region is $\Omega_1(K) = \{D_1 > K_{\{1\}}\}$, and the expected marginal value of capacity is $p\mathbb{P}(\Omega_1(K)) = p\mathbb{P}(D_1 > K_{\{1\}})$.

For the two-product setting ($N = 2$), for a given capacity vector K, Figure 3 depicts four mutually exclusive scenarios. Denoted by $\tilde{\Omega}_0$ is the scenario where capacity exceeds demand and no product experiences a shortage. The contingent marginal value of all three capacities is zero. In contrast, when demand falls in $\tilde{\Omega}_1$, there is a shortage of product 1, but abundant capacity to meet product 2 demand. Resources $\{1\}$ and $\{1, 2\}$ are contingent bottlenecks with marginal value p: an increase in the capacity of either resource will decrease the product 1 shortfall when demand falls in $\tilde{\Omega}_1$, but an increase in resource $\{2\}$ capacity would be valueless. The scenario $\tilde{\Omega}_2$ is the analog of scenario $\tilde{\Omega}_1$, where product 2 experiences a shortage but not product 1. Finally, both products experience a shortage in region $\tilde{\Omega}_3$ in the sense that an increase in any capacity decreases the total shortfall.

We can combine the above scenarios to obtain the shortage regions as follows: the shortage region of resource $\{1\}$ is $\Omega_1(K) = \tilde{\Omega}_1 \cup \tilde{\Omega}_3$, and the marginal value of dedicated resource 1 equals $p\mathbb{P}(\Omega_1)$. Similarly, we

Figure 3 The Demand Space for a Two-Product Setting



Notes. With two products, the demand space is partitioned into four regions: All demand can be satisfied in $\tilde{\Omega}_0$. There is a shortage of only product 1 in $\tilde{\Omega}_1$ and of only product 2 in $\tilde{\Omega}_2$; both products may experience shortages in $\tilde{\Omega}_3$

have $\Omega_2(K)=\tilde{\Omega}_2\cup\tilde{\Omega}_3$, and the marginal value of dedicated resource 2 equals $p\mathbb{P}(\Omega_2)$.

3.2. Characterizing the Optimal Portfolio

We use the shortage regions of the dedicated resources to characterize the optimal portfolio. There are two important steps in our approach.

First, we express the marginal value of investing in a level-$k>1$ flexible resource in terms of the shortage regions $\Omega_i(K)$, as follows. By definition, a marginal unit of a level-2 flexible resource that can produce products $\{i,j\}$ is equivalent to a marginal unit of dedicated resource i *or* j, where the choice of product i or j typically depends on the realized demand. Regardless of the product produced, the marginal value of level-2 capacity is $p\mathbb{P}(\Omega_i\cup\Omega_j)$. Given the symmetry, $p\mathbb{P}(\Omega_i\cup\Omega_j)=p\mathbb{P}(\Omega_1\cup\Omega_2)$. Similarly, the marginal value of any level-k flexible resource is $p\mathbb{P}(\bigcup_{i=1}^k\Omega_i)$ (cf. Lemma 3 in Appendix A) and the optimality Equations (5) and (6) simplify:

Proposition 1. *In a symmetric system, the optimal flexibility configuration $K^\star\geq 0$ solves*

$$K_k^\star\left(c_k-p\mathbb{P}\left(\bigcup_{i=1}^k\Omega_i(K^\star)\right)\right)=0\quad \text{for } 1\leq k\leq N, \tag{8}$$

$$p\mathbb{P}\left(\bigcup_{i=1}^k\Omega_i(K^\star)\right)\leq c_k\quad \text{for } 1\leq k\leq N. \tag{9}$$

Second, we use formal set algebra to establish that the marginal value of level-k flexible capacity is concave increasing in k. For example, consider a three-product system. Clearly,

$$\mathbb{P}(\Omega_1)\leq\mathbb{P}(\Omega_1\cup\Omega_2)\leq\mathbb{P}(\Omega_1\cup\Omega_2\cup\Omega_3),$$

so that the marginal value of a dedicated resource is less than that of a level-2 flexible resource, which at its turn is less than the marginal value of a fully flexible resource. In addition, the corresponding increments are decreasing:

$$\begin{aligned}0&\leq\mathbb{P}(\Omega_1\cup\Omega_2\cup\Omega_3)-\mathbb{P}(\Omega_1\cup\Omega_2)\\&\leq\mathbb{P}(\Omega_1\cup\Omega_2)-\mathbb{P}(\Omega_1),\end{aligned}\tag{10}$$

so that the marginal value of capacity is concave increasing in the level of flexibility k. This result that flexibility exhibits "decreasing returns" generalizes to symmetric newsvendor networks:

Proposition 2 (Decreasing Returns to Flexibility). *In a symmetric system, for any capacity vector K, the marginal value of capacity is concave increasing in the level of flexibility k; that is, we have*

$$\mathbb{P}\left(\bigcup_{i=1}^k\Omega_i(K)\right)\leq\mathbb{P}\left(\bigcup_{i=1}^{k+1}\Omega_i(K)\right)\quad \text{for } 1\leq k<N, \tag{11}$$

and

$$\begin{aligned}&\mathbb{P}\left(\bigcup_{i=1}^{k+2}\Omega_i(K)\right)-\mathbb{P}\left(\bigcup_{i=1}^{k+1}\Omega_i(K)\right)\\&\quad\leq\mathbb{P}\left(\bigcup_{i=1}^{k+1}\Omega_i(K)\right)-\mathbb{P}\left(\bigcup_{i=1}^{k}\Omega_i(K)\right)\quad\text{for } 1\leq k<N-1,\end{aligned}\tag{12}$$

where the inequality is strict if $K_k>0$.

Proof. Equation (11) is self-evident. For the second relationship, basic set theory yields that

$$\mathbb{P}\left(\bigcup_{i=1}^{k+2}\Omega_i\right)=\mathbb{P}(\Omega_{k+2})+\mathbb{P}\left(\bigcup_{i=1}^{k+1}\Omega_i\right)-\mathbb{P}\left(\Omega_{k+2}\cap\left(\bigcup_{i=1}^{k+1}\Omega_i\right)\right),\tag{13}$$

$$\mathbb{P}\left(\bigcup_{i=1}^{k+1}\Omega_i\right)=\mathbb{P}(\Omega_{k+1})+\mathbb{P}\left(\bigcup_{i=1}^{k}\Omega_i\right)-\mathbb{P}\left(\Omega_{k+1}\cap\left(\bigcup_{i=1}^{k}\Omega_i\right)\right).\tag{14}$$

Given system symmetry, $\mathbb{P}(\Omega_{k+2})=\mathbb{P}(\Omega_{k+1})$ and $\mathbb{P}(\Omega_{k+1}\cap(\bigcup_{i=1}^k\Omega_i))=\mathbb{P}(\Omega_{k+2}\cap(\bigcup_{i=1}^k\Omega_i))$. The observation that $\mathbb{P}(\Omega_{k+2}\cap(\bigcup_{i=1}^k\Omega_i))\leq\mathbb{P}(\Omega_{k+2}\cap(\bigcup_{i=1}^{k+1}\Omega_i))$ yields (12). Notice that for any i,j, $\Omega_i\cap\Omega_j$ cannot be a null set for any finite capacity portfolio because there will be some high demands exceeding the overall capacity. A similar argument yields that if $K_k>0$, then there exists a set of demand realizations with positive probability A such that $A\subseteq\Omega_{k+2}\cap\Omega_{k+1}$ but $A\cap(\Omega_{k+2}\cap\Omega_i)$ is empty for all $i\leq k$, which yields $\mathbb{P}(\Omega_{k+2}\cap(\bigcup_{i=1}^k\Omega_i))<\mathbb{P}(\Omega_{k+2}\cap(\bigcup_{i=1}^{k+1}\Omega_i))$ (the argument is analogous to that in the proof of the more general Proposition 5 and is omitted). □

The optimal investment problem can be graphically represented as in Figure 1. The marginal value of capacity is concave and increasing in the level of flexibility k, and strictly concave when $K_{k'}>0$ for some $1\leq k'<N-1$, whereas the marginal cost of capacity is linearly increasing in k. Thus, both can be equal for at most two levels of flexibility, which then must be adjacent. This yields our main result:

Proposition 3. *In a symmetric system, the optimal flexibility portfolio invests in at most two levels of flexibility. These two levels are always adjacent. Furthermore, it is never optimal to invest in the fully flexible resource alone: There exists a $k\in\{2,\ldots,N\}$ such that*

$$K^\star=(0,\ldots,0,K_{k-1}^\star,K_k^\star,0,\ldots,0).$$

If $K_N^\star>0$, then we must also have $K_{N-1}^\star>0$.

Proof. Suppose $K^\star_{k'} > 0$ for some $1 \le k' < N$. Then, flexibility has strictly decreasing returns (cf. Proposition 2). Combining this with the linear cost of flexibility, the result follows immediately. Now, consider the case $K_{k'} = 0$ for all $1 \le k' < N$. Then, all shortage regions are equal: for any dedicated resource i, its shortage region is $\Omega_i = \{\sum_{j=1}^{N} D_j > K_N\}$, and the marginal value of any level-k resource is independent of k. However, as the cost of capacity is strictly increasing in the amount of flexibility, this portfolio cannot satisfy the KKT conditions, and hence cannot be optimal. □

The above proposition proves that the optimal flexibility portfolio invests in at most two, adjacent levels of flexibility. This implies that there are only $N-1$ optimal flexibility configurations in a symmetric newsvendor network with N products. It is possible that an optimal configuration invests in no resources, or in only one level of flexibility, but this cannot be the fully flexible resource. For example, consider a four-product setting where the capacity portfolio can consist of four dedicated resources, six level-2 resources, four level-3 resources, and one fully flexible resource. Proposition 3 implies that there are only three optimal flexibility configurations that invest in flexibility levels 1 and 2 (that is, tailored pairing as shown in panel (a) in Figure 1), levels 2 and 3 (panel (b)), or levels 3 and 4 (panel (c)).

If it is optimal to invest in dedicated resources, we immediately obtain the optimality of tailored pairing.

