

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

Achal Bassamboo, Ramandeep S. Randhawa and Jan A. Van Mieghem\*

*Submitted:* June 21, 2009, *Revised:* November 3, 2010, October 24, 2011

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  $\lambda \rightarrow \infty$ . 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 in 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. We numerically investigate the accuracy and robustness of our results.

---

## 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

---

\*Bassamboo and Van Mieghem are at Northwestern University's Kellogg School of Management; Randhawa is at University of Southern California's Marshall School of Business. The authors thank the seminar participants at Columbia and Northwestern University, especially Seyed Iravani and Ward Whitt, and the anonymous reviewers.

extant literature on flexibility refers to the ability of a resource to process multiple types of products as *mix-* (Chod et al., 2008), *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 2-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 2-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  $k \in \{1, 2, \dots, N\}$  different product types. Then there are  $\binom{N}{k} = \frac{N!}{(N-k)!k!}$  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.

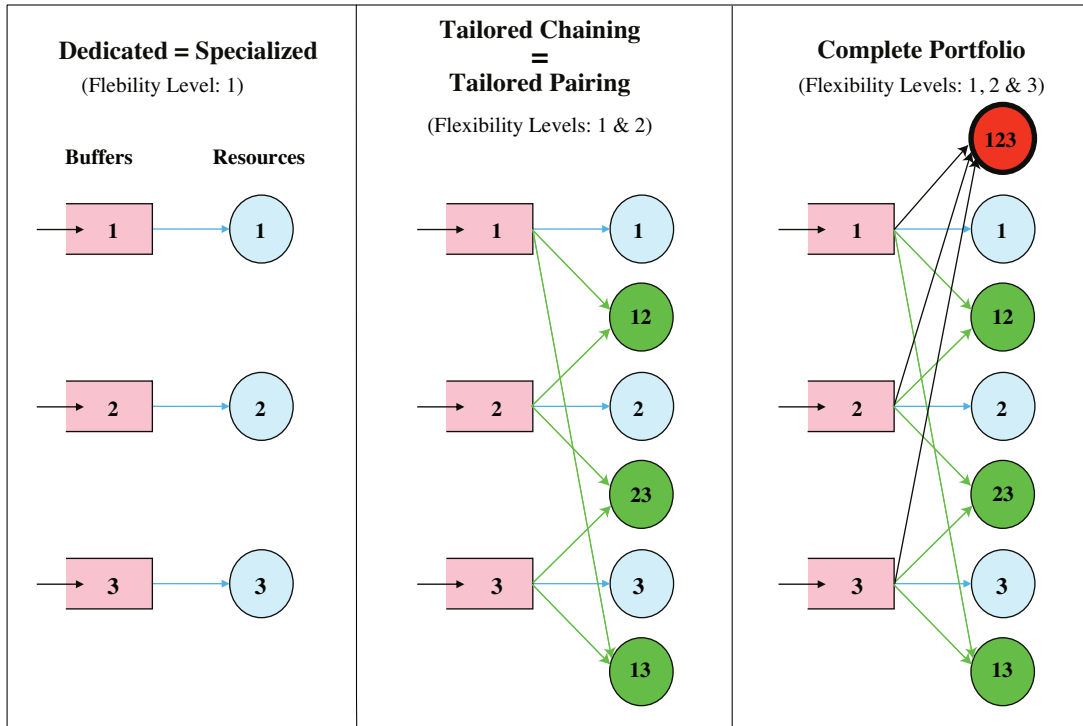
In their seminal paper, Jordan and Graves (1995) showed that “a little flexibility can achieve almost all the benefits of total flexibility” by using *only* level-2 flexible resources in a special configuration called chaining. Imagine a graph where product types are represented by rectangles and resources 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 pre-fixed 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 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 multi-stage 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.

While the chaining literature considered all servers capable of performing two tasks (i.e., all are level-2 flexible), this paper considers a configuration where only a small fraction of the servers is bi-flexible and the majority is specialized. In addition, we show that the bi-flexible capacity is proportional to the square root of the demand, playing a similar role to safety stock in inventory systems.

To formulate our contributions more precisely, 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. While our system is not amenable to exact analysis, we characterize analytically



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

the optimal amount, level, and configuration of flexibility for high-volume systems where the arrival rate  $\lambda \rightarrow \infty$ . 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, in contrast to chaining of all bi-flexible servers, “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<sup>1</sup> with mild economies of scope<sup>2</sup>, we prove that level-2 flexibility is all that is needed. 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

<sup>1</sup> 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.

<sup>2</sup> We analytically derive the (almost logarithmic) sufficient flexibility cost frontier.

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 while 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. Yet 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 chaining is asymptotically optimal for  $N = 3$  in a queuing setting, which differs from the newsvendor setting studied by Jordan and Graves (1995). Indeed, chaining and 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 has been the natural assumption in the growing literature on flexible queuing networks in heavy traffic and obviously leads to excellent waiting time performance; see for example Van Mieghem (1995), Harrison (1998), Harrison and Lopez (1999), Williams (2000), Stolyar (2004), Mandelbaum and Stolyar (2004), Ata and Kumar (2005), and references therein. 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 multi-dimensional reflected Brownian motion with state-dependent drift (arising from the longest queue scheduling). In other words, while 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  $i = 1, 2, \dots, N$  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.) Let  $\sigma_a^\lambda$  denote the standard deviation of the inter-arrival times. Each arriving job has a service requirement that is independent and identically distributed across all the customers with mean  $m$  and variance  $\sigma_s^2$ . The coefficient of variation of service times is denoted by  $c_s = \sigma_s/m$ , while that of the inter-arrival times is  $c_a = \lambda\sigma_a^\lambda$ . We assume that  $c_a$  is a constant, independent of the rate  $\lambda$ , 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  $\mu_1^\lambda$  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  $\binom{p}{q} = \frac{p!}{(p-q)!q!}$  if  $p \geq q$ , 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  $\mu_k^\lambda$ . (Note that capacities scale the actual average service time, i.e., if a service rate of  $\mu$  is allocated to a job, its average service time is  $m/\mu$  and

its variance is  $\sigma_s^2/\mu^2$ .) 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 capacity costs are linear in size. The cost rate of capacity size  $\mu_k$  of a level- $k$  flexible resource is  $c_k\mu_k^\lambda$ , where  $c_k = c(1 + \Delta_k)$  with  $\Delta_k$  denoting the flexibility premium for level- $k$  flexible resources and we have  $\Delta_k \geq \Delta_{k-1}$  for  $k \geq 2$  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  $\mu^\lambda = (\mu_1^\lambda, \mu_2^\lambda, \dots, \mu_N^\lambda)$  as<sup>3</sup>

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

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

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

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:  $\Pi^\lambda(\mu^{\lambda*}) \leq \min_{\mu_1^\lambda \geq 0} \Pi^\lambda(\mu_1^\lambda, 0, \dots, 0)$ . Using only dedicated servers results in  $N$  independent  $GI/G/1$  queues so that

$$\Pi^\lambda(\mu_1^\lambda, 0, \dots, 0) = N \left( h\mathbb{E}Q_1^\lambda + c\mu_1^\lambda \right) \leq N \left( h \left[ \sigma^2 \frac{\mu_1^\lambda}{\mu_1^\lambda - m\lambda} + 1 \right] + c\mu_1^\lambda \right),$$

using Kingman's bound (cf. Kingman, 1962).<sup>4</sup> The right hand side is convex in  $\mu_1^\lambda$  and

<sup>3</sup> Note that without loss of generality, we could assume  $h = 1$  and rescale the cost of capacity  $c_k$  appropriately.

<sup>4</sup> Kingman's bound implies that the expected steady-state time in queue is bounded above by the term  $\left( \frac{c^2}{\lambda^2} + \frac{c_s^2 m^2}{(\mu_1^\lambda)^2} \right) \frac{\lambda \mu_1^\lambda}{2(\mu_1^\lambda - m\lambda)}$ , which is further bounded by  $\sigma^2 \frac{\mu_1^\lambda}{\lambda(\mu_1^\lambda - m\lambda)}$ . Thus, the expected steady-state number of jobs in the system is bounded above by  $\left( \sigma^2 \frac{\mu_1^\lambda}{\mu_1^\lambda - m\lambda} + 1 \right)$ .

reaches a minimum at  $\tilde{\mu}_1^\lambda = m\lambda + \sigma\sqrt{\frac{h}{c}m\lambda}$ , which yields an exact upper bound:  $\Pi^\lambda(\mu^{\lambda*}) \leq \min_{\mu_1^\lambda \geq 0} \Pi^\lambda(\mu_1^\lambda, 0, \dots, 0) \leq \bar{\Pi}^\lambda + Nh(\sigma^2 + 1)$ , where

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

The upper bound also bounds the capacity cost and directly shows how the optimal capacities depend on the volume  $\lambda$ , which is key to our analysis:  $\mu^{\lambda*}$  cannot grow faster than a term proportional to  $\lambda$  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  $c$ . Hence, the 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 multi-queue, multi-server system. In heavy traffic, the Kingman bound is tight and, using (2) for a single queue with arrival rate  $N\lambda$ , yields as an asymptotic lower bound:  $\Pi^{\lambda*} \geq \underline{\Pi}^\lambda + o(\sqrt{\lambda})$ , where

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

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

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

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

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

where

$$\tilde{\mu}_1^{\lambda*} = m\lambda + \hat{\mu}_1\sqrt{\lambda}, \quad \text{and} \quad (5)$$

$$\tilde{\mu}_k^{\lambda*} = \hat{\mu}_k\sqrt{\lambda} \quad \text{for } k \geq 2, \quad (6)$$

for some  $\hat{\mu}_1, \dots, \hat{\mu}_N \in \mathbb{R}$  with  $\hat{\mu}_k \geq 0$  for  $k \geq 2$  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  $\lambda$ . 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  $\underline{\Pi}^\lambda = \overline{\Pi}^\lambda$  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  $\lambda m$  that matches the average arriving workload plus safety capacity  $\sigma\sqrt{\frac{hm\lambda}{c}}$  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  $m + \sigma\sqrt{\frac{hm}{c\lambda}}$ , where the safety capacity per unit decreases in  $\lambda$ , 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  $\lambda \rightarrow \infty$ .) 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  $(m\lambda + \hat{\mu}_1\sqrt{\lambda}, \hat{\mu}_2\sqrt{\lambda}, \dots, \hat{\mu}_N\sqrt{\lambda})$  to characterize an approximate solution to (1) for large volume systems, where  $\hat{\mu} \in M := \{\hat{\mu} : \hat{\mu}_k \geq 0 \text{ for } k \geq 2 \text{ and } \sum_{k=1}^N \binom{N}{k} \hat{\mu}_k > 0\}$ . 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  $\mu^\lambda$ , i.e.,  $N\lambda < \sum_{k=1}^N \binom{N}{k} \frac{\mu_k^\lambda}{m}$ . 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:

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

Although we can solve this optimization problem for any finite  $\lambda$  through simulation, to derive structural insights, we will consider an analytical asymptotic analysis that is accurate when the arrival rate  $\lambda \rightarrow \infty$ . 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  $(m\lambda + \hat{\mu}_1^*\sqrt{\lambda}, \hat{\mu}_2^*\sqrt{\lambda}, \dots, \hat{\mu}_N^*\sqrt{\lambda})$  as our approximate solution to (1) for a system with (finite) arrival rate  $\lambda$ .

To illustrate our mode of analysis, we begin by considering the  $N = 2$  type setting. In particular, we will demonstrate the novel 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 Section 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  $\sigma$ -field, we say that as  $n \rightarrow \infty$ ,  $P^n \Rightarrow P$ , i.e.,  $P^n$  weakly converges to  $P$ , if and only if  $\int_A f dP^n \rightarrow \int_A f dP$  for all bounded, continuous real-valued functions  $f$  on  $A$ . Further, if  $X^n$  and  $X$  are random elements of this space such that  $P^n$  and  $P$  are the probability measures associated with  $X^n$  and  $X$  respectively, then  $X^n \Rightarrow X$  if and only if  $P^n \Rightarrow P$ .

*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. That is, if the portfolio invests in a level- $k$  flexible resource, then it invests in all  $\binom{N}{k}$  level- $k$  resources, and we do not optimize further on the number of level- $k$  resources that should be invested in. Given the symmetry in the problem parameters, one expects such a symmetric portfolio to be optimal. That is, we expect identical capacity investments in all resources at the same level of flexibility. 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). While these results suggest that convexity should extend to our setting, we have not been able to prove this conjecture in general. Though such a proof is a worthy endeavor in itself, it is beyond the scope of this paper. Hence, we focus on symmetric capacity portfolios which 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 Section 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. In addition to capacity levels, we need to prescribe a routing rule that determines the order in which jobs are allocated to the different resources. 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.<sup>5</sup> 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

---

<sup>5</sup> 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  $t \geq 0$  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 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.

the multi-dimensional 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  $O(\sqrt{\lambda})$ ) than 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  $\lambda \rightarrow \infty$ , if  $\frac{Q^\lambda(0)}{\sqrt{\lambda}} \Rightarrow \hat{Q}(0)$ , then  $\frac{Q^\lambda(\cdot)}{\sqrt{\lambda}} \Rightarrow \hat{Q}(\cdot)$ , where*

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

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

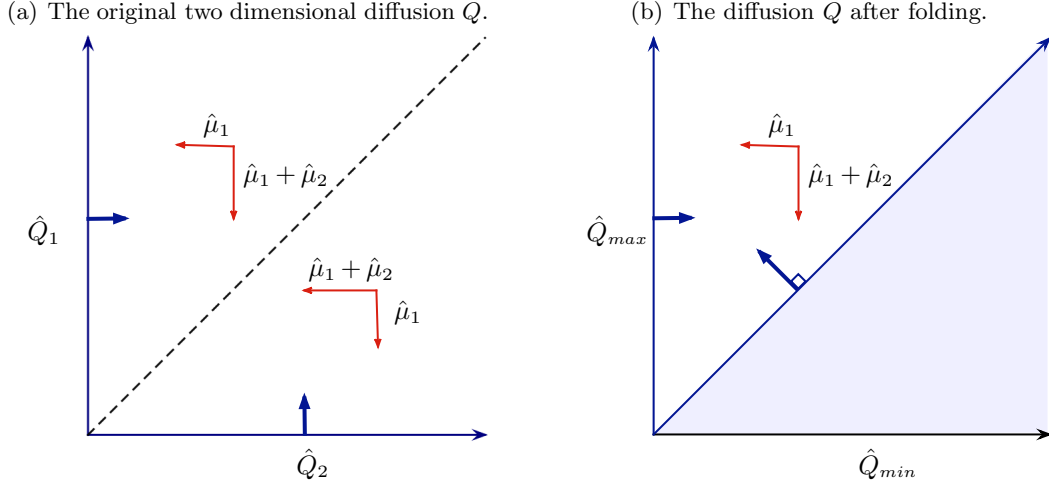
The limiting diffusion characterized in (8) is not directly amenable to analysis because the drift of the reflected Brownian motion  $(\hat{Q}_1, \hat{Q}_2)$  is not continuous. This discontinuity stems from the LQ routing policy under which the flexible resource serves the longer queue in a preemptive 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  $\lambda \rightarrow \infty$ , if  $\frac{Q^\lambda(0)}{\sqrt{\lambda}} \Rightarrow \hat{Q}(0)$ , then  $\left(\frac{Q_{max}^\lambda(\cdot)}{\sqrt{\lambda}}, \frac{Q_{min}^\lambda(\cdot)}{\sqrt{\lambda}}\right) \Rightarrow \hat{Q}(\cdot)$ , 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 non-decreasing continuous processes such that  $Y_1(0) = Y_2(0) = 0$ , and  $Q_{max}(t) \geq Q_{min}(t) \geq 0$ ,  $\int_0^t (\hat{Q}_{max}(s) - \hat{Q}_{min}(s)) dY_1(s) = 0$  and  $\int_0^t \hat{Q}_{min}(s) dY_2(s) = 0$  for all  $t \geq 0$ .



