OPERATIONS RESEARCH
Vol. 60, No. 6, November–December 2012, pp. 1423–1435
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.1120.1107

A Little Flexibility Is All You Need: On the Asymptotic Value of Flexible Capacity in Parallel Queuing Systems

Achal Bassamboo
Kellogg School of Management, Northwestern University, Evanston, Illinois 60203,
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 60203,
vanmieghem@kellogg.northwestern.edu

We analytically study optimal capacity and flexible technology selection in parallel queuing systems. We consider N stochastic arrival streams that may wait in N queues before being processed by one of many resources (technologies) that differ in their flexibility. A resource's ability to process k different arrival types or classes is referred to as level-k flexibility. We determine the capacity portfolio (consisting of *all* resources at *all* levels of flexibility) that minimizes linear capacity and linear holding costs in high-volume systems where the arrival rate $\lambda \to \infty$. We prove that "a little flexibility is all you need": the optimal portfolio invests $O(\lambda)$ in specialized resources and only $O(\sqrt{\lambda})$ in flexible resources and these optimal capacity choices bring the system into heavy traffic. Further, considering symmetric systems (with type-independent parameters), a novel "folding" methodology allows the specification of the asymptotic queue count process for any capacity portfolio under longest-queue scheduling in closed form that is amenable to optimization. This allows us to sharpen "a little flexibility is all you need": the asymptotically optimal flexibility configuration for symmetric systems with mild economies of scope invests a lot in specialized resources but only a little in flexible resources and only in level-2 flexibility, but effectively nothing ($o(\sqrt{\lambda})$) in level-$k > 2$ flexibility. We characterize "tailored pairing" as the theoretical benchmark configuration that maximizes the value of flexibility when demand and service uncertainty are the main concerns.

Subject classifications: flexibility; capacity optimization; queueing network; diffusion approximation.
Area of review: Manufacturing, Service, and Supply Chain Operations.
History: Received June 2009; revisions received November 2010, October 2011, May 2012; accepted July 2012.

1. Introduction

Deciding on the appropriate amount and configuration of flexibility is a classic management problem: should different types of products or customers be processed or served with specialized or flexible capacity? And how much flexibility is needed to effectively match demand and supply? The extant literature on flexibility refers to the ability of a resource to process multiple types of products as *mix-* (Chod et al. 2010), *process-* (Sethi and Sethi 1990), *product-* (Fine and Freund 1990) or *scope-flexibility* (Van Mieghem 2008). Substantial progress has been made in our understanding of flexibility over the last 20 years. One important insight is that the choice between specialization and flexibility is not an "all-or-nothing" proposition. The literature has advanced two different interpretations of this insight that are most relevant to our paper: tailoring and chaining.

Van Mieghem (1998) showed that it is typically optimal to invest in a portfolio of two specialized and one flexible resource in a two-product newsvendor network with a linear cost structure. The dedicated resources act as base capacity and the flexible resource serves as an optimal cost/benefit response to demand variability. We will refer to such a portfolio approach of fitting or optimizing the amounts and levels of flexibility to demand profiles as *tailored flexibility*. While tailored flexibility is well understood in a two-product setting, finding desirable flexible processing systems for $N > 2$ products is much more difficult because the capacity portfolio can now consist of $2^N - 1$ different resources, and hence grows exponentially in N. Recently, Bassamboo et al. (2010) analyzed such a system in a newsvendor setting. To describe their key result, let "level-k flexibility" refer to the ability to process $k \in \{1, 2, \ldots, N\}$ different product types. Then 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 $k = N$. Bassamboo et al. (2010) shows that, if the flexibility premiums are linear in the flexibility level, then the optimal capacity portfolio invests in at most two adjacent levels of flexibility. In this paper, we expand this result in a parallel queuing network and show that, with mild economies of scope (i.e., as long as capacity costs are not too concave in flexibility), the investment is in levels 1 and 2.

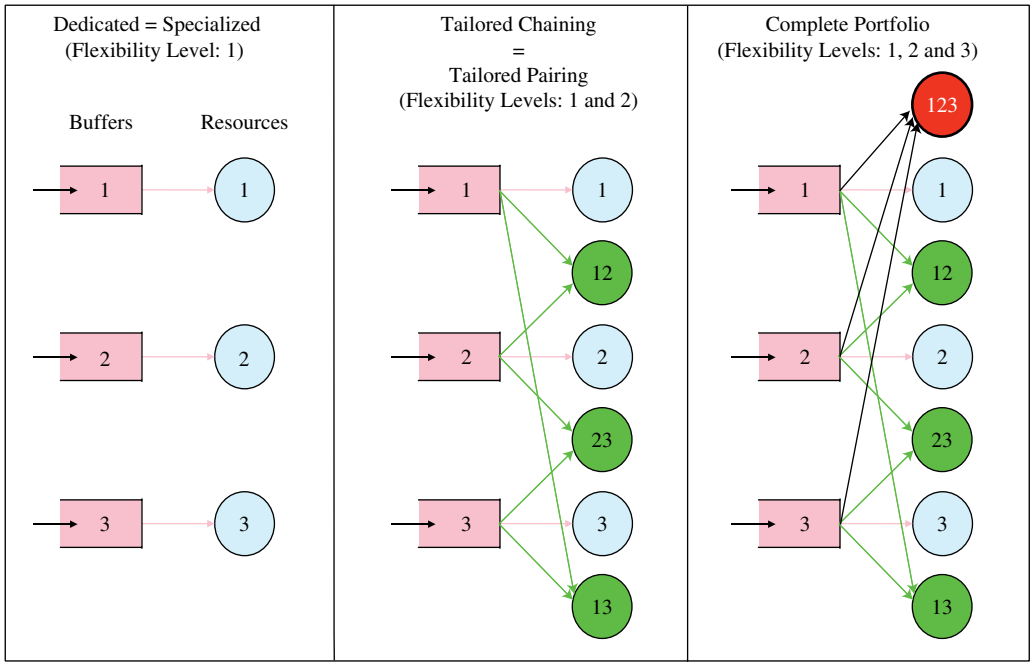
In their seminal paper, Jordan and Graves (1995) showed that "a little flexibility can achieve almost all the benefits of total flexibility" by using *only* level-2 flexible resources in a special configuration called chaining. Imagine a graph where product types are represented by rectangles and resources by circles, such as in Figure 1 for $N = 3$ product types. An arc from a rectangle to a circle then represents a possible product-resource assignment and thus that resource's flexibility. Chaining represents any flexibility configuration of N level-2 flexible resources that are connected, directly or indirectly, 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. Jordan and Graves consider a single-period newsvendor network model where random demand is allocated ex post to prefixed capacity. Excess demand is assumed lost and the allocation objective is to minimize the corresponding shortfall. Using simulation and providing some analytical justification, 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. Hopp et al. (2004) generalized these chaining configurations that utilize level-2 flexible resources to D-skilled chains that consist of level-D flexible resources and showed that these configurations perform well in serial production lines. In recent work, Chou et al. (2008) used the concept of graph expansion to construct flexible configurations that work well in newsvendor networks.

In this paper, we consider a processing system with N stochastic arrival streams, each requiring a different type of stochastic service. Type i arrivals wait in buffer i before processing and incur holding costs. The system manager can invest in a portfolio of $2^N - 1$ different resources that differ in their flexibility. The trade-off is simple: higher levels of flexibility reduce holding costs more but come at a higher investment cost. Indeed, in addition to the holding costs, the system incurs a capacity cost rate that is linear in capacity size and depends on the flexibility level. Although our system is not amenable to exact analysis, we characterize analytically the optimal amount, level, and configuration of flexibility for high-volume systems where the arrival rate $\lambda \to \infty$. The key contributions of this paper are:

1. We prove that the optimal portfolio invests $O(\lambda)$ in specialized resources and only $O(\sqrt{\lambda})$ in flexible resources when costs are linear in the flexibility level. In other words, "a little flexibility is all you need" in any high-volume, parallel queuing system. We also show that economic capacity optimization brings the queuing system in heavy traffic.

2. For symmetric systems[1] with mild economies of scope,[2] we prove that level-2 flexibility is all that is needed.

Figure 1. Flexibility configurations for $N = 3$ product types.



Specifically, the asymptotically optimal flexibility configuration invests $O(\lambda)$ or a lot in dedicated resources, $O(\sqrt{\lambda})$ or a little in level-2 flexibility, but $o(\sqrt{\lambda})$ or effectively nothing in level-$k > 2$ flexibility. This sharpens "a little flexibility is all you need"—not only the amount but, also the level of flexibility is small—and refines the findings in Bassamboo et al. (2010).

3. We provide analytical expressions for the symmetric capacity portfolio for $N = 2$ and $N = 3$ and for the maximal asymptotic value of flexibility. This expression corresponds to the performance of the asymptotically optimal symmetric configuration called "tailored pairing" (cf. Bassamboo et al. 2010). Tailored pairing uses a dedicated resource for each arrival stream to serve the base demand, and a level-2 flexible resource for each pair of arrival streams to serve the variable demand. Dedicated capacity is sized proportional to expected demand, whereas level-2 flexible capacity is proportional to the square root of demand. Because pairing requires too many $(N(N-1)/2)$ servers, its practical appeal diminishes quickly as N grows. However, it serves two important purposes: (i) it provides an upper bound on the value of flexibility against which other configurations can be "benchmarked"; (ii) it allows us to provide the first analytic proof that tailored chaining is asymptotically optimal for $N = 3$ in a queuing setting, which differs from the newsvendor setting studied by Jordan and Graves (1995). Indeed, tailored chaining and tailored pairing are identical configurations for $N = 3$ and thus dominate dedicated or fully flexible configurations, as shown in Figure 1.