Corollary 1 (Optimality of Tailored Pairing). *If the optimal flexibility portfolio for a symmetric system invests in dedicated resources, then there will be no investment in level-*$k > 2$ *flexible resources.*

Given that for $N = 3$, pairing is equivalent to chaining, we establish that tailored chaining then is optimal for three product systems. Recall that, as discussed in the introduction, for $N > 3$ tailored chaining is an asymmetric configuration, and therefore suboptimal in our setting. With economies of scale, however, chaining quickly becomes more attractive than pairing, as we shall discuss in §5.2.

Although we have characterized the structure of the optimal flexibility configuration, determining the actual capacity levels typically is done numerically. Given that the optimal solution invests at most in two consecutive levels of flexibility, one computational strategy to solve the problem is to restrict the optimization over two adjacent levels of flexibility; that is, we optimize (1) over capacity portfolios of the form $\{0, \ldots, 0, K_{k-1}, K_k, 0, \ldots, 0\}$, where $1 < k \le N$. Note that there are $N-1$ such restrictions. If we find a solution to one of these restrictions with $K^\star_{k-1}, K^\star_k > 0$, the fact that the marginal value of an increase in flexibility is decreasing, along with the fact that the KKT conditions uniquely characterize the optimal solution, immediately implies that it cannot be optimal to invest in flexibility levels other than $k-1$ and k in the unrestricted problem; that is, the optimal capacity portfolio is in fact $\{0, \ldots, 0, K^\star_{k-1}, K^\star_k, 0, \ldots, 0\}$.

Next we will investigate the key drivers of the optimal flexibility configuration. It will be useful to quantify the value of flexibility as follows. Let Π_d denote the optimal total cost when only dedicated resources can be used. Thus, Π_d is the cost of a zero-flexibility strategy. The value of flexibility then is the relative decrease in cost when using the optimal flexibility strategy:

$$\text{value of flexibility } V^\star = \frac{\Pi_d - \Pi^\star}{\Pi_d}.$$

We expect that the system performance will deteriorate as demand is more variable or more correlated. Indeed, given that $\pi(K, D)$ is supermodular in D (cf. Corollary 3 in Appendix A), Proposition 3 in Van Mieghem and Rudi (2002) yields the following:

Proposition 4. *Let* D *be normally distributed with any mean vector* μ *and covariance matrix* Σ. *The optimal costs* $\Pi^\star$ *and* Π_d *are increasing in any (co)variance term. Given that* Π_d *is independent of correlation, the value of flexibility is decreasing in any pairwise demand correlation.*

We will illustrate this general property with a numerical study in the next section.

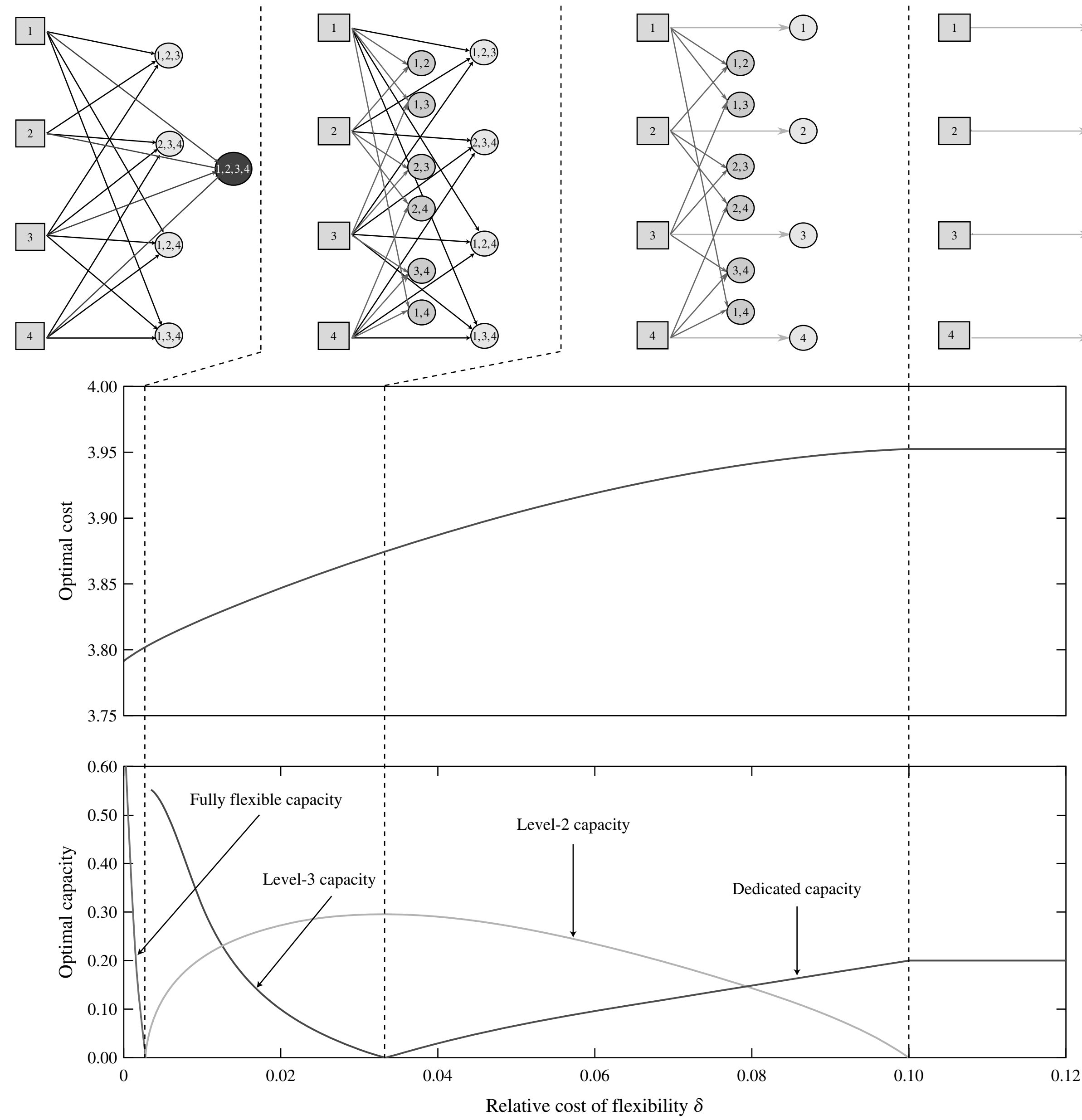
4. Key Drivers of the Optimal Flexibility Configuration

In this section, we study how the structure of the optimal capacity portfolio depends on model parameters via a numerical study for $N = 4$ products. In our first study, the demand for each product is uniformly distributed on the interval $[0, 2]$ and is independent of the demand for other products. In addition to unit mean demand, we normalize the shortage cost to $p = 1$ and fix the marginal cost of level-1 resources at $c_1 = 0.9$. As a benchmark, it is useful to consider the optimal no-flexibility strategy, which would invest in four specialized resources, each having a 90% shortage probability. Thus, each optimal dedicated capacity is 0.2 with a capacity cost of $0.2 \times 0.9 = 0.18$ and an expected shortage cost of 90% × (0.9 units short on average) × $1 = 0.81$. Summing the total capacity cost of 0.72 and the total shortage cost of 3.24 yields a total cost of 3.96 for the optimal zero-flexibility strategy, which must be an upper bound for the flexible configurations.

4.1. The Substitution Impact of the Flexibility Cost Premium δ

As the marginal cost of flexibility increases, we expect to invest less in higher levels of flexibility and the

Figure 4 Optimal Configurations, Cost, and Capacities vs. the Flexibility Premium for Demand Uniformly Distributed on $[0, 2]$

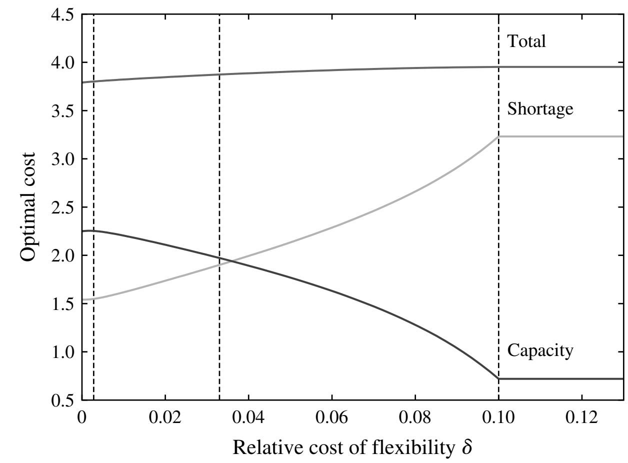


total cost to rise. This is confirmed by Figure 4, which shows the optimal configuration and cost as a function of the flexibility premium δ. The lower panel shows the optimal level of investment in each level of flexibility as a function of the flexibility premium.

When flexibility is almost costless (δ near 0), the optimal portfolio invests mainly in full flexibility and a small level of level-3 resources as expected. As the premium increases, the optimal portfolio "rebalances" its investment by reducing the fully flexible capacity and increasing the level-3 capacity. At a premium around 0.003, the optimal investment in the fully flexible resource is zero, and the flexibility configuration changes toward levels 2 and 3. As the premium increases, the portfolio again rebalances by substituting the higher level flexible capacity for lower level flexible capacity until the next flexibility cost threshold of $\delta = 0.032$ is reached and the optimal configuration changes again to a lower level of flexibility. This substitution repeats itself until $\delta = 0.1$, beyond which point a no-flexibility configuration is optimal.