**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).

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  $G_2 = \{(x, y) \in \mathbb{R}_+^2 : x \geq y\}$ , 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*

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

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

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):

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

The following proposition presents the results.

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

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

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

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

Notice that  $(\hat{\mu}_1^*, \hat{\mu}_2^*) \frac{1}{\sigma} \sqrt{\frac{c}{hm}}$  depends only on the flexibility premium  $\Delta_2$ . Hence, for a fixed  $\Delta_2$  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  $\lambda$  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

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

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.<sup>6</sup>

### 3.2. Discussion of results: amount and level of flexibility

All graphs and numerical results in this paper will normalize the scale factor  $\sigma \sqrt{\frac{hm}{c}} = 1$  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.,  $\Delta_2 \geq 0.5$ .

<sup>6</sup> Note that the absolute optimality gap is  $o(\sqrt{\lambda})$ , and hence may in fact be increasing with  $\lambda$ .

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  $\lambda$ , and thus the flexibility is used for maintaining the stability of the system as well.

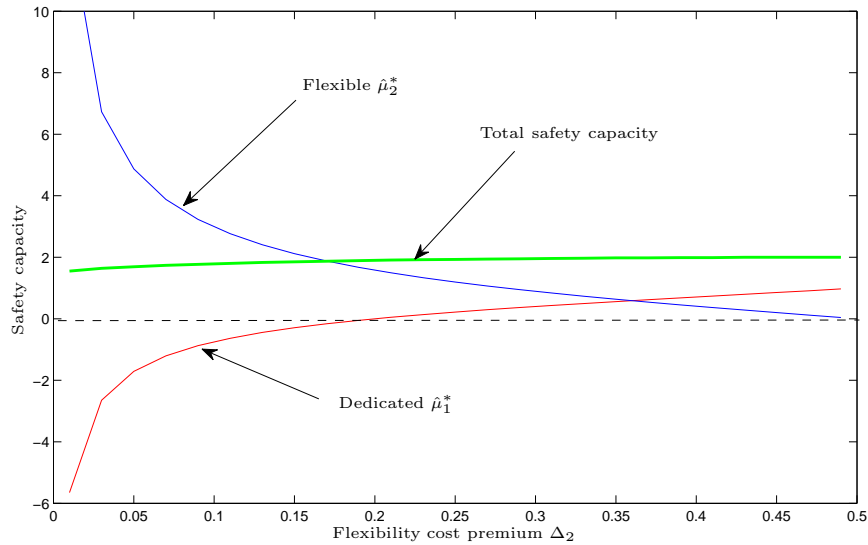
Figure 3(b) shows how the investment cost in flexible and total capacity varies with the flexibility cost premium  $\Delta_2$ . As expected, an increase in the premium leads to an increase in the total capacity cost and a decrease in the investment in flexible capacity. The latter entails lesser pooling 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 over-substitute and the total safety and, hence, the total capacity increases as a function of  $\Delta_2$ . 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  $\sigma\sqrt{\frac{hm}{c}}$ , 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  $\hat{\mu}_2$ .

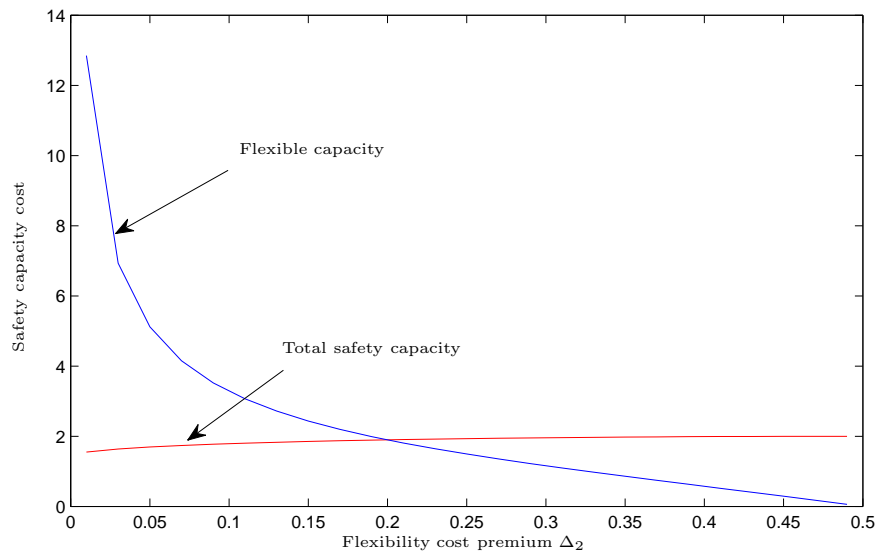
#### 4. Generalization to $N$ types

In this section, we generalize our analysis to symmetric processing systems of  $N$  customer types. As described in Section 2, the system can invest in a portfolio of level- $k$  flexible resources ( $1 \leq k \leq N$ ). As before, such systems are intractable so we resort to an approximate analysis for large arrival

(a) Optimal safety capacity levels.



(b) Optimal safety capacity costs.



**Figure 3** The optimal capacity portfolio (top) and investment cost (bottom) as a function of the flexibility premium  $\Delta_2$ .

rates  $\lambda$  that is asymptotically correct when  $\lambda \rightarrow \infty$ . 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  $Q_{[1]}^\lambda(t) := (Q_{[1]}^\lambda(t), \dots, Q_{[N]}^\lambda(t))$  be the order statistics for the number of customers of various types, where  $Q_{[1]}^\lambda(t) \geq Q_{[2]}^\lambda(t) \geq \dots \geq Q_{[N]}^\lambda(t)$ . Under the LQ policy, the longest queue  $Q_{[1]}^\lambda$  is served

by all resources that can process it, and hence is processed at rate  $\mu_1 + (N-1)\mu_2 + \dots + (N-1)\mu_{N-1} + \mu_N$ . Note that this rate is feasible only if the number of jobs in this queue exceeds the number of resources that can process it. As 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  $k \leq N - i + 1$ , 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  $\lambda \rightarrow \infty$ , if  $\frac{Q^\lambda(0)}{\sqrt{\lambda}} \Rightarrow \hat{Q}(0)$ , then  $\frac{Q^\lambda(\cdot)}{\sqrt{\lambda}} \Rightarrow \hat{Q}(\cdot)$ , where  $\hat{Q}$  is given by*

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

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

Defining  $G_N = \{x \in \mathbb{R}_+^N : x_1 \geq x_2 \geq \dots \geq x_N\}$ , we can characterize the steady-state distribution of the  $\hat{Q}_{[1]}$  process as follows.

PROPOSITION 6. *The steady-state distribution of the process  $\hat{Q}_{[1]}(\cdot)$  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  $i = 1, \dots, N$ , 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

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

The diffusion-scale optimization problem is then

$$\min_{\{\hat{\mu}: \sum_k \binom{N}{k} \hat{\mu}_k > 0, \hat{\mu}_k \geq 0 \forall k \geq 2\}} \hat{\Pi}(\hat{\mu}). \quad (16)$$

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

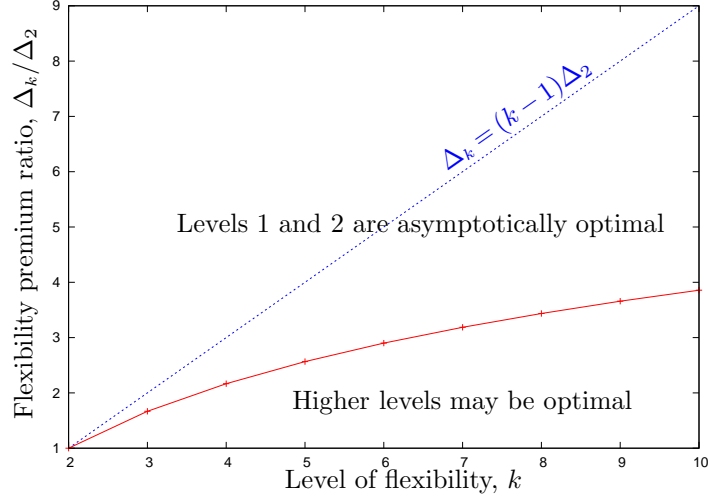
**THEOREM 2.** *The capacity portfolio  $\mu^* = (m\lambda, 0, \dots, 0) + \hat{\mu}^* \sqrt{\lambda}$ , where  $\hat{\mu}^*$  denotes an optimizer of (16), is asymptotically optimal for the optimization problem (1) in the sense that*

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

The first order conditions that characterize the optimal solution to the diffusion-scale optimization problem entail solving a polynomial of order  $N + 1$ . So, it follows that there is an explicit closed form solution for the capacity portfolio only when the number of types  $N \leq 3$ . (Obviously, these conditions are easily solved numerically for any parameter values.) Yet, 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  $\Delta_k/\Delta_2 \geq \sum_{j=2}^k 2/j$  for  $k \geq 3$ , then any solution to the asymptotic optimization problem  $\min_{\hat{\mu}} \hat{\Pi}(\hat{\mu})$  has  $\hat{\mu}_k^* = 0$  for  $2 < k \leq N$ , 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



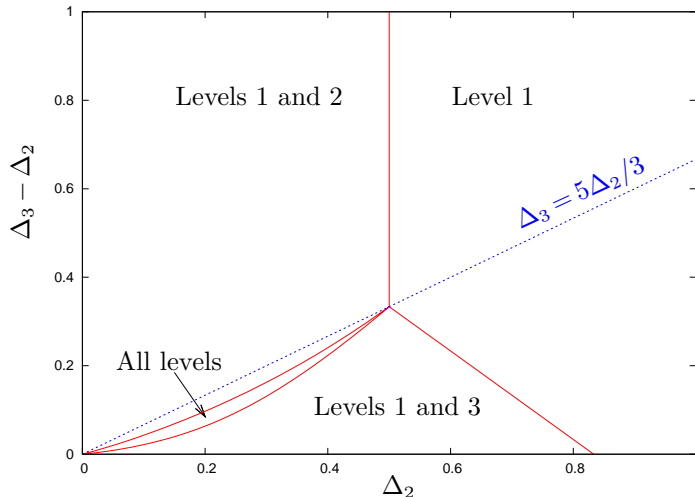
**Figure 4** Investing in levels 1 and 2 only is asymptotically optimal for flexibility premiums above the thresholds computed in Theorem 3, which are depicted by the red curve in the plot. The juxtaposition with the blue line illustrates how concave this frontier is. Note that this figure is derived from the diffusion limit and is independent of all system parameters.

(as in Proposition EC.1), but are intricate and depend on  $N$ . While 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 (the red curve), 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  $\Delta_k \geq \sum_{j=2}^k 1/j$  for  $k \geq 2$ , it is asymptotically optimal to only use dedicated capacity, i.e.,  $\hat{\mu}_k^* = 0$  for all  $2 \leq k \leq N$ .*



**Figure 5** The optimal capacity portfolio as a function of the flexibility premium. This characterization depends only on the flexibility premiums and is independent of all other system parameters.

*Explicit solution for a 3-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  $\Delta_2$  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 optimal for the largest set 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 (in a chaining configuration) 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, chaining of level-2 resources (as suggested by Jordan and Graves, 1995) together with dedicated resources is the asymptotically optimal flexibility configuration for symmetric queueing systems with  $N = 3$ . Theorem 3 provides the following simple sufficient condition on the flexibility premiums for tailored

chaining to be optimal for this setting.

