Staffing Tandem Queues with Impatient Customers – Application in Financial Service Operations

Jianfu Wang[1], Hossein Abouee-Mehrizi[2], Opher Baron[3], Oded Berman[3]
[1]: Nanyang Business School, Nanyang Technological University, Singapore
[2]: Department of Management Sciences, University of Waterloo, Canada
[3]: Rotman School of Management, University of Toronto, Canada

We study a Markovian two-station tandem queueing network with impatient customers, applying it to the financial service process in investment banks. Since the 2008 Financial Crisis, deals negotiated by the front office are required by regulations to be reviewed internally by a control function to control risk taking, and deals may be called off by clients at any time. We study the staffing policy of financial service operations using the service throughput as its performance measure. Queueing networks with abandonment are common in many industries, e.g., call centers and healthcare. Therefore, their management has received much attention. However, the resulting queueing model is a level-dependent quasi-birth-and-death (LDQBD) process - a model considered intractable because previous numerical methods for solving LDQBD processes may not converge to the correct value. We analyze an equivalent last-come-first-serve system to develop a recursive relation in our LDQBD process, reducing the problem to solving quadratic matrix equations, where efficient and exact numerical methods exist. We further simplify the analysis by combining the recursive renewal reward theorem with Queueing and Markov chain decomposition, so that only one quadratic matrix equation must be solved. We develop an exact numerical method to calculate the steady state probability distribution of a tandem queueing network with abandonment. We provide the first exact analysis of performance measures of queueing networks with abandonment. For the financial service application, we find the optimal staffing policy with the minimum number of staff required to achieve a service throughput target; we show that if the service rate of the control function is below a cutoff point, banks can reduce the total staff needed by assigning more staff to the front office than the benchmark rule of assigning identical capacity to both stations. If the service rate is at or above the cutoff point, the benchmark assignment rule is close to optimal, and assigning more staff to the control function may slightly reduce the head count required. Our results provide insights and guidelines for financial service operations. Our method is applicable to the analysis of queueing networks with abandonment under settings with diverse features and in various service disciplines.

Key words: financial service operations, tandem queue, impatient customers, abandonment, staffing

1. Introduction

We study the staffing problem of a two-station tandem queueing network with abandonment. Tandem queueing networks, where customers need to visit several stations in sequence, abound in the modern economy. Examples range from call centers, where customers talk to general call-takers before being transferred to specialists (see, e.g., Gans et al. 2003), to hospital emergency rooms, where patients are admitted by triage nurses before going on to have a number of medical tests and procedures (see, e.g., Zayas-Cabán et al. 2013), to cost-efficient blood screenings, where a less sensitive but inexpensive test is conducted before a more sensitive and expensive one (see, e.g., Bar-Lev et al. 2013). In these applications, abandonment (i.e., leaving the queue before or while being served) is an important phenomenon. For instance, in call centers, customers may hang up before reaching an agent. In hospital emergency rooms, patients may leave because they can obtain treatment elsewhere; in extreme cases, they may die during the wait. And in medical tests, blood samples need to be processed within a certain time before they perish.

There are clearly many other possibilities; in this paper we apply a tandem queueing network with abandonment to financial service operations.

Application in Financial Service Operations: The 2008 Financial Crisis was a wake-up call to central banks, investment banks, and banking regulators. To make banks more resilient and restore confidence in the banking system, the governments of many countries have revised their regulations. The more stringent now require investment banks to clearly separate business and control functions. The control function is empowered to conduct independent assessments of business decisions and to continuously monitor risk taking at both transaction and portfolio levels.

The following financing transaction illustrates the process. A client has mandated a bank to arrange financing for a project. The bank's front office will negotiate trade details, such as pricing and loan covenants, with the client and conduct primary due diligence on the client. Once the trade details are drafted, the transaction will be presented to the control function for an internal review, a process including profitability analysis, risk assessment, compliance, legal documentation, etc. If

the deal's estimated return is commensurate to the risk portfolio of the bank, the transaction is approved by the control function; at that point but not before, it may be executed. If not approved by the control function, the transaction will be turned down. This process includes a tandem service structure, with the front office acting as the upstream station and the control function acting as the downstream station.

Clients can cancel trade requests at any time during the process for reasons such as: (i) other banks offer more competitive pricing or faster execution; (ii) clients pull out for strategic reasons; or (iii) uncontrollable events, like natural disasters, occur. These cancellation causes are independent of the bank's service operations and thus, are independent of the stage the trade is at in the system. Furthermore, canceled deals do not affect the bank's relationship with the client. In the first situation, as long as the project is financed, the client is satisfied. In the other two situations, the causes for cancelling deals are out of the bank's control. Thus, the bank has no reasons to associate canceled deals with any costs, such as loss of goodwill.

Of course, canceled deals will not generate profit for the bank, and in a market with abundant liquidity, banks must compete with each other. Being able to execute a transaction quickly is a competitive advantage, so the speed of the bank's service process matters. This means the optimization of both the staffing in the front office and the control function is critical. The two functions need to work together to maximize their efficiency in conducting due diligence of transactions while not undermining the rigor of review. Accordingly, the service throughput, i.e., the number of deals reviewed by the control function, becomes a main focus of the bank. Both approved and rejected deals are considered contributions to the bank's risk adjusted profitability.

To develop insights into this type of service system, we model it as a two-station Markovian tandem queueing network with abandonment. Deals arrive randomly at the bank, and both upstream, front office, and downstream, control function, are modeled as multi-server queues. Clients may make independent deal cancellation decisions at any time. We model the upstream station with an infinite buffer. For tractability, and as is common in many applications, we consider a finite buffer

size between the two stations. To capture a system with infinite buffer size between the stations, we would have to allow the buffer to be so large that it has no economic effect on the system. A bank's competitiveness in the market determines its clients' abandonment rate – abandonment occurs more often when the bank's competitors are able to offer a lower interest rate sooner. The service level measure of interest is the system's *service throughput* (ST), defined as the number of deals successfully negotiated and reviewed, not including canceled deals.

The staffing policy in a tandem service system is composed of two elements: (i) the total number of servers in the system, and (ii) the *assignment rule* that assigns the staff to the stations. We look at and compare two assignments rules. The first is an easily calculable benchmark assignment rule that assigns identical capacities to both stations, and the second is the optimal rule that maximizes ST for the same staffing level. In what follows, we investigate how the corresponding staffing policies change with the demand rate for the system.

The specific managerial questions we consider are:

1. Given that there are $N \geq 2$ servers available, how can we assign them to a two-station tandem queueing network with abandonment to maximize the ST?

2. What is the minimum number of servers needed to achieve a ST target in such a network?

Methodology: We develop an exact numerical method to derive various service level measures of general tandem queueing systems with impatient customers. Because the abandonment rate depends on the number of customers in the queue, the system falls into the category of level-dependent quasi-birth-and-death (LDQBD) processes (see, e.g., Kharoufeh 2011, and references therein), hitherto considered intractable. It is computationally formidable to solve LDQBD processes using matrix analytic methods (see, e.g., Chapter 12 in Latouche and Ramaswami 1999), and the numerical method may not converge to the correct value. Following a different line of thinking, we develop a *recursive relation* by analyzing an equivalent last-come-first-serve system, so that the problem boils down to solving quadratic matrix equations, and the standard techniques of matrix analytic methods can be applied to make an exact analysis of the system. We can further simplify

the derivation of various service level measures by combining queueing and Markov chain decomposition (see, e.g., Abouee-Mehrizi et al. 2012, Wang et al. 2015) with a renewal reward theorem based approach, thereby extending the recursive renewal reward technique (see, e.g., Gandhi et al. 2014) to more general Markov chains.

Broadly stated, quantitative models in general and queueing models more specifically are analyzed using one of three methods: exact solutions, approximations, or simulations. An example of exact analysis is the closed form solution proposed by Jackson (1963) for product form queueing networks. Both transform analysis and matrix analytic methods that are common tools in analysis of queueing systems typically lead to exact numerical solutions. An important advantage of exact analysis methods is that they work for all system parameter choices, including a large or small number of servers, large or small buffer sizes, fast or slow services. An example of approximations is Baron and Milner's (2009) analysis of the $M/M/n+M$ model. Common approximations methods in queueing are fluid and diffusion approximations; both are typically more accurate for systems with many servers or with a heavy load. Finally, analysis using simulations for queueing systems is numerical and is often done when the system is too complex to be amenable to exact or approximated solutions (i.e., operating room scheduling in hospitals). Simulations methods are typically time consuming and provide less guarantee of accuracy (especially as the systems in question are complex). Each method and solution has its place and offers certain advantages which must be weighed before selecting one. Complicating the issue, some systems, for example, the $M/M/n+M$, are often analyzed using different methods; for example, a simulation may be used to demonstrate the accuracy of an approximation.

Within this context, our paper provides the first exact analysis of a queueing network model with abandonment. As such systems are likely too complex to have a closed form solution, several authors suggest using approximations. For example, Zychlinski et al. (2017) use fluid approximation to analyze tandem queues with blocking, while Armony et al. (2017) use diffusion approximation to analyze a tandem healthcare system with flexible servers and abandonment. Our theoretical

contribution is to develop an efficient numerical exact solution for a Markovian two-station tandem queues with abandonment and blocking. We discuss how our solution can accommodate different assumptions of the network and its structure and demonstrate its applications in Section 5.

Results: Using the exact numerical method developed, we provide insight into a bank's financial service operations with a focus on its ST performance. First, we establish an upper bound of the system's ST under any assignment rule. This upper bound can be shown to be a piecewise concave function of the number of staff in the front office. Second, we define an easily calculable benchmark assignment rule that assigns identical capacities to both stations, and prove its optimality when the abandonment rate is small. We then search for the optimal assignment rule of N servers that results in the highest ST over all possible assignments of N servers to both stations. We observe that the ST as a function of the number of staff in the front office is concave; hence, the optimal assignment is unique.

To answer the first managerial question, we perform an exhaustive search over all possible assignments of N servers in two stations, to find the optimal assignment with the highest ST. Next, under the assumption of a fixed head count, N, we compare the optimal assignment rule with the benchmark assignment rule. We observe three operational regions in the downstream control function's service rate μ_2. When μ_2 is at a critical cutoff point, the benchmark assignment rule is optimal. When μ_2 is lower than the cutoff point, the optimal assignment rule assigns more servers to the upstream front office, improving the system's ST by up to 9% in our numerical results. When μ_2 is greater than the cutoff point, the benchmark assignment rule deviates slightly from the optimal one and its ST is close to that of the optimal assignment rule.

To answer the second managerial question, we find the minimum total number of servers needed to achieve a given percentage (we use 95%) of the ST upper bound, under both the optimal and the benchmark assignment rules. We compare the staffing policies based on two assignment rules, again finding three operational regions in the control function's service rate. When μ_2 is at a cutoff point, the staffing policy based on the benchmark assignment rule is optimal. When μ_2 is lower

than the cutoff point, the optimal assignment rule can save the bank a significant number of servers, compared to the benchmark assignment rule, by assigning more staff to the front office. When μ_2 is greater than the cutoff point, the optimal assignment rule can save a few servers compared to the benchmark assignment rule (by assigning more to the downstream control function). However, this saving seems less significant.

For firms operating a tandem queueing system with abandonment, our numerical results provide useful guidelines. When the control function's processing rate μ_2 is at or above the cutoff point, the benchmark assignment rule is almost optimal and provides great operational simplicity. When μ_2 is below this cutoff point, it is critical for the firms to identify the optimal assignment rule, as this will save a significant amount of staffing cost.