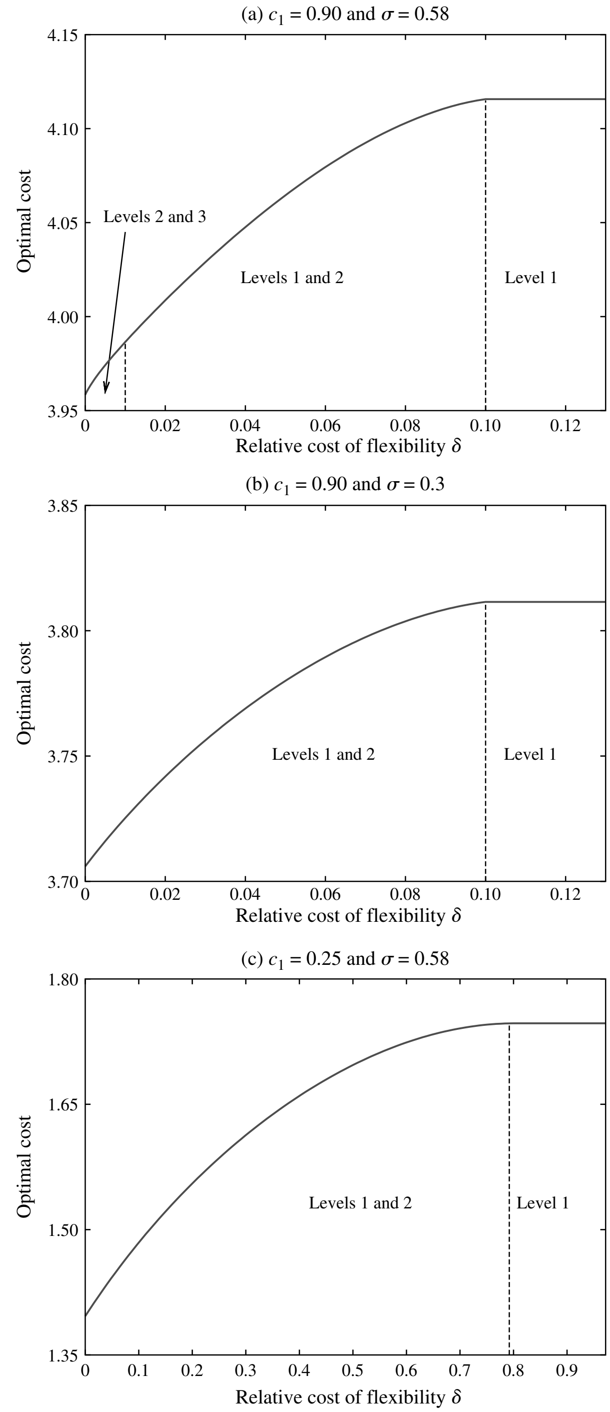
A similar substitution pattern was proved in Van Mieghem (1998, Proposition 3) for a two-product system where the substitution toward the dedicated capacity was monotone. With more products, the capacity levels are *not* monotone in the flexibility cost premium, but the substitution is always from a higher level of flexibility to a lower level. Indeed, the

Figure 5 **Decomposition of Total Optimal Cost Into the Capacity and Shortage Costs**



Note. As flexibility becomes more expensive, total cost and shortage cost rise while capacity cost falls.

Figure 6 **Changing the Demand Distribution from Uniform to Normal and Keeping Identical Cost c_1 and Demand Mean and Standard Deviation σ Leads to Optimality of Lower Levels of Flexibility (a); Tailored Pairing Is Always Optimal When the Variability Decreases (b), or When the Cost of Dedicated Resources Decreases (c)**



capacity investments in level-2 and level-3 resources rise and then fall as the flexibility cost premium δ increases. Furthermore, the investment in flexibility level-k peaks exactly when the investment in level-$(k+1)$ first reaches zero.

Figure 5 shows the decomposition of the total optimal cost into capacity and shortage costs. As the flexibility premium increases, there is lower investment in flexible capacity and the shortage cost of lost demand increases. It is worth noting that the change in the cost components over the range of the premium is about 50%–100%, as opposed to the total cost, which only varies about 5% for the chosen parameter values.

4.2. The Normal vs. the Uniform Distribution

Given that the newsvendor optimality conditions involve the entire demand distribution, we expect the functional form of the demand distribution to affect the optimal flexibility configuration. This is indeed observed when we repeat our first study but only replace the uniform demand distribution with a normal demand distribution with the same mean of 1 and standard deviation $\sigma = 0.58$. (We truncate the normal distribution to eliminate negative demand values.) Figure 6(a) shows the optimal cost as a function of the flexibility premium. Investing in flexibility levels 1 and 2 (i.e., tailored pairing) is optimal for $\delta \in [0.01, 0.1]$, which is larger than the region $[0.032, 0.1]$ for the uniform distribution (see Figure 4). Furthermore, the region where it is optimal to invest in levels 3 and 4 is now too small to discern.

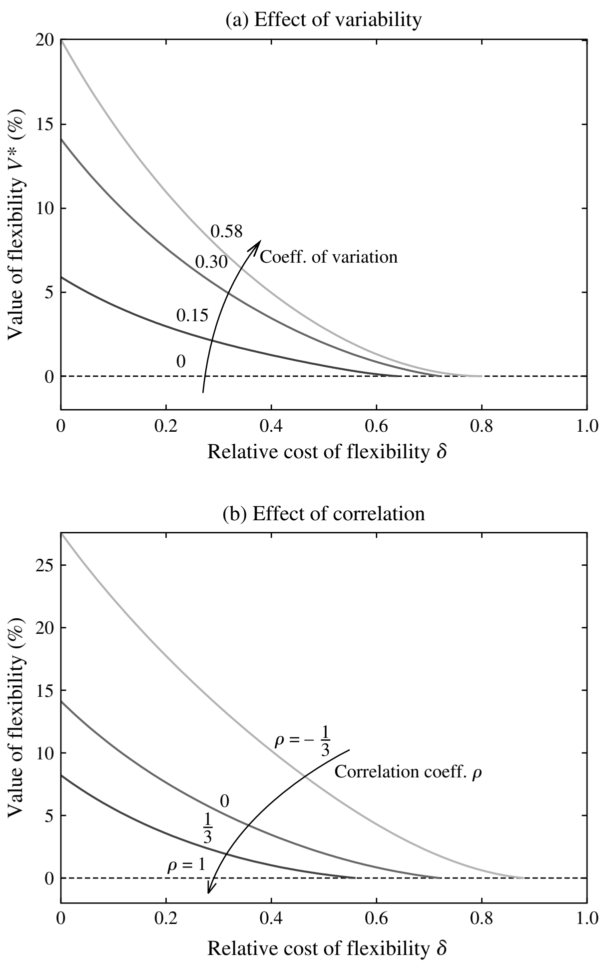
Summarizing, even when the first two moments are matched, the normal distribution seems to favor the lower levels of flexibility of the tailored pairing configuration compared to the uniform distribution. This effect is even magnified when we lower the demand variability or the cost of dedicated capacity, as shown in Figure 6, panels (a) and (b). Panel (b) lowers the demand standard deviation from 0.58 to 0.3. Tailored pairing is now optimal for all discernible value of the flexibility premium. Panel (c) lowers the cost of dedicated capacity from $c_1 = 0.9$ to 0.25, which leads to a similar observation. These observations suggest that tailored pairing is a desirable flexibility configuration for most reasonable cases of cost parameters and demand variability. Higher levels of flexibility will be valuable only with high demand variability or a high relative cost of dedicated capacity (compared to the penalty cost).

4.3. The Impact of Demand Variability and Correlation on the Value of Flexibility

Figure 7(a) shows the value of flexibility for our first study with normal demand. The plot shows the value as a function of the cost of flexibility and as a function of the coefficient of variation of the demand distribution (which equals the standard deviation given that the mean is 1). Tailored pairing was the optimal flexibility configuration for all investigated parameter values. As expected, the value of flexibility increases as the demand variability increases. Note that this does not follow from Proposition 4 because both $\Pi^\star$ and Π_d are affected by variance. The numerical finding implies that, as expected, $\Pi^\star$ is less affected by the variance because of the pooling benefits of flexibility. Even with a relatively low coefficient of variation of 0.15, the maximal value of flexibility is about 6%. This figure increases to 20% when the coefficient of variation increases to 0.58. As expected, the value of flexibility decreases in the relative cost of flexibility δ. Notice that the threshold flexibility premium beyond which there is no investment in flexibility is also increasing in variability.

Next, we study the impact of correlation. We do so by varying the pairwise correlation coefficient ρ over the interval $[-1/3, 1]$.[1] Figure 7(b) verifies that the value of flexibility decreases as the correlation coefficient ρ increases, as proved in Proposition 4.

Figure 7 The Value of Flexibility Increases When the Demand Variability Increases But Decreases When the Correlation Between the Different Product Demands Increases



5. Extensions: Asymmetric Products, Economies of Scale, and Scope

5.1. Asymmetric Products

In this section, we consider products with potentially different demand distributions and financial parameters, and prove that our main results (Proposition 3 and Corollary 1) continue to hold in a slightly generalized manner. Obviously, the optimal capacity portfolio $K^\star$ will no longer be symmetric and we will specify it as described in §2. In particular, the portfolio K now consists of the capacity of resource F for all $F \subseteq \{1, 2 \ldots, N\}$, where F denotes the set of products that can be processed by resource F. Note that now different resources with the same level of flexibility can have different capacities depending on the products they can process. This asymmetry in the solution increases the number of candidate flexibility configurations (measured as the number of product–resource allocation graphs) from 2^N to 2^{2^N-1} configurations. This doubly exponential relationship is so strong that even small problems quickly become computationally infeasible. For example, with only four products, there can only be 16 candidate flexibility configurations with demand symmetry. However, with general demand distributions the number of candidate configurations increases to 32,768. Thus, this problem is much more complex than with symmetric demand.

As before, our solution approach follows two key steps. First, we describe the marginal value of a flexible resource F in terms of only the shortage regions. Second, we prove that the marginal value of a resource is concave increasing in the level of flexibility. However, given the asymmetry in the products,

[1] For values of $\rho < -1/3$, the covariance matrix is no longer positive semidefinite. Indeed, one cannot have four products with perfectly negatively pairwise correlations.

we need to modify our approach slightly. In particular, to characterize the marginal value of a resource, we must know the product type that benefits from additional capacity as well as the shortage region. Thus, we refine the original N shortage regions into $N \times N$ shortage regions, depending on the product type benefitting from additional capacity in the resources. We denote $\Omega_{i,j}$ for $i, j = 1, \ldots, N$ as the region where there is a shortage of product j because of insufficient capacity of resource $\{i\}$. Indeed, in a parallel network, an infinitesimal increase in the capacity of resource $\{i\}$ reduces the shortage of product j by the same amount for demand realizations that lie in $\Omega_{i,j}$. In the following, we assume that the products are labeled in decreasing order of shortage penalty costs, i.e., $p_1 \geq p_2 \geq \cdots \geq p_N \geq p_{N+1} \equiv 0$.