4. The above analytic characterizations follow from two methodological novelties:

• Our "folding" methodology allows us to specify the asymptotic queue count process for symmetric systems with a general capacity portfolio under dynamic longest-queue scheduling in closed form that is amenable to optimization. This technique involves folding the state-space and studying the order statistics of the limiting queue-length. This ordered queue-length process behaves as a reflected Brownian motion in a wedge. For symmetric systems, we can then use the results in Williams (1987) to specify the stationary distribution and expected holding costs and optimize capacity analytically. To our knowledge, we present the first closed-form analytical expressions for the stationary queue-length distribution and asymptotically optimal capacities for symmetric parallel queueing networks.

• We also show that it is not economical to invest in the sufficient amount of flexibility that leads to so-called complete resource pooling (CRP). CRP amounts to assuming that the resources have sufficiently overlapping flexibility and that they work collectively to the extent that they act as a single "super-server" in the heavy traffic limit. That is, processing capacities of the various resources are completely exchangeable in the heavy traffic limit and single-dimensional dynamics result. Complete resource pooling as introduced in Harrison and López (1999) has been the natural assumption in the growing literature on flexible queuing networks in heavy traffic and obviously leads to excellent waiting time performance. In contrast, CRP is suboptimal in our setting, given that we prove the optimal amount of level-2 flexibility to be $O(\sqrt{\lambda})$, which results in a truly multidimensional reflected Brownian motion with state-dependent drift (arising from the longest-queue scheduling). In other words, although CRP could be obtained using level-2 flexibility only, it would require more capacity than is optimal.

2. Model Primitives and Basic Setup for Flexibility

We denote types by $i = 1, 2, \ldots, N$ and the number of type i customer or job arrivals by time t by $A_i^\lambda(t)$. We assume that all arrival processes are independent renewal processes with common rate $\lambda > 0$. A general model is presented in Appendix EC.2. An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/opre.1120.1107. Let σ_a^λ denote the standard deviation of the interarrival times. Each arriving job has a service requirement that is independent and identically distributed across all the customers with mean m and variance σ_s^2. The coefficient of variation of service times is denoted by $c_s = \sigma_s/m$, whereas that of the interarrival times is $c_a = \lambda\sigma_a^\lambda$. We assume that c_a is a constant, independent of the rate λ, and will henceforth denote $\sigma^2 = (c_a^2 + c_s^2)/2$.

Unless we explicitly mention otherwise, we will assume that our system is completely symmetric (i.e., all model parameters are type independent), and we consider only symmetric capacity assignments. That is, we assume that each type has a dedicated resource assigned to it that operates at a fixed deterministic rate μ_1^λ that is the same for each type. Further, note that each level-k flexible resource can handle precisely one of $\binom{N}{k}$ different subsets of types. (We use the notation $\binom{p}{q} = p!/((p-q)!q!)$ if $p \geqslant q$, and 0 otherwise.) Thus, there are a total of $\sum_{k=1}^{N}\binom{N}{k} = 2^N - 1$ different resources in the system. Due to the symmetry in the system, each of the $\binom{N}{k}$ level-k flexible resources are assumed to have the same capacity, which we will denote by μ_k^λ. (Note that capacities scale the actual average service time, i.e., if a service rate of μ is allocated to a job, its average service time is m/μ and its variance is σ_s^2/μ^2.) Note that we assume that capacity can be sized continuously by varying the service rate of a given portfolio of resources.

The system incurs two types of costs: a holding cost h that is incurred per job per unit of time spent in the system (waiting and service) and a capacity cost rate that depends on capacity size and flexibility level. We assume that capacity costs are linear in size. The cost rate of capacity size μ_k of a level-k flexible resource is $c_k\mu_k^\lambda$, where $c_k = c(1+\Delta_k)$, with Δ_k denoting the flexibility premium for level-k flexible resources and we have $\Delta_k \geqslant \Delta_{k-1}$ and $\Delta_k > 0$ for $k \geqslant 2$ and

$\Delta_1 = 0$. Notice that this includes concave flexibility costs or economies of scope.

Let $Q_i^\lambda(t)$ denote the number of customers of type i in the system at time t and $\mathbb{E}\,Q_i^\lambda(\infty)$ its steady-state expected value. Using the holding cost of h per job per unit time, we obtain the total cost rate of a symmetric capacity portfolio $\mu^\lambda = (\mu_1^\lambda, \mu_2^\lambda, \ldots, \mu_N^\lambda)$ as

$$\Pi^\lambda(\mu^\lambda) = \sum_{i=1}^{N} h\,\mathbb{E}\,Q_i^\lambda(\infty) + \sum_{k=1}^{N} \binom{N}{k} c_k \mu_k^\lambda.$$

Given that optimal capacities will lead to a stable system where all jobs eventually get served, expected steady-state revenues are independent of μ^λ, and we seek the capacity portfolio $\mu^{\lambda*}$ that minimizes costs:

$$\Pi^{\lambda*} = \Pi^\lambda(\mu^{\lambda*}) = \min_{\mu \geqslant 0} \Pi^\lambda(\mu). \tag{1}$$

Given that our system involves $GI/G/1$ queue dynamics, its stationary queue-length distribution cannot be solved analytically in general. We can, however, obtain a useful upper bound on the optimal cost as follows. Observe that the optimal cost is bounded by the minimal cost when using only dedicated servers: $\Pi^\lambda(\mu^{\lambda*}) \leqslant \min_{\mu_1^\lambda \geqslant 0} \Pi^\lambda(\mu_1^\lambda, 0, \ldots, 0)$. Using only dedicated servers results in N independent $GI/G/1$ queues so that

$$\begin{aligned}\Pi^\lambda(\mu_1^\lambda, 0, \ldots, 0) &= N(h\,\mathbb{E}\,Q_1^\lambda + c\mu_1^\lambda) \\ &\leqslant N\left(h\left[\sigma^2 \frac{\mu_1^\lambda}{\mu_1^\lambda - m\lambda} + 1\right] + c\mu_1^\lambda\right),\end{aligned}$$

using Kingman's bound (cf. Kingman 1962).[3] The right-hand side is convex in μ_1^λ and reaches a minimum at $\tilde{\mu}_1^\lambda = m\lambda + \sigma\sqrt{(h/c)m\lambda}$, which yields an exact upper bound: $\Pi^\lambda(\mu^{\lambda*}) \leqslant \min_{\mu_1^\lambda \geqslant 0} \Pi^\lambda(\mu_1^\lambda, 0, \ldots, 0) \leqslant \bar{\Pi}^\lambda + Nh(\sigma^2 + 1)$, where

$$\bar{\Pi}^\lambda = Ncm\lambda + 2N\sigma\sqrt{chm\lambda}. \tag{2}$$

The upper bound also bounds the capacity cost and directly shows how the optimal capacities depend on the volume λ, which is key to our analysis: $\mu^{\lambda*}$ cannot be larger than a term proportional to λ plus a term that is $O(\lambda^{1/2})$, which is exactly the condition to bring the system into heavy traffic.

A lower bound stems from considering a system where all customer types are pooled into a single queue served by a single server that costs only c. This lower bound is similar to having a fully flexible server at the cost of a dedicated server. Such a totally pooled system never experiences any server idleness while jobs are waiting and thus dominates the original multiqueue, multiserver system. In heavy traffic, the Kingman's bound is tight and, using (2) for a single queue with arrival rate $N\lambda$, yields as an asymptotic lower bound $\Pi^{\lambda*} \geqslant \underline{\Pi}^\lambda + o(\sqrt{\lambda})$, where

$$\underline{\Pi}^\lambda = Ncm\lambda + 2\sigma\sqrt{chm\lambda N}. \tag{3}$$

The following result summarizes these results and is the justification for solving this optimization problem asymptotically when λ is large.

THEOREM 1. *The optimal cost* $\Pi^\lambda(\mu^{\lambda*})$ *is bounded*:

$$\underline{\Pi}^\lambda + o(\sqrt{\lambda}) \leqslant \Pi^\lambda(\mu^{\lambda*}) \leqslant \bar{\Pi}^\lambda + o(\sqrt{\lambda}), \tag{4}$$

and any optimal solution $(\mu_1^{\lambda*}, \ldots, \mu_N^{\lambda*})$ *to the optimization problem* (1) *satisfies* $\mu^{\lambda*} = \tilde{\mu}^\lambda + o(\sqrt{\lambda})$, *where*

$$\tilde{\mu}_1^{\lambda*} = m\lambda + \hat{\mu}_1\sqrt{\lambda}, \quad \textit{and} \tag{5}$$

$$\tilde{\mu}_k^{\lambda*} = \hat{\mu}_k\sqrt{\lambda} \quad \textit{for } k \geqslant 2, \tag{6}$$

for some $\hat{\mu}_1, \ldots, \hat{\mu}_N \in \mathbb{R}$ *with* $\hat{\mu}_k \geqslant 0$ *for* $k \geqslant 2$ *and* $\sum_{k=1}^{N} \binom{N}{k} \hat{\mu}_k > 0$.

We call $\tilde{\mu}^\lambda$ the "prescription" for a system with arrival rate λ. This theorem, which holds for asymmetric systems as well (see Appendix EC.2 for details), has two important implications. First, the optimal dedicated resources are sized on the order of the mean demand or the arrival rate and will serve the majority of the jobs. In contrast, the flexible capacities are much smaller and only proportional to the standard deviation of the demand, which is $O(\sqrt{\lambda})$. Additional insight is found by considering a single class system for which $\underline{\Pi}^\lambda = \bar{\Pi}^\lambda$ and the asymptotically optimal capacity and cost are:

$$\begin{aligned}\mu_1^{\lambda*} &= \tilde{\mu}_1^\lambda + o(\sqrt{\lambda}) = m\lambda + \sigma\sqrt{\frac{h}{c}m\lambda} + o(\sqrt{\lambda}), \\ \Pi^{\lambda*} &= \underline{\Pi}^\lambda + o(\sqrt{\lambda}) = cm\lambda + 2\sigma\sqrt{chm\lambda} + o(\sqrt{\lambda}).\end{aligned}$$

The asymptotically optimal capacity prescription $\tilde{\mu}_1^\lambda$ is the sum of two parts: base capacity λm that matches the average arriving workload plus safety capacity $\sigma\sqrt{(hm\lambda)/c}$ that accommodates variability in the arriving workload. The optimal safety capacity increases linearly with standard deviation $\sigma\sqrt{\lambda}$, as earlier observed (e.g., Kleinrock 1976, p. 331), and exhibits economies of scale. Indeed, the capacity per unit of demand rate is $m + \sigma\sqrt{(hm)/(c\lambda)}$, where the safety capacity per unit decreases in λ, as does the optimal cost per unit. Notice that these expressions are similar to results for capacity sizing in a newsvendor setting with normal demand.