We note that our methodology enables derivations of various relevant service level measures in similar queueing networks applications with diverse features, like different abandonment rates during waiting and service, direct departure from the upstream station, external arrival at the downstream station, cross trained servers, etc. We discuss the applicability of the methodology in other real-world problems in Section 5.

Brief Literature Review: Staffing problems in service systems, such as call centers, have been studied extensively in the literature, but most previous work has either focused on network systems ignoring abandonment (see, e.g., Chen and Yao 2001, and references therein) or considered abandonment in a single-stage queue. For example, queueing models with abandonment have been modeled using stochastic calculus (see, e.g., Boxma et al. 2014, and references therein), and asymptotic analysis (see, e.g., Ward and Glynn 2005, Baron and Milner 2009, and references therein). This work focuses on single-stage queueing systems that are not capable of representing the complex queueing systems in today's service industry. Staffing a single-stage queue only requires the total number of servers, but staffing a queueing network requires assigning these servers to different stations.

Armony et al. (2017) use diffusion approximation to study a healthcare system with an upstream Intensive Care Unit (ICU) and a downstream Step Down Unit (SDU), where critical patients

arriving at the ICU may abandon and the SDU has no waiting room for semi-critical patients who have been served by the ICU. In this application, the ICU beds are flexible, so semi-critical patients can be served in ICU, but critical patients have preemptive priority over semi-critical customers in the ICU. We will show that our model can be modified to cover the ICU-SDU system and our method can be directly applied to an exact analysis of their system.

The paper proceeds as follows. In Section 2, we define the model and provide some preliminary information. We demonstrate our method of developing a recursive relation in Section 3, using the busy period of a multi-server queue with impatient customers, where customers abandon during waiting or service, as an example. The staffing problems and managerial insights of banks' financial service operations based on the numerical results are discussed in Section 4. We discuss the applicability of our model to other real-world problems in Section 5. All proofs not in the text are in the Appendix.

2. Model

In this section, we define the Markov chain (MC) of our model, discuss notation, explain the distribution of the number of customers at Station 1, and construct a level-dependent one-step transition matrix for the MC.

2.1. Model Description and Preliminaries

We consider a two-station tandem queue with Station 1 as the upstream station and Station 2 as the downstream one as depicted in Figure 1. Each station is a multi-server queue. Let n_i be the number of servers at Station i, and μ_i be the service rate of Station i's servers, for $i = 1, 2$. Similar to Reed and Yechiali (2013), we assume Station 1 has a waiting room of infinite size, and Station 2 has a waiting room with $m < \infty$ spots. Note that by letting m approach ∞, we can approximate a general two-station tandem queue with infinite waiting rooms in both stations.

Customers arrive at Station i following a Poisson process with rate λ_i, for $i = 1, 2$. We call customers in Station i, whether waiting or receiving service, Station i customers, for $i = 1, 2$. Upon a customer's arrival at Station 1, if any of the n_1 servers of Station 1 is free, the arriving customer

immediately enters service. Otherwise, she waits at Station 1. Customers who finish service in Station 1 go to Station 2 with probability (w.p.) p. When there is either an external arrival or an arrival at Station 2 from Station 1, if Station 2 has a server available, the customer enters service immediately. Otherwise, if all servers are busy but Station 2's waiting room has some spots available, she will wait there for a Station 2 server; if the waiting room is full, this customer balks (i.e., leaves without joining the waiting room).

The customers are considered impatient. We model customers' patience in Station i's waiting room, Θ_{iw}, as an exponentially distributed random variable with parameter θ_{iw}, for $i = 1, 2$. Once a customer's waiting time passes her patience threshold Θ_{iw}, she abandons (i.e., leaves without being served). Moreover, customers may abandon while in service. Similar to customer patience in the waiting room, we model customer patience during service in Station i, Θ_{is}, as an exponentially distributed random variable with parameter θ_{is}, for $i = 1, 2$. Once a customer's service time passes Θ_{is}, she abandons service in Station i.

Here, we define different abandonment rates from different parts of the tandem queueing system to allow flexibility in modeling the abandonment behavior. When $\theta_{1s} = \theta_{2s} = \theta_{1w} = \theta_{2w} = 0$, all customers have infinite patience, and they wait until obtaining service, so the system operates like a tandem queue without abandonment. Then, the departure process from Station 1 is Poisson in stationarity (see, e.g., Adan and Resing 2001), and this makes the tandem queueing system a relatively simple Jackson network. When $\theta_{1s} = \theta_{2s} = 0$ and $\theta_{1w} = \theta_{2w} = \infty$, this tandem queueing system operates as a loss system with no waiting room - an arriving customer leaves immediately if no servers are available, as represented by a finite two-dimensional MC whose exact solution is straightforward.

Let $Q_i(t)$, $i = 1, 2$ be the random variable denoting the total number of customers in Station i at time t. Given the number of servers at both stations, the process $(Q_1(t), Q_2(t))$ is a continuous time MC. Let π_{q_1,q_2} denote the steady state probability that this MC is in state (q_1, q_2). Figure 2 illustrates the MC of the tandem queueing system with $n_1 = n_2 = 2$ and $m = 1$.

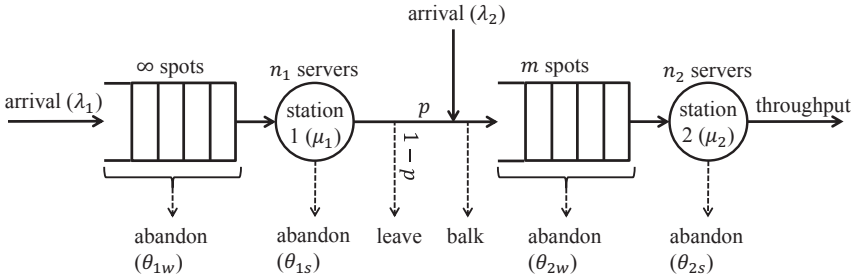

Figure 1 **Tandem queueing system with abandonment and balking.**

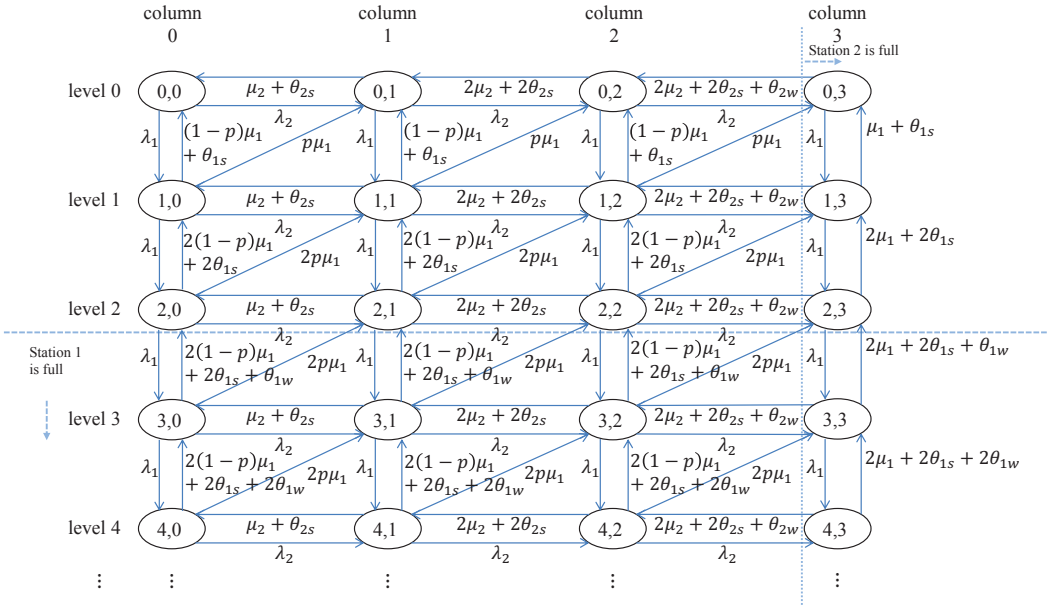

Figure 2 **The (q_1, q_2) MC of the general tandem queueing system with $n_1 = n_2 = 2$ and $m = 1$.**

Note that the assumption of a finite waiting room with size $m < \infty$ at Station 2 is essential to form a one-dimensional infinite MC, like the one in Figure 2, instead of a more difficult to analyze two-dimensional infinite MC.

For convenience, we call states $\{(q_1, q_2) | q_1 = i\}$ states at *level* i and let $\eta_i = \sum_{q_2=0}^{n_2+m} \pi_{i,q_2}$ be the steady state probability that the system is at level i of the MC. We call states $\{(q_1, q_2) | q_2 = j\}$ in the MC states at *column* j, and we let $\kappa_j = \sum_{q_1=0}^{\infty} \pi_{q_1,j}$ be the steady state probability that the system is at column j of the MC. Because Station 2 has a finite waiting room, as illustrated in Figure 2, the MC has a finite number of columns and an infinite number of levels.

2.2. Distribution of the Number of Station 1 Customers

Deriving the distribution of Station 1 customers Q_1 is immediate from the following observation.

Observation 1 *From an outside observer's point of view, Station 1 is a multi-server queue with impatient customers abandoning the waiting room at rate θ_{1w} and abandoning service at rate θ_{1s}.*

Observation 1 can be understood from Figure 2; every state (q_1, q_2) of the MC has a transition rate of λ_1 to the upper level q_1+1, and a transition rate of $\theta_{1w}\max(q_1-n_1,0)+(\mu_1+\theta_{1s})\min(q_1,n_1)$ to the lower level q_1-1; both are independent of q_2. We can now write the detailed balance equations between two adjacent levels as

$$\lambda_1 \sum_{q_2=0}^{n_2+m} \pi_{q_1-1,q_2} = (\theta_{1w}\max(q_1-n_1,0)+(\mu_1+\theta_{1s})\min(q_1,n_1)) \sum_{q_2=0}^{n_2+m} \pi_{q_1,q_2} \text{ for } q_1 = 1,2,\ldots. \quad (1)$$

Substituting the probability of having i Station 1 customers in the system $\eta_i = \sum_{q_2=0}^{n_2+m} \pi_{i,q_2}$ into (1) results in the balance equations of a multi-server queue with impatient customers abandoning the waiting room at rate θ_{1w} and abandoning service at rate θ_{1s}. The steady state probability distribution of such system is

$$\eta_i = \eta_0 \prod_{k=1}^{i} \frac{\lambda_1}{\theta_{1w}\max(k-n_1,0)+(\mu_1+\theta_{1s})\min(k,n_1)} \text{ for } i = 0,1,2,\ldots, \quad (2)$$

where

$$\eta_0 = \left(1+\sum_{i=1}^{\infty}\prod_{k=1}^{i}\frac{\lambda_1}{\theta_{1w}\max(k-n_1,0)+(\mu_1+\theta_{1s})\min(k,n_1)}\right)^{-1}.$$

Note that a special case of the multi-server queue with impatient customers, where customers do not abandon during service, i.e., $\theta_{1s}=0$, has been studied recently in the asymptotic region when the arrival rate and the number of servers grow to infinity; see, e.g., Garnett et al. (2002), Whitt (2004), and Baron and Milner (2009).