Using the shortage regions, we can thus compute the marginal value of a dedicated resource $\{i\}$ as

$$V(\{i\}) = \sum_{k=1}^{N} p_k \mathbb{P}(\Omega_{i,k}) = \sum_{k=1}^{N} (p_k - p_{k+1}) \mathbb{P}\left(\bigcup_{j=1}^{k} \Omega_{i,j} \right).$$

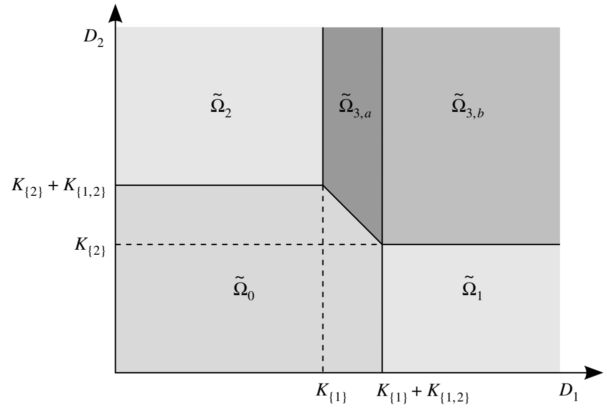
The second equality follows by noting that we first attribute the lowest shortage penalty p_N to the demand realizations that lie in the shortage region of at least one product, i.e., the set $\bigcup_{j=1}^{N} \Omega_{i,j}$. We then add the incremental benefit $(p_{N-1} - p_N)$ over the demand realizations that lie in the shortage regions of at least one of the products $\{1, 2, \ldots, N-1\}$. We continue in this manner until all the shortage penalty costs are accounted for. Extending this logic to a resource F, we obtain its marginal value

$$V(F) = \sum_{k=1}^{N} (p_k - p_{k+1}) \mathbb{P}\left(\bigcup_{j=1}^{k} \bigcup_{i \in F} \Omega_{i,j} \right).$$

5.1.1. Shortage Regions in a Two-Product Setting. For illustration we consider the case $N = 2$, which is studied in Van Mieghem (1998). Figure 8 displays the five mutually exclusive scenarios analogous to Figure 3 for a capacity portfolio K. The main differences are that as $p_1 \geq p_2$, the scenario $\tilde{\Omega}_3$ can be subdivided into two further regions $\tilde{\Omega}_{3,a}$, in which any increase in capacity is allocated to product 2, and $\tilde{\Omega}_{3,b}$, in which the additional capacity is allocated to product 1. Thus, we obtain our shortage regions $\Omega_{1,1} = \tilde{\Omega}_1 \cup \tilde{\Omega}_3$, $\Omega_{1,2} = \tilde{\Omega}_{3,a}$, $\Omega_{2,1} = \{\}$, and $\Omega_{2,2} = \tilde{\Omega}_2 \cup \tilde{\Omega}_{3,a} \cup \tilde{\Omega}_{3,b}$. Then, using the above analysis, the marginal values for the three resources equals

$$\begin{aligned}
V(\{1\}) &= p_2 \mathbb{P}(\tilde{\Omega}_1 \cup \tilde{\Omega}_{3,a} \cup \tilde{\Omega}_{3,b}) + (p_1 - p_2) \mathbb{P}(\tilde{\Omega}_1 \cup \tilde{\Omega}_{3,b}) \\
&= p_1 \mathbb{P}(\tilde{\Omega}_1 \cup \tilde{\Omega}_{3,b}) + p_2 \mathbb{P}(\tilde{\Omega}_{3,a}), \\
V(\{2\}) &= p_2 \mathbb{P}(\tilde{\Omega}_2 \cup \tilde{\Omega}_{3,a} \cup \tilde{\Omega}_{3,b}), \\
V(\{1,2\}) &= p_2 \mathbb{P}(\tilde{\Omega}_1 \cup \tilde{\Omega}_2 \cup \tilde{\Omega}_{3,a} \cup \tilde{\Omega}_{3,b}) \\
&\quad + (p_1 - p_2) \mathbb{P}(\tilde{\Omega}_1 \cup \tilde{\Omega}_{3,b}) \\
&= p_1 \mathbb{P}(\tilde{\Omega}_1 \cup \tilde{\Omega}_{3,b}) + p_2 \mathbb{P}(\tilde{\Omega}_2 \cup \tilde{\Omega}_{3,a}).
\end{aligned}$$

Figure 8 With Two Products Such That $p_1 > p_2$, the Demand Space Is Partitioned Into Five Regions



These expressions are identical to those in Van Mieghem (1998).

The decreasing returns to flexibility result we obtain in this asymmetric setting is slightly different from before in the sense that it is product specific: given a resource F, the marginal value of increasing flexibility by adding a product $r \notin F$ decreases in the cardinality of F:

Proposition 5 (Decreasing Returns to Flexibility). *For any resource $F \subseteq \{1, \ldots, N\}$ in a capacity portfolio K and any products $q, r \notin F$, we have*

$$V(F \cup \{q, r\}) - V(F \cup \{q\}) \leq V(F \cup \{r\}) - V(F). \quad (15)$$

Further, the inequality is strict if $K_F > 0$.

Given the key result that the marginal value of flexibility remains concave increasing, our main result continues to hold but becomes product specific:

Proposition 6. *With a general demand distribution, the optimal flexibility portfolio is such that for any two resources F and F' it invests in, if $F \subset F'$, then F and F' must be at adjacent levels, i.e., $|F'| = |F| + 1$.*

This result leads to a significant reduction in the number of configurations that needs to be considered. For example, in a four-product system, we only need to consider 2,256 configurations out of a total of 32,768. Proposition 6 immediately leads to the following result that demonstrates the optimality of tailored pairing portfolios.

Corollary 2 (Optimality of Tailored Pairing).
1. *If the optimal portfolio invests in a dedicated resource for a product, then there will be no investment in level-$k > 2$ resources to process this product.*
2. *Thus, if the optimal portfolio invests in dedicated resources for all products, there will be no investment in any level-$k > 2$ resource.*

Table 1 Optimal Solution to the Asymmetric Problem and the Performance of the Heuristic Capacity Portfolio Obtained by Assuming $p_1 = p_2 = 1$

CV (%)	p_2	Optimal cost	Optimality gap of heuristic (%)
	0.8	0.55	0
10	0.5	0.54	1
	0.2	0.47	15
	0.8	0.59	0
20	0.5	0.57	2
	0.2	0.49	18
	0.8	0.64	0
30	0.5	0.61	3
	0.2	0.51	21

Notes. The unit shortage penalty cost for product 1 is set to $p_1 = 1$. The coefficient of variation (CV) of demand and the shortage penalty for product 2, p_2, are varied.

5.1.2. A Numerical Example with Two Products: Benchmarking Symmetric Heuristics. The preceding analysis illustrates the complexity involved in dealing with asymmetric products. Using a two-product example, we use the optimal solution to benchmark the performance of a heuristic that computes the capacity portfolio assuming the products are symmetric. We assume the two products have independent and identically distributed, normally distributed demand with unit mean. The cost of dedicated capacity is $c_1 = 0.25$, and that of the flexible resource is $c_2 = 0.275$ (using $c_2 = c_1(1+\delta)$ and $\delta = 0.1$). The unit penalty cost for product 1 is $p_1 = 1$, and we varied the value of p_2 and the coefficient of variation of the demand of the two products. In each case, we computed the optimal solution numerically and then compared its cost to that obtained from the capacity portfolio that is optimal for the symmetric problem with $p_1 = p_2 = 1$. The results are displayed in Table 1. Surprisingly, even when p_2 is 0.5, this heuristic works extremely well (with optimality gap less than 3%). Only when p_2 is extremely low compared to p_1 does the performance deteriorate. Less surprisingly, the performance also deteriorates when the variability increases.

5.2. Economies of Scale: Nonlinear Capacity Sizing Cost Structure

Our theoretical analysis so far has assumed that the investment cost $C_F(K_F)$ for resource F is linear in its capacity size K_F. In practice, investment costs often exhibit scale economies, meaning that one resource with capacity $2K_F$ is cheaper than two resources each having capacity K_F. Clearly, scale economies induce a firm to invest in fewer but bigger resources. Mathematically, scale economies imply that the investment cost $C_F(K_F)$ is concave in the capacity size K_F. Unfortunately, the capacity optimization problem then is no longer guaranteed to be convex, and the first-order conditions are no longer sufficient. In this section, we generalize our analysis to a particular but important form of scale economies where any positive investment incurs a fixed setup cost that may depend on the resource's level of flexibility, whereas marginal cost of capacity remains constant. Specifically, there are N positive setup costs $s_k > 0$, and the investment cost $C_F(K_F)$ for resource F becomes

$$C_F(K_F) = s_{|F|} 1_{\{K_F>0\}} + c_{|F|} K_F \quad \text{for } K_F \geq 0.$$

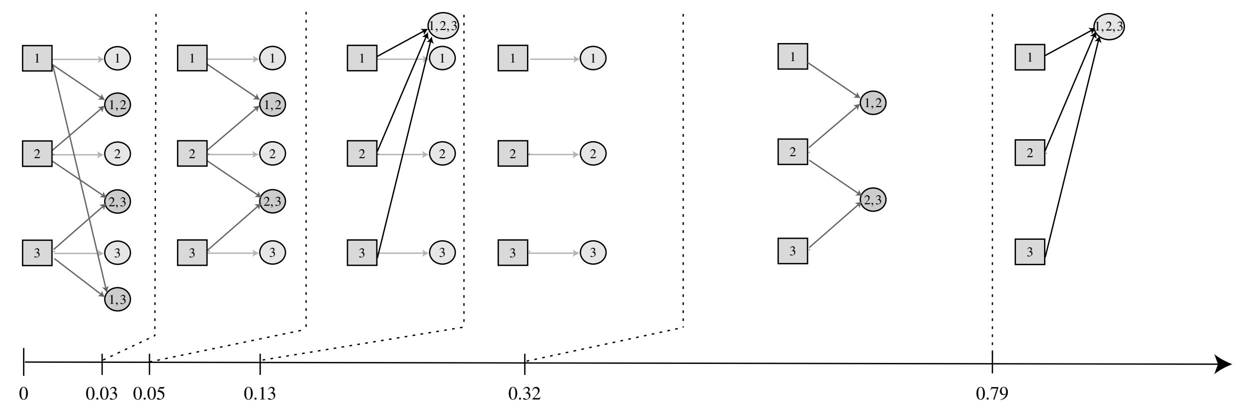
Including fixed costs results in a much harder combinatorial problem where we first must decide which resources to invest in, and then solve our earlier optimization problem (1) restricted to these resources. Note that with fixed cost, the optimal capacity portfolio for even a symmetric system is no longer guaranteed to be symmetric. Fortunately, as the marginal cost of capacity remains linear once the setup cost has been incurred, our main result continues to hold in slightly modified form and allows us to reduce the computational complexity:

Proposition 7. *With general demand distribution and setup costs, the optimal flexibility portfolio is such that for any three resources F, F', and F'' it invests in, we cannot have $F \subset F' \subset F''$.*

Thus, the main result that only two levels of flexibility should be invested in per product is robust to the addition of setup costs, but these two levels need no longer be adjacent. Nevertheless, this result still leads to a significant reduction in the number of configurations that needs to be considered. For example, in a four-product system, we only need to consider 3,771 configurations out of a total of 32,768. Unfortunately, our method does not allow a finer characterization of the optimal portfolio in this setting, and we thus resort to two numerical studies to understand the effect of setup costs.