Second, the theorem proves that economic optimization naturally brings the system into a parameter regime called "heavy traffic." (Loosely speaking, this means that the dedicated resources are heavily utilized. Indeed, the optimal dedicated utilization $\mu_1^{\lambda*}/\lambda \simeq 1 - \hat{\mu}_1/\sqrt{\lambda}$ tends to 100% as $\lambda \to \infty$.) The theoretical significance of the theorem is that heavy traffic is not assumed, but the proved result of capacity optimization. It also proves that configurations that satisfy the so-called CRP condition, which are widely studied in literature, are suboptimal. Under CRP, for all practical purposes the capacities of all resources can be thought of as being pooled together into one super-server that can process all types. CRP leads to state-space collapse and results in a single-dimensional limiting system. In contrast, we shall prove that the optimal capacity configuration only exhibits

partial resource pooling and results in an N-dimensional limiting system. In other words, the optimal flexible capacity is too small to lead to CRP.

Diffusion-scale optimization problem. Theorem 1 guarantees that we need only consider capacity portfolios of the form $(m\lambda + \hat{\mu}_1\sqrt{\lambda}, \hat{\mu}_2\sqrt{\lambda}, \ldots, \hat{\mu}_N\sqrt{\lambda})$ to characterize an approximate solution to (1) for large-volume systems, where $\hat{\mu} \in M := \{\hat{\mu}\colon \hat{\mu}_k \geqslant 0 \text{ for } k \geqslant 2 \text{ and } \sum_{k=1}^{N} \binom{N}{k}\hat{\mu}_k > 0\}$. The latter condition is essential for stability as it ensures that the total demand rate $N\lambda$ does not exceed total capacity of the portfolio μ^λ, i.e., $N\lambda < \sum_{k=1}^{N} \binom{N}{k}(\mu_k^\lambda/m)$. Equivalently, stability requires that we have positive safety capacity $\sum_{k=1}^{N} \binom{N}{k}\hat{\mu}_k > 0$. The corresponding resource cost is $Nc_1(m\lambda + \hat{\mu}_1^\lambda\sqrt{\lambda}) + \sum_{k=2}^{N} c_k\binom{N}{k}\hat{\mu}_k\sqrt{\lambda}$. Focusing on this regime, we can rewrite the optimization problem (1) as

$$\min_{\hat{\mu}\in M} Ncm\lambda + \sqrt{\lambda}\left(h\sum_{i=1}^{N}\mathbb{E}\,Q_i^\lambda(\infty)/\sqrt{\lambda} + \sum_{k=1}^{N} c_k\binom{N}{k}\hat{\mu}_k\right).$$

This optimization problem is equivalent to the following optimization problem that we refer to as the diffusion-scale optimization problem:[4]

$$\min_{\hat{\mu}\in M}\left\{\hat{\Pi}^\lambda(\hat{\mu}) := h\sum_{i=1}^{N}\mathbb{E}\,Q_i^\lambda(\infty)/\sqrt{\lambda} + \sum_{k=1}^{N} c_k\binom{N}{k}\hat{\mu}_k\right\}. \quad (7)$$

Although we can solve this optimization problem for any finite λ through simulation, to derive structural insights, we will consider an analytical asymptotic analysis that is accurate when the arrival rate $\lambda \to \infty$. Indeed, we shall prove that the function $\hat{\Pi}^\lambda(\hat{\mu})$ converges to the limiting function $\hat{\Pi}(\hat{\mu})$, which we will be able to specify in closed form. Moreover, we will characterize the optimal scaled capacity $\hat{\mu}^*$ that minimizes the limiting cost $\hat{\Pi}$ and use that solution to construct the prescription $(m\lambda + \hat{\mu}_1^*\sqrt{\lambda}, \hat{\mu}_2^*\sqrt{\lambda}, \ldots, \hat{\mu}_N^*\sqrt{\lambda})$ as our approximate solution to (1) for a system with (finite) arrival rate λ.

To illustrate our mode of analysis, we begin by considering the $N = 2$ type setting. In particular, we will demonstrate the folding approach that allows tractability, and even closed-form solutions. The general N case will be analyzed in a similar manner and the detailed treatment is presented in §4.

To formalize the mode of analysis, the following terminology will be useful. All random elements in this paper are defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Further, we assume all stochastic processes to lie in the space of functions that are right continuous and possess left limits. For a collection of probability measures P^n and P defined on $(A, \mathcal{A})$, where A is a general metric space and $\mathcal{A}$ its Borel σ-field, we use the notation $P^n \Rightarrow P$ as $n \to \infty$ to denote the weak convergence of P^n to P (cf. Whitt 2002).

A note on the use of symmetric capacity portfolios. To characterize the asymptotically optimal capacity investments, we restrict attention to symmetric capacity portfolios that invest equally in all resources at the same level of flexibility. Given the symmetry in the problem parameters, one expects such a symmetric portfolio to be optimal. This optimality follows if the objective function is convex in the capacity levels. Such a convexity is straightforward to show in the newsvendor setting of Jordan and Graves (1995) (see Van Mieghem 1998 and Bassamboo et al. 2010). In a queueing setting, however, this amounts to showing that the sum of the N queue-lengths is convex in the entire $2^N - 1$-dimensional capacity portfolio. Proving such convexity statements in queueing systems is not easy and, to the best of our knowledge, has only been done for single-class systems and for parallel server systems with queue-length independent routing (see, for example, Neely and Modiano 2005). Although these results suggest that convexity should extend to our setting, we have not been able to prove this conjecture in general. Hence, we focus on symmetric capacity portfolios that were shown to be optimal in flexible newsvendor systems (cf. Bassamboo et al. 2010). These portfolios were also found to be optimal in numerical experiments that we conducted (see §5 for details). We would like to point out that we are able to prove our first result that the asymptotically optimal capacity portfolio invests $O(\lambda)$ in dedicated resources and $O(\sqrt{\lambda})$ in flexible resources without the symmetry assumption, that is, for any number of resources at each level of flexibility with potentially different capacity investments (see Appendix EC.2 for details).

3. A Two-Type Symmetric Model: Asymptotically Optimal Flexibility

In this section, we analyze the optimal system configuration in a symmetric system with two types of incoming jobs. Such systems can use two dedicated resources and one flexible resource that can serve either type. We will restrict attention to "longest-queue (LQ)" policies with a preemptive feature described as follows: When a dedicated resource completes a service request, it next processes any job in the system of its own type; if there is no such job, it idles. Each flexible resource serves the type with the longer queue preemptively, where the remainder of the service time of the preempted job is taken up by the server, which resumes processing this job. This method of preemption and the use of longest queue in symmetric system has been studied in Zipkin (1995). LQ policies have also been studied in Zheng and Zipkin (1990), Menich and Serfozo (1991), and Van Mieghem (2003), and shown to be optimal in specific settings. We expect this policy to be optimal in our setting. However, proving this claim is beyond the scope of the current treatment.[5] In numerical and simulation studies, Sheikhzadeh et al. (1998) and Jordan et al. (2004) compare the LQ policy with other reasonable policies and find that it always outperforms these policies, even for asymmetric systems.

3.1. The Folding Method

Asymptotically, we expect the scaled queue-length processes to behave as diffusions. Much of the literature has

shown that flexibility in such systems can result in complete resource pooling where the multidimensional state-space collapses in the limit to a single-dimensional state-space. Such collapse requires more flexible capacity (i.e., at a scale greater than $O(\sqrt{\lambda})$) that is optimal for our system. Indeed, we now show that the limiting system behavior remains a bona fide two-dimensional diffusion process:

LEMMA 1. *As* $\lambda \to \infty$, *if* $Q^\lambda(0)/\sqrt{\lambda} \Rightarrow \hat{Q}(0)$, *then* $Q^\lambda(\cdot)/\sqrt{\lambda} \Rightarrow \hat{Q}(\cdot)$, *where*

$$\begin{aligned}\hat{Q}_1(t) &= \hat{Q}_1(0) - \frac{1}{m}\int_0^t (\hat{\mu}_1 + 1\{\hat{Q}_1(s) \geqslant \hat{Q}_2(s)\}\hat{\mu}_2)\,ds \\ &\quad + \sigma\sqrt{2}B_1(t) + L_1(t) \\ \hat{Q}_2(t) &= \hat{Q}_2(0) - \frac{1}{m}\int_0^t (\hat{\mu}_1 + 1\{\hat{Q}_2(s) > \hat{Q}_1(s)\}\hat{\mu}_2)\,ds \\ &\quad + \sigma\sqrt{2}B_2(t) + L_2(t),\end{aligned} \tag{8}$$

where B_1 *and* B_2 *are two standard independent Brownian motions,* L_i *are nondecreasing, continuous processes such that* $L_1(0) = L_2(0) = 0$, *and* $\hat{Q}_i(t) \geqslant 0$ *and* $\int_0^t \hat{Q}_i(s)\,dL_i(s) = 0$ *for all* $t \geqslant 0$.