**COROLLARY 1 (Asymptotic optimality of tailored chaining).** *If the flexibility premiums are such that  $\Delta_3 \geq 5\Delta_2/3$ , then the asymptotically optimal flexibility portfolio with  $N = 3$  never invests in the fully flexible resource, that is, tailored chaining is asymptotically optimal.*

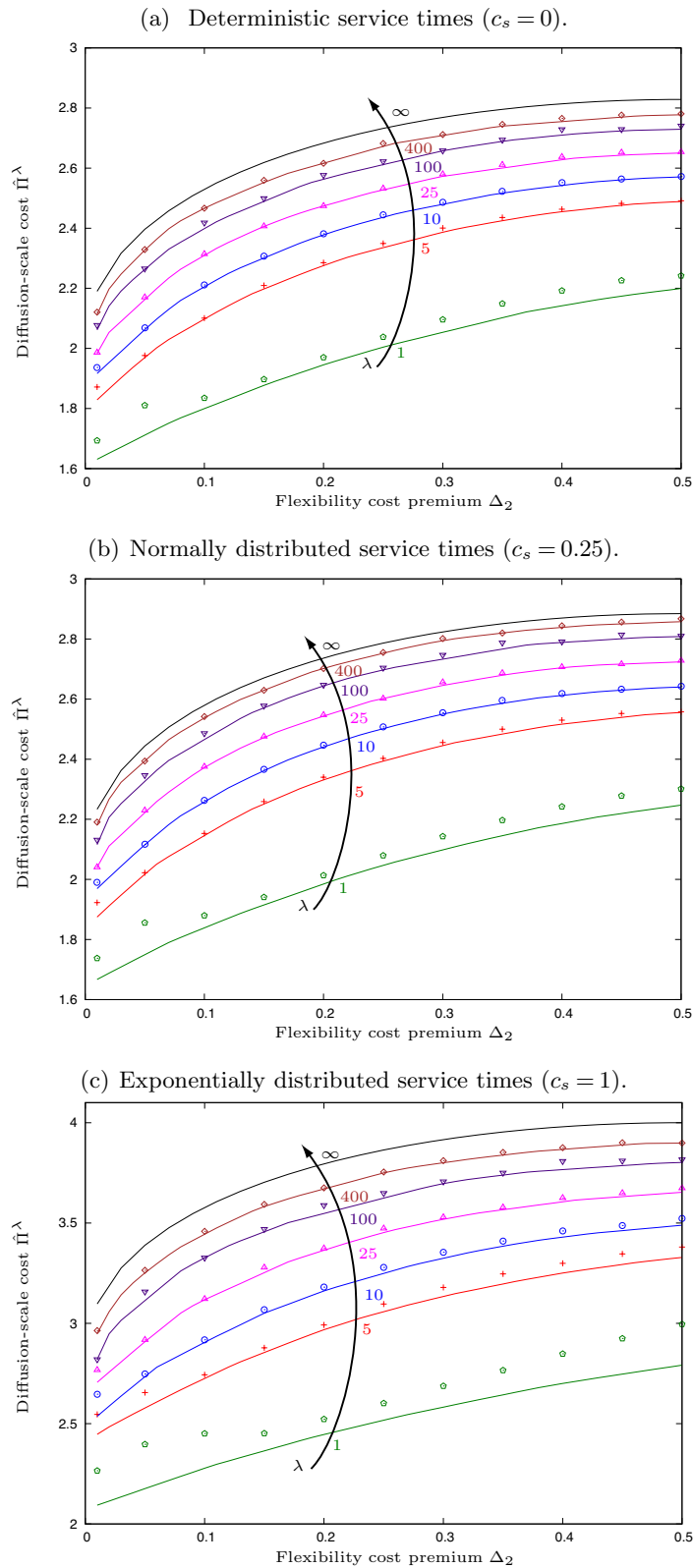
As the flexibility premium  $\Delta_3 - \Delta_2$  decreases to a level lower than  $\Delta_2$ , the marginal cost of the fully flexible resource decreases, and the optimal portfolio invests in all three types of resources. This extreme capacity portfolio is optimal only for a small set of parameters 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 sub-optimal. 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

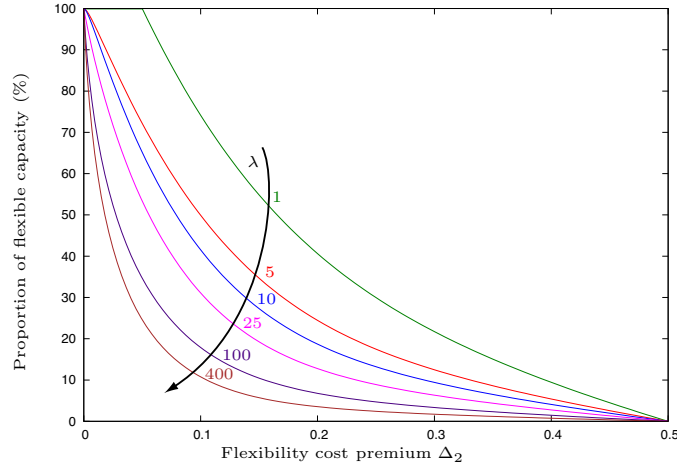
In this section, we investigate the robustness of our results. In Section 5.1, we numerically investigate the accuracy of the asymptotically optimal flexibility portfolio over a wide range of arrival rates. In Section 5.2, we analytically compute the worst case performance of tailored pairing when the 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 non-preemptive manner. To study the effect of variability in service times, we study three different service time distributions: deterministic, normal (standard deviation=0.25), and exponential. In each case, we compare the optimal cost with the expected total cost of the



**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.

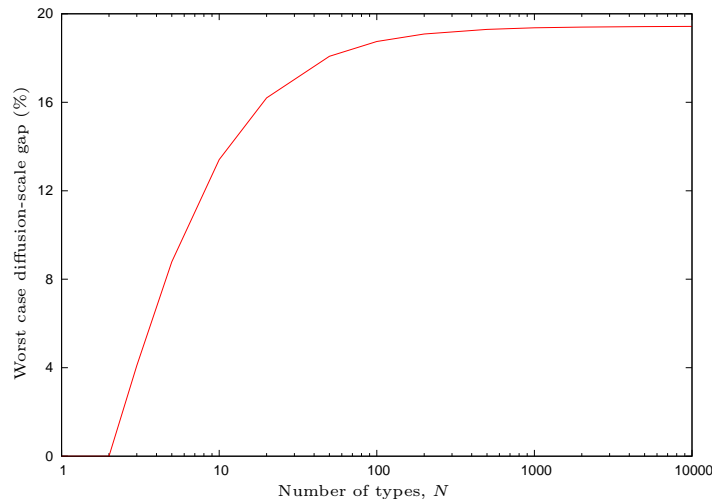


**Figure 7** Proportion of flexible capacity in the prescribed portfolio as a function of the flexibility cost premium for different arrival rates.

system when operating with our capacity prescription. The optimal cost is derived via simulation and discrete search over a capacity grid for  $(\mu_1, \mu_2)$ . 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  $\Delta_2$ . 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.,  $\Delta_2$  approaches zero, 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).



**Figure 8** The worst-case suboptimality on the diffusion scale of investing in level-1 and 2 flexible resources. The horizontal axis is plotted on a logarithmic scale.

## 5.2. Worst case suboptimality of investing in level-1 and 2 flexible resources

Theorem 3 gives us sufficient conditions for the optimality of investing in level-1 and 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 sub-optimality that can be incurred by investing only in level-1 and 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 sub-optimality 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  $\Delta_2$ , 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 2 flexible resources would lead to a suboptimality of 20% on the diffusion-scale. This analytical optimality gap is consistent with the observations in Sheikhzadeh et al. (1998), Jordan et al. (2004) and Chou et al. (2010). Noting that higher levels of flexibility indeed entail some premium, the actual suboptimality would typically be much lower.

## 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 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  $N \leq 3$ . 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 multi-dimensional 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 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, Saghaian et al., 2011).

Finally, while 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)).

## References

- Andradóttir, S., Ayhan, H. and Down, D. (2001), ‘Server assignment policies for maximizing the steady-state throughput of finite queueing systems’, *Management Science* **47**(10), 1421–1439.
- Ata, B. and Kumar, S. (2005), ‘Heavy traffic analysis of open processing networks with complete resource pooling: Asymptotic optimality of discrete review policies’, *Annals of Applied Probability* **15**, 331–391.
- Atar, R., Budhiraja, A. and Dupuis, P. (2001), ‘On positive recurrence of constrained diffusion processes’, *The Annals of Probability* **29**(2), 979–1000.
- Banner, A. D. and Ghomrasni, R. (2008), ‘Local times of ranked continuous semimartingales’, *Stochastic Processes and their Applications* **118**, 1244–1253.
- Bassamboo, A., Randhawa, R. S. and Van Mieghem, J. A. (2010), ‘Optimal Flexibility Configurations in Newsvendor Networks: Going Beyond Chaining and Pairing’, *Management Science* **56**(8), 1285–1303.
- Chen, H. and Zhang, H. (2000), ‘Diffusion approximations for some multiclass queueing networks with fifo service disciplines’, *Mathematics of Operations Research* **25**(4), 679–707.
- Chod, J., Rudi, N. and Van Mieghem, J. A. (2008), Operational flexibility and financial hedging: Complements or substitutes? Working paper.
- Chou, M. C., Chua, G. A., Teo, C.-P. and Zheng, H. (2010), ‘Design for process flexibility: Efficiency of the long chain and sparse structure’, *Operations Research* **58**(1), 43–58.
- Chou, M., Teo, C. and Zheng, H. (2008), ‘Process flexibility: design, evaluation, and applications’, *Flexible Services and Manufacturing Journal* **20**(1), 59–94.
- Fine, C. H. and Freund, R. M. (1990), ‘Optimal investment in product-flexible manufacturing capacity’, *Management Science* **36**(4), 449–466.
- Graves, S. and Tomlin, B. (2003), ‘Process flexibility in supply chains’, *Management Science* **49**, 907–919.
- Gurvich, I. and Whitt, W. (2010), ‘Service-level differentiation in many-server service systems via queue-ratio routing’, *Operations Research* **58**(2), 316–328.
- Halfin, S. and Whitt, W. (1981), ‘Heavy-traffic limits for queues with many exponential servers’, *Operations Research* **29**, 567–588.
- Harrison, J. M. (1998), ‘Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete review policies’, *Annals of Applied Probability* pp. 822–848.

- 
- Harrison, J. M. and Lopez, M. J. (1999), ‘Heavy traffic resource pooling in parallel-server systems’, *Queueing Systems* **33**, 339–368.
- Hopp, W. J., Tekin, E. and Van Oyen, M. P. (2004), ‘Benefits of skill chaining in serial production lines with cross-trained workers’, *Management Science* **50**, 83–98.
- Hopp, W. and Oyen, M. (2004), ‘Agile workforce evaluation: A framework for cross-training and coordination’, *IIE Transactions* **36**(10), 919–940.
- Iglehart, D. (1973), ‘Weak convergence in queueing theory’, *Advances in Applied Probability* **5**(3), 570–594.
- Jordan, W. and Graves, S. C. (1995), ‘Principles on the benefits of manufacturing process flexibility’, *Management Science* **41**(4), 577–594.
- Jordan, W., Inman, R. and Blumenfeld, D. (2004), ‘Chained cross-training of workers for robust performance’, *IIE Transactions* **36**(10), 953–967.
- Karatzas, I. and Shreve, S. E. (1991), *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York.
- Kingman, J. F. C. (1962), ‘Some inequalities for the queue GI/G/1’, *Biometrika* **49**, 315–324.
- Kleinrock, L. (1976), *Queueing Systems: Volume 2: Computer Applications*, John Wiley and Sons, New York.
- Mandelbaum, A. and Stolyar, A. (2004), ‘Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized  $c\mu$ -rule’, *Operations Research* **52**(6), 836–855.
- Menich, R. and Serfozo, R. (1991), ‘Optimality of routing and servicing in dependent parallel processing systems’, *Queueing Systems* **9**(4), 403–418.
- Neely, M. J. and Modiano, E. (2005), ‘Convexity in queues with general inputs’, *IEEE Transactions on Information Theory* **51**(2), 706–714.
- Ostolaza, J., McClain, J. and Thomas, J. (1990), ‘The use of dynamic (state-dependent) assembly-line balancing to improve throughput’, *Journal of Manufacturing and Operations Management* **3**(2), 105–133.
- Rockafeller, R. T. (1970), *Convex Analysis*, Princeton University Press, Princeton, New Jersey.
- Saghafian, S., Oyen, M. V. and Kolfal, B. (2011), ‘The “W” network and the dynamic control of unreliable flexible servers’, *IIE Transactions* **12**(43), 893–907.
- Sethi, A. K. and Sethi, S. P. (1990), ‘Flexibility in manufacturing: A survey’, *International Journal of Flexible Manufacturing Systems* **2**(4), 289–328.

- Sheikhzadeh, M., Benjaafar, S. and Gupta, D. (1998), ‘Machine sharing in manufacturing systems: Total flexibility versus chaining’, *International Journal of Flexible Manufacturing Systems* **10**(4), 351–378.
- Stolyar, A. L. (2004), ‘Maxweight scheduling in a generalized switch: State space collapse and equivalent workload minimization under complete resource pooling’, *Annals of Applied Probability* **14**(1), 1–53.
- Van Mieghem, J. A. (1995), ‘Dynamic scheduling with convex delay costs: The generalized  $c\mu$  rule’, *Annals of Applied Probability* **5**(3), 808–833.
- Van Mieghem, J. A. (1998), ‘Investment strategies for flexible resources’, *Management Science* **44**(8), 1071–1078.
- Van Mieghem, J. A. (2003), ‘Due date scheduling: Asymptotic optimality of generalized longest queue and generalized largest delay rules’, *Operations Research* **51**(1), 113–122.
- Van Mieghem, J. A. (2008), *Operations Strategy: Principles and Practice*, Dynamic Ideas, Charlestown, MA.
- Van Oyen, M., Gel, E. and Hopp, W. (2001), ‘Performance opportunity for workforce agility in collaborative and noncollaborative work systems’, *IIE Transactions* **33**(9), 761–777.
- Wallace, R. and Whitt, W. (2005), ‘A staffing algorithm for call centers with skill-based routing’, *Manufacturing and Service Operations Management* **7**(4), 276–294.
- Williams, R. J. (1987), ‘Reflected brownian motion with skew symmetric data in a polyhedral domain’, *Probability Theory and Related Fields* **75**, 459–485.
- Williams, R. J. (2000), On dynamic scheduling of a parallel server system with complete resource pooling in analysis of communication networks: Call centres, traffic and performance, D.R. McDonald and S.R.E. Turner (eds.), in ‘Field Institute Communications’, Vol. 28, American Mathematical Society, Providence, RI, pp. 49–71.
- Zavadlav, E., McClain, J. and Thomas, L. (1996), ‘Self-buffering, self-balancing, self-flushing production lines’, *Management Science* **42**(8), 1151–1164.
- Zheng, Y. and Zipkin, P. (1990), ‘A queueing model to analyze the value of centralized inventory information’, *Operations Research* **38**(2), 296–307.
- Zipkin, P. (1995), ‘Performance analysis of a multi-item production-inventory system under alternative policies’, *Management Science* **41**(4), 690–703.

# E-Companion to A Little Flexibility is All You Need: On the Asymptotic Value of Flexible Capacity in Parallel Queuing Systems

## EC.1. Proofs of results in the main text

### EC.1.1. Proofs of results in Section 2

Theorem 1 follows from the general version of the result, Theorem EC.1, which is proved in Appendix EC.2.

### EC.1.2. Proof of results in Section 3

Most of the results in this section are specific cases of the general results in Section 4, and we omit their proofs. Specifically, Lemma 1 follows from Lemma EC.1. Proposition 1 follows from Proposition 5, Proposition 2 follows from Proposition 6, and Proposition 4 from Theorem 2. We do provide a proof of Proposition 3.