First, we consider a three-product symmetric system with common setup costs for all levels of flexibility: $s_1 = s_2 = s_3 \equiv s$. The demand for the products is independent and normally distributed with unit mean and standard deviation $\sigma = 0.3$. We fix $p = 1$, $c_1 = 0.25$, and the flexibility premium $\delta = 0.25$ and study the optimal portfolio as the setup cost s varies. Figure 9 shows that the optimal configuration is tailored chaining, which is equivalent to tailored pairing for $N = 3$, without fixed costs ($s = 0$). As s increases, the number of resources in the optimal portfolio begins to decrease, as expected. For sufficiently large setup costs, the scale economies are so high that the optimal portfolio invests only in the fully flexible resource. Notice that for $s/c_1 \in [0.05, 0.13]$, the optimal configuration invests in level-1 and fully flexible resources, demonstrating that adjacency is no

Figure 9 The Optimal Capacity Portfolio as a Function of the Setup Cost

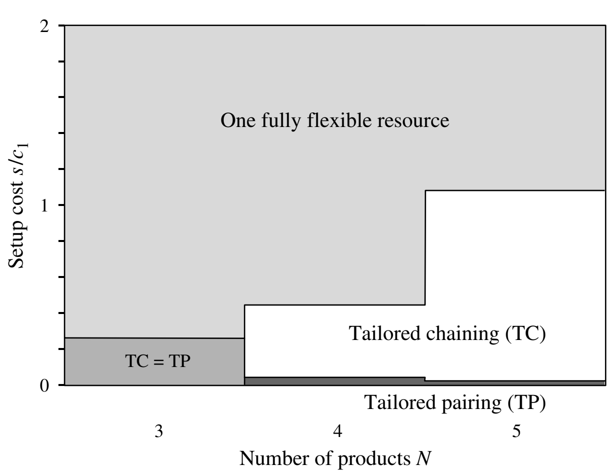


Notes. For no setup costs, tailored pairing (which is equivalent to tailored chaining) is optimal. As the setup cost increases, the optimal portfolio consists of fewer resources. For large enough setup costs, investing in the fully flexible resource alone is optimal.

longer optimal under economies of scale. Note that because of economies of scale, one can also obtain nonsymmetric configurations to be optimal as well. For instance, when $s/c_1 \in [0.03, 0.05]$, the optimal configuration invests in two out of the three level-2 flexible resources.

Our second study investigates the relative performance of three flexibility configurations that have been studied in the literature under scale economies: tailored pairing, tailored chaining, and the fully flexible configuration. We consider symmetric systems with $N = 3, 4,$ and 5 products. All other parameters are the same as in the first study. Figure 10 shows the regions where each configuration dominates the other two as a function of the number of products N (on the horizontal axis) and the setup cost s (on the vertical axis). For $s = 0$, tailored pairing is the optimal configuration and dominates tailored chaining and full flexibility. As the setup costs increase, tailored chaining becomes more cost effective than tailored pairing because it uses $N(N-3)/2$ fewer resources. Finally, for large setup costs, full flexibility dominates. Notice that the region where tailored chaining dominates grows as the number of products N increases: chaining saves on increasingly more resources over pairing (explaining the downward sloping boundary between the two) and full flexibility becomes increasingly more expensive than level-2 flexibility (explaining the upward sloping boundary). This provides additional evidence of the attractiveness of tailored chaining in practice.

Figure 10 Comparison of Tailored Pairing, Tailored Chaining, and One Fully Flexible Resource with Setup Costs



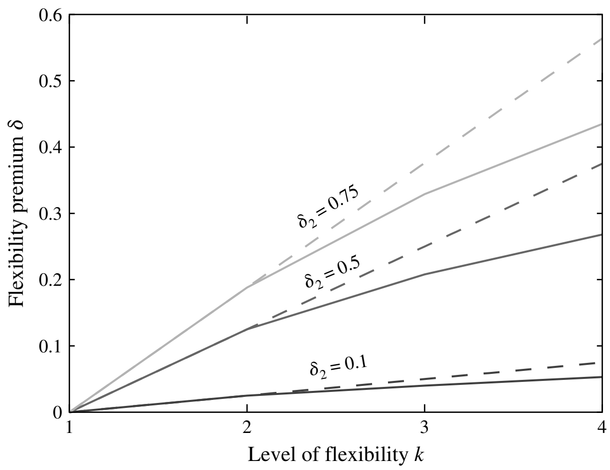
Note. The regions show which configuration dominates the other two.

5.3. Economies of Scope: Nonlinear Flexibility Cost Structure

Our main results (Proposition 3 and Corollary 1) require capacity costs to be affine in the level of flexibility. Assuming that the cost of one unit of capacity of a level-k resource is $c_k = c_1[1 + \delta(k-1)]$ means that the marginal cost of flexibility is constant and equal to δ. Remarkably, under this affine cost structure, the results hold independent of the magnitude of δ, as long as it is positive. In this section, we investigate the robustness of our result for nonaffine flexibility cost structures. Clearly, when the capacity cost is convex in the level of flexibility ($c_k > c_{k-1} + \delta c_1$), higher levels of flexibility become even less attractive and our main result holds. So let us investigate how the optimal flexibility configuration changes when there are economies of scope so that the marginal cost of flexibility is *concave* increasing in the level of flexibility.

We established earlier that the marginal *value* of flexibility is concave increasing in the level of flexibility k. If now the marginal *cost* is concave increasing as well, the optimal solution depends on the relative curvature of the two curves: with slightly concave cost, our main result continues to hold. As economies of scope increase, the curvature increases and the optimal

Figure 11 **Our Main Result Holds Even When the Cost Structure Is Fairly Concave in Flexibility**



Notes. The figure shows the affine cost structure (dashed lines) that is assumed for our main analytic result for $N = 4$ and when $c_1 = 0.25$ and δ_2 takes on three values: 0.1, 0.5, 0.75. The solid lines show the corresponding maximal concavity in flexibility costs for which tailored pairing continues to be optimal.

levels of flexibility increase. Eventually, investment in only the fully flexible resource becomes optimal.

Noting that for most reasonable parameters tailored pairing is the optimal configuration, we investigate how concave the flexibility cost structure can be for tailored pairing to remain optimal. Let δ_k denote the marginal cost to increase the level of flexibility of one capacity unit from level-$k-1$ to k. Then, $c_k = c_1[1 + \sum_{j=2}^{k} \delta_j]$. We perform a numerical study in a four-product setting where the demand for each product is normally distributed with mean one and variance 0.3^2 and is independent of the demand of other products ($\rho = 0$). We set $c_1 = 0.25$ and consider some fixed values of $\delta_2 = 0.1, 0.5,$ and 0.75. At these cost values, under the affine cost structure, tailored pairing is the optimal portfolio. For each fixed value of δ_2, we solve for the smallest marginal cost of flexibility values δ_3 and δ_4 for which it remains optimal not to invest in level-$k > 2$ flexibility. The results are displayed in Figure 11. Notice that our main result continues to hold for any concave flexibility cost structure above the solid frontiers. This suggests our results are robust to the cost structure choice to some extent.

6. Other Applications: Substitution and Queuing

In this section, we demonstrate how the methods we have developed can be applied to characterize optimal flexibility portfolios in other systems beyond the models considered thus far. In §6.1, we add the feature of product substitution to our model and characterize the optimal portfolio, and in §6.2 we consider the case of flexible make-to-order queuing systems with dynamic customer arrivals and stochastic production times.

6.1. Product Substitution

We demonstrate how our methodology can be used to handle the case of product substitution. Consider the case where products are vertically differentiated with $p_i \geq p_{i+1}$ for $i = 1, 2, \ldots, N-1$, and product i can be substituted for $i+1$. We assume that each substitution entails a cost of s, with $s < p_N$. (This is similar to the model in Bassok et al. (1999), but with a single level of downward product substitution and flexible capacity portfolio.) In this case, the firm's optimization problem is

$$\Pi^\star = \min_K \left\{ \mathbb{E}_D \pi(K, D) + \sum_{F \subseteq \{1, 2, \ldots, N\}} c_F K_F \right\}, \tag{16}$$

where $\pi(K, D)$ is the optimal contingent operating profit:

$$\pi(K, D) = \min_{x, y \geq 0} \sum_{i=1}^{N} \left[p_i \left(D_i - \sum_{\{F: i \in F\}} x_{i,F} - \sum_{\{F: i \in S(F)\}} y_{i,S(F)} \right) + s \sum_{\{F: i \in S(F)\}} y_{i,S(F)} \right] \tag{17}$$

$$\text{s.t.} \quad \sum_{\{F: i \in F\}} x_{i,F} + \sum_{\{F: i \in S(F)\}} y_{i,S(F)} \leq D_i, \tag{18}$$

$$\sum_{i \in F} x_{i,F} + \sum_{i \in S(F)} y_{i,S(F)} \leq K_F. \tag{19}$$

For mathematical convenience, we assume $p_N > p_2 - s$, so that any available capacity is first used to satisfy direct customer demand, before any substitution attempts to meet excess demand.