The limiting diffusion characterized in (8) is not directly amenable to analysis because the drift of the reflected Brownian motion $(\hat{Q}_1, \hat{Q}_2)$ is not continuous. This discontinuity stems from the LQ routing policy under which the flexible resource serves the longer queue in a preemptive fashion. This causes the drift of the diffusion to change when a queue switches from being the longer to shorter, or vice versa, as depicted in Figure 2(a).

Luckily, we can transform the diffusion $\hat{Q}$ into one with constant drift and recover analytic tractability by monitoring the order statistics of the queue-length processes and "folding" the state-space. Given that we consider symmetric systems, we only need $\hat{Q}_1(t) + \hat{Q}_2(t)$, which equals $\hat{Q}_{\max}(t) + \hat{Q}_{\min}(t)$, where $\hat{Q}_{\max}(t) = \max(\hat{Q}_1(t), \hat{Q}_2(t))$ and $\hat{Q}_{\min}(t) = \min(\hat{Q}_1(t), \hat{Q}_2(t))$. The benefit of considering the maximum and minimum queue-lengths is that the drifts of these ordered queues are constant, which allows the simpler dynamics of Proposition 1.

PROPOSITION 1. *As* $\lambda \to \infty$, *if* $Q^\lambda(0)/\sqrt{\lambda} \Rightarrow \hat{Q}(0)$, *then* $(Q^\lambda_{\max}(\cdot)/\sqrt{\lambda}, Q^\lambda_{\min}(\cdot)/\sqrt{\lambda}) \Rightarrow \hat{Q}(\cdot)$, *where*

$$\begin{aligned}\hat{Q}_{\max}(t) &= \hat{Q}_{\max}(0) - \frac{\hat{\mu}_1 + \hat{\mu}_2}{m}t + \sigma\sqrt{2}B_1(t) + Y_1(t) \\ \hat{Q}_{\min}(t) &= \hat{Q}_{\min}(0) - \frac{\hat{\mu}_1}{m}t + \sigma\sqrt{2}B_2(t) - Y_1(t) + Y_2(t),\end{aligned} \tag{9}$$

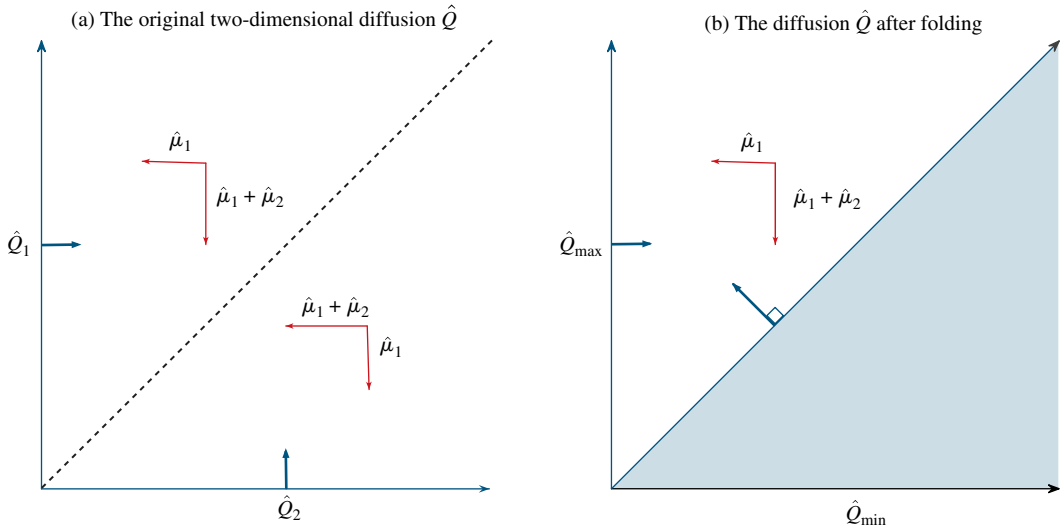
where B_1 *and* B_2 *are two standard independent Brownian motions,* Y_1, Y_2 *are two nondecreasing continuous processes such that* $Y_1(0) = Y_2(0) = 0$, *and* $Q_{\max}(t) \geqslant Q_{\min}(t) \geqslant 0$, $\int_0^t (\hat{Q}_{\max}(s) - \hat{Q}_{\min}(s))\,dY_1(s) = 0$ *and* $\int_0^t \hat{Q}_{\min}(s)\,dY_2(s) = 0$ *for all* $t \geqslant 0$.

We can now compute the steady-state distribution of the process $(\hat{Q}_{\max}, \hat{Q}_{\min})$ by "unfolding" the state-space and considering the process with constant drift on the entire positive quadrant. Given that this process then simplifies to two independent Brownian motions in a quadrant, its steady-state distribution is a simple product form of exponentials. When "folding" the state-space into the upper triangle (or wedge) in Figure 2(b), owing to the normal reflection, we still obtain a product form of exponentials. Defining $G_2 = \{(x, y) \in \mathbb{R}^2_+ : x \geqslant y\}$, we characterize the steady-state distribution of the process $(\hat{Q}_{\max}, \hat{Q}_{\min})$ in the following result.

PROPOSITION 2. *The steady-state distribution of the process* $(\hat{Q}_{\max}, \hat{Q}_{\min})$ *on* G_2 *has the density*

$$\pi(x, y) = \alpha \exp\left(-\left(\frac{\hat{\mu}_1 + \hat{\mu}_2}{\sigma^2 m}\right)x - \frac{\hat{\mu}_1}{\sigma^2 m}y\right),$$

Figure 2. A pictorial representation of the drifts of the limiting queueing dynamics $\hat{Q}$ (left). The order statistics $(\hat{Q}_{\min}, \hat{Q}_{\max})$ live in the folded state space with constant drift (right).



where

$$\alpha = \left(\int_{G_2} \exp\left(-\left(\frac{\hat{\mu}_1 + \hat{\mu}_2}{m\sigma^2} \right) x - \frac{\hat{\mu}_1}{m\sigma^2} y \right) dx\, dy \right)^{-1}$$

is a normalizing constant. Further, the corresponding expected queue-lengths are

$$\mathbb{E}\,\hat{Q}_{\min}(\infty) = \frac{1}{2\hat{\mu}_1 + \hat{\mu}_2} \sigma^2 m$$

and $\mathbb{E}\,\hat{Q}_{\max}(\infty) = \mathbb{E}\,\hat{Q}_{\min}(\infty) + (1/(\hat{\mu}_1 + \hat{\mu}_2))\sigma^2 m$.

Using this steady-state characterization, we can compute the diffusion-scale cost $\hat{\Pi}$ and characterize its optimal capacities $\hat{\mu}^*$ by solving the diffusion-scale optimization problem (7):

$$\min_{\{(\hat{\mu}_1, \hat{\mu}_2):\, \hat{\mu}_2 \geqslant 0,\, 2\hat{\mu}_1 + \hat{\mu}_2 > 0\}} \hat{\Pi}(\hat{\mu}_1, \hat{\mu}_2)$$
$$\equiv \left(\frac{2}{2\hat{\mu}_1 + \hat{\mu}_2} + \frac{1}{\hat{\mu}_1 + \hat{\mu}_2} \right) \sigma^2 hm + (2c\hat{\mu}_1 + c(1+\Delta_2)\hat{\mu}_2). \tag{10}$$

The following proposition presents the results.

Proposition 3. *For $N = 2$, the optimal safety capacity that solves* (10) *is*

$$(\hat{\mu}_1^*, \hat{\mu}_2^*) = \begin{cases} \sigma\sqrt{\dfrac{hm}{c}}(-\psi^*, -\gamma^*\psi^*) & \text{if } 0 \leqslant \Delta_2 < 0.2, \\ \sigma\sqrt{\dfrac{hm}{c}}\left(0, \sqrt{\dfrac{3}{(1+\Delta_2)}}\right) & \text{if } \Delta_2 = 0.2, \\ \sigma\sqrt{\dfrac{hm}{c}}(\psi^*, \gamma^*\psi^*) & \text{if } 0.2 < \Delta_2 < 0.5, \\ \sigma\sqrt{\dfrac{hm}{c}}(1, 0) & \text{if } 0.5 \leqslant \Delta_2, \end{cases} \tag{11}$$

where $\psi^* = \sqrt{(3 + 1/(1+\gamma^*))/((2+\gamma^*)(2+\gamma^*(1+\Delta_2)))}$ *and* γ^* *is defined as follows*:

$$\gamma^* = \begin{cases} 2\dfrac{(1 - 3\Delta_2 + \sqrt{\Delta_2(1-\Delta_2)})}{5\Delta_2 - 1} \\ \quad \text{if } 0 \leqslant \Delta_2 < 0.5,\ \Delta_2 \neq 0.2, \\ 0 \quad \text{if } 0.5 \leqslant \Delta_2. \end{cases} \tag{12}$$

Notice that $(\hat{\mu}_1^*, \hat{\mu}_2^*)(1/\sigma)\sqrt{c/(hm)}$ depends only on the flexibility premium Δ_2. Hence, for a fixed Δ_2 value, the optimal safety capacities scale with the standard deviation as expected. At the optimal solution, the safety capacity cost $c\hat{\mu}_1^* + c(1+\Delta_2)\hat{\mu}_2^*$ equals the holding cost $h\mathbb{E}[\hat{Q}_1(\infty) + \hat{Q}_2(\infty)]$ (this is similar to the properties of the classical Economic Order Quantity (EOQ) model). Using the solution to the limiting problem, we can construct a capacity prescription for a system with finite arrival rate λ that is asymptotically optimal.