*Proof of Proposition 3.* Substituting  $\hat{\mu}_2 = \hat{\mu}_1\gamma$  in (10), we can rewrite the cost function as

$$\tilde{\Pi}(\hat{\mu}_1, \gamma) = \frac{1}{\hat{\mu}_1} \left( \frac{2}{2+\gamma} + \frac{1}{1+\gamma} \right) \sigma^2 hm + \hat{\mu}_1 (2c + \gamma c(1 + \Delta_2)). \quad (\text{EC.1})$$

Note that the feasible region for the pair  $(\hat{\mu}_1, \gamma)$  is the set  $\{(x, y) : x > 0, y \geq 0\} \cup \{(x, y) : x < 0, y < -2\}$ . (We consider the case  $\hat{\mu}_1 = 0$  separately.) For each fixed value of  $\gamma$ , the necessary and sufficient first order condition  $\frac{\partial \tilde{\Pi}(\hat{\mu}_1, \gamma)}{\partial \hat{\mu}_1} = 0$  must be satisfied for optimality, and this gives us

$$\hat{\mu}_1^* = \pm \sigma \sqrt{\frac{hm}{c}} \sqrt{\frac{\left(3 + \frac{1}{1+\gamma}\right)}{(2+\gamma)(2+\gamma(1+\Delta_2))}}. \quad (\text{EC.2})$$

Thus, we can rewrite the optimization problem (10) as

$$\min \left( \inf_{\gamma \in \mathbb{R} \setminus (0, -2]} f(\gamma), 2\sigma \sqrt{3hmc(1 + \Delta_2)} \right) \quad (\text{EC.3})$$

where  $f(\gamma) = 2\sigma \sqrt{\frac{\left(3 + \frac{1}{1+\gamma}\right)hm(2c + \gamma c(1 + \Delta_2))}{2+\gamma}}$  corresponds to the case  $\hat{\mu}_1^* \neq 0$ , and the optimal cost corresponding to the case  $\hat{\mu}_1^* = 0$  can be computed in a straightforward fashion from (10) as

$2\sigma\sqrt{3hmc(1+\Delta_2)}$ . Note that  $f(\gamma) \rightarrow 2\sigma\sqrt{3hmc(1+\Delta_2)}$ , as  $\gamma \rightarrow \infty$ , which allows us to restate (EC.3) as

$$\min_{\gamma \in \mathbb{R} \setminus (0, -2]} f(\gamma).$$

Using the first order conditions of optimality it is easy to see that the optimizer of this problem is

$$\gamma^* = \begin{cases} 2 \frac{(1-3\Delta_2 + \sqrt{\Delta_2 - \Delta_2^2})}{5\Delta_2 - 1} & \text{if } 0 \leq \Delta_2 < 0.5, \Delta_2 \neq 0.2, \\ 0 & \text{if } \Delta_2 \geq 0.5. \end{cases}$$

Note that as  $\Delta_2 \rightarrow 0.2$ , we have  $|\gamma^*| \rightarrow \infty$ . It can be easily verified that for the case  $\Delta_2 = 0.2$ , we have  $f(\gamma) \geq 2\sqrt{3hmc(1+\Delta_2)}$  for all  $\gamma \in \mathbb{R}/[-2, 0)$ , and thus the optimal solution here corresponds to  $\hat{\mu}_1^* = 0$ . Combining this with the expression for  $\hat{\mu}_1^*$  in (EC.2), the result follows. *Q.E.D.*

#### EC.1.2.1. Proof of results in Section 4

*Proof of Proposition 5.* Let  $\mathcal{N} = 2^{\{1,2,\dots,N\}} \setminus \phi$ , where  $\phi$  is the null set, denote the collection of all the skill sets. Then, any set  $F \in \mathcal{N}$  corresponds to a level- $|F|$  flexible resource which can process types  $i \in F$ ;  $|F|$  denotes the cardinality of the set  $F$ . We refer to this resource as the resource with skill set  $F$ . It will also be useful to define the indicator function  $\mathbb{I}: \mathcal{N} \times \{1, 2, \dots, N\} \times \mathbb{R}_+^N \rightarrow \{0, 1\}$  as

$$\mathbb{I}(F, i, Q) = 1\{i \in F, Q_i > Q_j, \text{ for } j \in F, j < i\} 1\{i \in F, Q_i \geq Q_j, \text{ for } j \in F, j > i\}.$$

That is  $\mathbb{I}(F, i, Q(t)) = 1$  at time  $t$  if type  $i$  queue is the longest among all types that can be processed by the resource with skill set  $F$  where ties are broken in favor of the type with the lowest index.

We begin by proving the following result:

LEMMA EC.1. *As  $\lambda \rightarrow \infty$ , if  $\frac{Q^\lambda(0)}{\sqrt{\lambda}} \Rightarrow \hat{Q}(0)$ , then  $\frac{Q^\lambda(\cdot)}{\sqrt{\lambda}} \Rightarrow \hat{Q}(\cdot)$ , where  $\hat{Q}$  is given by*

$$\hat{Q}_i(t) = \hat{Q}_i(0) - \frac{1}{m} \sum_{F \in \mathcal{N}} \int_0^t \mathbb{I}(F, i, \hat{Q}(t)) \hat{\mu}_{|F|} dt + \sigma \sqrt{2} B_i(t) + L_i(t) \quad (\text{EC.4})$$

where  $B_i$ ,  $i = 1, \dots, N$  are  $N$  independent standard Brownian motions,  $L_i$  are non-decreasing, continuous processes such that  $L_i(0) = 0$ , and  $\hat{Q}_i(t) \geq 0$  and  $\int_0^t \hat{Q}_i(s) dL_i(s) = 0$  for all  $t \geq 0$ .

*Proof.* For  $F \in \mathcal{N}$ ,  $S_{F,i}^\lambda(\cdot)$  denote the mutually independent renewal processes corresponding to departures from the resource with skill set  $F$  for jobs of type  $i$  if the resource worked all the time. Further, let  $T_{F,i}^\lambda(t)$  be the amount of time that this flexible resource spends processing type  $i$  jobs in  $[0, t]$ . For all  $F$  with  $|F| > 1$ , define  $M_{F,i}^\lambda(t) = S_{F,i}^\lambda(T_{F,i}^\lambda(t))$  which denotes the number of service completions of type  $i$  jobs by the flexible resource with skill set  $F$  in  $[0, t]$ .

Let  $D_{\mathbb{R}_+^N}[0, \infty)$  denote the space of right continuous processes with left limits that take values in  $\mathbb{R}_+^N$ . Define the mapping  $\Phi : D_{\mathbb{R}_+^N}[0, \infty) \rightarrow D_{\mathbb{R}_+^N}[0, \infty)$  by

$$\Phi(X)(t) = X(t) + \left( \sup_{0 \leq s \leq t} X_1(s)^-, \sup_{0 \leq s \leq t} X_2(s)^-, \dots, \sup_{0 \leq s \leq t} X_N(s)^- \right)'. \quad (\text{EC.5})$$

Then, an algebraic manipulation along with the definition  $X^\lambda \equiv A^\lambda - D^\lambda - \frac{\mu_1^\lambda}{m} \left( t - T_{\{i\},i}^\lambda(t) \right)$  with  $D_i^\lambda(t) \equiv S_{\{i\},i}^\lambda(T_{\{i\},i}^\lambda(t))$  denoting the departure processes corresponding to the dedicated resources, allows us to rewrite the queue-length process as follows:

LEMMA EC.2. *We can write  $Q^\lambda$  as*

$$Q^\lambda(t) = \Phi(Z^\lambda)(t), \quad (\text{EC.6})$$

$$Z_i^\lambda(t) = Q_i^\lambda(0) + X_i^\lambda(t) - \sum_{F \in \mathcal{N}, |F| \geq 2} M_{F,i}^\lambda(t), \quad (\text{EC.7})$$

for  $i = 1, \dots, N$ .

*Proof.* We can write the queue-length process as

$$Q_i^\lambda(t) = Q_i^\lambda(0) + X_i^\lambda(t) - \sum_{F \in \mathcal{N}, |F| \geq 2} M_{F,i}^\lambda(t) + \frac{\mu_1^\lambda}{m} \left( t - T_{\{i\},i}^\lambda(t) \right).$$

Thus, we have  $Q_i^\lambda(t) = Z_i^\lambda(t) + Y_i^\lambda(t)$ , where  $Z^\lambda$  is given in (EC.7) and  $Y_i^\lambda(t) = \frac{\mu_1^\lambda}{m} \left( t - T_{\{i\},i}^\lambda(t) \right)$ . Note that  $Y_i^\lambda$  is a non-decreasing process and increases only when  $Q_i^\lambda = 0$ . Thus, we must have  $Y_i^\lambda(t) = \sup_{0 \leq s \leq t} Z_i^\lambda(s)^-$  (see Iglehart, 1973). This completes the proof. *Q.E.D.*

The mapping  $\Phi$  possesses the following useful property (that is straightforward to verify, and hence the proof is omitted).

LEMMA EC.3. *The mapping  $\Phi$  is Lipschitz continuous, i.e., there exists  $K > 0$  such that  $\|\Phi(X) - \Phi(Y)\|_T \leq K \|X - Y\|_T$  for any  $X, Y \in D_{\mathbb{R}_+^N}[0, \infty)$ .*

We now scale the parameters by  $\sqrt{\lambda}$ , and use the notation  $(\hat{Q}^\lambda, \hat{Z}^\lambda, \hat{M}^\lambda, \hat{X}^\lambda, \hat{Q}^\lambda(0),)$  to denote  $(\frac{Q^\lambda}{\sqrt{\lambda}}, \frac{Z^\lambda}{\sqrt{\lambda}}, \frac{M^\lambda}{\sqrt{\lambda}}, \frac{X^\lambda}{\sqrt{\lambda}}, \frac{Q^\lambda(0)}{\sqrt{\lambda}})$ . Lemmas EC.4-EC.6 below will be useful in completing the proof.

LEMMA EC.4.  $\hat{X}^\lambda \Rightarrow W$ , where  $W_i(t) = -\hat{\mu}_1 t + \sigma\sqrt{2}B_i(t)$  for  $i = 1, \dots, N$ .

*Proof.* Fix any  $\bar{t} > 0$ . Applying the functional strong law of large numbers and the fact that  $\|T_{\{i\},i}^\lambda(t) - t\|_{\bar{t}} \rightarrow 0$  as  $\lambda \rightarrow \infty$  we obtain the following almost sure convergences for the renewal processes:

$$\begin{aligned} \left\| \frac{A_i^\lambda(\cdot)}{\lambda} - \tau(\cdot) \right\|_{\bar{t}} &\rightarrow 0, \\ \left\| \frac{D_i^\lambda(\cdot)}{\lambda} - \tau(\cdot) \right\|_{\bar{t}} &\rightarrow 0, \end{aligned}$$

as  $\lambda \rightarrow \infty$ , where  $\tau(t) = t$  for  $t \geq 0$ . Thus, applying the functional central limit theorem, we obtain

$$\frac{A_i^\lambda - D_i^\lambda}{\sqrt{\lambda}} \Rightarrow -\hat{\mu}_1(\cdot) + \sigma\sqrt{\frac{2}{m}}B_i(\cdot), \quad (\text{EC.8})$$

where  $\hat{\mu}_1(t) = \hat{\mu}_1 t$  for all  $t \geq 0$ . The result then follows by noting the mutual independence of the renewal processes. *Q.E.D.*

LEMMA EC.5.  $\{(\hat{X}^\lambda, \hat{M}^\lambda, \hat{Z}^\lambda, \hat{Q}^\lambda)\}$  is jointly C-tight.

*Proof.* We shall use the same notion of C-tightness used in Chen and Zhang (2000). Let  $\{R^n\}$  be a sequence of stochastic processes such that  $R^n \in D_{\mathbb{R}^m}[0, \infty)$ . Then,  $\{R^n\}$  is C-tight if  $\{R^n(0)\}$  is tight, and for any  $T > 0$ , for each  $\epsilon, \eta > 0$ , there exist  $\delta > 0$  and  $M < \infty$  such that for all  $n > M$

$$\mathbb{P}\left(\sup_{0 \leq s, t \leq T, |s-t| < \delta} \|R^n(s) - R^n(t)\| \geq \epsilon\right) \leq \eta.$$

Note that using this notion of C-tightness, to prove that  $\{(\hat{X}^\lambda, \hat{M}^\lambda, \hat{Z}^\lambda, \hat{Q}^\lambda)\}$  is C-tight, it suffices to prove that the individual sequences  $\{\hat{X}^\lambda\}$ ,  $\{\hat{M}^\lambda\}$ ,  $\{\hat{Z}^\lambda\}$ , and  $\{\hat{Q}^\lambda\}$  are C-tight.

The weak convergence of  $\hat{X}^\lambda$  established in Lemma EC.4 gives us that  $\{\hat{X}^\lambda\}$  is C-tight by applying Lemma 4.2(ii) of Chen and Zhang (2000).