As before, we let $\Omega_{i,j}$ denote the shortage region of resource i for product j. Considering resource $\{i\}$, its marginal value consists of two components: (a) the value from direct consumption, which equals $V_D(\{i\}) = \sum_{j=1}^{N} p_j \mathbb{P}(\Omega_{i,j})$, and (b) the value from substitution, which is realized only if there is no value from direct consumption, and equals $V_S(\{i\}) = \sum_{j=1}^{N} (p_j - s) \cdot \mathbb{P}(\Omega_{i+1,j} \setminus \bigcup_{j=1}^{N} \Omega_{i,j})$. (Note that substitution essentially transforms a unit of resource $\{i\}$ into that of resource $\{i+1\}$ with a cost of s.) Thus, we obtain the marginal value of resource $\{i\}$ is

$$V(\{i\}) = V_D(\{i\}) + V_S(\{i\}) = \sum_{j=1}^{N} p_j \mathbb{P}(\Omega_{i,j}) + \sum_{j=1}^{N} (p_j - s) \mathbb{P}\left(\Omega_{i+1,j} \setminus \bigcup_{j=1}^{N} \Omega_{i,j} \right).$$

Similarly, we obtain the marginal value of resource F,

$$V(F) = \sum_{k=1}^{N} (p_k - p_{k+1}) \mathbb{P}\left(\bigcup_{j=1}^{k} \bigcup_{i \in F} \Omega_{i,j} \right) + \sum_{j=1}^{N} (p_j - s) \mathbb{P}\left(\bigcup_{i \in F} \left(\Omega_{i+1,j} \setminus \bigcup_{j=1}^{N} \Omega_{i,j} \right) \right).$$

Using this characterization of marginal value of the various resources, we obtain that flexibility once again has diminishing returns.

Proposition 8. *For a system with downward product substitution, flexibility continues to exhibit diminishing returns; that is, for $F \subseteq \{1, 2, \ldots, N\}$ and $q, r \notin F$, we have*

$$V(F \cup \{q, r\}) - V(F \cup \{q\}) \leq V(F \cup \{r\}) - V(F).$$

Furthermore, the inequality is strict if $K_F > 0$.

Given that our key result that the marginal value of flexibility remains concave increasing, our main result continues to hold and remains identical to Proposition 6.

Proposition 9. *For asymmetric products with downward substitution, if the optimal flexibility portfolio contains two resources F and F' with $F \subset F'$, then F and F' must be at adjacent levels, i.e., $|F'| = |F| + 1$.*

Thus, the set theoretic methodology extends to this setting with substitution, and we obtain similar results for the optimal flexibility portfolio; that is, if the firm optimizes on its capacity portfolio, it should invest only in adjacent flexible resources. The purpose of this section was to demonstrate this robustness via an illustration considering only "one step" downward product substitution. However, the method described can be extended to consider the case of general substitutions.

6.2. Flexible Queuing Systems

Flexibility in services modeled as queuing systems has received ample attention. Two connections are worth making. First, consider traditional queuing systems with known arrival rates. The recent work by Bassamboo et al. (2010a) (henceforth abbreviated as BRV) proves that tailored pairing is the asymptotically optimal flexibility configuration in symmetric queuing systems with large arrival rates. A numerical study suggests that a similar observation holds in asymmetric settings. As we shall explain next, the traditional assumptions that the mean interarrival times are known is not innocuous. Statistical pooling implies that the variance of the stationary queue count process is of a smaller order, actually $O(\sqrt{\lambda})$, than the arrival rate. Consequently, BRV prove that economic optimization leads to a portfolio investing mostly in dedicated capacity to serve the base demand λ and only in a small amount $O(\sqrt{\lambda})$ of minimal level-2 flexibility to serve the variable demand. This finding that *a little flexibility is all you need* is consistent with our results for newsvendor systems, where investing in dedicated resources leads to optimality of tailored pairing portfolios. Yet, the optimal configurations in newsvendor systems are richer and go beyond tailored chaining and pairing.

Realizing that typical asymptotic queuing analysis relegates uncertainty to a second-order effect, recent studies have investigated queuing systems with arrival rate uncertainty, e.g., Harrison and Zeevi (2005) and Bassamboo et al. (2006). When the first moment is uncertain, variability is elevated to a first-order phenomenon, and capacity decisions asymptotically reduce to a newsvendor network problem of the type studied here. To be precise, consider a queuing system where N customer classes can be served by resources that differ in their level of flexibility. As before, there are $2^N - 1$ different resources or pools of servers. A server pool that can process customers of classes $F \subseteq \{1, \ldots, N\}$ is referred to as pool F. Assuming that all servers process work at a unit, deterministic rate, the capacity portfolio can be denoted by $K = \{K_F\colon F \subseteq \{1, \ldots, N\}\}$, where K_F is the number of servers in pool F.

Customers of class i arrive according to a Poisson process with rate Λ_i, which is a random variable; arrival rates can be correlated across customer classes. Class i customers have exponentially distributed services requirements with mean $1/\mu$; that is, analogous to our newsvendor model, customers have identical service requirements. Customers of any given class will abandon their queue if forced to wait too long for the commencement of service. Specifically, each class i customer is endowed with an exponentially distributed "impatience" random variable with mean $1/\gamma_i$, independent of the impatience random variables characterizing other customers, and independent of service times and arrival processes.

Similar to before, each level-k server costs c_k per unit time and each customer abandonment costs p. The firm's optimization problem is to select the optimal capacity portfolio to minimize the average system cost rate, which includes cost of capacity and cost of customer abandonment.

With large arrival rates, the asymptotically optimal portfolio selection entails solving a "higher level" problem that ignores the lower level queuing. This higher level problem is the fluid approximation to the actual system and is exactly our newsvendor network problem (1)–(4), where D_i is replaced by the uncertain arrival rate Λ_i. Bassamboo et al. (2010b) show that the newsvendor network solution is asymptotically optimal in a very strong sense: not only does its relative error with the exactly optimal capacity portfolio tend to zero as $\mathbb{E}\Lambda \to \infty$, its absolute error does not grow with the arrival rate. This makes the newsvendor network solution, for all practical purposes, an optimal prescription to the queuing system if arrival rates are uncertain.

Given the equivalence between the asymptotic optimization problem of queuing systems with uncertain arrival rates and our newsvendor network problem, Proposition 6 holds here as well. Thus, the asymptotically optimal flexibility portfolio will invest in resources at adjacent levels. Furthermore, analogous to Corollary 2, if the asymptotically optimal portfolio invests in dedicated servers for each class, then there will be no investment in servers of level-$k > 2$ flexibility. Recall that in queuing systems without arrival rate uncertainty, it is always asymptotically optimal to invest a lot of capacity in dedicated servers, and the corollary remains consistent with the asymptotic optimality of tailored pairing in symmetric queuing systems established in BRV.

7. Summary

This paper has studied the classic problem of capacity and flexible technology selection with a newsvendor network model of resource portfolio investment. We have presented an exact analytic methodology and characterized the optimal flexibility configuration for newsvendor networks with N products. This simple abstract approach is sufficiently powerful to prove that (i) flexibility exhibits decreasing returns, and (ii) the optimal portfolio will invest in at most two, adjacent levels of flexibility in symmetric systems. The analytic results were shown to extend to asymmetric demand distributions, financial parameters, scale economies with setup costs, and economies of scope. The characterization of the optimal portfolio allows a significant reduction in computational complexity. Finally, we showed that our results can be applied to product substitution scenarios and flexible queuing systems with arrival rate uncertainty.

Appendix A. Supporting Results

A.1. Equivalence to the Transportation Problem

First, note that our optimization problem (2)–(4) is equivalent to the following transportation problem that minimizes the *unmet* demand per product u_i:

$$\pi(K,D)=\min_{x,\,u\geq 0}\sum_{i=1}^{N}p_i u_i, \tag{A1}$$

$$\sum_{\{F:\,i\in F\}} x_{i,F}+u_i\geq D_i \quad \text{for all } i=1,\dots,N, \tag{A2}$$

$$\sum_{i\in F} x_{i,F}\leq K_F \quad \text{for all } F\subseteq\{1,\dots,N\}. \tag{A3}$$

The equivalence follows by noting that for any optimal solution of the above problem, the constraint (A2) is binding. Theorem 3.4.1 of Topkis (1998) states that the transportation problem is supermodular in D. Hence, we directly have the following corollary:

Corollary 3. *For any $K, D\in\mathbb{R}_+^N$, $\pi(K,D)$ is supermodular in D.*

A.2. The Dual Problem and Definition of $\Omega_{i,j}$

We return to our original optimization problem (2)–(4) and note that the first term $p\sum_{i=1}^N D_i$ is constant, and hence the objective can be re-expressed as maximizing $p\sum_{F\subseteq\{1,\dots,N\}}\sum_{i=1}^N x_{i,F}$. Given this equivalence, and denoting λ_F and μ_i to be the dual variables associated with the capacity constraint for resource F and demand constraint for product i respectively, the dual problem can be written as

$$\min_{\lambda,\,\mu\geq 0}\left\{\sum_{F\subseteq\{1,\dots,N\}} K_F\lambda_F+\sum_{i=1}^N D_i\mu_i\right\} \tag{A4}$$

$$\text{s.t.}\quad \lambda_F+\mu_i\geq p_i, \quad \text{for } F\subseteq\{1,\dots,N\} \text{ and } i\in F. \tag{A5}$$

Using the dual problem, we can define the shortage regions as follows:

Definition 1. The shortage regions for dedicated resource i are defined as

$$\Omega_{i,j}(K)\equiv\{D\colon \exists \text{ an optimal solution to (A4)–(A5)}, \ (\lambda^\star,\mu^\star) \text{ with } \lambda^\star_{\{i\}}(K,D)=p_j\}, \text{ for } j=1,2,\dots,N.$$

The following result characterizes the dual variables:

Lemma 1. *There exist optimal dual variables $(\lambda^\star,\mu^\star)$ such that*
1. *$\lambda^\star_F\in\{p_{N+1},p_N,\dots,p_1\}$, for all $F\subseteq\{1,\dots,N\}$;*
2. *for any $F, F'\subseteq\{1,\dots,N\}$, $\lambda^\star_{F\cup F'}=\max\{\lambda^\star_F,\lambda^\star_{F'}\}$.*

Proof of Lemma 1. Noting that the dual variables solve a linear program, they must be extreme, or corner, points of the constraint set:

$$\Delta=\big\{(\lambda,\mu)\colon\ \lambda,\mu\geq 0,\ \lambda_F+\mu_i\geq p_i, \quad \text{for } F\subseteq\{1,\dots,N\} \text{ and } i\in F\big\},$$

where we use the following definition of an extreme point:

Definition 2. $y\in S$ is an extreme point of the convex set S, if there do not exist feasible points $y', y''\in S$ with $y'\neq y''$, and $\alpha\in(0,1)$ such that $y=\alpha y'+(1-\alpha)y''$.