Proposition 4. *The capacity portfolio* $(m\lambda + \hat{\mu}_1^*\sqrt{\lambda}, \hat{\mu}_2^*\sqrt{\lambda})$, *with* $\hat{\mu}_1^*$, $\hat{\mu}_2^*$ *given by* (11), *is asymptotically optimal for the optimization problem* (1) *in the sense that*

$$\lim_{\lambda\to\infty} \frac{\Pi^\lambda(m\lambda + \hat{\mu}_1^*\sqrt{\lambda}, \hat{\mu}_2^*\sqrt{\lambda}) - \Pi^{\lambda *}}{\sqrt{\lambda}} = 0. \tag{13}$$

This result states that the loss in optimality incurred by using the prescription $(m\lambda + \hat{\mu}_1^*\sqrt{\lambda}, \hat{\mu}_2^*\sqrt{\lambda})$ is negligible at the $O(\sqrt{\lambda})$ scale.

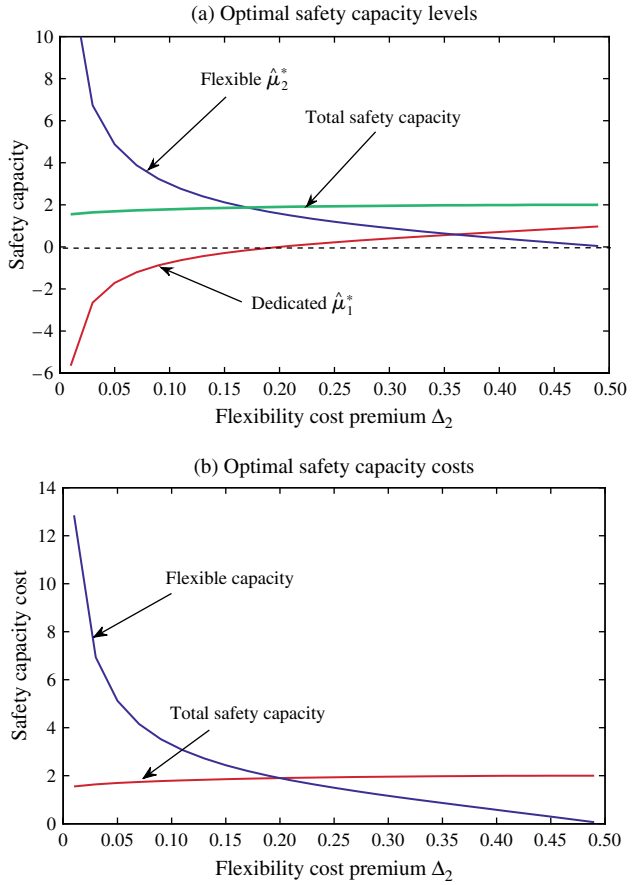
3.2. Discussion of Results: Amount and Level of Flexibility

All graphs and numerical results in this paper will normalize the scale factor $\sigma\sqrt{(hm)/c} = 1$ and the cost of the dedicated resource $c = 1$. The explicit characterization of the asymptotic solution yields some interesting insights. Figure 3(a) depicts the optimal safety capacities. Proposition 3 prescribes that it is *never* optimal to use any flexibility if the flexibility premium exceeds 50%, i.e., $\Delta_2 \geqslant 0.5$. As the flexibility premium decreases, it becomes optimal to use flexibility, and the corresponding flexible capacity increases as expected. When the premium falls below 20%, we obtain $\hat{\mu}_1^* < 0$, which implies that the optimal dedicated capacity is less than the nominal level λ, and thus the flexibility is used for maintaining the stability of the system as well.

Figure 3(b) shows how the investment cost in flexible and total capacity varies with the flexibility cost premium Δ_2. As expected, an increase in the premium leads to an increase in the total capacity cost and a decrease in the investment in flexible capacity. The latter entails lesser pooling benefits, and hence an increase in the total safety capacity needed as depicted in Figure 3(a). We observe that as the flexibility premium increases, the optimal flexible capacity decreases and is substituted by dedicated capacity. However, this substitution is not perfect: as shown in the figure, we oversubstitute and the total safety and, hence, the total capacity, increases as a function of Δ_2. Though similar sizing substitution effects have been observed (see, for example, Van Mieghem 1998), the benefit of our analysis is that we find these sizing results analytically, which cannot be done in newsvendor models.

The dependence of the prescription on the variability and holding cost is also worth pointing out. We can think of the solution $(m\lambda + \hat{\mu}_1^*\sqrt{\lambda}, \hat{\mu}_2^*\sqrt{\lambda})$ as the analog of a safety capacity refinement around the mean demand in a standard newsvendor problem with normal demand. Our safety capacity $(\hat{\mu}_1^*\sqrt{\lambda}, \hat{\mu}_2^*\sqrt{\lambda})$ is also proportional to the underlying standard deviation $\sigma\sqrt{\lambda}$. As the safety capacity cost is equal to the holding cost similar to the economic order quantity (EOQ) model, we also obtain that the safety capacities are proportional to $\sigma\sqrt{(hm)/c}$, in particular to the square root of the holding cost. Thus, as the variability in the system (or the holding cost) increases, one requires higher dedicated safety capacity $\hat{\mu}_1$ and higher flexible capacity $\hat{\mu}_2$.

Figure 3. The optimal capacity portfolio (top) and investment cost (bottom) as a function of the flexibility premium Δ_2.



4. Generalization to N Types

In this section, we generalize our analysis to symmetric processing systems of N customer types. As described in §2, the system can invest in a portfolio of level-k flexible resources ($1 \leqslant k \leqslant N$). As before, such systems are intractable, so we resort to an approximate analysis for large arrival rates λ that is asymptotically correct when $\lambda \to \infty$. We assume that an LQ policy is used to route jobs to different servers. Specifically, any flexible resource serves the type with the largest number of customers in the system among the types it can serve.

Let $Q^\lambda_{[\,]}(t) := (Q^\lambda_{[1]}(t), \ldots, Q^\lambda_{[N]}(t))$ be the order statistics for the number of customers of various types, where $Q^\lambda_{[1]}(t) \geqslant Q^\lambda_{[2]}(t) \geqslant \cdots \geqslant Q^\lambda_{[N]}(t)$. Under the LQ policy, the longest queue $Q^\lambda_{[1]}$ is served by all resources that can process it, and hence is processed at rate $\mu_1 + (N-1)\mu_2 + \cdots + (N-1)\mu_{N-1} + \mu_N$. Note that this rate is feasible only if the number of jobs in this queue exceeds the number of resources that can process it. Because our goal is an asymptotic analysis, the likelihood that the number of jobs is less than the number of resources is so small that we can ignore it. Now, consider type $[i]$ with $i > 1$. We can compute the number of level-k flexible resources that will serve this type in the following manner. A level-k flexible resource will serve type $[i]$ only if it has the longest queue-length among all types than can be handled by the resource. Thus, a level-k flexible resource will not serve type $[i]$ if $k > N - i + 1$. However, if $k \leqslant N - i + 1$, the level-k flexible resources for which type $[i]$ is the longest queue will serve it. This is simply the number obtained by selecting $k-1$ types from the N types removing the top i ranked types, i.e., $\binom{N-i}{k-1}$. Hence, the total processing rate for type $[i]$ equals $\sum_{k=1}^{N-i+1}\binom{N-i}{k-1}\mu_k$.

Proposition 5. *As $\lambda \to \infty$, if $Q^\lambda(0)/\sqrt{\lambda} \Rightarrow \hat{Q}(0)$, then $Q^\lambda_{[\,]}(\cdot)/\sqrt{\lambda} \Rightarrow \hat{Q}(\cdot)$, where $\hat{Q}$ is given by*

$$\hat{Q}_{[i]}(t) = \hat{Q}_{[i]}(0) - \frac{1}{m}\sum_{k=1}^{N-i+1}\binom{N-i}{k-1}\hat{\mu}_k t + \sigma\sqrt{2}B_i(t) - Y_{i-1}(t) + Y_i(t), \tag{14}$$

for $i = 1, \ldots, N$, where B_i are N standard independent Brownian motions, $Y_0 \equiv 0$, Y_i are nondecreasing continuous processes such that $Y_i(0) = 0$, and $\hat{Q}_{[1]}(t) \geqslant \hat{Q}_{[2]}(t) \geqslant \cdots \geqslant \hat{Q}_{[N]}(t) \geqslant 0$, $\int_0^t (\hat{Q}_{[i]}(s) - \hat{Q}_{[i+1]}(s))\,dY_i(s) = 0$, and $\int_0^t \hat{Q}_{[N]}(s)\,dY_N(s) = 0$ for all $t \geqslant 0$.

Defining $G_N = \{x \in \mathbb{R}^N_+ \colon x_1 \geqslant x_2 \geqslant \cdots \geqslant x_N\}$, we can characterize the steady-state distribution of the $\hat{Q}_{[\,]}$ process as follows.

Proposition 6. *The steady-state distribution of the process $\hat{Q}_{[\,]}(\cdot)$ on G_N has density*

$$\pi(x) = \alpha\prod_{i=1}^{N}\exp\left(-\left(\frac{\sum_{k=1}^{N-i+1}\binom{N-i}{k-1}\hat{\mu}_k}{\sigma^2 m}\right)x_i\right),$$

where

$$\alpha = \left(\int_{G_N}\prod_{i=1}^{N}\exp\left(-\left(\frac{\sum_{k=1}^{N-i+1}\binom{N-i}{k-1}\hat{\mu}_k}{\sigma^2 m}\right)x_i\right)dx\right)^{-1}$$

is the normalizing constant. Further, for $i = 1, \ldots, N$, we have

$$\mathbb{E}\,\hat{Q}_{[i]}(\infty) = \frac{N-i+1}{\sum_{k=1}^{N}\sum_{j=\max(i-1,k-1)}^{N-1}\binom{j}{k-1}\hat{\mu}_k}\sigma^2 m.$$

Proposition 6 allows us to express the diffusion-scale cost as a function of dedicated and flexible resources as

$$\hat{\Pi}(\hat{\mu}) = \sum_{i=1}^{N}\frac{N-i+1}{\sum_{k=1}^{N}\sum_{j=\max(i-1,k-1)}^{N-1}\binom{j}{k-1}\hat{\mu}_k}\sigma^2 hm + \sum_{k=1}^{N}\binom{N}{k}\hat{\mu}_k c(1+\Delta_k). \tag{15}$$

The diffusion-scale optimization problem is then

$$\min_{\{\hat{\mu}\colon \sum_k\binom{N}{k}\hat{\mu}_k > 0,\ \hat{\mu}_k \geqslant 0\ \forall k \geqslant 2\}} \hat{\Pi}(\hat{\mu}). \tag{16}$$

The formal optimality property similar to that in Proposition 4 then follows.