We now prove that  $\{\hat{M}^\lambda\}$  is C-tight. Fix any  $|F| \geq 2$ . We will prove the result for  $\{\hat{M}_{F,i}^\lambda\}$  for any  $i \in F$ , the others follow similarly. We define  $S_F^\lambda := \sum_{i \in F} S_{F,i}^\lambda$ . We begin by noting that by

the functional strong law of large numbers, we have  $\| \frac{S_F^\lambda(\cdot)}{\sqrt{\lambda}} - \mu_{|F|}(\cdot) \|_{\bar{t}} \rightarrow 0$  a.s., as  $\lambda \rightarrow \infty$ , where  $\mu_{|F|}(t) = \mu_{|F|}t$  for  $t \geq 0$ . Using this convergence and the following bound for  $0 \leq s < t \leq T$  proves the desired tightness

$$|\hat{M}_{F,i}^\lambda(t) - \hat{M}_{F,i}^\lambda(s)| \leq \frac{S_F^\lambda(t) - S_F^\lambda(s)}{\sqrt{\lambda}}.$$

The tightness of  $\hat{Z}^\lambda$  follows from the definition of  $\hat{Z}_n$ ,  $\hat{M}_n$  and  $\hat{X}_n$  and the relationship stated in (EC.7). The tightness of  $\hat{Q}^\lambda$  then follows by noting that

$$\| \hat{Q}^\lambda(t) - \hat{Q}^\lambda(s) \| \leq K \sup_{s \leq u \leq t} \| \hat{Z}^\lambda(t) - \hat{Z}^\lambda(u) \|$$

for  $0 \leq s \leq t \leq T$  and some constant  $K > 0$  (using Lemma 4.3 of Chen and Zhang, 2000 and the fact that  $\hat{Q}^\lambda = \Phi(\hat{Z}^\lambda)$ ). *Q.E.D.*

LEMMA EC.6. *If  $(\hat{X}, \hat{M}, \hat{Z}, \hat{Q}, \hat{Q}(0))$  is the limit of any weakly convergent subsequence of  $\{\hat{X}^\lambda, \hat{M}^\lambda, \hat{Z}^\lambda, \hat{Q}^\lambda, \hat{Q}(0)\}$ , then  $\hat{Q}$  satisfies (EC.4).*

*Proof.* As  $\hat{Q}^\lambda(0)$  converges by assumption, and  $\hat{X}^\lambda$  converges as per Lemma EC.4, if we prove the convergence of  $\hat{M}^\lambda$  to the appropriate process, we obtain the convergence of  $\hat{Z}^\lambda$ . Then using the Lipschitz continuity of  $\Phi$ , the result follows. As our sequence is C-tight, the limiting processes are continuous, and thus appealing to the Skorohod representation theorem, we can restrict attention to almost sure convergence in the uniform topology. Let  $\{\lambda_k\}$  denote the index of the converging subsequence.

Fix any  $F \in \mathcal{N}$  with  $|F| \geq 2$  and any  $i \in F$ . As before, we will only prove the result for the case  $\hat{M}_{F,i}^{\lambda_k}$  for any fixed  $F \in \mathcal{N}$  and  $i \in F$ . We need to prove  $\hat{M}_{F,i}^{\lambda_k} \rightarrow \frac{1}{m} \int_0^\cdot \mathbb{I}(F, i, \hat{Q}(s)) \hat{\mu}_{|F|} ds$ . It is easy to see the following:

$$\begin{aligned} & \| \hat{M}_{F,i}^{\lambda_k}(\cdot) - \frac{1}{m} \int_0^\cdot \mathbb{I}(F, i, \hat{Q}(s)) \hat{\mu}_{|F|} ds \|_{\bar{t}} \\ &= \| \hat{S}_{F,i}^{\lambda_k}(T_{F,i}^{\lambda_k}(\cdot)) - \frac{1}{m} \int_0^\cdot \mathbb{I}(F, i, \hat{Q}(s)) \hat{\mu}_{|F|} ds \|_{\bar{t}} \\ &\leq \| \hat{S}_{F,i}^{\lambda_k}(T_{F,i}^{\lambda_k}(\cdot)) - \frac{1}{m} T_{F,i}^{\lambda_k}(\cdot) \hat{\mu}_{|F|} \|_{\bar{t}} + \frac{\hat{\mu}_{|F|}}{m} \| T_{F,i}^{\lambda_k}(t) - \int_0^\cdot \mathbb{I}(F, i, \hat{Q}(s)) ds \|_{\bar{t}} \end{aligned} \tag{EC.9}$$

The functional strong law of large numbers implies that

$$\sup_{0 \leq t \leq \bar{t}} |\hat{S}_{F,i}^{\lambda_k}(t) - \mu_{|F|} t| \rightarrow 0, \text{ a.s.},$$

and hence the first term on the right hand side converges to zero. For the second term, we prove that

$$\|T_{F,i}^{\lambda_k}(\cdot) - \int_0^\cdot \mathbb{I}(F, i, \hat{Q}(s)) ds\|_{\bar{t}} \rightarrow 0 \text{ a.s.}, \text{ as } \lambda \rightarrow \infty. \text{ Note that } T_{F,i}^{\lambda_k}(t) = \int_0^t 1\{\hat{Q}_i(s) > 0\} \mathbb{I}(F, i, \hat{Q}(s)) ds,$$

and

$$\begin{aligned} \int_0^t 1\{\hat{Q}_i(s) > 0\} 1\{\hat{Q}_i(s) > \max_{j \in F \setminus \{i\}} \hat{Q}_j(s)\} ds &\leq \liminf_{k \rightarrow \infty} \int_0^t 1\{\hat{Q}_i(s) > 0\} \mathbb{I}(F, i, \hat{Q}^{\lambda_k}(s)) ds \\ &\leq \limsup_{k \rightarrow \infty} \int_0^t 1\{\hat{Q}_i(s) > 0\} \mathbb{I}(F, i, \hat{Q}^{\lambda_k}(s)) ds \\ &\leq \int_0^t 1\{\hat{Q}_i(s) \geq \max_{j \in F} \hat{Q}_j(s)\} ds. \end{aligned}$$

Further,

$$\int_0^t 1\{\hat{Q}_i(s) > 0\} 1\{\hat{Q}_i(s) > \max_{j \in F \setminus \{i\}} \hat{Q}_j(s)\} ds \leq \int_0^t \mathbb{I}(F, i, \hat{Q}(s)) ds \leq \int_0^t 1\{\hat{Q}_i(s) \geq \max_{j \in F} \hat{Q}_j(s)\} ds.$$

(EC.10)

Now, note that each element of  $\hat{M}$  is continuous and increasing, and hence absolutely continuous. Thus,  $\hat{M}_{F,i}$  is differentiable almost everywhere on  $[0, \bar{t}]$ , and we can write  $\hat{M}_{F,i}(t) = \int_0^t \dot{M}_{F,i}(s) ds$  for  $0 \leq t \leq \bar{t}$  for some  $\dot{M}_{F,i}$ . Then, using the Girsanov's theorem (see for example, Theorem 5.1 of Karatzas and Shreve, 1991), we obtain the existence of a measure  $\mathbb{P}'$  under which  $\hat{Q}_i$  is a reflected Brownian motion. Thus, we obtain  $\int_0^t 1\{\hat{Q}_i(s) > 0\} 1\{\hat{Q}_i(s) > \max_{j \in F \setminus \{i\}} \hat{Q}_j(s)\} ds = \int_0^t 1\{\hat{Q}_i(s) \geq \max_{j \in F} \hat{Q}_j(s)\} ds$ , a.s. Using the relation (EC.10) then completes the proof. *Q.E.D.*

We now use the fact that (EC.4) has solutions that are unique in law, which follows from the fact that under a suitable probability measure, we can consider the case where (EC.4) has no drift terms (using Girsanov's theorem), and thus (EC.4) becomes a collection of  $N$  independent, non-degenerate, one dimensional reflected Brownian motions. This uniqueness along with Lemma EC.6 and the tightness established in Lemma EC.5 completes the proof of Lemma EC.1. *Q.E.D.*

Proposition 5 then follows by using the relation for the order statistics of a set of continuous semi-martingales as in Corollary 2.6 of Banner and Ghomrasni (2008). The intuition behind the

result is that focusing on the order statistics implies that the corresponding drifts can be identified precisely without the need for the indicator functions. Further, to ensure the ranking order, the  $k^{\text{th}}$  order-statistic for  $1 < k < n$  is associated with two local times. *Q.E.D.*

*Proof of Proposition 6.* The distribution corresponding to  $\pi$  is easily seen to be invariant for  $\hat{Q}$  by applying Theorem 1.2 of Williams (1987) and noting that the reflection directions are normal to the corresponding faces, which implies that the skew symmetry condition is trivially satisfied, and scaling the solution appropriately (the result in Williams, 1987 is with a unit diffusion coefficient). The positive recurrence derived in Theorem 2.2 of Atar et al. (2001) proves that this is indeed the steady-state distribution. The expressions for the expected queue-lengths follow by a straightforward integration. *Q.E.D.*

*Proof of Theorem 2.* The optimality of  $\Pi^{\lambda^*}$  ensures that  $\liminf_{\lambda \rightarrow \infty} \frac{\Pi^\lambda(\mu^{\lambda^*}) - \Pi^{\lambda^*}}{\sqrt{\lambda}} \geq 0$ . Let  $\lambda_k$  denote the index of a subsequence such that

$$\liminf_{k \rightarrow \infty} \frac{\Pi^{\lambda_k}(\lambda_k + \hat{\mu}_1^* \sqrt{\lambda_k}, \hat{\mu}_2^* \sqrt{\lambda_k}, 0, \dots, 0) - \Pi^{\lambda_k^*}}{\sqrt{\lambda_k}} > 0.$$

By Theorem 1, we must have a further subsequence of any optimal solution to (1),  $(\mu_1^{\lambda_{\ell^*}}, \mu_2^{\lambda_{\ell^*}}, \dots)$  such that the  $\frac{\mu_i^{\lambda_{\ell^*}}}{\sqrt{\lambda_{\ell^*}}}$  is convergent for  $i \geq 2$ . But, then we obtain a contradiction to the fact that  $(\lambda + \hat{\mu}_1^* \sqrt{\lambda}, \hat{\mu}_2^* \sqrt{\lambda}, 0, \dots, 0)$  solves (7). *Q.E.D.*

*Proof of Theorem 3.* It suffices to prove the result for the setting where  $\Delta_k = \sum_{j=2}^k 2/j \Delta_2$ . Define  $\alpha_\ell(\hat{\mu}) \equiv [\sum_{k=1}^N \sum_{j=\max(\ell-1, k-1)}^{N-1} \binom{j}{k-1} \hat{\mu}_k]^{-1}$ . Then, we can write the centered and scaled cost function  $\hat{\Pi}(\hat{\mu}) = \sum_{k=1}^N c(1 + 2(H(k) - 1)\Delta_2) \binom{N}{k} \hat{\mu}_k + \sigma^2 hm \sum_{\ell=1}^N (N - \ell + 1) \alpha_\ell(\hat{\mu})$ , where  $H(k) = \sum_{j=1}^k 1/j$  for all  $k = 1, 2, \dots$  denotes the Harmonic sequence.

Further, we can write

$$\frac{\partial \hat{\Pi}}{\partial \hat{\mu}_k} = c(1 + 2(H(k) - 1)\Delta_2) \binom{N}{k} - \sigma^2 hm \sum_{\ell=1}^N (N - \ell + 1) \alpha_\ell^2(\hat{\mu}) \sum_{j=\ell-1}^{N-1} \binom{j}{k-1}. \quad (\text{EC.11})$$

Thus, we have

$$\frac{\partial \hat{\Pi}}{\partial \hat{\mu}_1} = cN - \sigma^2 hm \sum_{\ell=1}^N (N - \ell + 1)^2 \alpha_\ell^2(\hat{\mu}), \quad \text{and} \quad (\text{EC.12})$$

$$\frac{\partial \hat{\Pi}}{\partial \hat{\mu}_2} = c(1 + \Delta_2) \binom{N}{2} - \sigma^2 hm \sum_{\ell=1}^N (N - \ell + 1) \alpha_\ell^2(\hat{\mu}) \binom{N-1}{\sum_{j=\ell-1} j}. \quad (\text{EC.13})$$

We will now compute a weighted sum of  $\frac{\partial \hat{\Pi}}{\partial \hat{\mu}_k}$ ,  $\frac{\partial \hat{\Pi}}{\partial \hat{\mu}_2}$ , and  $\frac{\partial \hat{\Pi}}{\partial \hat{\mu}_1}$  for each  $k \geq 3$ , with the weights chosen to ensure that the sum can be expressed entirely in terms of  $\alpha_\ell^2$ . Consider the following relation:

$$\begin{aligned} D_k(\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_N) &\equiv \frac{\partial \hat{\Pi}}{\partial \hat{\mu}_k} - \frac{\binom{N}{k}}{\binom{N}{2}} 2(H(k) - 1) \frac{\partial \hat{\Pi}}{\partial \hat{\mu}_2} - \binom{N}{k} (3 - 2H(k)) / N \frac{\partial \hat{\Pi}}{\partial \hat{\mu}_1} \\ &= \sigma^2 hm \sum_{\ell=1}^N T(N, k, \ell) \alpha_\ell^2(\hat{\mu}), \end{aligned} \quad (\text{EC.14})$$

where

$$\begin{aligned} T(N, k, \ell) &= -\frac{(N - \ell + 1)}{N(N - 1)} \left( -(N - 1)N \binom{\ell - 1}{k} + (\ell - 1)(3N - 2\ell + 1) \binom{N}{k} \right. \\ &\quad \left. - 2(\ell - 1)(N - \ell + 1) \binom{N}{k} H(k) \right). \end{aligned}$$

The outline of the remainder of the proof is as follows. Lemmas EC.7-EC.9 state and prove some auxiliary results, using which Lemma EC.10 demonstrates that the above expression evaluated under the condition  $\hat{\mu}_i = 0$  for  $i \geq 3$  is strictly positive for  $k \geq 3$ . Lemma EC.11 then proves that the first order optimality conditions for level- $k \geq 3$  resources cannot hold with equality, i.e., they cannot have an interior solution.

LEMMA EC.7. *If  $a, b, c, d, e, f > 0$ ,  $a/b = d/e$  and  $b/c < e/f$ , then*

$$\frac{a}{b\hat{\mu}_1 + c\hat{\mu}_2} < \frac{d}{e\hat{\mu}_1 + f\hat{\mu}_2}.$$

for all  $\hat{\mu}_1 \in \mathbb{R}$  and  $\hat{\mu}_2 \in \mathbb{R}_+$  such that  $b\hat{\mu}_1 + c\hat{\mu}_2 > 0$  and  $e\hat{\mu}_1 + f\hat{\mu}_2 > 0$ .