The result then follows using properties 2 and 4 of the following lemma.

Lemma 2. *Any extreme point of the set Δ has the following properties*:
1. $\lambda_{\{i\}}\leq p_i$,
2. $\lambda_F=\sup_{i\in F}\lambda_{\{i\}}$,
3. $\lambda_{\{i\}}+\mu_i=p_i$,
4. $\lambda_{\{i\}}\in\{p_{N+1},p_N,\dots,p_i\}$.

Proof of Lemma 2. We prove the result by using a contradiction argument. We first prove property 1. Suppose there exist j such that $\lambda_{\{j\}}>p_j$. Then set (λ',μ') and (λ'',μ'') as follows: $\lambda'_F=\lambda''_F=\lambda_F$ for all F except $F=j$, $\lambda'_j=p_j$, and $\lambda''_j=2\lambda_{\{j\}}-p_i$ and $\mu'_i=\mu''_i=\mu_i$ for all i. Thus, we have $0.5(\lambda',\mu')+0.5(\lambda'',\mu'')$. Furthermore, (λ',μ') and (λ'',μ'') lie in the set Δ. Thus, (λ,μ) is not an extreme point if $\lambda_{\{i\}}>p_i$ for some $i\in\{1,\dots,N\}$.

For property 2, suppose we have some $(\lambda,\mu)\in\Delta$ with $\lambda_{\{j\}}\leq p_j$ for all j. Then, if there exists $\lambda_G\neq\sup_{i\in G}\lambda_{\{i\}}$, then it must be the case that $\lambda_G>\sup_{i\in G}\lambda_{\{i\}}$. Set (λ',μ') and (λ'',μ'') as follows: $\lambda'_F=\lambda''_F=\lambda_F$ for all F except $F=G$, $\lambda'_G=\sup_{i\in G}\lambda_{\{i\}}$, and $\lambda''_G=2\lambda_G-\lambda'_G$ and $\mu'_i=\mu''_i=\mu_i$ for all i.

Thus, we have $0.5(\lambda', \mu') + 0.5(\lambda'', \mu'')$. Furthermore, (λ', μ') and (λ'', μ'') lie in the set Δ. Thus, (λ, μ) is not an extreme point $\lambda_F > \sup_{i\in F} \lambda_{\{i\}}$ for some $i \in \{1, \dots, N\}$.

For property 3, suppose we have $(\lambda, \mu) \in \Delta$ with $\lambda_{\{j\}} \le p_j$ for all j. Then, if there exists j, $\lambda_{\{j\}} + \mu_j \neq p_j$, then it must be the case that $\lambda_{\{j\}} + \mu_j \ge p_j$. Set (λ', μ') and (λ'', μ'') as follows: $\lambda'_F = \lambda''_F = \lambda_F$ for all F, $\mu'_i = \mu''_i = \mu_i$ for all $i \neq j$, $\mu'_j = p_j - \lambda_{\{j\}}$, and $\mu''_j = 2\mu_j - \mu'_j$. Thus, we have $0.5(\lambda', \mu') + 0.5(\lambda'', \mu'')$. Furthermore, (λ', μ') and (λ'', μ'') lie in the set Δ. Thus, (λ, μ) is not an extreme point $\lambda_{\{i\}} + \mu_i \neq p_i$ for some $i \in \{1, \dots, N\}$.

Finally, for property 4, assume we have $(\lambda, \mu) \in \Delta$ with properties 1, 2, and 3, but there exists $\lambda_{\{j\}} \notin \{p_{N+1}, p_N, \dots, p_j\}$. Then, by property 3, we must have $\mu_j > 0$. Define sets

$$I = \{i\colon \lambda_{\{i\}} = \lambda_{\{j\}}\},$$
$$\mathcal{F} = \{F\colon \lambda_F = \lambda_{\{j\}}\}.$$

Set $\delta := \min\{\min_{k\notin I} |\lambda_{\{j\}} - \lambda_{\{k\}}|, \min_{k=1,2,\dots,N+1} |\lambda_{\{j\}} - p_k|, \mu_j\}$. Now, define two feasible points (λ', μ') and (λ'', μ'') as follows: $\lambda'_F = \lambda_F + \delta/2$ for $F \in \mathcal{F}$, and $\lambda'_F = \lambda_F$ for $F \notin \mathcal{F}$, $\mu'_i = \mu_i - \delta/2$ for $i \in I$ and $\mu'_i = \mu_i$ for $i \notin I$; $\lambda''_F = \lambda_F - \delta/2$ for $F \in \mathcal{F}$, and $\lambda''_F = \lambda_F$ for $F \notin \mathcal{F}$, $\mu''_i = \mu_i + \delta/2$ for $i \in I$ and $\mu''_i = \mu_i$ for $i \notin I$. Then, we have $0.5(\lambda', \mu') + 0.5(\lambda'', \mu'') = (\lambda, \mu)$, and thus (λ, μ) cannot be an extreme point if $\lambda_j \notin \{p_{N+1}, p_N, \dots, p_j\}$ for all $j = 1, 2, \dots, N$. □ □

Lemma 1 and (7) directly lead to the following result:

Lemma 3. *For any capacity portfolio K, the marginal value of resource F, $\mathbb{E}_D[\lambda^\star_F(K, D)]$, equals $\sum_{j=1}^N p_j \mathbb{P}(\bigcup_{i\in F} \Omega_{i,j}(K))$.*

A.3. Symmetric Products

For a system with symmetric products, i.e., where the demand for the products is independent and identically distributed, and with equal financial parameters $p_i = p$ for $i = 1, 2, \dots, N$, the results derived thus far are simplified. In this case, the dual problem (A4)–(A5) can be rewritten with $p_i = p$ for all i. The corresponding shortage regions can now be defined as

$$\Omega_i(K) \equiv \big\{D\colon \exists \text{ an optimal solution to (A4)–(A5)}, \\ (\lambda^\star, \mu^\star) \text{ with } \lambda^\star_{\{i\}}(K, D) = p\big\}.$$

Appendix B. Proof of Results

Proof of Proposition 5. The result follows if we demonstrate that

$$\mathbb{P}\Bigg(\bigcup_{j=1}^k \bigcup_{i\in F\cup\{q,r\}} \Omega_{i,j}\Bigg) - \mathbb{P}\Bigg(\bigcup_{j=1}^k \bigcup_{i\in F\cup\{q\}} \Omega_{i,j}\Bigg) \\ \le \mathbb{P}\Bigg(\bigcup_{j=1}^k \bigcup_{i\in F\cup\{r\}} \Omega_{i,j}\Bigg) - \mathbb{P}\Bigg(\bigcup_{j=1}^k \bigcup_{i\in F} \Omega_{i,j}\Bigg), \qquad \text{(B1)}$$

for $k = 1, 2, \dots, N$. Defining $\tilde{\Omega}_i = \bigcup_{j=1}^k \Omega_{i,j}$, (B1) is equivalent to

$$\mathbb{P}\Bigg(\bigcup_{i\in F\cup\{q,r\}} \tilde{\Omega}_i\Bigg) - \mathbb{P}\Bigg(\bigcup_{i\in F\cup\{q\}} \tilde{\Omega}_i\Bigg) \\ \le \mathbb{P}\Bigg(\bigcup_{i\in F\cup\{r\}} \tilde{\Omega}_i\Bigg) - \mathbb{P}\Bigg(\bigcup_{i\in F} \tilde{\Omega}_i\Bigg). \qquad \text{(B2)}$$

We can write the right-hand side of (B2) as

$$\mathbb{P}\Bigg(\bigcup_{i\in F\cup\{r\}} \tilde{\Omega}_i\Bigg) - \mathbb{P}\Bigg(\bigcup_{i\in F} \tilde{\Omega}_i\Bigg) = \mathbb{P}\Bigg(\tilde{\Omega}_r \cup \Bigg(\bigcup_{i\in F} \tilde{\Omega}_i\Bigg)\Bigg) - \mathbb{P}\Bigg(\bigcup_{i\in F} \tilde{\Omega}_i\Bigg) \\ = \mathbb{P}(\tilde{\Omega}_r) - \mathbb{P}\Bigg(\tilde{\Omega}_r \cap \Bigg(\bigcup_{i\in F} \tilde{\Omega}_i\Bigg)\Bigg). \qquad \text{(B3)}$$

Similarly, the left-hand side of (B2) can be written as

$$\mathbb{P}\Bigg(\bigcup_{i\in F\cup\{q,r\}} \tilde{\Omega}_i\Bigg) - \mathbb{P}\Bigg(\bigcup_{i\in F\cup\{q\}} \tilde{\Omega}_i\Bigg) \\ = \mathbb{P}\Bigg(\tilde{\Omega}_r \cup \Bigg(\bigcup_{i\in F\cup\{q\}} \tilde{\Omega}_i\Bigg)\Bigg) - \mathbb{P}\Bigg(\bigcup_{i\in F} \tilde{\Omega}_i\Bigg) \\ = \mathbb{P}(\tilde{\Omega}_r) - \mathbb{P}\Bigg(\tilde{\Omega}_r \cap \Bigg(\bigcup_{i\in F\cup\{q\}} \tilde{\Omega}_i\Bigg)\Bigg). \qquad \text{(B4)}$$

Comparing (B3) and (B4) and using the fact that $\mathbb{P}(\tilde{\Omega}_r \cap (\bigcup_{i\in F} \tilde{\Omega}_i)) \le \mathbb{P}(\tilde{\Omega}_r \cap (\bigcup_{i\in F\cup\{q\}} \tilde{\Omega}_i))$, (B2) follows.