Deriving the distribution of the number of Station 2 customers Q_2 is more complicated than deriving the distribution of Q_1. In Section 2.3, we first construct the one-step transition matrix - a preliminary step in the Matrix Analytic Method, a building block of our method. We then introduce our method by characterizing the distribution of Q_2, i.e., $P\{Q_2=j\}=\kappa_j$, as an example in Section 3. In Appendix A1.6, we establish that our method works for a broad range of service level measures.

2.3. One-step Transition Matrices

Let $v(q_1, q_2)$ be the total rate at which the system leaves state (q_1, q_2). Then,

$$v(q_1,q_2) = \lambda_1 + \lambda_2 + \mu_1 \min(q_1, n_1) + \mu_2 \min(q_2, n_2) + \theta_{1s} \min(q_1, n_1) + \theta_{1w} \max(q_1 - n_1, 0)$$
$$+\theta_{2s} \min(q_2, n_2) + \theta_{2w} \max(q_2 - n_2, 0) \text{ for } q_1 = 0, 1, \ldots \text{ and } q_2 = 0, 1, \ldots, n_2 + m, \quad (3)$$

After spending an $\exp(v(q_1, q_2))$ time in state (q_1, q_2), the system will move to one of the adjacent states with a certain probability. Consider state $(3, 1)$ in Figure 2, for example. After spending an $\exp(\lambda_1 + 2\mu_1 + \mu_2 + 2\theta_{1s} + \theta_{1w} + \theta_{2s})$ time in this state, the system will move to state $(4, 1)$, if an arrival occurs at Station 1, w.p. $\frac{\lambda_1}{v(3,1)}$; state $(3, 2)$, if an arrival occurs at Station 2, w.p. $\frac{\lambda_2}{v(3,1)}$; state $(3, 0)$ if a Station 2 customer finishes service, w.p. $\frac{\mu_2}{v(3,1)}$, or abandons, w.p. $\frac{\theta_{2s}}{v(3,1)}$; state $(2, 2)$ if a Station 1 customer finishes service and needs service from Station 2, w.p. $\frac{2p\mu_1}{v(3,1)}$; or state $(2, 1)$ if any Station 1 customer finishes service and leaves directly, w.p. $\frac{2(1-p)\mu_1}{v(3,1)}$, or abandons, w.p. $\frac{2\theta_{1s}+\theta_{1w}}{v(3,1)}$. These are the *one-step transition probabilities*.

We now express the one-step transition probabilities in matrix form. Note that transitions from level q_1 can only bring the system into levels $q_1 + 1$ (after arrivals), q_1 (after service completion at or upon abandonment from Station 2), or $q_1 - 1$ (after service completion at or upon abandonment from Station 1). Let matrices $\mathbf{A}_0^{(q_1)}$, $\mathbf{A}_1^{(q_1)}$, and $\mathbf{A}_2^{(q_1)}$ represent, respectively, the one-step transition matrices at level q_1 $(q_1 = 0, 1, 2, \ldots)$ due to (i) arrivals, (ii) service completion at or abandonment from Station 2, (iii) service completion at or abandonment from Station 1. Also note that at any level q_1, there are $n_2 + m + 1$ states, so the transition matrices from level q_1 (to levels $q_1 - 1$, q_1, or $q_1 + 1$) are all of size $(n_2 + m + 1) \times (n_2 + m + 1)$. In any matrix $\mathbf{X}$, we use $[\mathbf{X}]_{ij}$ to represent the entry in its i^{th} row and j^{th} column. Then,

$$\left[\mathbf{A}_0^{(q_1)}\right]_{ij} = \begin{cases} \frac{\lambda_1}{v(q_1,i-1)} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}, \quad \left[\mathbf{A}_1^{(q_1)}\right]_{ij} = \begin{cases} \frac{\mu_2 \min(i-1,n_2) + \theta_{2s} \min(i-1,n_2) + \theta_{2w} \max(i-1-n_2,0)}{v(q_1,i-1)} & \text{if } i-1 = j \\ \frac{\lambda_2}{v(q_1,i-1)} & \text{if } i+1 = j \\ 0 & \text{otherwise} \end{cases},$$

and

$$\left[\mathbf{A}_2^{(q_1)}\right]_{ij} = \begin{cases} \frac{(1-p)\mu_1 \min(q_1,n_1)+\theta_{1s}\min(q_1,n_1)+\theta_{1w}\max(q_1-n_1,0)}{v(q_1,i-1)} & \text{if } i < n_2+m+1 \text{ and } i=j \\ \frac{p\mu_1\min(q_1,n_1)}{v(q_1,i-1)} & \text{if } i < n_2+m+1 \text{ and } i+1=j \\ \frac{\mu_1\min(q_1,n_1)+\theta_{1s}\min(q_1,n_1)+\theta_{1w}\max(q_1-n_1,0)}{v(q_1,i-1)} & \text{if } i=n_2+m+1 \text{ and } i=j \\ 0 & \text{otherwise} \end{cases}.$$

Note that $\mathbf{A}_0^{(q_1)}$, $\mathbf{A}_1^{(q_1)}$, and $\mathbf{A}_2^{(q_1)}$ are all one-step transition probability matrices from level q_1, so the sum of each row of all three matrices is 1; i.e., $\mathbf{A}_0^{(q_1)}\overrightarrow{1}+\mathbf{A}_1^{(q_1)}\overrightarrow{1}+\mathbf{A}_2^{(q_1)}\overrightarrow{1}=\overrightarrow{1}$, where $\overrightarrow{1}$ is a one column vector of size n_2+m+1.

3. Level-Dependent Quasi-Birth-and-Death Process

The analysis for Station 2 customers is challenging because the MC in Figure 2 is a level-dependent quasi-birth-and-death (LDQBD) process. The common approach to solving LDQBD processes involves deriving the *first passage probability matrix* from level q_1 to level q_1-1, $\mathbf{G}^{(q_1)}$, where $\left[\mathbf{G}^{(q_1)}\right]_{ij}$ represents the probability that given the MC starts from state (q_1,i), it first reaches level q_1-1 at state (q_1-1,j), for $q_1=0,1,2,\ldots$. $\mathbf{G}^{(q_1)}$ needs to be derived iteratively from $\mathbf{G}^{(q_1+1)}$ using

$$\mathbf{G}^{(q_1)} = \left(\mathbf{I}-\mathbf{A}_1^{(q_1)}-\mathbf{A}_0^{(q_1)}\mathbf{G}^{(q_1+1)}\right)^{-1}\mathbf{A}_2^{(q_1)} \tag{4}$$

and $\mathbf{G}^{(\infty)}$ is approximated by $\mathbf{G}^{(M)}$ for a large M. This approach is, as Latouche and Ramaswami (1999) put it, "somewhat arbitrary." In this section, we establish a recursive relation for $\mathbf{G}^{(q_1)}$ for $q_1 > n_1$, so that $\mathbf{G}^{(q_1)}$ can be expressed as the solution of a quadratic matrix equation. This derivation leads to one of our paper's main contributions: developing a feasible way to solve an LDQBD process, hitherto considered intractable.

In Section 3.1, we use the example of the first passage time of the MC to level n_1, starting from level n_1+1, to introduce our method. Then, in Section 3.2, we show how this method can be used to derive $\mathbf{G}^{(q_1)}$ for $q_1 > n_1$.

3.1. Method of Analysis: Creating a Recursive Relation

Let $T^{(q_1)}$ denote the time periods after the MC leaves level q_1 until it reaches level $q_1 - 1$, for $q_1 > n_1$. Then, $T^{(n_1+1)}$ represents the time period the MC stays in the subspace $\{(q_1, q_2) \,|q_1 > n_1\}$. Let $L_{T^{(n_1+1)}}$ denote the length of $T^{(n_1+1)}$. Let $F_{T^{(n_1+1)}}(x)$ be the cumulative distribution function of $L_{T^{(n_1+1)}}$ and $\tilde{L}_{T^{(n_1+1)}}(s)$ be its Laplace Transform (LT). While $\tilde{L}_{T^{(n_1+1)}}(s)$ has been derived by Jouini and Roubos (2014), we derive it here to illustrate the general idea of our method: generating a *recursive relation* using the Last-Come-First-Serve scheduling rule in the system. Importantly, due to the memoryless property of the exponentially distributed patience, the order of service does not change the length of $T^{(n_1+1)}$.

PROPOSITION 1. *The Laplace Transform of $L_{T^{(n_1+1)}}$, $\tilde{L}_{T^{(n_1+1)}}(s)$, is the solution of*

$$\begin{aligned}\tilde{L}_{T^{(n_1+1)}}(s) &= \frac{\theta_{1w} + n_1(\mu_1 + \theta_{1s})}{\lambda_1 + \theta_{1w} + n_1(\mu_1 + \theta_{1s}) + s} \\ &+ \frac{\lambda_1}{\lambda_1 + \theta_{1w} + n_1(\mu_1 + \theta_{1s}) + s} \int_0^\infty P\{\Theta_{1w} \le L_{T^{(n_1+1)}} \,|L_{T^{(n_1+1)}} = x\}\, e^{-sx} dF_{T^{(n_1+1)}}(x) \\ &+ \frac{\lambda_1}{\lambda_1 + \theta_{1w} + n_1(\mu_1 + \theta_{1s}) + s} \int_0^\infty P\{\Theta_{1w} > L_{T^{(n_1+1)}} \,|L_{T^{(n_1+1)}} = x\}\, e^{-sx} \tilde{L}_{T^{(n_1+1)}}(s)\, dF_{T^{(n_1+1)}}(x).\end{aligned} \tag{5}$$

Proof of Proposition 1 *Say a $T^{(n_1+1)}$ starts; i.e., the MC is at level $n_1 + 1$. All servers at Station 1 are busy and a customer (call her "A") is waiting. After an $\exp(\lambda_1 + \theta_{1w} + n_1(\mu_1 + \theta_{1s}))$ time interval (with Laplace Transform $\frac{\lambda_1 + \theta_{1w} + n_1(\mu_1+\theta_{1s})}{\lambda_1 + \theta_{1w} + n_1(\mu_1+\theta_{1s}) + s}$), three events can happen at Station 1: arrival, abandonment, or completion 1. If the next event is abandonment or completion 1 (w.p. $\frac{\theta_{1w} + n_1(\mu_1+\theta_{1s})}{\lambda_1 + \theta_{1w} + n_1(\mu_1+\theta_{1s})}$), the MC enters level n_1, and the $T^{(n_1+1)}$ ends (the LT is 1 in this case). Therefore, in this instance, the first line in (5) gives the LT of the $T^{(n_1+1)}$.*

If the next event is an arrival (w.p. $\frac{\lambda_1}{\lambda_1 + \theta_{1w} + n_1(\mu_1+\theta_{1s})}$), a new customer (call her "B") joins the queue at Station 1, so there are now two customers waiting for Station 1. We use a technique similar to that used to derive the busy period of an $M/G/1$ queue, and we use the Last-Come-First-Serve rule to establish a recursive relation. According to the Last-Come-First-Serve rule, customer A will be considered when no other customers are waiting. We imagine putting customer A into a separate room and temporarily ignoring her until no other customers are waiting. The important

observation is that the time period from now (customer B is the only waiting customer in the queue when customer A is ignored) until no other customers are waiting has exactly the same distribution as $T^{(n_1+1)}$, which also starts with one customer waiting and ends when no other customers are waiting. We call this time period the $T^{(n_1+1)}$ initiated by customer B.