*Proof.* The result follows immediately by cross-multiplying and using the properties stated; we omit the details. *Q.E.D.*

LEMMA EC.8. *For  $3 \leq k \leq N$ , we have  $D_k(\hat{\mu}_1, 0, \dots, 0) = 0$ .*

*Proof.* Fix  $3 \leq k \leq N$ . We have

$$D_k(\hat{\mu}_1, 0, \dots, 0) = \sigma^2 hm \left( \frac{\binom{N}{k} H(k)}{\hat{\mu}_1^2} - \sum_{\ell=1}^n \frac{-(\ell - k) \binom{\ell - 1}{k - 1} + (N - k + 1) \binom{N}{k - 1}}{k(N - \ell + 1) \hat{\mu}_1^2} \right).$$

Noting that

$$\frac{-(\ell - k)\binom{\ell-1}{k-1} + (N - k + 1)\binom{N}{k-1}}{\ell(N - \ell + 1)\hat{\mu}_1^2} = \frac{\sum_{j=\ell-1}^{N-1} \binom{j}{k-1}}{(N - \ell + 1)\hat{\mu}_1^2},$$

it suffices to prove that  $\sum_{\ell=1}^N \frac{\sum_{j=\ell-1}^{N-1} \binom{j}{k-1}}{(N - \ell + 1)} = \binom{N}{k} H(k)$ . We can write

$$\begin{aligned} \sum_{\ell=1}^N \frac{\sum_{j=\ell-1}^{N-1} \binom{j}{k-1}}{(N - \ell + 1)} - \binom{N}{k} H(k) &= \sum_{j=1}^N \sum_{\ell=1}^j \frac{\binom{j-1}{k-1}}{N - \ell + 1} - \binom{N}{k} H(k) \\ &= \sum_{j=k}^N \binom{j-1}{k-1} (H(N) - H(N - j)) - \binom{N}{k} H(k) \\ &= \binom{N}{k} (H(N) - H(k)) - \sum_{j=k}^N \binom{j-1}{k-1} H(N - j). \end{aligned}$$

Further, we have

$$\begin{aligned} \sum_{j=k}^N \binom{j-1}{k-1} H(N - j) &= \sum_{j=k}^N \binom{j-1}{k-1} \sum_{\ell=1}^{N-j} 1/\ell \\ &= \sum_{\ell=1}^N \sum_{j=k}^{N-\ell} \binom{j-1}{k-1} 1/\ell \\ &= \sum_{\ell=1}^N \frac{1}{\ell} \binom{k}{N - \ell} \\ &= \binom{N}{k} [H(N) - H(k)]. \end{aligned}$$

This completes the proof. *Q.E.D.*

LEMMA EC.9. *For any  $k \geq 3$ , there exists  $N > \bar{\ell} > 1$  such that  $T(N, k, \ell) \geq 0$  for  $\ell \leq \bar{\ell}$  and  $T(N, k, \ell) \leq 0$  for  $N \geq \ell > \bar{\ell}$ .*

*Proof.* We can write  $T(N, k, \ell) = -\frac{(N-\ell+1)}{N(N-1)}(U_{N,k}(\ell) - V_{N,k}(\ell))$ , where

$$\begin{aligned} U_{N,k}(\ell) &= (\ell - 1) \binom{N}{k} (1 - 2\ell + 3N - 2(N - \ell + 1)H(k)) \\ V_{N,k}(\ell) &= (N - 1)N \binom{\ell-1}{k}. \end{aligned}$$

Note that  $U'_{N,k}(\ell)$  is an affine function. Further, we can write  $V'_{N,k}(\ell) = g_{N,k}(\ell)h_{N,k}(\ell)$  where  $g_{N,k}(\ell) = N(N - 1)\binom{\ell-1}{k}$  and  $h_{N,k}(\ell) = H(\ell) - H(\ell - k)$  are non-negative, convex increasing functions (where we consider the extension of the functions to the real line by replacing factorials by the Gamma function and the Harmonic function by the Polygamma function). Thus,  $V'_{N,k}(\ell)$  is

also convex increasing. This implies that  $U'_{N,k} - V'_{N,k} = 0$  can have at most two roots, which gives us that  $U_{N,k}(\ell) - V_{N,k}(\ell) = 0$  can have at most three roots. Further, noting that  $U_{N,k}(1) - V_{N,k}(1) = U_{N,k}(N+1) - V_{N,k}(N+1) = 0$ , it follows that  $T(N, k, \cdot)$  can have at most two sign changes. Also, noting that

$$T(N, k, 2) = 2 \frac{N-1}{N} \binom{N}{k} (H(k) - 3/2) > 0, \quad (\text{EC.15})$$

because  $H(k) > 3/2$  for  $k \geq 3$ , we obtain that  $T(N, k, 2) > 0$ . Appealing to Lemma EC.8, we obtain the existence of  $\bar{\ell}$ . *Q.E.D.*

LEMMA EC.10. *For every  $3 \leq k \leq N$ ,  $D_k(\hat{\mu}_1, \hat{\mu}_2, 0, \dots, 0) > 0$ .*

*Proof.* Applying Lemmas EC.8 and EC.9, we can write

$$D_k(\hat{\mu}_1, \hat{\mu}_2, 0, \dots, 0) = \sigma^2 h m \sum_{\ell=1}^{\bar{\ell}} \sum_{j=\bar{\ell}+1}^N (u_{\ell,j} \alpha_{\ell}^2(\hat{\mu}_1, \hat{\mu}_2, 0, \dots, 0) - v_{\ell,j} \alpha_j^2(\hat{\mu}_1, \hat{\mu}_2, 0, \dots, 0)),$$

where  $\sum_{j=\bar{\ell}+1}^N u_{\ell,j} = T(N, k, \ell)$  and  $\sum_{\ell=1}^{\bar{\ell}+1} v_{\ell,j} = T(N, k, j)$  and  $u_{\ell,j} \alpha_{\ell}^2(\hat{\mu}_1, 0, \dots, 0) = v_{\ell,j} \alpha_j^2(\hat{\mu}_1, 0, \dots, 0)$ .

Noting that  $\alpha_{\ell}(\hat{\mu}_1, \hat{\mu}_2, \dots, 0) = \frac{1}{(N-\ell+1)\hat{\mu}_1 + \binom{N}{2} - \binom{\ell-1}{2}\hat{\mu}_2}$ . Combining this with the fact that  $(N-\ell)/(\binom{N}{2} - \binom{\ell-1}{2})$  is decreasing in  $\ell$ , and applying Lemma EC.7, we obtain

$$\sqrt{u_{\ell,j}} \alpha_{\ell}(\hat{\mu}_1, \hat{\mu}_2, 0, \dots, 0) > \sqrt{v_{\ell,j}} \alpha_j(\hat{\mu}_1, \hat{\mu}_2, 0, \dots, 0).$$

This completes the proof. *Q.E.D.*

The following result then completes the proof of the theorem.

LEMMA EC.11. *Any optimal solution  $\hat{\mu}^*$  must have  $\frac{\partial \hat{\Pi}(\hat{\mu}^*)}{\partial \hat{\mu}_k} > 0$  for  $k \geq 3$ .*

*Proof.* Using Theorem 5.7 of Rockafeller (1970), we obtain that  $\hat{\Pi}(\hat{\mu})$  is convex in  $\hat{\mu}$ . Thus, the KKT conditions are sufficient for optimality. Since  $\hat{\mu}_1^* \in (-\infty, \infty)$ , we must have  $\frac{\partial \hat{\Pi}(\hat{\mu}^*)}{\partial \hat{\mu}_1} = 0$ . Combining this with (EC.14) and Lemma EC.10, we obtain that  $\frac{\partial \hat{\Pi}(\hat{\mu}^*)}{\partial \hat{\mu}_k} > 2(H(k) - 1) \frac{\binom{N}{k}}{\binom{N}{2}} \frac{\partial \hat{\Pi}(\hat{\mu}^*)}{\partial \hat{\mu}_2}$  for  $k \geq 3$ . Noting that  $\frac{\partial \hat{\Pi}(\hat{\mu}^*)}{\partial \hat{\mu}_2} \geq 0$  for optimality, we obtain that  $\frac{\partial \hat{\Pi}(\hat{\mu}^*)}{\partial \hat{\mu}_k} > 0$ , and thus we must have  $\hat{\mu}_k^* = 0$ . *Q.E.D.*

Thus, any optimal solution must have  $\hat{\mu}_k^* = 0$  for  $k \geq 3$ , and the proof of Theorem 3 is complete. *Q.E.D.*

*Proof of Proposition 7.* It suffices to prove the result for the setting where  $\Delta_k = \sum_{j=2}^k 1/j$  for  $k \geq 2$ . For  $\Delta_2 = 1/2$ , Theorem 3 implies that if  $\Delta_k \geq \sum_{j=2}^k 1/j$ , then  $\hat{\mu}_k^* = 0$  for  $k \geq 3$ . We next prove that  $\hat{\mu}_2^* = 0$ , which would complete the proof.

As  $\hat{\mu}_1^* \in \mathbb{R}$  and  $\hat{\mu}_2^* \geq 0$ , the first order condition for optimality implies that  $\frac{\partial \hat{\Pi}(\hat{\mu}_1^*, \hat{\mu}_2^*)}{\partial \hat{\mu}_1} = 0$ ,  $\frac{\partial \hat{\Pi}(\hat{\mu}_1^*, \hat{\mu}_2^*)}{\partial \hat{\mu}_2} \geq 0$  and  $\hat{\mu}_2^* \frac{\partial \hat{\Pi}(\hat{\mu}_1^*, \hat{\mu}_2^*)}{\partial \hat{\mu}_2} = 0$ . Using the expressions for the derivatives of the scaled cost function in (EC.12) and (EC.13), for  $\hat{\mu}_2^*$  to be strictly positive we must have

$$cN = \sigma^2 hm \sum_{k=1}^N (N-k+1)^2 \alpha_k^2 \quad (\text{EC.16})$$

$$c(1 + \Delta_2) \frac{N(N-1)}{2} = \sigma^2 hm \sum_{k=1}^N (N-k+1) \alpha_k^2 \left( \sum_{j=k-1}^{N-1} j \right), \quad (\text{EC.17})$$

where  $\alpha_k = [(N-k+1)\hat{\mu}_1^* + \sum_{j=k-1}^{N-1} j\hat{\mu}_2^*]^{-1}$ . Consider the following optimization problem that maximizes the right hand side term of (EC.17) subject to (EC.16), i.e., maximizes the value of level-2 flexibility:

$$\begin{aligned} & \max_{\{\hat{\mu}_1^* \in \mathbb{R}, \hat{\mu}_2^* \geq 0, N\hat{\mu}_1^* + \binom{N}{2}\hat{\mu}_2^* > 0\}} \sigma^2 hm \sum_{k=1}^N (N-k+1) \alpha_k^2 \left( \sum_{j=k-1}^{N-1} j \right), \\ & \text{s.t. } cN = \sigma^2 hm \sum_{k=1}^N (N-k+1)^2 \alpha_k^2. \end{aligned}$$

As the objective function is convex in the arguments, this problem must have a corner solution, i.e.,  $\hat{\mu}_2^* = 0$ . Using this value for  $\hat{\mu}_2^*$ , and solving for  $\Delta_2$  in (EC.16-EC.17), we obtain  $\Delta_2 = 0.5$ , and the result follows. *Q.E.D.*

## EC.2. General asymmetric systems: A little flexibility is all you need

In this section, we extend Theorem 1 to general asymmetric systems. We first generalize the notation to cater to asymmetric systems. Denote types by  $i = 1, 2, \dots, N$  and the arrival process of type  $i$  customers or jobs by  $A_i^\lambda(t)$ . We assume that all arrival processes are independent renewal processes with type  $i$  customers arriving at rate  $\alpha_i \lambda > 0$ , where we normalize  $\alpha$  so that  $\sum_{i=1}^N \alpha_i / N = 1$ . Let  $\sigma_{\alpha, i}^\lambda$  denote the standard deviation of the inter-arrival times of type  $i$  customers. Each arriving

job of type  $i$  has a service requirement that is independent (across all customers) and identically distributed (for customers of the same type) with mean  $m_i$  and variance  $\sigma_{s,i}^2$ . The coefficient of variation of service times is denoted by  $c_{s,i} = \sigma_{s,i}/m_i$ , while that of the inter-arrival times is  $c_{a,i} = \lambda\sigma_{a,i}^\lambda$ . We assume that  $c_{a,i}$  is a constant independent of the rate  $\lambda$  and will henceforth denote  $\sigma_i^2 = (c_{a,i}^2 + c_{s,i}^2)/2$ .

We assume that each type has a dedicated resource assigned to it that operates at a fixed deterministic rate  $\mu_{\{i\}}^\lambda$ . Let  $\mu_F$  denote the capacity of resource that can serve customers of types that lie in set  $F$ , where  $F \subseteq \{1, \dots, N\}$ . Note that  $\mu_F = 0$  represents the case in which resource- $F$  has zero capacity or equivalently is not available. As before, capacities scale the actual average service time, i.e., if a service rate of  $\mu$  is allocated to a type  $i$  job, its average service time is  $m_i/\mu$  and its variance is  $\sigma_{s,i}^2/\mu^2$ .

The system incurs holding costs and a capacity cost rate that depends on capacity size and flexibility level. We assume capacity costs are linear in size. The cost rate of capacity size  $\mu_F$  of flexible resource- $F$  (that has level- $|F|$  flexibility) is  $\mu_F^\lambda c(1 + \Delta_F)$ , where  $\Delta_F$  is the flexibility premium for resource- $F$ . We set  $\Delta_{\{i\}} = 0$ , for all  $i = 1, \dots, N$  so that  $c$  is the per unit cost of dedicated capacity. Note that while we allow resources at the same level of flexibility to have different flexibility premiums, we assume that all dedicated resources have identical per unit costs.

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  $h_i$  per job per unit time, we obtain the total cost rate of a capacity portfolio  $\mu^\lambda = \{\mu_F^\lambda \geq 0\}$  as

$$\Pi^\lambda(\mu^\lambda) = \sum_{i=1}^N h_i \mathbb{E}Q_i^\lambda(\infty) + \sum_{F \subseteq \{1, 2, \dots, N\}} c_F \mu_F^\lambda.$$

Given that all jobs eventually get served, revenues are independent of  $\mu^\lambda$  and we seek the capacity portfolio  $\mu^{\lambda*}$  that minimizes costs:

$$\Pi^{\lambda*} = \Pi^\lambda(\mu^{\lambda*}) = \min_{\mu \geq 0} \Pi^\lambda(\mu). \quad (\text{EC.18})$$

Defining

$$\bar{\Pi}^\lambda = \sum_{i=1}^N \left( cm_i \alpha_i \lambda + 2\sigma_i \sqrt{ch_i m_i \alpha_i \lambda} \right). \quad (\text{EC.19})$$

and

$$\underline{\Pi}^\lambda = \sum_{i=1}^N cm_i \alpha_i \lambda + 2\sqrt{\sum_{i=1}^N \sigma_i^2 ch_i m_i \alpha_i \lambda}, \quad (\text{EC.20})$$

where  $\underline{h} = \min_i h_i$ , we have the following result for general asymmetric systems that is analogous to Theorem 1.