If $K_F > 0$, then we prove that

$$\mathbb{P}\Bigg(\bigcup_{j=1}^N \Omega_{r,j} \cap \Bigg(\bigcup_{i\in F} \bigcup_{j=1}^N \Omega_{i,j}\Bigg)\Bigg) \\ < \mathbb{P}\Bigg(\bigcup_{j=1}^N \Omega_{r,j} \cap \Bigg(\bigcup_{i\in F\cup\{q\}} \bigcup_{j=1}^N \Omega_{i,j}\Bigg)\Bigg),$$

and thus (B2) holds with strict inequality for $k = N$. To see why this inequality holds, pick any $i \in F$ and a demand realization along with a corresponding optimal allocation $x^\star$ such that resource F has some excess capacity left after all allocations, resources $\{q\}$ and $\{r\}$ are completely depleted after allocations, whereas there is shortfall in products q and r, i.e., $K_F - \sum_{j\in F} x^\star_{j,F} > 0$, $K_{\{q\}} - x^\star_{q,\{q\}} = 0$, $K_{\{r\}} - x^\star_{r,\{r\}} = 0$, $D_q > \sum_{F'\subseteq\{1,\dots,N\}} x^\star_{q,F'}$ and $D_r > \sum_{F'\subseteq\{1,\dots,N\}} x^\star_{r,F'}$. Thus, this demand realization lies in $\Omega_{q,q}$ and $\Omega_{r,r}$, but not in $\bigcup_{i\in F} \bigcup_{j=1}^N \Omega_{i,j}$. Then, using the fact that the demand distribution has a positive density on $\mathbb{R}^N_+$, we obtain the existence of a set A of demand realizations of positive measure (along with an optimal allocation) on which resource F still has excess capacity left after allocations, resource $\{j\}$ remains exhausted, and demand of product j continues to have a shortfall. Clearly, $A \subseteq \Omega_{q,q} \cap \Omega_{r,r}$ and $A \cap (\bigcup_{i\in F} \bigcup_{j=1}^N \Omega_{i,j})$ is empty, and thus

$$\mathbb{P}\Bigg(\bigcup_{j=1}^N \Omega_{r,j} \cap \Bigg(\bigcup_{i\in F} \bigcup_{j=1}^N \Omega_{i,j}\Bigg)\Bigg) \\ < \mathbb{P}\Bigg(\bigcup_{j=1}^N \Omega_{r,j} \cap \Bigg(\bigcup_{i\in F\cup\{q\}} \bigcup_{j=1}^N \Omega_{i,j}\Bigg)\Bigg). \quad \square$$

Proof of Proposition 7. Consider any portfolio such that there are resources $F \subset F' \subset F''$ with $K_F, K_{F'}, K_{F''} > 0$. We shall show that such a portfolio cannot be optimal. The optimality conditions requires that the optimal portfolio must satisfy the following KKT conditions:

$$V(F) = c_1[1 + (|F| - 1)\delta], \quad V(F') = c_1[1 + (|F'| - 1)\delta], \\ V(F'') = c_1[1 + (|F''| - 1)\delta].$$

Applying Proposition 5 (which applies to this setting without change), there are diminishing returns to flexibility, and using $K_F > 0$ we obtain

$$\frac{V(F'') - V(F')}{|F''| - |F'|} < \frac{V(F') - V(F)}{|F'| - |F|}.$$

This leads to a contradiction and the result follows. □

Proof of Proposition 8. Consider any capacity portfolio K. We can rewrite the marginal value of an increase in the capacity of resource $\{i\}$ as

$$\begin{aligned} V(\{i\}) &= \sum_{k=1}^{N} (p_k - p_{k+1}) \mathbb{P}\left(\bigcup_{j=1}^{k} \Omega_{i,j}\right) \\ &\quad + \sum_{k=1}^{N} (s_k - s_{k+1}) \mathbb{P}\left(\bigcup_{j=1}^{k} \Omega_{i+1,j} \Big\backslash \bigcup_{j=1}^{N} \Omega_{i,j}\right) \\ &= \sum_{k=1}^{N} (s_k - s_{k+1}) \mathbb{P}\left(\bigcup_{j=1}^{k} \Omega_{i,j} \cup \Omega_{i+1,j}\right) + s \sum_{k=1}^{N} \mathbb{P}\left(\bigcup_{j=1}^{k} \Omega_{i,j}\right), \end{aligned}$$

where $p_{N+1} = 0$, $s_i = p_i$ for $i = 1, 2, \ldots, N$, and $s_{N+1} = s$. Similarly, the marginal value of resource F, where $F \subseteq \{1, 2, \ldots, N\}$, can be written as

$$\begin{aligned} V(F) &= \sum_{k=1}^{N} (s_k - s_{k+1}) \mathbb{P}\left(\bigcup_{j=1}^{k} \left(\bigcup_{i \in F} (\Omega_{i,j} \cup \Omega_{i+1,j})\right)\right) \\ &\quad + s \sum_{k=1}^{N} \mathbb{P}\left(\bigcup_{j=1}^{k} \bigcup_{i \in F} \Omega_{i,j}\right), \end{aligned}$$

where $\Omega_{N+1,j} = \phi$ is the null set for $j = 1, \ldots, N$. The rest of the proof follows analogous to that of Proposition 5. □

References

Bassamboo, A., J. M. Harrison, A. Zeevi. 2006. Design and control of a large call center: Asymptotic analysis of an LP-based method. *Oper. Res.* **54**(3) 419–435.

Bassamboo, A., R. S. Randhawa, J. A. Van Mieghem. 2010a. A little flexibility is all you need: Asymptotic optimality of tailored chaining and pairing in queuing systems. Working paper, Northwestern University, Evanston, IL.

Bassamboo, A., R. S. Randhawa, A. Zeevi. 2010b. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Sci.* Forthcoming.

Bassok, Y., R. Anupindi, R. Akella. 1999. Single-period multiproduct inventory models with substitution. *Oper. Res.* **47**(4) 632–642.

Bish, E. K., Q. Wang. 2004. Optimal investment strategies for flexible resources, considering pricing and correlated demands. *Oper. Res.* **52**(6) 954–964.

Chod, J., N. Rudi. 2005. Resource flexibility with responsive pricing. *Oper. Res.* **53**(3) 532–549.

Dong, L., N. Rudi. 2004. Who benefits from transshipment? Exogenous vs. endogenous wholesale prices. *Management Sci.* **50**(5) 645–657.

Fine, C. H., R. M. Freund. 1990. Optimal investment in product-flexible manufacturing capacity. *Management Sci.* **36**(4) 449–466.

Gerchak, Y., M. Henig. 1989. Component commonality in assemble-to-order systems: Models and properties. *Naval Res. Logist.* **36**(1) 61–68.

Goyal, M., S. Netessine. 2007. Strategic technology choice and capacity investment under demand uncertainty. *Management Sci.* **53**(2) 192–207.

Graves, S. C., B. T. Tomlin. 2003. Process flexibility in supply chains. *Management Sci.* **49**(7) 907–919.

Gupta, D., Y. Gerchak, J. A. Buzacott. 1992. The optimal mix of flexible and dedicated manufacturing capacities: Hedging against demand uncertainty. *Internat. J. Production Econom.* **28**(3) 309–319.

Gurumurthi, S. S. Benjaafar. 2004. Modeling and analysis of flexible queueing systems. *Naval Res. Logist.* **51**(5) 755–782.

Gurvich, I., W. Whitt. 2010. Service-level differentiation in many-server service systems via queue-ratio routing. *Oper. Res.* **58**(2) 316–328.

Harrison, J. M., J. A. Van Mieghem. 1999. Multi-resource investment strategies: Operational hedging under demand uncertainty. *Eur. J. Oper. Res.* **113**(1) 17–29.

Harrison, J. M., A. Zeevi. 2005. A method for staffing large call centers using stochastic fluid models. *Manufacturing Service Oper. Management* **7**(1) 20–36.

Iravani, S. M., M. P. Van Oyen, K. T. Sims. 2005. Structural flexibility: A new perspective on the design of manufacturing and service operations. *Management Sci.* **51**(2) 151–166.

Jordan, W. C., S. C. Graves. 1995. Principles on the benefits of manufacturing process flexibility. *Management Sci.* **41**(4) 577–594.

Netessine, S., G. Dobson, R. A. Shumsky. 2002. Flexible service capacity: Optimal investment and the impact of demand correlation. *Oper. Res.* **50**(2) 375–388.

Sheikhzadeh, M., S. Benjaafar, D. Gupta. 1998. Machine sharing in manufacturing systems: Total flexibility versus chaining. *Internat. J. Flexible Manufacturing Systems* **10**(4) 351–378.

Tomlin, B. 2006. On the value of mitigation and contingency strategies for managing supply chain disruption risks. *Management Sci.* **52**(5) 639–657.

Tomlin, B., Y. Wang. 2005. On the value of mix flexibility and dual sourcing in unreliable newsvendor networks. *Manufacturing Service Oper. Management* **7**(1) 37–57.

Topkis, D. M. 1998. *Supermodularity and Complementarity*. MIT Press, Cambridge, MA.

Van Mieghem, J. A. 1998. Investment strategies for flexible resources. *Management Sci.* **44**(8) 1071–1078.

Van Mieghem, J. A. 2004. Commonality strategies: Value drivers and equivalence with flexible capacity and inventory substitution. *Management Sci.* **50**(3) 419–424.

Van Mieghem, J. A., N. Rudi. 2002. Newsvendor networks: Inventory management and capacity investment with discretionary activities. *Manufacturing Service Oper. Management* **4**(4) 313–335.

Wallace, R. B., W. Whitt. 2005. A staffing algorithm for call centers with skill-based routing. *Manufacturing Service Oper. Management* **7**(4) 276–294.

Whitt, W. 2006. Staffing a call center with uncertain arrival rate and absenteeism. *Production Oper. Management* **15**(1) 88–102.