THEOREM 2. *The capacity portfolio* $\mu^* = (m\lambda, 0, \ldots, 0) + \hat{\mu}^*\sqrt{\lambda}$, *where* $\hat{\mu}^*$ *denotes an optimizer of* (16), *is asymptotically optimal for the optimization problem* (1) *in the sense that*

$$\lim_{\lambda\to\infty} \frac{\Pi^\lambda(\mu^*) - \Pi^{\lambda*}}{\sqrt{\lambda}} = 0. \tag{17}$$

The first-order conditions that characterize the optimal solution to the diffusion-scale optimization problem entail solving a polynomial of order $N+1$. Therefore, it follows that there is an explicit closed-form solution for the capacity portfolio only when the number of types $N \leqslant 3$. (Obviously, these conditions are easily solved numerically for any parameter values.) However, we can obtain following property of the solution $\hat{\mu}^*$ to the diffusion-scale optimization problem.

THEOREM 3 (ASYMPTOTIC OPTIMALITY OF INVESTING ONLY IN LEVELS 1 AND 2). *If the flexibility premiums satisfy* $\Delta_k/\Delta_2 \geqslant \sum_{j=2}^{k} 2/j$ *for* $k \geqslant 3$, *then any solution to the asymptotic optimization problem* $\min_{\hat{\mu}} \hat{\Pi}(\hat{\mu})$ *has* $\hat{\mu}_k^* = 0$ *for* $2 < k \leqslant N$, *that is, investing only in levels* 1 *and* 2 *is asymptotically optimal for such symmetric queueing systems.*

This result provides sufficient conditions on the flexibility premiums for the asymptotic optimality of investing only in levels 1 and 2. These conditions are only sufficient to ensure this optimality, and there may be other parameters at which investing only in levels 1 and 2 is asymptotically optimal. The necessary and sufficient conditions for this optimality can be computed analytically (as in Proposition EC.1), but are intricate and depend on N. Although Theorem 3 is a relaxation of these conditions, it provides a simple sufficient condition that is independent of N.

Figure 4 illustrates Theorem 3. If the flexibility premiums for level-$k > 2$ resources are above the threshold, then investing only in levels 1 and 2 is asymptotically optimal. However, if the flexibility premiums are below this threshold, then it may be optimal to invest in higher levels of flexibility. The figure also plots the linear flexibility premium curve, in which each level of flexibility incurs the same additional premium, to illustrate that this threshold is quite concave so that even with strong economies of scope it is sufficient to only use level-1 and level-2 flexible resources regardless of the number of customer types.

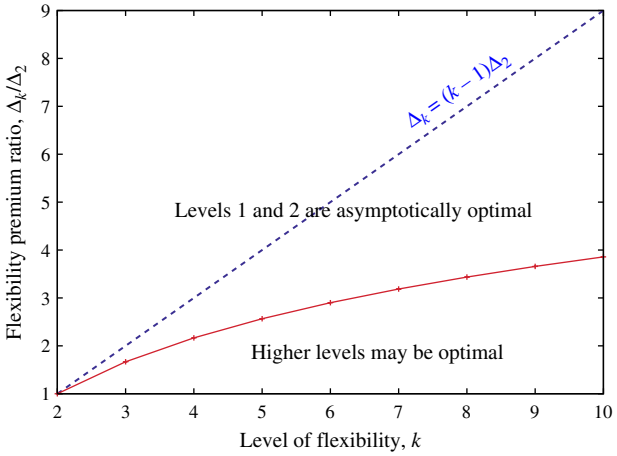
Further, we can characterize the maximum flexibility premium beyond which it is never optimal to invest in flexible resources for any N:

PROPOSITION 7. *For flexibility premiums* $\Delta_k \geqslant \sum_{j=2}^{k} 1/j$ *for* $k \geqslant 2$, *it is asymptotically optimal to only use dedicated capacity, i.e.,* $\hat{\mu}_k^* = 0$ *for all* $2 \leqslant k \leqslant N$.

Explicit solution for a three-type system. A three-type system has the following resources: three dedicated resources (level-1), three level-2 resources that can process any pair of types, and one fully flexible resource (level-3) that can process any type. Proposition EC.1 in Appendix EC.3 characterizes the exact asymptotically optimal capacity portfolio, and Figure 5 depicts the structure of this portfolio as a function of the flexibility premiums Δ_2 and $\Delta_3 - \Delta_2$ (the incremental premium of level-3 resources as compared with level-2 resources). The regions depicted in the figure *do not* depend on any other primitive data, and thus this figure is representative of the solution for any set of parameters.

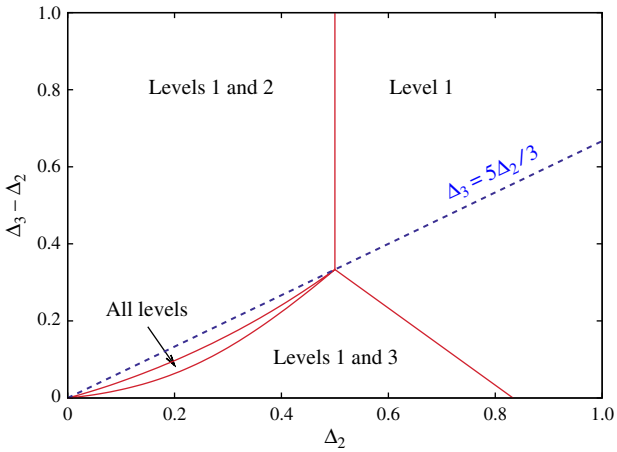
Notice that among all the flexible portfolios, investing in levels 1 and 2 is asymptotically optimal for the largest set

Figure 4. Investing in levels 1 and 2 only is asymptotically optimal for flexibility premiums above the thresholds computed in Theorem 3.



Note. This figure is derived from the diffusion limit and is independent of all system parameters.

Figure 5. The optimal capacity portfolio as a function of the flexibility premium.



Note. This characterization depends only on the flexibility premiums and is independent of all other system parameters.

of parameters. In this region, the proposition proves that the firm can achieve asymptotically optimal performance by using *only* dedicated and level-2 flexible resources and not using the fully flexible resource at all. This implies that the marginal benefit from having a fully flexible resource is less than its marginal cost when the firm can invest in level-2 flexible resources. Thus, this result proves that for suitable flexibility premiums, a tailored chaining configuration that utilizes dedicated and level-2 resources is asymptotically optimal for symmetric queueing systems with $N = 3$. Theorem 3 provides the following simple sufficient condition on the flexibility premiums for tailored chaining to be asymptotically optimal for this setting.

COROLLARY 1 (ASYMPTOTIC OPTIMALITY OF TAILORED CHAINING). *If the flexibility premiums are such that* $\Delta_3 \geqslant 5\Delta_2/3$, *then the asymptotically optimal flexibility portfolio with* $N = 3$ *never invests in the fully flexible resource, that is, tailored chaining is asymptotically optimal.*

As the flexibility premium $\Delta_3 - \Delta_2$ decreases to a level lower than Δ_2, the marginal cost of the fully flexible resource decreases, and the optimal portfolio invests in all three types of resources. This extreme capacity portfolio is optimal only for a small set of parameters and, as $\Delta_3 - \Delta_2$ decreases further, it becomes optimal to not invest in level-2 resources at all, and the optimal portfolio consists only of three dedicated and one fully flexible resources. Finally, note that for high-flexibility premiums, as expected, investing in flexibility is suboptimal. Specifically, if $\Delta_3 > 5/6$ and $\Delta_2 > 1/2$, investing in dedicated resources alone is asymptotically optimal.

5. Accuracy and Robustness of Results

In this section, we investigate the robustness of our results. In §5.1, we numerically investigate the accuracy of the asymptotically optimal flexibility portfolio over a wide range of arrival rates. In §5.2, we analytically compute the worst-case performance of tailored pairing when the flexibility premiums are below the thresholds of Theorem 3.

5.1. Accuracy of Capacity Prescriptions

To study the accuracy of the asymptotically optimal capacity prescriptions derived in the paper, we consider the case of $N = 2$ types and compare the capacity prescription presented in Proposition 4 with the optimal capacities derived via simulation and discrete search for a given arrival rate. Specifically, we consider Poisson arrivals with rates $\lambda = 1$, 5, 10, 25, 100, 400 and mean service time $m = 1$, unit dedicated capacity cost $c = 1$, and holding cost $h = 1$. Further, we implement the longest-queue-first policy in a nonpreemptive manner. To study the effect of variability in service-times, we study three different service-time distributions: deterministic, normal (standard deviation $= 0.25$, truncated), and exponential. In each case, we compare the optimal cost with the expected total cost of the system when operating with our capacity prescription. The optimal cost is derived via simulation and discrete search over a capacity grid for (μ_1, μ_2). For each capacity level in this grid, we used a simulation run length of 100,000 time units to estimate the expected queue length of the system. A grid search then allows us to compute the optimal total expected cost for $\Delta_2 \in (0, 0.5]$.