**THEOREM EC.1 (A little flexibility is all you need).** *The optimal cost  $\Pi^\lambda(\mu^{\lambda*})$  is bounded:*

$$\underline{\Pi}^\lambda + o(\sqrt{\lambda}) \leq \Pi^\lambda(\mu^{\lambda*}) \leq \bar{\Pi}^\lambda + o(\sqrt{\lambda}), \quad (\text{EC.21})$$

and any optimal solution  $\mu^{\lambda*}$  to the optimization problem (EC.18) satisfies

$$\begin{aligned} \mu_{\{i\}}^{\lambda*} &= m_i \alpha_i \lambda + \hat{\mu}_{\{i\}} \sqrt{\lambda} + o(\sqrt{\lambda}), \quad \text{and} \\ \mu_F^{\lambda*} &= \hat{\mu}_F \sqrt{\lambda} + o(\sqrt{\lambda}) \quad \text{for } F \subseteq \{1, 2, \dots, N\}, |F| \geq 2, \end{aligned}$$

for some  $\hat{\mu}_F \in \mathbb{R}$  for  $F \subseteq \{1, 2, \dots, N\}$  with  $\hat{\mu}_F \geq 0$  for  $|F| \geq 2$  and  $\sum_{F \subseteq \{1, 2, \dots, N\}} \hat{\mu}_F > 0$ .

*Proof.* We can obtain an upper bound on the optimal cost by noting that the optimal cost dominates the minimal cost when using only dedicated servers:  $\Pi^\lambda(\mu^{\lambda*}) \leq \min_{\mu = \{\mu_{\{i\}}\}_{i=1, \dots, N}} \Pi^\lambda(\mu)$ .

Using Kingman's bound, we obtain the following bound on each queue's cost:

$$\Pi_i^\lambda(\mu_{\{i\}}^\lambda) = h_i \mathbb{E}Q_i + c\mu_{\{i\}}^\lambda \leq h_i \left( \sigma_i^2 \frac{\mu_{\{i\}}^\lambda}{\mu_{\{i\}}^\lambda - m_i \alpha_i \lambda} + 1 \right) + c\mu_{\{i\}}^\lambda$$

The right hand side is convex in  $\mu_{\{i\}}^\lambda$  and reaches a minimum at  $\tilde{\mu}_{\{i\}}^\lambda = m_i \alpha_i \lambda + \sigma_i \sqrt{\frac{h_i}{c} m_i \alpha_i \lambda}$ .

Combining these bounds for all  $N$  queues, we obtain an exact upper bound  $\Pi^\lambda(\mu^{\lambda*}) \leq \sum_{i=1}^N \min_{\mu_{\{i\}}^\lambda} \Pi_i^\lambda(\mu_{\{i\}}^\lambda) = \bar{\Pi}^\lambda + \sum_{i=1}^N h_i (\sigma_i^2 + 1)$ .

A lower bound stems from considering a system where all customer types are pooled into a single queue which is served by a single server. In heavy traffic, the Kingman bound is tight and using (EC.19) for the single queue yields an asymptotic lower bound:  $\Pi^{\lambda*} \geq \underline{\Pi}^\lambda + o(\sqrt{\lambda})$ . The upper bound on the cost in (EC.19) implies that it is sufficient to restrict attention to capacity

portfolios with  $\sum_{F \subseteq \{1, \dots, N\}} \mu_F^\lambda \leq \sum_{i=1}^N m_i \alpha_i \lambda + O(\sqrt{\lambda})$ , as the cost of resources must be bounded above by  $\bar{\Pi}^\lambda$  and the cost of resources are bounded below by  $c$  per unit. For any such portfolio, we can lower bound the expected queue-length by pooling all customers together in to a single type and all resources into a single pool so that there is no inefficiency in the system. That is, a resource will idle only when there is no waiting customer. We can obtain a further lower bound by replacing the pool of resources by a single super-server that has a capacity equal to that of the pool and using the minimum holding cost rate per customer per unit time. This gives us the bound  $\Pi^\lambda(\mu^\lambda) \geq \sum_{F \subseteq \{1, 2, \dots, N\}} c_F \mu_F^\lambda + O\left(\frac{\sum_{i=1}^N m_i \alpha_i \lambda}{\sum_{F \subseteq \{1, 2, \dots, N\}} \mu_F^\lambda - \sum_{i=1}^N m_i \alpha_i \lambda}\right)$ . Thus, we obtain  $\sum_{F \subseteq \{1, 2, \dots, N\}} \mu_F^\lambda = \sum_{i=1}^N m_i \alpha_i \lambda + \Theta(\sqrt{\lambda})$ .<sup>7</sup> Combining this with the upper bound on the cost in (EC.21), we obtain  $\mu_F^{\lambda*} = o(\lambda)$  for all  $F \subseteq \{1, 2, \dots, N\}$  such that  $|F| \geq 2$ . To see why this is the case, suppose that for some  $F \subseteq \{1, 2, \dots, N\}$  such that  $|F| \geq 2$ , we have  $\mu_F = \Theta(\lambda)$ . Then, we would have  $\lim_{\lambda \rightarrow \infty} \sum_{F \subseteq \{1, 2, \dots, N\}} c_F \mu_F / \lambda > \sum_{i=1}^N c m_i \alpha_i = \lim_{\lambda \rightarrow \infty} \bar{\Pi}^\lambda / \lambda$ , which contradicts the fact that  $\bar{\Pi}^\lambda$  is an upper bound to the optimal cost. Thus, it follows that  $\sum_{i=1}^N \mu_{\{i\}}^{\lambda*} = \sum_{i=1}^N m_i \alpha_i \lambda + o(\lambda)$  and further for stability of each class, we must have  $\mu_{\{i\}}^{\lambda*} = m_i \alpha_i \lambda + o(\lambda)$ .

Next, we establish that  $\mu_F^{\lambda*} = O(\sqrt{\lambda})$ . Suppose,  $\mu_F^{\lambda*} = f(\lambda)$  for some  $F \subseteq \{1, 2, \dots, N\}$  such that  $|F| \geq 2$ , where  $f(\lambda)/\sqrt{\lambda} \rightarrow \infty$  as  $\lambda \rightarrow \infty$ , then we must also have  $\mu_{\{i\}}^{\lambda*} = m_i \lambda_i - g(\lambda)$  for some  $i$ , where  $g(\lambda) = \Theta(f(\lambda))$ . However, as  $\sum_{F \subseteq \{1, 2, \dots, N\}} \mu_F^{\lambda*} = \sum_{i=1}^N m_i \alpha_i \lambda + \Theta(\sqrt{\lambda})$ , and flexible resources are more expensive than dedicated ones, we obtain  $\Pi^\lambda(\mu^{\lambda*}) = \sum_{i=1}^N m_i \alpha_i c \lambda + \Theta(f(\lambda))$ , which contradicts the bound in (EC.21). Thus, we must have  $\mu_{\{i\}}^{\lambda*} = m_i \alpha_i \lambda + \hat{\mu}_{\{i\}}^* \sqrt{\lambda} + o(\sqrt{\lambda})$ , and  $\mu_F^{\lambda*} = \hat{\mu}_F^* \sqrt{\lambda} + o(\sqrt{\lambda})$  for  $F \subseteq \{1, 2, \dots, N\}$  such that  $|F| \geq 2$ . *Q.E.D.*

### EC.3. Explicit solution for a 3-class system

Using Proposition 6, we can compute

$$\mathbb{E}\hat{Q}_{min}(\infty) = \frac{1}{3\hat{\mu}_1 + 3\hat{\mu}_2 + \hat{\mu}_3} \sigma^2 m \quad (\text{EC.22})$$

$$\mathbb{E}\hat{Q}_{mid}(\infty) = \left( \mathbb{E}\hat{Q}_{min}(\infty) + \frac{1}{2\hat{\mu}_1 + 3\hat{\mu}_2 + \hat{\mu}_3} \right) \sigma^2 m \quad (\text{EC.23})$$

<sup>7</sup> For two functions  $a, b: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ , we use the notation  $a(x) = \Theta(b(x))$  to denote the existence of two positive constants  $C_1$  and  $C_2$  such that  $C_1 b(x) \leq a(x) \leq C_2 b(x)$  for all  $x \geq 0$ .

$$\mathbb{E}\hat{Q}_{max}(\infty) = \left( \mathbb{E}\hat{Q}_{mid}(\infty) + \frac{1}{\hat{\mu}_1 + 2\hat{\mu}_2 + \hat{\mu}_3} \right) \sigma^2 m. \quad (\text{EC.24})$$

Thus, we can write the total expected steady-state cost as a function of the choice of dedicated and flexible resources as

$$\begin{aligned} \hat{\Pi}(\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = & \left( \frac{3}{3\hat{\mu}_1 + 3\hat{\mu}_2 + \hat{\mu}_3} + \frac{2}{2\hat{\mu}_1 + 3\hat{\mu}_2 + \hat{\mu}_3} + \frac{1}{\hat{\mu}_1 + 2\hat{\mu}_2 + \hat{\mu}_3} \right) \sigma^2 hm \\ & + (3c\hat{\mu}_1 + 3c(1 + \Delta_2)\hat{\mu}_2 + c(1 + \Delta_3)\hat{\mu}_3). \end{aligned} \quad (\text{EC.25})$$

The optimization problem can be written as

$$\min_{\{\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3 : \hat{\mu}_2, \hat{\mu}_3 \geq 0, 3\hat{\mu}_1 + 3\hat{\mu}_2 + \hat{\mu}_3 > 0\}} \hat{\Pi}(\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3). \quad (\text{EC.26})$$

We now characterize the solution of this optimization problem.

PROPOSITION EC.1. *The solution to (EC.26) for any  $\Delta_2 > 0$  has the following property*

$$(\hat{\mu}_2^*, \hat{\mu}_3^*) = \begin{cases} (\hat{\mu}_2^* = 0, \hat{\mu}_3^* > 0) & \text{if } \Delta_3 \leq \phi_{13}(\Delta_2) \\ (\hat{\mu}_2^* > 0, \hat{\mu}_3^* > 0) & \text{if } \phi_{13}(\Delta_2) < \Delta_3 < \phi_{12}(\Delta_2) \\ (\hat{\mu}_2^* > 0, \hat{\mu}_3^* = 0) & \text{if } \phi_{12}(\Delta_2) \leq \Delta_3 < \phi_1(\Delta_2) \\ (\hat{\mu}_2^* = 0, \hat{\mu}_3^* = 0) & \text{if } \Delta_3 \geq \phi_1(\Delta_2), \end{cases} \quad (\text{EC.27})$$

where

$$(\phi_1, \phi_{12}, \phi_{13})(x) = \begin{cases} (\infty, g(x), h(x)) & x \leq 1/2, \\ (5/6, 5/6, 5/6) & 1/2 \leq x \leq 5/6, \\ (x, x, x) & x > 5/6. \end{cases} \quad (\text{EC.28})$$

where  $g$  and  $h$  are functions such that  $0 \leq h(x) \leq g(x) \leq 5x/3$  and are characterized in (EC.46)-(EC.47) in the Appendix (the thresholds are displayed in Figure 5). Further, the optimal investment levels for each of the above cases are:

1. Invest in levels 1 and 3, i.e.,  $(\hat{\mu}_2^* = 0, \hat{\mu}_3^* > 0)$ :

$$(\hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\mu}_3^*) = \begin{cases} (-\zeta^*(\eta_{min}), 0, -\eta_{min}\zeta^*(\eta_{min})) & \text{if } 0 \leq \Delta_3 < \frac{2}{7}, \\ \left(0, 0, \sigma \sqrt{\frac{6hm}{c(1+\Delta_3)}}\right) & \text{if } \Delta_3 = \frac{2}{7}, \\ (\zeta^*(\eta_{max}), 0, \eta_{max}\zeta^*(\eta_{max})) & \text{if } \frac{2}{7} < \Delta_3 < \frac{5}{6}, \end{cases} \quad (\text{EC.29})$$

where  $\zeta^*(\eta) = \sigma \sqrt{\frac{hm}{c}} \sqrt{\frac{2(9+11\eta+3\eta^2)}{(1+\eta)(2+\eta)(3+\eta)(3+\eta(1+\Delta_3))}}$  and  $\eta_{min}$  and  $\eta_{max}$  respectively denote the smallest and largest real root of

$$\Delta_3 = \frac{45 + 84\eta + 59\eta^2 + 18\eta^3 + 2\eta^4}{54 + 132\eta + 121\eta^2 + 48\eta^3 + 7\eta^4}. \quad (\text{EC.30})$$

2. Invest in all levels, i.e.,  $(\hat{\mu}_2^* > 0, \hat{\mu}_3^* > 0)$ :

$$\hat{\mu}_1^* = \sigma \sqrt{\frac{hm}{3c}} \left( -\frac{\sqrt{2}}{\sqrt{(2\Delta_2 - \Delta_3)}} + \frac{3}{\sqrt{(1 - 3\Delta_2 + \Delta_3)}} \right), \quad (\text{EC.31})$$

$$\hat{\mu}_2^* = \sigma \sqrt{\frac{hm}{3c}} \left( \frac{2\sqrt{2}}{\sqrt{(2\Delta_2 - \Delta_3)}} - \frac{1}{\sqrt{\Delta_3 - \Delta_2}} - \frac{3}{\sqrt{(1 - 3\Delta_2 + \Delta_3)}} \right), \quad (\text{EC.32})$$

$$\hat{\mu}_3^* = \sigma \sqrt{\frac{3hm}{c}} \left( -\frac{\sqrt{2}}{\sqrt{(2\Delta_2 - \Delta_3)}} + \frac{1}{\sqrt{\Delta_3 - \Delta_2}} + \frac{1}{\sqrt{(1 - 3\Delta_2 + \Delta_3)}} \right). \quad (\text{EC.33})$$