When the $T^{(n_1+1)}$ initiated by customer B ends, we need to consider customer A again. At this moment, customer A may have abandoned, if the length of the $T^{(n_1+1)}$ initiated by customer B is longer than customer A's patience. If so, the waiting room is empty, and the $T^{(n_1+1)}$ ends. Note that the probability of customer A staying until the end of the $T^{(n_1+1)}$ initiated by customer B is correlated with the length of this $T^{(n_1+1)}$. Thus, in this case, the Laplace Transform can be calculated as $\int_0^\infty P\{\Theta_{1w} \leq L_{T^{(n_1+1)}} | L_{T^{(n_1+1)}} = x\} e^{-sx} dF_{T^{(n_1+1)}}(x)$. This gives the second line of (5).

Alternatively, if the length of the $T^{(n_1+1)}$ initiated by customer B is less than customer A's patience, she will be waiting at Station 1. From the memoryless property of Markovian systems, we know the time from now until no customers are waiting is distributed as a $T^{(n_1+1)}$. In this case, the LT is $\int_0^\infty P\{\Theta_{1w} > L_{T^{(n_1+1)}} | L_{T^{(n_1+1)}} = x\} e^{-sx} \tilde{L}_{T^{(n_1+1)}}(s)\, dF_{T^{(n_1+1)}}(x)$, giving the expression in the third line of (5). This completes the proof. □

We know that $P\{\Theta_{1w} > L_{T^{(n_1+1)}} | L_{T^{(n_1+1)}} = x\} = e^{-\theta_{1w}x}$ and $P\{\Theta_{1w} \leq L_{T^{(n_1+1)}} | L_{T^{(n_1+1)}} = x\} = 1 - e^{-\theta_{1w}x}$, so simplifying (5) gives

Corollary 1. *The Laplace Transform of $L_{T^{(n_1+1)}}$, $\tilde{L}_{T^{(n_1+1)}}(s)$, is the solution of*

$$\tilde{L}_{T^{(n_1+1)}}(\theta+s) = \frac{(\theta_{1w} + n_1(\mu_1+\theta_{1s}) + s)\,\tilde{L}_{T^{(n_1+1)}}(s) - \theta_{1w} - n_1(\mu_1+\theta_{1s})}{\lambda_1 \tilde{L}_{T^{(n_1+1)}}(s) - \lambda_1}, \tag{6}$$

with the boundary condition $\tilde{L}_{T^{(n_1+1)}}(0) = 1$.

We could solve (6) for $\tilde{L}_{T^{(n_1+1)}}(s)$, but it is out of the scope of this paper, so we no longer discuss it.

From Observation 1, we know that the $T^{(n_1+1)}$ is distributed in the same way as the busy period of an $M/M/1+M$ system with arrival rate λ_1, service rate $\theta_{1w} + n_1(\mu_1+\theta_{1s})$, and abandonment rate θ_{1w}. Boxma et al. (2014) gives $E[L_{T^{(n_1+1)}}]$ as (see, e.g., (3.20) of Boxma et al. 2014):

$$E[L_{T^{(n_1+1)}}] = \sum_{k=0}^{\infty} \frac{\lambda_1^k}{(n_1(\mu_1+\theta_{1s})+\theta_{1w})(n_1(\mu_1+\theta_{1s})+2\theta_{1w})\cdots(n_1(\mu_1+\theta_{1s})+(1+k)\theta_{1w})}. \tag{7}$$

Note that $\tilde{L}_{T^{(n_1+1)}}(\theta_{1w})$ can be considered the probability of having no arrivals from a Poisson process with rate θ_{1w} during $T^{(n_1+1)}$ (see, e.g., p59, Buzacott and Shanthikumar 1993). From the relation of the Poisson process and the exponential distribution, this is also the probability that an $\exp(\theta_{1w})$ random variable is greater than the length of $T^{(n_1+1)}$, $P\{\Theta_{1w} > L_{T^{(n_1+1)}}\}$. Following (6), we have

COROLLARY 2. *The probability that a customer's patience exceeds the length of* $T^{(n_1+1)}$ *is*

$$P\{\Theta_{1w} > L_{T^{(n_1+1)}}\} = \frac{\theta_{1w} + n_1(\mu_1 + \theta_{1s})}{\lambda_1} - \frac{1}{\lambda_1 E[L_{T^{(n_1+1)}}]}, \tag{8}$$

where $E[L_{T^{(n_1+1)}}]$ *is given in (7).*

Thus, we can substitute (7) into (8) to calculate $P\{\Theta_{1w} > L_{T^{(n_1+1)}}\} = \tilde{L}_{T^{(n_1+1)}}(\theta_{1w})$, which is, of course, identical to $\tilde{L}_{T^{(n_1+1)}}(\theta_{1w})$ calculated by using (8) in Jouini and Roubos (2014) or by iterating (3.17) in Boxma et al. (2014).

An important observation is that the analysis to create the recursive relation in Proposition 1 can be applied to other metrics of interest. For example, considering $E[L_{T^{(n_1+1)}}]$, we get

$$\begin{aligned} E[L_{T^{(n_1+1)}}] = & \frac{1}{\lambda_1 + \theta_{1w} + n_1(\mu_1 + \theta_{1s})} \\ & + \frac{\lambda_1}{\lambda_1 + \theta_{1w} + n_1(\mu_1 + \theta_{1s})} \left(\int_0^\infty x P\{\Theta_{1w} \le L_{T^{(n_1+1)}} | L_{T^{(n_1+1)}} = x\} dF_{T^{(n_1+1)}}(x) \right. \\ & \left. + \int_0^\infty (x + E[L_{T^{(n_1+1)}}]) P\{\Theta_{1w} > L_{T^{(n_1+1)}} | L_{T^{(n_1+1)}} = x\} dF_{T^{(n_1+1)}}(x) \right), \end{aligned}$$

the simplification of which gives

$$E[L_{T^{(n_1+1)}}] = \frac{1}{\lambda_1 + \theta_{1w} + n_1(\mu_1 + \theta_{1s})} + \frac{\lambda_1}{\lambda_1 + \theta_{1w} + n_1(\mu_1 + \theta_{1s})} \left(E[L_{T^{(n_1+1)}}] + P\{\Theta_{1w} > L_{T^{(n_1+1)}}\} E[L_{T^{(n_1+1)}}] \right). \tag{9}$$

Of course, (9) is equivalent to (8). Moreover, the fact that (9) does not consider the correlation between the probability of customer A staying until the end of the $T^{(n_1+1)}$ initiated by customer B and the length of this $T^{(n_1+1)}$ logically follows because expectations are additive, even among correlated random variables.

3.2. First Passage Probability Matrix $\mathbf{G}^{(q_1)}$

Using the method for creating a recursive relation, we can solve for the first passage probability matrix $\mathbf{G}^{(n_1+1)}$ from:

PROPOSITION 2. *The first passage probability matrix during $T^{(n_1+1)}$, $\mathbf{G}^{(n_1+1)}$, satisfies*

$$\mathbf{G}^{(n_1+1)} = \mathbf{A}_2^{(n_1+1)} + \left(\mathbf{A}_1^{(n_1+1)} + P\left\{\Theta_{1w} \le L_{T^{(n_1+1)}}\right\}\mathbf{A}_0^{(n_1+1)}\right)\mathbf{G}^{(n_1+1)} + P\left\{\Theta_{1w} > L_{T^{(n_1+1)}}\right\}\mathbf{A}_0^{(n_1+1)}\left(\mathbf{G}^{(n_1+1)}\right)^2, \tag{10}$$

where $P\{\Theta_{1w} > L_{T^{(n_1+1)}}\}$ is given by Corollary 2.

Further, it is straightforward to generalize Proposition 2 to $\mathbf{G}^{(n_1+i)}$, $i = 1, 2, \ldots$.

COROLLARY 3. *The first passage probability matrix during $T^{(n_1+i)}$, $\mathbf{G}^{(n_1+i)}$, satisfies*

$$\mathbf{G}^{(n_1+i)} = \mathbf{A}_2^{(n_1+i)} + \left(\mathbf{A}_1^{(n_1+i)} + P\left\{\Theta_{1w} \le L_{T^{(n_1+i)}}\right\}\mathbf{A}_0^{(n_1+i)}\right)\mathbf{G}^{(n_1+i)} + P\left\{\Theta_{1w} > L_{T^{(n_1+i)}}\right\}\mathbf{A}_0^{(n_1+i)}\left(\mathbf{G}^{(n_1+i)}\right)^2$$
$$\text{for } i = 1, 2, \ldots. \tag{11}$$

where

$$P\left\{\Theta_{1w} > L_{T^{(n_1+i)}}\right\} = \frac{i\theta_{1w} + n_1(\mu_1 + \theta_{1s})}{\lambda_1} - \frac{1}{\lambda_1 E\left[L_{T^{(n_1+i)}}\right]}$$

and

$$E\left[L_{T^{(n_1+i)}}\right] = \sum_{k=0}^{\infty} \frac{\lambda_1^k}{(n_1(\mu_1+\theta_{1s}) + i\theta_{1w})(n_1(\mu_1+\theta_{1s}) + (i+1)\theta_{1w})\cdots(n_1(\mu_1+\theta_{1s}) + (i+k)\theta_{1w})} \text{ for } i = 1, 2, \ldots.$$

The Last-Come-First-Serve rule provides an explicit quadratic matrix equation for each $\mathbf{G}^{(n_1+i)}$. Although closed form solutions for quadratic matrix equations are hard to obtain, we have an efficient and exact numerical method to derive $\mathbf{G}^{(n_1+i)}$ from (11). Note that $\left(\mathbf{A}_2^{(n_1+i)} + \left(\mathbf{A}_1^{(n_1+i)} + P\left\{\Theta_{1w} \le L_{T^{(n_1+i)}}\right\}\mathbf{A}_0^{(n_1+i)}\right) + P\left\{\Theta_{1w} > L_{T^{(n_1+i)}}\right\}\mathbf{A}_0^{(n_1+i)}\right)\overrightarrow{1} = \overrightarrow{1}$, so we can consider $\mathbf{A}_2^{(n_1+i)}$, $\mathbf{A}_1^{(n_1+i)} + P\left\{\Theta_{1w} \le L_{T^{(n_1+i)}}\right\}\mathbf{A}_0^{(n_1+i)}$ and $P\left\{\Theta_{1w} > L_{T^{(n_1+i)}}\right\}\mathbf{A}_0^{(n_1+i)}$ as one-step transition matrices falling under the matrix analytic methods, while $\mathbf{G}^{(n_1+i)}$ can be derived using Algorithm 8.1 in Latouche and Ramaswami (1999). This numerical algorithm is efficient and exact, and is given in Appendix A2.4. Once $\mathbf{G}^{(n_1+i)}$, $i = 1, 2, \ldots$, is calculated, $\mathbf{G}^{(q_1)}$ for $q_1 \le n_1$ can be derived using (4).

We note that Corollary 3 is one of our main contributions. Compared to the iterative approximation using (4), our method greatly improves the accuracy of the derivation of $\mathbf{G}^{(q_1)}$.