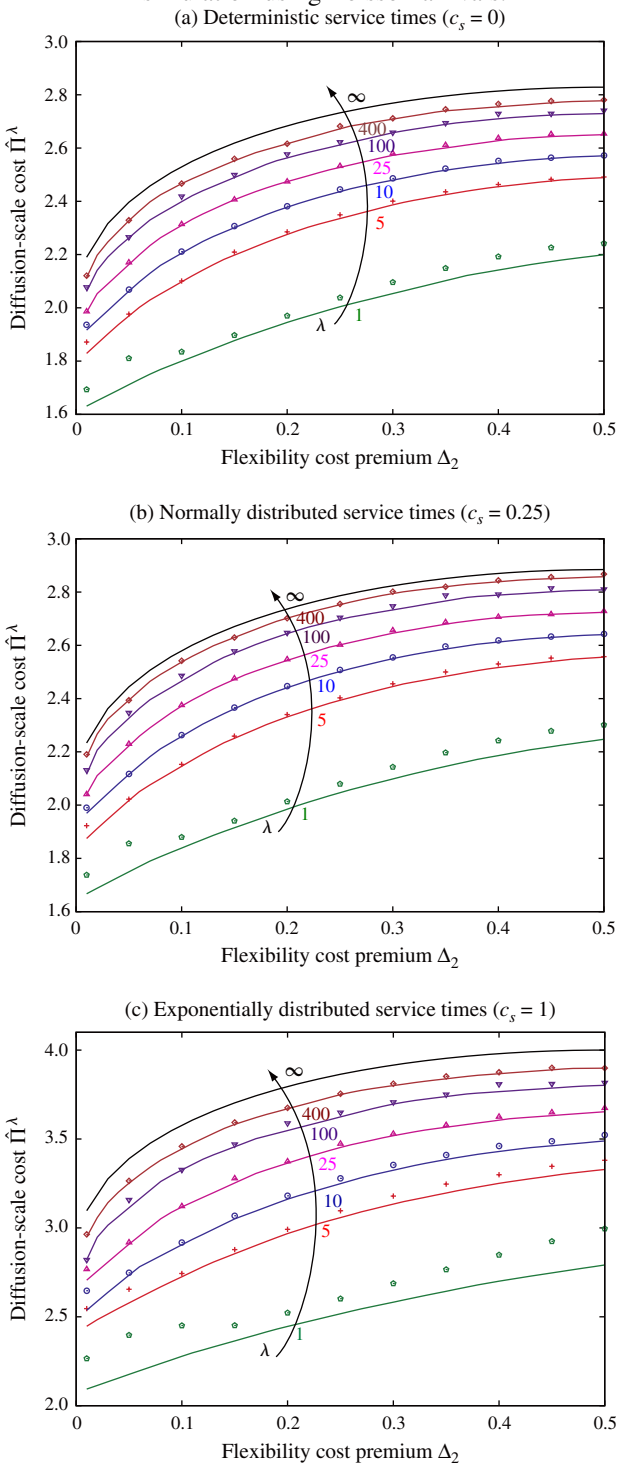
Figures 6(a)–6(c) show the diffusion-scale cost as a function of flexibility premium Δ_2. The markers depict the cost using the capacity prescription while the solid lines represent the optimal cost obtained via simulation. Observe three facts: First, the cost when using the capacity prescription is quite close to the optimal cost for all cases considered. Second, as expected, all simulated costs (both the optimal and the cost when using the prescription) converge to the asymptote $\hat{\Pi}^* = \hat{\Pi}(\hat{\mu}^*)$, which we have characterized analytically. Finally, total costs increase as variability increases from (a) to (b) to (c).

For normally distributed service times, Figure 7 plots the proportion of flexible capacity installed in the prescribed portfolio for the arrival rates $\lambda = 1, 5, 10, 15, 25, 100$ and 400. Note that as flexibility becomes costless, i.e., Δ_2 approaches 0, the prescribed portfolio invests only in flexible capacity. Further, the proportion of flexible capacity decreases as the arrival rate increases, which is consistent with our main result that the optimal capacity portfolio invests $O(\lambda)$ in dedicated resources and $O(\sqrt{\lambda})$ in flexible capacity (notice that for $\lambda = 1$ both are of the same order).

5.2. Worst-Case Suboptimality of Investing in Level-1 and Level-2 Flexible Resources

Theorem 3 gives us sufficient conditions for the optimality of investing in level-1 and level-2 flexible resources. Clearly, if higher levels of flexibility are cheap, it would be optimal to invest in them. In this section, we investigate the maximal suboptimality that can be incurred by investing only in level-1 and level-2 flexible resources. To do this, using the analytical expressions for the steady-state of the diffusion limit, we numerically compare the optimal tailored pairing configuration with the optimal tailored fully flexible solution (investing in level-1 and level-N) under the conservative assumption that the cost of the fully flexible resource is identical to that of the level-2 resource, i.e., $\Delta_N = \Delta_2$. This assumption yields the maximal suboptimality possible of the tailored pairing configuration. Figure 8 plots this optimality gap on the diffusion scale in percentage versus the number of types, N. For each N, the optimality gap is maximized over Δ_2, so that the plot is independent of *all* system parameters. We note that for small values of N, the optimality gap is very small and increases as N increases. However, the gap seems to asymptote below 20%. Thus, in the worst case, investing in level-1 and level-2 flexible resources would lead to a suboptimality of 20% on the diffusion scale. This analytical optimality gap is consistent with the observations in

Figure 6. The accuracy of the capacity prescriptions was investigated by comparing its simulated scaled cost (markers) to the optimal cost (solid lines) found through optimization by simulation using Poisson arrivals.



Sheikhzadeh et al. (1998), Jordan et al. (2004), Chou et al. (2010a,b). Noting that higher levels of flexibility indeed entail some premium, the actual suboptimality would typically be much lower.[6]

Figure 7. Proportion of flexible capacity in the prescribed portfolio as a function of the flexibility cost premium for different arrival rates.

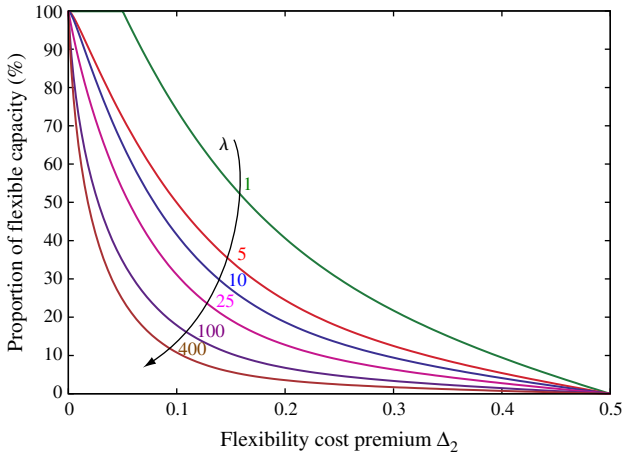
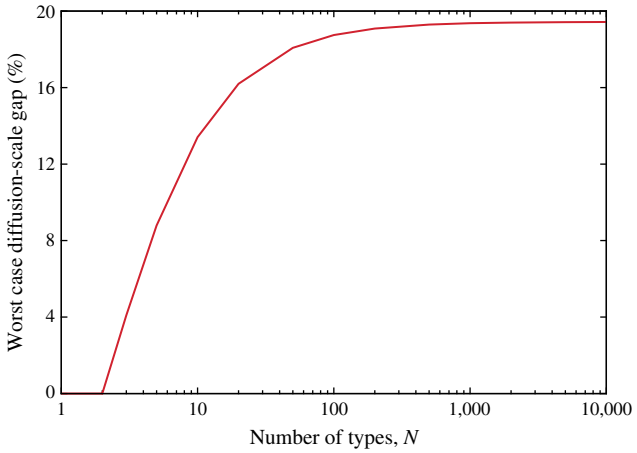


Figure 8. The worst-case suboptimality on the diffusion scale of investing in level-1 and level-2 flexible resources.



Note. The horizontal axis is plotted on a logarithmic scale.

6. Conclusion, Limitations, and Extensions

This paper studies the asymptotically optimal amount, level, and configuration of flexibility for symmetric queueing systems. Focusing on symmetric systems with linear costs, we analytically prove that the asymptotically optimal flexibility configuration invests a lot in dedicated resources, and a little in flexible resources. The literature has indicated that "a little flexibility can achieve almost all benefits of total flexibility" (Jordan and Graves 1995) in the sense that chained configurations of only level-2 flexible resources perform quite well. We find sufficient conditions on the cost of flexibility for the asymptotic optimality of investing in level-1 and level-2 flexible resources in symmetric queueing systems. We prove that these configurations are

asymptotically optimal even for fairly high economies of scope. Further, in the extreme case where additional levels of flexibility are costless, the maximum drop in performance (at the diffusion scale) in using these configurations is 20%.

To the best of our knowledge, this is the first analytic proof that a mix of dedicated resources with chained level-2 flexible resources is asymptotically optimal for symmetric queueing systems with $N \leqslant 3$. The main limitation of our analysis, however, is the assumption that capacity costs are linear in size. It is obvious that our results will break down with strong scale economies for which it is optimal to have fewer resources and often higher levels of flexibility (potentially even total flexibility) than our results predict. Investigating robustness to economies of scale requires a substantially different setup and is a future research topic.

From a methodological perspective, our analysis is based on Brownian approximations of a queueing system where the so-called complete resource-pooling condition is not satisfied at optimality. This leaves us with a multidimensional Brownian motion with discontinuous drifts. We analyze this process using a novel folding technique that studies the order statistics of the queue-length process and allows us to derive closed-form expressions for the expected queue-lengths, which in turn gives us a closed-form asymptotic characterization of the optimal resource capacities. Up until now, no closed-form expressions seem to exist, not even for simple static newsvendor models.