3. Invest in levels 1 and 2, i.e.,  $(\hat{\mu}_2^* > 0, \hat{\mu}_3^* = 0)$ :

$$(\hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\mu}_3^*) = \begin{cases} (-\xi^*(\gamma_{min}), -\gamma_{min}\xi^*(\gamma_{min}), 0) & \text{if } 0 \leq \Delta_2 < \frac{17}{61}, \\ \left( 0, \sigma \sqrt{\frac{13hm}{18c(1+\Delta_2)}}, 0 \right) & \text{if } \Delta_2 = \frac{17}{61}, \\ (\xi^*(\gamma_{max}), \gamma_{max}\xi^*(\gamma_{max}), 0) & \text{if } \frac{17}{61} < \Delta_2 < \frac{1}{2}. \end{cases} \quad (\text{EC.34})$$

where  $\xi^*(\gamma) = \sigma \sqrt{\frac{hm}{c}} \sqrt{\frac{18+54\gamma+39\gamma^2}{(1+2\gamma)(2+3\gamma)(3+3\gamma)(3+3(1+\Delta_2)\gamma)}}$  and  $\gamma_{min}$  and  $\gamma_{max}$  respectively denote the smallest and largest real root of

$$\Delta_2 = \frac{6 + 32\gamma + 63\gamma^2 + 54\gamma^3 + 17\gamma^4}{12 + 72\gamma + 162\gamma^2 + 162\gamma^3 + 61\gamma^4}. \quad (\text{EC.35})$$

4. Invest in level 1 only, i.e.,  $(\hat{\mu}_2^* = 0, \hat{\mu}_3^* = 0)$ :

$$(\hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\mu}_3^*) = \left( \sigma \sqrt{\frac{hm}{c}}, 0, 0 \right). \quad (\text{EC.36})$$

*Proof of Proposition EC.1.* We can write the total cost as

$$\begin{aligned} & \hat{\Pi}(\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) \\ & \equiv \mathbb{E}(\hat{Q}_{[1]}(\infty) + \hat{Q}_{[2]}(\infty) + \hat{Q}_{[3]}(\infty))h + (3\hat{\mu}_1c + 3\hat{\mu}_2c(1 + \Delta_2) + \hat{\mu}_3c(1 + \Delta_3)) \\ & = \left( \frac{3}{3\hat{\mu}_1 + 3\hat{\mu}_2 + \hat{\mu}_3} + \frac{2}{2\hat{\mu}_1 + 3\hat{\mu}_2 + \hat{\mu}_3} + \frac{1}{\hat{\mu}_1 + 2\hat{\mu}_2 + \hat{\mu}_3} \right) \sigma^2 hm + (3\hat{\mu}_1c + 3\hat{\mu}_2c(1 + \Delta_2) + \hat{\mu}_3c(1 + \Delta_3)). \end{aligned} \quad (\text{EC.37})$$

Our objective is to optimize  $\hat{\Pi}(\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3)$  over  $\hat{\mu}_1 \in \mathbb{R}$ ,  $\hat{\mu}_2 \geq 0$  and  $\hat{\mu}_3 \geq 0$ . We consider the following four cases that arise based on whether  $\hat{\mu}_2^*$  and  $\hat{\mu}_3^*$  are both positive, only one of them is positive, or both are zero. We first characterize the optimal solution in each of these cases and then discuss when each case is applicable based on the problem primitives.

**Case 1: Investing in levels 1 and 3:** We set  $\hat{\mu}_2 = 0$  and  $\hat{\mu}_3 = \eta\hat{\mu}_1$ . As  $\hat{\mu}_3 \geq 0$ , we must have  $\eta \geq 0$  if  $\hat{\mu}_1 \geq 0$ , and  $\eta \leq 0$  if  $\hat{\mu}_1 \leq 0$ . Now, optimizing on  $\hat{\mu}_1$  we obtain

$$\hat{\mu}_1^* = \pm \sigma \sqrt{\frac{hm}{c}} \sqrt{\frac{2(9 + 11\eta + 3\eta^2)}{(1 + \eta)(2 + \eta)(3 + \eta)(3 + \eta(1 + \Delta_3))}}$$

Thus, the optimization problem is equivalent to

$$\min \left( \inf_{\eta < -3} a(\eta), \inf_{\hat{\mu}_3 \geq 0} \hat{\Pi}(0, 0, \hat{\mu}_3), \inf_{\eta \geq 0} a(\eta) \right),$$

where  $a(\eta) = 2\sigma\sqrt{2hmc}\sqrt{\frac{(9+\eta(11+3\eta))(3+\eta+\eta\Delta_3)}{(1+\eta)(2+\eta)(3+\eta)}}$ , corresponding to the regions  $\hat{\mu}_1 < 0$ ,  $\hat{\mu}_1 = 0$ , and  $\hat{\mu}_1 > 0$  respectively. Using the continuity of the objective function we can equivalently rewrite this optimization problem as

$$\inf_{\eta \in \mathbb{R} \setminus [-3, 0)} a(\eta). \quad (\text{EC.38})$$

The first order optimality condition can then be written as

$$\Delta_3 = \frac{45 + 84\eta + 59\eta^2 + 18\eta^3 + 2\eta^4}{54 + 132\eta + 121\eta^2 + 48\eta^3 + 7\eta^4}.$$

Solving for  $\eta$ , we obtain the optimal investments for this case.

**Case 2: Investing in levels 1, 2 and 3:** The first order conditions yield the following set of equations:

$$3c + h \left( -\frac{1}{(\hat{\mu}_1 + 2\hat{\mu}_2 + \hat{\mu}_3)^2} - \frac{4}{(2\hat{\mu}_1 + 3\hat{\mu}_2 + \hat{\mu}_3)^2} - \frac{9}{(3(\hat{\mu}_1 + \hat{\mu}_2) + \hat{\mu}_3)^2} \right) = 0 \quad (\text{EC.39})$$

$$3c(1 + \Delta_2) + h \left( -\frac{2}{(\hat{\mu}_1 + 2\hat{\mu}_2 + \hat{\mu}_3)^2} - \frac{6}{(2\hat{\mu}_1 + 3\hat{\mu}_2 + \hat{\mu}_3)^2} - \frac{9}{(3(\hat{\mu}_1 + \hat{\mu}_2) + \hat{\mu}_3)^2} \right) = 0 \quad (\text{EC.40})$$

$$c(1 + \Delta_3) + h \left( -\frac{1}{(\hat{\mu}_1 + 2\hat{\mu}_2 + \hat{\mu}_3)^2} - \frac{2}{(2\hat{\mu}_1 + 3\hat{\mu}_2 + \hat{\mu}_3)^2} - \frac{3}{(3(\hat{\mu}_1 + \hat{\mu}_2) + \hat{\mu}_3)^2} \right) = 0 \quad (\text{EC.41})$$

Thus, multiplying (EC.41) by three and subtracting (EC.40) from it gives:

$$-3\Delta_2 + 3\Delta_3 = \frac{h}{c(\hat{\mu}_1 + 2\hat{\mu}_2 + \hat{\mu}_3)^2}. \quad (\text{EC.42})$$

Similarly, multiplying (EC.40) by two and subtracting (EC.39) combined with two times (EC.41)

from it we obtain

$$-3(\Delta_3 - 2\Delta_2) = \frac{2h}{c(2\hat{\mu}_1 + 3\hat{\mu}_2 + \hat{\mu}_3)^2}. \quad (\text{EC.43})$$

Lastly, adding (EC.39) and (EC.41) and subtracting (EC.40) from it gives

$$(1 + \Delta_3) - 3\Delta_2 = \frac{3h}{c(3(\hat{\mu}_1 + \hat{\mu}_2) + \hat{\mu}_3)^2}. \quad (\text{EC.44})$$

Solving (EC.42)-(EC.44), we obtain (EC.31)-(EC.33).

**Case 3: Investing in levels 1 and 2:** We set  $\hat{\mu}_2 = \gamma\hat{\mu}_1$  and  $\hat{\mu}_3 = 0$ . As  $\hat{\mu}_2 \geq 0$ , we must have  $\gamma \geq 0$  if  $\hat{\mu}_1 \geq 0$ , and  $\gamma \leq 0$  if  $\hat{\mu}_1 \leq 0$ . Now, optimizing on  $\hat{\mu}_1$  we obtain

$$\hat{\mu}_1^* = \pm \sigma \sqrt{\frac{hm}{c} \frac{\sqrt{18 + 54\gamma + 39\gamma^2}}{3\sqrt{(1+2\gamma)(2+3\gamma)(1+\gamma)(1+(1+\Delta_2)\gamma)}}}.$$

Thus, the optimization problem is equivalent to

$$\min \left( \inf_{\gamma < -1} b(\gamma), \inf_{\hat{\mu}_2 \geq 0} \hat{\Pi}(0, \hat{\mu}_2, 0), \inf_{\gamma \geq 0} b(\gamma) \right),$$

where  $b(\gamma) = 2\sigma\sqrt{hmc} \sqrt{\frac{(3+3(1+\Delta_2)\gamma)(18+54\gamma+39\gamma^2)}{(1+2\gamma)(2+3\gamma)}}$ , corresponding to the regions  $\hat{\mu}_1 < 0$ ,  $\hat{\mu}_1 = 0$ , and  $\hat{\mu}_1 > 0$  respectively. Using the continuity of the objective function we can equivalently rewrite this optimization problem as

$$\inf_{\gamma \in \mathbb{R} \setminus [-1, 0]} b(\gamma). \quad (\text{EC.45})$$

The first order optimality condition can then be written as

$$\Delta_2 = \frac{6 + 32\gamma + 63\gamma^2 + 54\gamma^3 + 17\gamma^4}{12 + 72\gamma + 162\gamma^2 + 162\gamma^3 + 61\gamma^4}.$$

Solving for  $\gamma$ , we obtain the optimal investments for this case.

**Case 4: Investing in level 1 only:** Setting  $\hat{\mu}_2 = \hat{\mu}_3 = 0$  in the objective function, the first order conditions imply that  $\hat{\mu}_1^* = \sigma\sqrt{\frac{hm}{c}}$ , which is the optimal investment for this case.

**Conditions for optimality of the above cases.** Consider the parameter regime in which the capacity portfolio of Case 2 is real-valued and non-negative. In this case, the convexity of the objective function implies that this must be the optimal solution. Thus, if the solution characterized by equations (EC.42)-(EC.44) is such that  $\hat{\mu}_2^* \geq 0$  and  $\hat{\mu}_3^* \geq 0$  then the solution in Case 2 is optimal. Using these values, for any  $\Delta_2 < 1/2$ , define the function  $h(\Delta_2)$  to be the value of  $\Delta_3$  such that

$\hat{\mu}_2^* = 0$  and the function  $g(\Delta_2)$  to be the value of  $\Delta_3$  such that  $\hat{\mu}_3^* = 0$ . That is,  $h(x)$  is defined as the unique solution to

$$\frac{2}{\sqrt{x - \frac{h(x)}{2}}} - \frac{3}{\sqrt{1 - 3x + h(x)}} - \frac{1}{\sqrt{-x + h(x)}} = 0 \quad (\text{EC.46})$$

and  $g(x)$  is defined as the unique solution to

$$\frac{1}{\sqrt{x - \frac{g(x)}{2}}} - \frac{1}{\sqrt{1 - 3x + g(x)}} - \frac{1}{\sqrt{-x + g(x)}} = 0. \quad (\text{EC.47})$$

Further, note that for fixed  $\Delta_2$ ,  $\hat{\mu}_2^*$  is increasing in  $\Delta_3$ , and  $\hat{\mu}_3^*$  is decreasing in  $\Delta_3$ . It thus follows that for  $\Delta_3 > g(\Delta_2)$ ,  $\hat{\mu}_3^* = 0$  and  $\hat{\mu}_2^* > 0$ , so that Case 3 is optimal. Further, if  $h(\Delta_2) < \Delta_3 < g(\Delta_2)$ ,  $\hat{\mu}_3^* > 0$  and  $\hat{\mu}_2^* > 0$ , so that Case 2 is optimal, and if  $\Delta_3 < h(\Delta_2)$ ,  $\hat{\mu}_3^* > 0$  and  $\hat{\mu}_2^* = 0$ , so that Case 1 is optimal. Noting that for a fixed  $\Delta_3$ ,  $\hat{\mu}_2^*$  is decreasing in  $\Delta_2$  and  $h(1/2) = g(1/2) = 5/6$ , we obtain that for a fixed  $\Delta_3 \leq 5/6$  if  $\Delta_2 > h^{-1}(\Delta_3)$  then Case 1 is optimal.

Further, we observe that if  $\Delta_3 = 5/6$  and  $\Delta_2 = 1/2$  then  $\hat{\mu}_3^* = 0$  and  $\hat{\mu}_2^* = 0$ . Thus, we obtain that for  $\Delta_2 \geq 1/2$  and  $\Delta_3 \geq 5/6$ , Case 4 is optimal.

Next, we prove that  $h(x) \leq g(x)$ . It is easy to see that for  $x \leq 1/2$  and  $y > x$ , we have  $\frac{1}{\sqrt{1-3x+y}} \leq \frac{1}{\sqrt{-x+y}}$ . Thus, we obtain

$$2 \left( \frac{1}{\sqrt{x - \frac{h(x)}{2}}} - \frac{1}{\sqrt{1 - 3x + h(x)}} - \frac{1}{\sqrt{-x + h(x)}} \right) \leq \frac{2}{\sqrt{x - \frac{h(x)}{2}}} - \frac{3}{\sqrt{1 - 3x + h(x)}} - \frac{1}{\sqrt{-x + h(x)}} = 0.$$

Noting that the left hand side of the expression above is increasing in  $h(x)$  and noting that  $g(x)$  solves (EC.47), we obtain that  $h(x) \leq g(x)$ .

Finally, we prove that  $g(x) \leq 5/3x$ . We can rewrite  $g(x) = 5/3x - \delta_g(x)$ , thus  $\delta_g(x)$  solves

$$\sqrt{3} \left( \frac{1}{\sqrt{3 - 3\delta_g(x) - 4x}} - \frac{\sqrt{2}}{\sqrt{3\delta_g(x) + x}} + \frac{1}{\sqrt{-3\delta_g(x) + 2x}} \right) = 0.$$

We observe that for a fixed  $x$  the left-hand-side is increasing in  $\delta_g(x)$ . Further, we observe

$$\frac{d}{dx} \sqrt{3} \left( \frac{1}{\sqrt{3 - 4x}} - \frac{\sqrt{2}}{\sqrt{x}} + \frac{1}{\sqrt{2x}} \right) = \frac{1}{4} \sqrt{3} \left( \frac{8}{(3 - 4x)^{3/2}} + \frac{\sqrt{2}}{x^{3/2}} \right) > 0, \text{ for all } x \in (0, 1/2).$$

Thus, noting that  $\delta_g(1/2) = 0$ , we obtain that  $\delta_g(x) > 0$  for  $x < 1/2$ , which implies that  $g(x) \leq 5x/3$ .

*Q.E.D.*