Let $\pi_i^{(q_1)}$ denote the steady state probability the MC is in state (q_1, i), and let $\pi^{(q_1)} = \left[\pi_0^{(q_1)}, \ldots, \pi_{n_2+m}^{(q_1)}\right]$ denote the steady state probability vector the MC is at level q_1. When $\mathbf{G}^{(q_1)}$ is derived, we can apply the results in Section 12.1 in Latouche and Ramaswami (1999) to numerically solve $\pi^{(q_1)}$ from

$$\pi^{(q_1)} = \pi^{(q_1-1)}\mathbf{A}_0^{(q_1-1)}\left(I - \mathbf{A}_1^{(q_1)} - \mathbf{A}_0^{(q_1)}\mathbf{G}^{(q_1+1)}\right)^{-1}$$

normalized by

$$\pi^{(0)}\sum_{n=0}^{\infty}\prod_{k=1}^{n}\mathbf{A}_0^{(k-1)}\left(I - \mathbf{A}_1^{(k)} - \mathbf{A}_0^{(k)}\mathbf{G}^{(k+1)}\right)^{-1}\overrightarrow{1} = 1.$$

However, no easily computable analytic expression is available for the infinite sum in the normalization step. The standard practice is to truncate the MC at level M and solve

$$\pi^{(0)}\sum_{n=0}^{M}\prod_{k=1}^{n}\mathbf{A}_0^{(k-1)}\left(I - \mathbf{A}_1^{(k)} - \mathbf{A}_0^{(k)}\mathbf{G}^{(k+1)}\right)^{-1}\overrightarrow{1} = 1.$$

Let $\pi^{(0)}(M)$ denote the solution of this normalization equation truncated at level M. We can choose an M large enough so that $\left|\pi^{(0)}(M+1) - \pi^{(0)}(M)\right| < \epsilon$, where $|\bullet|$ is the L^2-norm and ϵ is the error tolerance. If we choose a smaller ϵ, the accuracy of this method improves. However, the computation of $\mathbf{G}^{(q_1)}$ up to level M requires a large computational effort.

In Appendix A1, we further simplify the analysis of Station 2 customers by showing that only $\mathbf{G}^{(n_1+1)}$ is required for the exact derivation of the service level measures of Station 2 customers. The idea is to consider every time the system moves from level n_1 to level n_1+1 as a renewal point. We call the time period between two renewal points a cycle. Using renewal reward theorem (see, e.g., Ross 2007), we can write any measures of interest as the expected reward earned in a cycle divided by the expected cycle length; see, e.g., Theorem A1 in Appendix A1.5. Then, by selecting corresponding reward functions, we can derive other service level measures of interest, like the abandonment and balking rates from Station 2; see, e.g., Appendix A1.6.

4. Numerical Results and Insights

We now consider the staffing problem in banks' tandem financial service system. This is a special case of our general model presented in Section 2, when (i) all customers request service from both stations i.e., $p = 1$; (ii) there are no external arrivals at Station 2, i.e., $\lambda_2 = 0$; and (iii) deals' abandonment rate stays the same in different parts of the system, i.e., $\theta_{1s} = \theta_{1w} = \theta_{2s} = \theta_{2w} = \theta$.

A staffing policy in a tandem queueing system is composed of two elements: (i) the total number of servers in the system N and (ii) the *assignment rule* that assigns these servers to the two stations. For a total number of servers, $N \geq 2$, suppose n_1 $(1 \leq n_1 \leq N-1)$ servers are assigned to Station 1, and the rest, $n_2 = N - n_1$, are assigned to Station 2. When no confusion arises, we use n_1, instead of (n_1, n_2), to represent an assignment rule.

Recall that by using a large waiting room of size m in Station 2, we can derive the performance of a tandem queueing system under any staffing policy with infinite waiting room in both stations. Note that when other parameters are kept the same, a larger waiting room (of size m) will lead to less balking at Station 2. When m approaches ∞, balking at Station 2 will diminish. Thus, we can evaluate a general tandem queueing system with an infinite waiting room by choosing an m large enough to ensure the balking rate at Station 2 is smaller than a selected error tolerance. In the following discussion, we use an error tolerance of 10^{-10}. Because of deals' abandonment, it is sufficient to use a moderate m (˜200 in all our numerical tests) to ensure the balking rate at Station 2 is less than this error tolerance.

We consider two questions:

1. Given that there are $N \geq 2$ servers available, how can we assign them to a two-station tandem queueing network with abandonment to maximize the service throughput (ST)?

2. *What is the minimum number of servers needed to achieve a ST target in such a network?*

In our numerical study, we first focus on the assignment rule when the total number of servers in the system N is fixed. In Section 4.1, we establish an upper bound of the system's ST under any assignment rule n_1, introduce an easily calculable benchmark assignment rule, and identify the

optimal assignment rule. We then compare these two assignment rules to gain insights into bank's financial service operations. Specifically, in Section 4.2, we study how the difference between these two assignment rules changes with a bank's competitiveness or the downstream control function's processing rate. Finally, in Section 4.3, we investigate the staffing policy and compare the numbers of servers needed under the optimal and the benchmark rules to guarantee 95% of the ST upper bound.

4.1. Staffing Policies and Assignment Rules

We first establish an upper bound of the ST in an $M/M/n$ model with arrival rate λ, service rate μ, and abandonment rate θ. On the one hand, if the multi-server queue has abundant servers, so that any arriving deal immediately enters service without wait, this deal will go through with probability (w.p.) $\frac{\mu}{\mu+\theta}$ (abandon during service w.p. $\frac{\theta}{\mu+\theta}$). Then the ST is $\lambda\frac{\mu}{\mu+\theta}$. On the other hand, if the $M/M/n$ model has abundant demand, so that all servers' utilization is 100%, the system ST is at most $n\mu$. Thus, the ST of the system under demand λ and n servers is bounded from above by the above two cases.

Proposition 3. *An upper bound of the ST in a multi-server queue with abandonment is* $\min\left(\lambda\frac{\mu}{\mu+\theta}, n\mu\right)$.

By applying Proposition 3 to the bank's tandem queueing system, we get an upper bound of the ST for any assignment rule n_1 under $N \geq 2$ servers, $\overline{TP}(n_1)$.

Corollary 4. *The upper bound of the ST under assignment rule n_1 in the bank's tandem queueing system with abandonment is* $\overline{TP}(n_1) = \min\left(\frac{\mu_2}{\mu_2+\theta}\min\left(\lambda_1\frac{\mu_1}{\mu_1+\theta}, n_1\mu_1\right), (N-n_1)\mu_2\right)$.

By conditioning on the value of n_1, we can show that the ST upper bound $\overline{TP}(n_1)$ is a piecewise concave function of n_1. In the best case scenario, when each station has abundant servers to accommodate the arrival rate λ_1, the upper bound in Corollary 4 becomes $\lambda_1\frac{\mu_1}{\mu_1+\theta}\frac{\mu_2}{\mu_2+\theta}$, and this serves as a ST upper bound for *any* staffing policy.

We further note that in the same station, deals in the waiting room have a higher abandonment probability than deals in service, because the former may abandon during both the waiting and

service processes; i.e., they incur an extra probability of abandonment than those in service. This means that starting to serve deals sooner reduces the abandonment probability towards the lower bound, e.g., $\frac{\theta}{\mu_1+\theta}$ at Station 1, but not further. Hence, adding servers to the station where most deals enter service immediately upon arrival may not increase the service completion rate all that much.

Given the total number of servers N and service rates in the two stations (μ_1 & μ_2), we define $\dot{n}_1 = \mu_2 N/(\mu_1+\mu_2)$ as the *fractional* benchmark assignment rule, under which Stations 1 and 2 have identical capacities $\mu_1\mu_2 N/(\mu_1+\mu_2)$. Note that $\dot{n}_1$ is independent of θ and may be fractional. To avoid the fractional server assignment, we focus on the benchmark assignment rule (BnchAR) $[\dot{n}_1] = \arg\max_{n_1\in\{\lfloor\dot{n}_1\rfloor,\lceil\dot{n}_1\rceil\}} \min(n_1\mu_1, (N-n_1)\mu_2)$, where $\lfloor\bullet\rfloor$ and $\lceil\bullet\rceil$ are floor and ceiling functions, respectively. We call $[\dot{n}_1]\mu_1$ the benchmark capacity (BnchCap). We have the following intuitive proposition for the BnchAR when $\theta\to 0^+$.

PROPOSITION 4. *In a tandem queueing system with abandonment rate $\theta\to 0^+$, the BnchAR is the optimal assignment rule, and the maximum ST is* $\min(\lambda_1, [\dot{n}_1]\mu_1, (N-[\dot{n}_1])\mu_2)$.

We next investigate the ST for assignment rules under a given total number of servers, $N\geq 2$. On the one hand, when n_1 is small (i.e., $n_2 = N-n_1$ is large), Station 2 can start the review process for most deals immediately, but many deals abandon from Station 1, during either waiting or service, before reaching Station 2. In this case, we can assign some of Station 2's servers to Station 1 to increase the system's ST. On the other hand, when n_1 is large (i.e., n_2 is small), Station 1 is able to capture many deals before they abandon, but Station 2 does not have enough capacity to handle all the input from Station 1, causing many deals to abandon from Station 2; consequently, Station 1's work on these deals is wasted. Therefore, it is better to move some servers from Station 1 to Station 2 to assure a higher ST. From this discussion, we see that the ST is an increasing function of n_1 when n_1 is small and a decreasing function of n_1 when n_1 is large – i.e., close to N.

From communicating with our contact in an international investment bank, we conclude that reviewing deals takes less time than negotiating them. Thus, we assume that the downstream

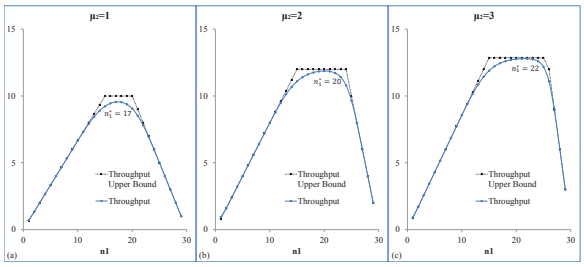

Figure 3 **Throughput and throughput upper bound as functions of** $n_1 \in \{1, \ldots, N-1\}$ **when** $\lambda_1 = 22.5$**,** $N = 30$**,** $\theta = 0.5$**, and** $\mu_2 \in \{1, 2,$ **and** $3\}$**.**

control function works at least as fast as the upstream front office whose service rate is normalized to 1, i.e., $\mu_2 \geq \mu_1 = 1$, throughout our numerical studies.

Let $TP(n_1)$ denote the ST of this tandem queueing system under assignment rule n_1. Using the method described in Section 3 and Appendix A1, for a series of systems with $n_1 = 1, 2, \ldots, N-1$ servers at Station 1, we calculate $TP(n_1)$ and find the optimal assignment rule (OptAR) $n_1^* = \arg\max_{1 \leq n_1 \leq N-1} TP(n_1)$ using enumeration. Figure 3 records the ST upper bound $\overline{TP}(n_1)$ and the ST $TP(n_1)$ as functions of n_1, and the OptAR n_1^*, for $\lambda_1 = 22.5$, $N = 30$, $\theta = 0.5$, and $\mu_2 \in \{1, 2,$ and $3\}$. We see from Figure 3 that our intuition is valid: the ST $TP(n_1)$ is, in fact, *concave* in n_1 and has a global maximum. The concavity of the ST holds for all other parameter settings we test. We summarize the observation below.