In this paper, we have assumed that capacity can be sized continuously by varying the service rate of a given portfolio of resources, which is the typical approach in capacity investment models. When capacity is indivisible or lumpy, however, capacity sizing is accomplished by varying the number of resources (each one with a fixed service rate) of a given level of flexibility. Our analysis does not apply to these settings and should be replaced by a many-server regime (see, for example, Halfin and Whitt 1981). This includes staffing in call centers, where, in addition to capacity being lumpy, multiple resources cannot pool their capacities to process an individual job. Here, the multiplicity of servers introduces other issues as well; for example, one needs to keep track of the type of each customer being processed by each server of each resource. This adds substantial complexity to the analysis and is left for potential future work. The following are two relevant papers that consider the problem of capacity planning in call centers to satisfy quality-of-service constraints: Wallace and Whitt (2005) develops a simulation-based iterative algorithm for staffing, and Gurvich and Whitt (2010) analytically derives asymptotically optimal capacity levels for a related problem.

Also note that we do not consider any constraints on the capacity portfolio. In practice, one might encounter constraints that prevent investing in a symmetric portfolio. For such configurations, the LQ policy may not perform well and alternate control policies may need to be considered. Such a situation may also occur even if there are no constraints on the capacity portfolio, but rather different classes have different holding costs (see, for instance, Saghafian et al. 2011).

Finally, whereas our model uses a holding cost criterion, it would be interesting to investigate a setup that minimizes capacity investment costs subject to quality-of-service constraints. Our characterization of the steady-state distribution of the queue-lengths allows us to compute the delay distribution (using a heavy-traffic version of Little's law) as a function of capacity. This is easily seen for the single-type system and may extend to N types. Another measure that would be worth investigating would be throughput (see, for instance, Ostolaza et al. 1990, Zavadlav et al. 1996, Andradóttir et al. 2001, Van Oyen et al. 2001, Hopp and Oyen 2004). Another alternative to our cost-based approach is in Iravani et al. (2005, 2011) which use structural and capacity flexibility methods, respectively, to rank different flexibility configurations.

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/opre.1120.1107.

Endnotes

1. In symmetric systems all model parameters are type independent: the arrival rates of all types are equal, and the capacity invested in resources at the same level of flexibility are equal. Hence, capacity decisions can only vary by flexibility level so that determining the capacity investment in $2^N - 1$ different resources reduces to optimizing N decision variables, one for each level of flexibility, to minimize the average holding and capacity cost rate.
2. We analytically derive the (almost logarithmic) sufficient flexibility cost frontier.
3. Kingman's bound implies that the expected steady-state time in queue is bounded above by the term $(c_a^2/\lambda^2 + (c_s^2 m^2)/(\mu_1^\lambda)^2)((\lambda\mu_1^\lambda)/(2(\mu_1^\lambda - m\lambda)))$, which is further bounded by $\sigma^2(\mu_1^\lambda/(\lambda(\mu_1^\lambda - m\lambda)))$. Thus, the expected steady-state number of jobs in the system is bounded above by $(\sigma^2(\mu_1^\lambda/(\mu_1^\lambda - m\lambda)) + 1)$.
4. Note that in this case, the first-order optimization, also known as fluid-scale optimization, is trivial and amounts to handling the base demand by the dedicated resources. However, this optimization ignores any queueing considerations, and thus we focus on the diffusion-scale optimization problem as is standard in asymptotic analysis of queueing systems (see, for instance, Chen and Yao 2001 and Whitt 2002 and references therein for more details).
5. The intuition behind this claim is as follows. The overall rate of departures from the system at any time t is given by $\sum_F \mu_F \mathbb{I}(\sum_{i\in F} Q_i^\lambda(t) > 0)$. Serving the longest queue first maximizes $\mathbb{I}(\sum_{i\in F} Q_i^\lambda(t) > 0)$ for all $t \geqslant 0$ over all scheduling policies. The number of departures from the system by time t equal $\int_0^t \sum_F \mu_F \mathbb{I}(\sum_{i\in F} Q_j^\lambda(t) > 0)\, dN_s$, where N_s is a unit rate Poisson process. Noting that this term is maximized by the LQ rule, using standard arguments to pass to the steady-state and taking expectations, it follows that the LQ has the highest aggregate departures among all scheduling policies. This translates to the LQ rule having the shortest aggregate queue-length. Thus, the LQ

rule should be optimal in our queuing system for any capacity portfolio.

6. Note that the unscaled costs would exhibit lesser suboptimality as compared with the diffusion-scale costs.

Acknowledgments

The authors thank the seminar participants at Columbia University and Northwestern University, especially Seyed Iravani and Ward Whitt, and the anonymous reviewers.

References

Andradóttir S, Ayhan H, Down DG (2001) Server assignment policies for maximizing the steady-state throughput of finite queueing systems. *Management Sci.* 47(10):1421–1439.

Bassamboo A, Randhawa RS, Van Mieghem JA (2010) Optimal flexibility configurations in newsvendor networks: Going beyond chaining and pairing. *Management Sci.* 56(8):1285–1303.

Chen H, Yao DD (2001) *Fundamentals of Queueing Networks* (Springer-Verlag, New York).

Chod J, Rudi N, Van Mieghem JA (2010) Operational flexibility and financial hedging: Complements or substitutes? *Management Sci.* 56(6):1030–1045.

Chou MC, Chua GA, Teo CP (2010a) On range and response: Dimensions of process flexibility. *Eur. J. Oper. Res.* 207(2):711–724.

Chou MC, Teo CP, Zheng H (2008) Process flexibility: Design, evaluation, and applications. *Flexible Services Manufacturing J.* 20(1):59–94.

Chou MC, Chua GA, Teo C-P, Zheng H (2010b) Design for process flexibility: Efficiency of the long chain and sparse structure. *Oper. Res.* 58(1):43–58.

Fine CH, Freund RM (1990) Optimal investment in product-flexible manufacturing capacity. *Management Sci.* 36(4):449–466.

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

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

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29:567–588.

Harrison JM, López MJ (1999) Heavy traffic resource pooling in parallel-server systems. *Queueing Systems* 33(4):339–368.

Hopp WJ, Oyen MP (2004) Agile workforce evaluation: A framework for cross-training and coordination. *IIE Trans.* 36(10):919–940.

Hopp WJ, Tekin E, Van Oyen MP (2004) Benefits of skill chaining in serial production lines with cross-trained workers. *Management Sci.* 50(1):83–98.

Iravani SM, Kolfal B, Van Oyen MP (2011) Capability flexibility: A decision support methodology for parallel service and manufacturing systems with flexible servers. *IIE Trans.* 43(5):363–382.

Iravani SMR, Van Oyen MP, Sims KT (2005) Structural flexibility: A new perspective on the design of manufacturing and service operations. *Management Sci.* 51(2):151–166.

Jordan WC, Graves SC (1995) Principles on the benefits of manufacturing process flexibility. *Management Sci.* 41(4):577–594.

Jordan WC, Inman RR, Blumenfeld DE (2004) Chained cross-training of workers for robust performance. *IIE Trans.* 36(10):953–967.

Kingman JFC (1962) Some inequalities for the queue GI/G/1. *Biometrika* 49:315–324.

Kleinrock L (1976) *Queueing Systems: Volume 2: Computer Applications* (John Wiley & Sons, New York).

Menich R, Serfozo RF (1991) Optimality of routing and servicing in dependent parallel processing systems. *Queueing Systems* 9(4):403–418.

Neely MJ, Modiano E (2005) Convexity in queues with general inputs. *IEEE Trans. Inform. Theory* 51(2):706–714.

Ostolaza J, McClain J, Thomas J (1990) The use of dynamic (state-dependent) assembly-line balancing to improve throughput. *J. Manufacturing Oper. Management* 3(2):105–133.

Saghafian S, Van Oyen MP, Kolfal B (2011) The "W" network and the dynamic control of unreliable flexible servers. *IIE Trans.* 12(43):893–907.

Sethi AK, Sethi SP (1990) Flexibility in manufacturing: A survey. *Internat. J. Flexible Manufacturing Systems* 2(4):289–328.

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

Van Mieghem JA (1998) Investment strategies for flexible resources. *Management Sci.* 44(8):1071–1078.

Van Mieghem JA (2003) Due date scheduling: Asymptotic optimality of generalized longest queue and generalized largest delay rules. *Oper. Res.* 51(1):113–122.

Van Mieghem JA (2008) *Operations Strategy: Principles and Practice* (Dynamic Ideas, Charlestown, MA).

Van Oyen MP, Gel EGS, Hopp WJ (2001) Performance opportunity for workforce agility in collaborative and noncollaborative work systems. *IIE Trans.* 33(9):761–777.

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

Whitt W (2002) *Stochastic-Process Limits* (Springer, New York).

Williams RJ (1987) Reflected Brownian motion with skew symmetric data in a polyhedral domain. *Probab. Theory Related Fields* 75: 459–485.

Zavadlav E, McClain JO, Thomas LJ (1996) Self-buffering, self-balancing, self-flushing production lines. *Management Sci.* 42(8):1151–1164.

Zheng YS, Zipkin P (1990) A queueing model to analyze the value of centralized inventory information. *Oper. Res.* 38(2):296–307.

Zipkin PH (1995) Performance analysis of a multi-item production-inventory system under alternative policies. *Management Sci.* 41(4):690–703.

Achal Bassamboo is a professor of managerial economics and decision sciences in the Kellogg School of Management at Northwestern University. His research focuses on using stochastic models to manage service operations, flexibility in service and production systems, and strategic information sharing in services and retail.

Ramandeep S. Randhawa is an associate professor in the Information and Operations Management Department in the Marshall School of Business at the University of Southern California. His research interests broadly lie in the areas of service management and revenue management. His current research focuses on designing flexible service systems, managing service operations under parameter uncertainty, studying the role of reputational incentives in decision making, and analyzing strategic customer behavior in retail settings.

Jan A. Van Mieghem is the Harold L. Stuart Professor of Managerial Economics and Professor of Operations Management at the Kellogg School of Management at Northwestern University, where he has served as senior associate dean and as department chair. His research and teaching focus on the strategic organization and tactical execution of product, service, and supply chain operations.