Observation 2 *For any fixed* λ_1*,* μ_1*,* μ_2*,* θ*, and* $N \geq 2$*, our numerical studies suggest that the ST* $TP(n_1)$ *is an initially increasing and then decreasing concave function of* n_1*, for* $n_1 = 1, \ldots, N-1$*.*

In the following sections, we compare the OptAR n_1^* to the BnchAR $[\dot{n}_1]$ to provide intuitions and guidelines for staffing in tandem queueing systems with impatient clients. These intuitions and guidelines were not previously available because of a lack of exact methods for such systems.

4.2. Optimal Assignment Rule vs. Benchmark Assignment Rule

In this section, we compare the OptAR n_1^* to the BnchAR $[\dot{n}_1]$, under a fixed head count N and different (i) bank competitiveness as captured via θ and (ii) service rate, μ_2, for the downstream control function. Recall that when the bank is relatively more competitive in the market, deals

abandon at a lower rate, i.e., θ becomes smaller. The opposite is also true. While the front office's deal negotiating rate μ_1 is relatively stable and it is difficult to speed up or slow down the deal negotiation process, the control function's deal reviewing rate μ_2 may be improved by simplifying the review procedure or worsened by adding new regulations which complicate the review process.

Similar to the discussion in Section 4.1, when $\theta > 0$, Station 1's capacity has two conflicting effects on Station 2; both dictate the differences between the OptAR and the BnchAR. For an arrival to contribute to the system's ST, Station 1 must complete the deal negotiation before abandonment and pass it on to Station 2 as input, and Station 2 must review the deal before abandonment. On the one hand, for Station 1 to capture enough deals before they abandon and thus maintain an adequate input rate at Station 2, there is pressure to assign more servers to Station 1. We call this the *deal-capturing effect*. On the other hand, part of Station 1's capacity is wasted on deals that eventually abandon from Station 2. Thus, there is a tendency to move servers from Station 1 to Station 2 to increase Station 2's capacity and reduce this undesirable abandonment. We call this the *deal-loss effect*. When θ increases, the deal-capturing and deal-loss effects are both strengthened, and the interplay of these two effects dictates the OptAR.

When θ is close to zero, from Proposition 4, we expect that the BnchAR is the optimal assignment rule that maximizes the system's ST for any value of μ_2, i.e., $n_1^* = [\dot{n}_1]$. Recall that, among two independent multi-server queues with identical arrival rates and capacities but different service rates, the one with a lower service rate has more servers and a longer expected sojourn time. In the presence of abandonment, a longer sojourn time leads to a higher probability of abandonment. Thus, as θ increases, in a tandem queueing system under BnchAR, the station with the lower service rate will have more abandonment. In this case, the deal-capturing effect is strengthened, and some servers from Station 2 should be moved to Station 1, unless the deal-loss effect caused by moving these servers dominates the deal-capturing effect. When servers in both stations work at similar rates under BnchAR, moving servers to Station 1 does not trigger a strong deal-loss effect, because deals served by Station 1 enter service in Station 2 almost immediately. In this instance,

the deal-capturing effect dominates the deal-loss effect, and it is favorable for the OptAR to assign more servers to Station 1 than the BnchAR would dictate. However, when μ_2 increases, the deal-loss effect is strengthened, because servers now work faster in Station 2 than in Station 1, and (recall from the discussion of Proposition 3) deals have a higher service completion probability when they are served by faster servers. Thus, when μ_2 increases under BnchAR, it becomes increasingly beneficial to keep servers in Station 2, or to even move some servers from Station 1 to Station 2, and there is a tendency for the OptAR to assign more servers to Station 2.

This intuition is verified in Figure 4, where $\Delta n_1 = n_1^* - [\dot{n}_1]$, $\%TP^* = TP\left(n_1^*\right)/\overline{TP}$, and $\%TP_{BM} = TP\left([\dot{n}_1]\right)/\overline{TP}$ are plotted as functions of abandonment rate $\theta \in \{0.1, 0.2, \ldots, 2\}$, when $N = 30$ and $\lambda_1 \in \{[\dot{n}_1]\,\mu_1, 1.5\,[\dot{n}_1]\,\mu_1, 2\,[\dot{n}_1]\,\mu_1\}$ for three representative values of $\mu_2 \in \{1, 3, 6\}$. Note that, in reality, the downstream control function is less likely to be five times faster than the upstream front office and deals' abandonment rate is not likely to reach 2 (i.e., all deals in the negotiation process are canceled w.p. 2/3), so the $\mu_2 = 6$ and $\theta = 2$ cases are of low practical significance. Nonetheless, to provide complete insights, we consider them here.

We see that there is a critical cutoff point in Station 2's service rate, i.e., $\mu_2 = 3$. At this point, the deal-capturing and deal-loss effects are of similar strength, so the OptAR n_1^* and the BnchAR $[\dot{n}_1]$ assigns an almost identical number of servers to Station 1; see, e.g., $\mu_2 = 3$ in Figure 4b. When μ_2 is smaller than this cutoff point, the deal-capturing effect dominates the deal-loss effect, so the OptAR n_1^* assigns more servers to Station 1 than the BnchAR $[\dot{n}_1]$; see, e.g., $\mu_2 = 1$ in Figure 4a. The difference between the ST under OptAR and BnchAR may reach as high as 9% of the ST upper bound $\overline{TP}$, when the demand rate is twice the benchmark capacity; see, e.g., Figure 4d. When μ_2 is greater than the cutoff point, the deal-loss effect dominates the deal-capturing effect, so the OptAR n_1^* assigns fewer servers to Station 1 than the BnchAR $[\dot{n}_1]$; see, e.g., $\mu_2 = 6$ in Figure 4c. However, in this case, the ST under both assignment rules is similar, and the difference between $\%TP^*$ and $\%TP_{BM}$ stays below 0.5% of $\overline{TP}$ for any θ; see, e.g., Figure 4f. We observe from the numerical results that the cutoff point $\mu_2 = 3$ applies for a wide range of λ_1 and θ of practical significance.

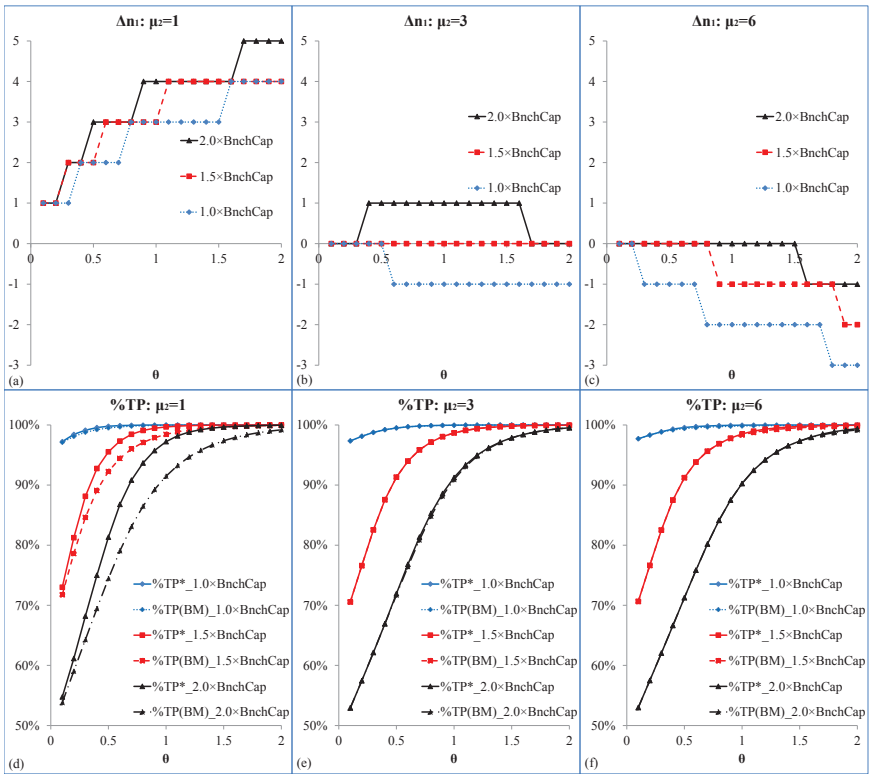

Figure 4 $\Delta n_1 = n_1^* - [\dot{n}_1]$, $\%TP^* = TP(n_1^*)/\overline{TP}$, **and** $\%TP_{BM} = TP([\dot{n}_1])/\overline{TP}$ **as functions of abandonment rate** θ, **when** $N = 30$ **and** $\lambda_{11} \in \{[\dot{n}_1]\mu_1, 1.5[\dot{n}_1]\mu_1, 2[\dot{n}_1]\mu_1\}$ **for** $\mu_2 \in \{1, 3, 6\}$ **cases.**

We further observe that when the abandonment rate, θ, is large, the difference between the system's ST under OptAR and BnchAR is small. For example, when $\theta = 2$, the difference between $\%TP^*$ and $\%TP_{BM}$ is typically less than 1% of $\overline{TP}$; see, e.g., Figure 4d-f. The intuition is that when θ is large, the abandonment probability is so high that servers' utilization is relatively low. Thus, almost all deals enter service immediately upon entering both stations. In this case, the system's ST under BnchAR is close to the ST upper bound for any staffing policies in Corollary 4. Although the OptAR may be significantly different from the BnchAR (see, e.g., Figure 4a-c), it does not improve the ST very much (see, e.g., Figure 4).

For banks, our results identify three operational regions in the downstream control function's service rate μ_2. (i) When μ_2 is less than the cutoff point and close to μ_1, banks should try their best to identify the optimal assignment rule. Doing so will increase the bank's ST by a significant amount and improve the bank's overall profitability. (ii) When μ_2 is at the cutoff point, an easily

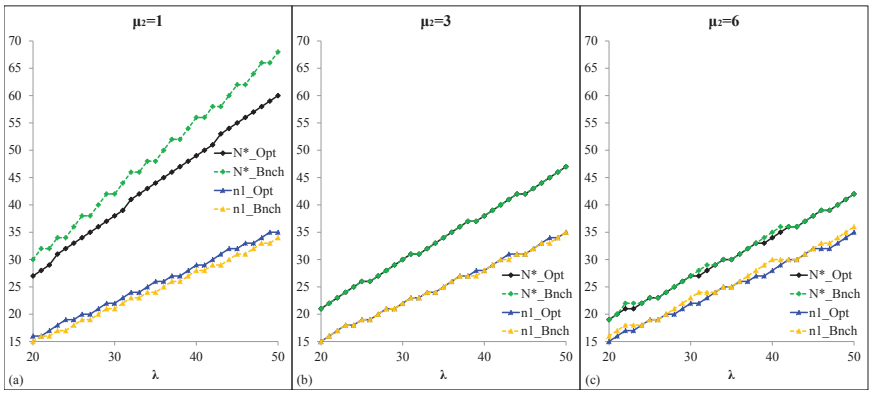

Figure 5 **Minimum number of staffs needed to reach 95% of the throughput upper bound under OptAR N^*_{Opt} and BnchAR N^*_{Bnch} and the corresponding assignment rules n_1^* and $[\dot{n}_1]$ as functions of arrival rate λ_1, when $\mu_1 = 1$, $\theta = 0.5$, for $\mu_2 \in \{1, 3, 6\}$ cases.**

calculable BnchAR is close to optimal. (iii) When μ_2 is greater than the cutoff point, although the BnchAR may deviate from the OptAR, the performance of the BnchAR is close to optimal, so banks can stick with it to keep their operations simple.

4.3. Staffing Policies

From Corollary 4, we know that the ST under any staffing policies in a tandem queueing system with abandonment is lower than $\lambda_1 \frac{\mu_1}{\mu_1+\theta} \frac{\mu_2}{\mu_2+\theta}$. Of course, to reach this upper bound, the bank has to spend a very large amount of its operating budget on staffing. We thus investigate the minimum number of staff needed to reach 95% of the ST upper bound under OptAR N^*_{Opt} and under BnchAR N^*_{Bnch} that can be quickly identified by using the enumeration of the total number of servers available N.

In Figure 5, when $\mu_1 = 1$ and the arrival rate λ_1 increases from 20 to 50, we plot the minimum numbers of staff needed to reach 95% of the ST upper bound N^*_{Opt} and N^*_{Bnch}, for the OptAR and BnchAR respectively, and the corresponding assignment rules n_1^* and $[\dot{n}_1]$, respectively, for three representative values of $\mu_2 \in \{1, 3, 6\}$. Note that we use $\theta = 0.5$ as a representative example, but the insights developed here hold for other values of θ.

Similar to Section 4.2, we observe a cutoff point (see Figure 5), at which the staffing levels under both OptAR and BnchAR, N^*_{Opt} and N^*_{Bnch}, respectively, are similar, as are the corresponding

assignment rules n_1^* and $[\dot{n}_1]$; see, e.g., $\mu_2 = 3$ in Figure 5b. If μ_2 is smaller than this cutoff point, there is a significant difference between the staffing policies under OptAR and BnchAR. For example, in Figure 5a, the difference between N_{Opt}^* and N_{Bnch}^* is 3, when $\lambda_1 = 20$, and can be up to 8 when $\lambda_1 = 50$. Moreover, although when it uses the OptAR, the system needs fewer servers than it does when it uses the BnchAR, more servers are assigned to Station 1 under OptAR than under BnchAR because the deal-capturing effect dominates the deal-loss effect. If μ_2 is greater than the cutoff point, the staffing level under OptAR is slightly lower than it is under BnchAR, and fewer servers are assigned to Station 1 under OptAR than BnchAR; see, e.g., Figure 5c. In this case, Station 2 is much more efficient than Station 1, so the deal-loss effect dominates the deal-capturing effect. It becomes beneficial to move some servers from Station 1 to Station 2. As shown by our numerical results, the cutoff point $\mu_2 = 3$ holds for a wide range of λ_1 and θ.

Our results provide useful guidelines for banks' financial service operations. As in Section 4.2, three operational regions are related to the downstream control function's service rate μ_2. (i) If μ_2 is less than the cutoff point, the OptAR saves the bank a significant head count compared to the BnchAR, by emphasizing the upstream front office and assigning more staff to there. (ii) If μ_2 is at the cutoff point, the simple BnchAR is optimal. (iii) If μ_2 is greater than the cutoff point, the OptAR can optimize the bank's head count. However, in this case, the optimal staffing policy puts more emphasis on the downstream control function, assigning more staff there, unlike in (i). Fortunately, our numerical method can easily find the optimal staffing policy for any deal arrival, processing, and abandonment rates.

5. Additional Applications

Our model in Section 2 goes beyond the financial services application specified here. In this section, we demonstrate its generalization to other applications with various features by incorporating flexible servers.

5.1. Flexible Servers

Berman and Sapna (2005) consider the management of homogeneous flexible servers in a two-station tandem queue. Andradottir and Ayhan (2005) consider a similar problem with finite intermediate buffer and heterogeneous flexible servers; i.e., different servers may have different service

rates at different stations. They aim to find the optimal policy to dynamically assign flexible servers to different stations to maximize long-run average throughput. However, the optimal policies are usually too complicated to describe and implement. In this section, we extend our model to consider flexible servers.

To focus on the effect of flexible servers, we study a simple tandem queueing system with identical abandonment rates in both stations' waiting rooms, i.e., $\theta_{1w} = \theta_{2w} = \theta$, without direct departure from Station 1, external arrivals at Station 2, or customer abandonment from service, i.e., $p = 1$, $\lambda_2 = 0$, and $\theta_{1s} = \theta_{2s} = 0$.

Suppose f flexible servers can work in both stations in addition to the dedicated n_1 and n_2 servers in Stations 1 and 2. We assume that when serving customers in Station i, the flexible servers have the same service rate as the dedicated servers in Station i, but they give preemptive priority to Station 1 customers over Station 2 customers. Thus, when all n_2 dedicated servers in Station 2 are busy and a flexible server is not needed by Station 1 customers, that server can serve Station 2 customers in Station 2 with a rate μ_2. Because of the preemptive priority of Station 1 customers, whenever a flexible server serving a Station 2 customer is needed by a Station 1 customer, the Station 2 customer will be preempted. The preempted Station 2 customer may wait in Station 2's waiting room, if it is not full; otherwise, she will be *bumped* out of the system; i.e., she will be lost.

The number of flexible servers that are available for Station 2 customers when there are q_1 Station 1 customers in the system is $\max\left(f - \max\left(q_1 - n_1, 0\right), 0\right)$, and the maximum number of Station 2 customers the system can have when there are q_1 Station 1 customers in the system is $n_2 + m + \max\left(f - \max\left(q_1 - n_1, 0\right), 0\right)$. Thus, the total rate at which the system leaves the state (q_1, q_2) in a system with f flexible servers is

$$\begin{aligned} v\left(q_1, q_2\right) = \lambda_1 &+ \theta\left(\max\left(q_1 - n_1 - f, 0\right) + \max\left(q_2 - n_2 - \max\left(f - \max\left(q_1 - n_1, 0\right), 0\right), 0\right)\right) \\ &+ \mu_1 \min\left(q_1, n_1 + f\right) + \mu_2 \min\left(q_2, n_2 + \max\left(f - \max\left(q_1 - n_1, 0\right), 0\right)\right) \end{aligned}$$

$$\text{for } q_1 = 0, 1, \ldots, \text{ and } q_2 = 0, 1, \ldots, n_2 + m + \max\left(f - \max\left(q_1 - n_1, 0\right), 0\right).$$

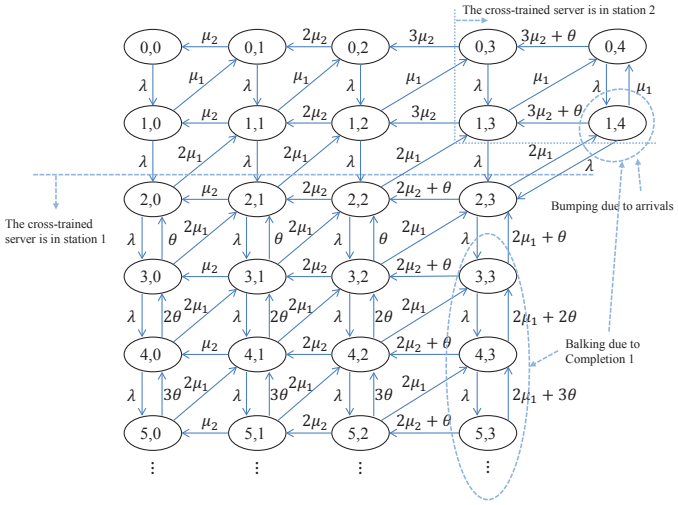

Figure 6 **The Markov Chain of tandem queueing system with $n_1 = 1$, $n_2 = 2$, $f = 1$, and $m = 1$.**

Figure 6 illustrates the MC of a system with one dedicated server in Station 1, two dedicated servers in Station 2, and one flexible server. Comparing Figures 2 and 6, we see that the MC in Figure 6 has two extra states $(0,4)$ and $(1,4)$, that are added because of the flexible server. When the number of Station 1 customers is greater than the total number of servers who can work in Station 1, i.e., $q_1 \geq n_1 + f$, the flexible servers act the same as dedicated servers in Station 1, because of the preemptive priority given to Station 1 customers. Introducing flexible servers changes the subspace $\{(q_1, q_2) \,|q_1 \leq n_1 + f\}$, but not the subspace $\{(q_1, q_2) \,|q_1 > n_1 + f\}$. Therefore, an analysis identical to that in Section 3 and Appendix A1 can be applied here to obtain the required service level measures.

5.2. Caring for Critical Patients

In the context of healthcare, Armony et al. (2017) use diffusion approximation to analyze a tandem queueing network with flexible servers serving impatient critical patients. They model a two-station tandem queueing system with an Intensive Care Unit (ICU) as upstream Station 1 and a Step Down Unit (SDU) as downstream Station 2. Patients in critical conditions arriving at ICU may go to other hospitals or, in extreme cases, they may die because of the long wait; in either case, they abandon the waiting line. Critical patients who receive service in ICU will become semi-critical patients and will visit Station 2 with probability (w.p.) p; otherwise, they are cured and can leave

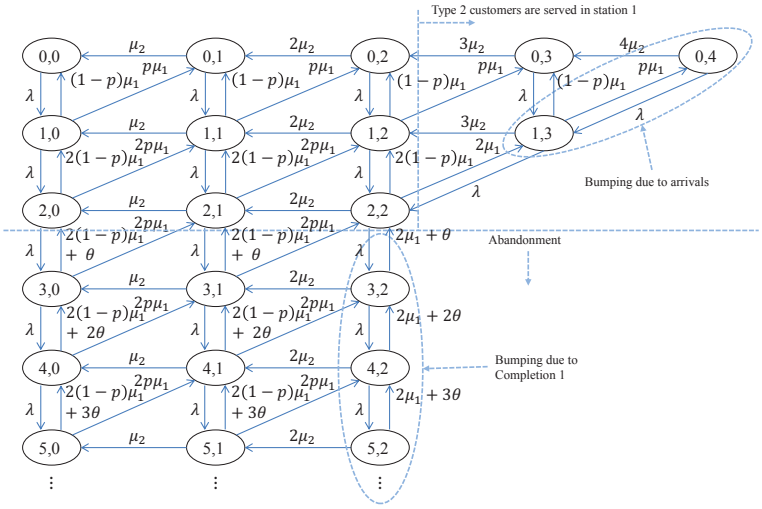

Figure 7 **The Markov Chain of Armony et al. (2017) with two servers in each station, which is equivalent to our model with $n_1 = 0$, $f = n_2 = 2$, and $m = 0$.**

immediately, w.p. $1-p$. If SDU is full and no critical patients are waiting for ICU whose service rate is μ_1, semi-critical patients can be served in ICU with rate μ_2, but critical patients can preempt semi-critical ones in ICU.

SDU has no waiting room; i.e., $m = 0$. Patients do not abandon while they are in service. Thus, we only see abandonment from the ICU's waiting line; i.e., $\theta_{1w} = \theta > 0$ and $\theta_{1s} = \theta_{2s} = 0$. Armony et al. (2017) assume there are no external arrivals at SDU; i.e., $\lambda_2 = 0$.

This is a special case of the model with flexible servers discussed in Section 5.1: there are no dedicated servers for Station 1, only flexible servers; i.e., $n_1 = 0$ and $f > 0$. Figure 7 illustrates the (q_1, q_2) MC of Armony et al.'s (2017) model with two servers in each station; this corresponds to our model where $n_1 = 0$, $f = n_2 = 2$, and $m = 0$.

Because introducing flexible servers only changes the finite part of the MC, e.g., Figure 7, it adds little complexity to the model and can be analyzed using a method similar to that in Section 3 and Appendix A1.

Notably, whereas diffusion approximation results are only valid when the workload is high, our method can provide *exact* analysis for any workload. Moreover, incorporating external arrivals into the downstream SDU, a practical element not captured by the model in Armony et al. (2017), is straightforward as per the discussion in Section 2.

5.3. No Abandonment during Service

Applications of tandem queueing system, where customers need to visit several stations in sequence, include call centers, where customers talk to general call-takers before being transferred to specialists, and hospital emergency rooms, where patients are admitted by triage nurses and then diagnosed by a doctor. In these applications, customers rarely abandon during service. Moreover, because of the waiting cost already incurred, customers waiting for the downstream station may abandon less often than those waiting for the upstream station. To adapt our general model to these applications, we simply use $\theta_{1s} = \theta_{2s} = 0$ and $\theta_{1w} \geq \theta_{2w} > 0$. In Online Appendix OA2, we carry out an initial numerical study in this direction with $\theta_{1w} = \theta_{2w} = \theta$. There are no abandonments during service, direct departures from Station 1, or external arrivals at Station 2, i.e., $p = 1$, $\lambda_2 = 0$ and $\theta_{1s} = \theta_{2s} = 0$. We develop managerial insights into the operations of such systems and suggest the need for additional studies in this direction.

6. Summary

In this paper, we study tandem queueing networks with impatient customers – a model with applications in a number of different industries. We provide the first exact analysis of these level-dependent quasi-birth-and-death (LDQBD) processes. In Proposition 1, we develop a technique to generate a recursive relation in this LDQBD process, so that the first passage matrices at different levels can be derived by solving quadratic matrix equations, using exact numerical methods from the literature. We simplify the derivation by jointly using the recursive renewal reward theorem and queueing and Markov chain decomposition. This simplification reduces the number of quadratic matrix equations we must solve to only one, greatly reducing the computational burden. We then provide an efficient exact numerical method to calculate different metrics for general tandem queueing systems with abandonment.

We use the numerical method to tackle the staffing problem in banks' financial service systems with the service throughput as the target measure. Our results point to useful guidelines for banks. When the control function's service rate is below a critical cutoff point, it is necessary to identify

the OptAR, as it will reduce the total number of servers needed by assigning more servers to the front office than the BnchAR (which assigns identical capacities to both stations). If the control function's service rate is at the cutoff point, it is optimal to use the easily calculable BnchAR. When the control function's processing rate is above the cutoff point, the OptAR may deviate from the BnchAR. However, the staffing policies based on the BnchAR may not be very different from those based on the OptAR.

This paper represents an initial study of tandem queueing systems with impatient customers, but our study goes beyond a basic examination. We demonstrate how to extend our model to include other features and achieve wider applicability. For example, it can be modified to contain tandem queueing networks with flexible servers; with flexible servers, it can optimally solve the ICU-SDU model studied by Armony et al. (2017).

References

Abouee-Mehrizi, H., B. Balcioglu, O. Baron (2012) Strategies for a Centralized Single Product Multi-Class M/G/1 Make-to-Stock Queue. Oper. Res. 60(4)803-812.

Adan, I., J. Resing (2002) Queueing Theory. Technische Universiteit Eindhoven.

Andradottir, S., H. Ayhan (2005) Throughput Maximization for Tandem Lines with Two Stations and Flexible Servers. Oper. Res. 53(3)516-531.

Armony, M., C.W. Chan, B. Zhu (2017) Critical Care Capacity Management: Understanding the role of a Step Down Unite. Production and Operations Management. *Forthcoming.* doi: 10.1111/poms.12825.

Bar-Lev, S., H. Blanc, O. Boxma, G. Janssen, D. Perry (2013) Tandem Queues with Impatient Customers for Blood Screening Procedures. Methodology and Computing in Applied Probability. 15(2)423-451.

Baron, O., J. Milner (2009) Staffing to Maximize Profit for Call Centers with Alternate Service Level Agreements. Oper. Res., 57(3)685-700.

Berman, O., K.P. Sapna-Isotupa (2005) Optimal Control of Servers in Front and Back Rooms with Correlated Work. IIE Transactions. 37:167-173.

Boxma, O., D. Perry, W. Stadje, S. Zacks (2014) The Busy Period of an M/G/1 Queue with Customer Impatience. Journal of Applied Probability. 47:130-145.

Buzacott, J., J. Shanthikumar (1993) Stochastic Models of Manufacturing Systems. Prentice Hall.

Chen, H., D. Yao (2001) Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization, Springer-Verlag, New York.

Gandhi, A., S. Doroudi, M. Harchol-Balter, A. Scheller-Wolf (2014) Exact Analysis of the M/M/k/setup Class of Markov Chains via Recursive Renewal Reward. Queueing Systems, 77(2)177-209.

Gans, N., G. Koole, A. Mandelbaum (2003) Telephone call centers:Tutorial, review and research prospects. Manufacturing and Service Operations Management, 5(2)79–141.

Garnett O., Mandelbaum A. and Reiman M. (2002) Designing a Call Center with Impatient Customers. Manufacturing and Service Operations Management, 4(3)208-227.

Gross, D., J. Shortle, J. Thompson, C. Harris. (2008) Fundamentals of Queueing Theory. Wiley & Sons.

Jackson, J. R. (1963). Jobshop-like Queueing Systems. Management Science. 10(1)131–142.

Jouini, O., A. Roubos (2014) On Multiple-Priority Multi-Server Queues with Impatience. Journal of the Operational Research Society. 65(5)616-632.

Kelly, F. (1979) Reversibility and Stochastic Networks. Wiley, New York.

Kharoufeh, J. (2011) Level-Dependent Quasi-Birth-and-Death Processes. Wiley Encyclopedia of Operations Research and Management Science.

Latouche, G., V. Ramaswami (1999) Introduction to Matrix Analytic Methods in Stochastic Modeling. SIAM.

Reed, J., U. Yechiali (2013) Queues in Tandem with Customer Deadlines and Retrials. Queueing System 73, 1-34.

Ross, S.M. (2007) Introduction to Probability Models. 9th Edition. ELSEVIER.

Wang, J., O. Baron, A. Scheller-Wolf (2015) M/M/c Queue with Two Priority Classes. Oper. Res., 63(3)733-749.

Ward, A., P. Glynn (2005) A Diffusion Approximation for a GI/GI/1 Queue with Balking or abandonment. Queueing System 50, 371-400.

Whitt, W. (2004) Efficiency-Driven Heavy-Traffic Approximations for Many-Server Queues with Abandonments. Management Science 50(10)1449-1461.

Zayas-Cabán G, Xie J, Green LV, Lewis ME (2013) Optimal control ofan emergency room triage and treatment process. Working paper, Cornell University, Ithaca, NY.

Zychlinski, N., A. Mandelbaum, P. Momčilović (2017) Tandem Queues with Blocking: Modeling, Analysis and Operational Insights via Fluid Models with Reflection. Working Paper, Technion – Israel Institute of Technology, Israel.

Appendix to
"Staffing Tandem Queues with Impatient Customers – Application in Financial Service Operations"

A1. Recursive Renewal Reward Extension

We now propose a method for the derivation of the service level measures of Station 2 deals' based on renewal reward theorem and QMCD. We use κ_j, the distribution of the number of Station 2 deals, to illustrate our method. Section A1.1 includes our second main theoretical contribution.

A1.1. QMCD and Renewal Reward Theorem

We focus on κ_j, the probability of having j Station 2 deals in the system. As the first step of QMCD, we decompose the Markov Chain into two subsystems, when $q_1 \leq n_1$ and when $q_1 > n_1$, and consider them separately.

Let $\sigma_0 = 0$ and assume that we start with an empty system. For $i = 1, 2, \ldots$, we define the stopping times $\tau_i = \inf\{t \,|Q_1(t) = n_1 + 1 \text{ and } t > \sigma_{i-1}\}$ and $\sigma_i = \inf\{t \,|Q_1(t) = n_1 \text{ and } t > \tau_i\}$; i.e., τ_i is the i^{th} time the system enters the subspace $\{(q_1, q_2) \,|q_1 > n_1\}$ from the subspace $\{(q_1, q_2) \,|q_1 \leq n_1\}$, and σ_i is the i^{th} time the system enters the subspace $\{(q_1, q_2) \,|q_1 \leq n_1\}$ from the subspace $\{(q_1, q_2) \,|q_1 > n_1\}$. Note that, due to abandonment, our system is stable, so that we have both $\tau_i < \infty$ and $\sigma_i < \infty$ for any $i < \infty$. From the definition of τ_i and σ_i, we have $\sigma_0 < \tau_1 < \sigma_1 < \tau_2 < \sigma_2 < \cdots < \tau_i < \sigma_i < \infty$. Clearly, from σ_i to τ_{i+1}, there are $q_1 \leq n_1$ Station 1 deals in the system, and from τ_i to σ_i, there are $q_1 > n_1$ Station 1 deals in the system.

From these definitions, τ_i and σ_i entirely depend on the number of Station 1 deals in the system. From Observation 1, then, the time periods from σ_i to τ_{i+1}, $i = 1, 2, \ldots$, are all independent and identically distributed (i.i.d). Since Station 1's waiting room is empty in these time periods, we call them "waiting room *empty periods*" (EP). We use a random variable, L_{EP}, to represent their length. Similarly, we use a random variable, L_{OP}, to represent the length of the i.i.d. time periods from τ_i to σ_i, $i = 1, 2, \ldots$, and we call these periods "waiting room *occupied periods*" (OP). As illustrated in the right side of Figure 2, EPs intertwine with OPs: once the system leaves the subspace $\{(q_1, q_2) \,|q_1 \leq n_1\}$, it enters the subspace $\{(q_1, q_2) \,|q_1 > n_1\}$, a ladder-like one dimensional infinite MC, and vice versa. Let $E[L_{EP}]$ and $E[L_{OP}]$ be the expected lengths of EP and OP, respectively. From the law of large numbers, we have $E[L_{EP}] = \lim_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k} (\tau_{i+1} - \sigma_i)$ and $E[L_{OP}] = \lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} (\sigma_i - \tau_i)$.

Gandhi et al. (2013) provide an innovative technique for solving ladder-like one dimensional infinite MCs using the renewal reward theorem (see, e.g., Ross 2007). The fundamental idea is to consider any quantity of interest as the "reward" earned per unit time in an MC, where the reward could be any function of the number of deals in the system. By the renewal reward theorem, the long-run average reward is the same as the expected reward earned over a cycle divided by the expected cycle length. For example, if the reward is 1 at any time t, the reward earned in a cycle is equal to the cycle length; if the reward equals the number of deals in the system at time t, the reward earned in a cycle is the accumulative number of deals in the system in the cycle, i.e., the expected number of deals in the system multiplied by the cycle length.

However, Gandhi et al.'s (2013) technique requires all "rung" transitions on the ladder to be uni-directional. Unfortunately, this special structure does not hold in many queueing networks, including the one we consider. For example, in Figure 2, the transitions between columns are bi-directional: a service completion or abandonment in Station 2 moves the system from column $i+1$ to i, while a service completion in Station 1 moves it from column i to $i+1$. Thus, the technique in Gandhi et al. (2013) cannot be applied directly here. We therefore extend their renewal reward theorem based approach to solve our MC as demonstrated below and call it the recursive renewal reward extension (RRRE).

We consider τ_i, $i=1,2,\ldots$, i.e., every time the system moves from level n_1 to level n_1+1, as a *renewal point*. We call the time period between τ_i and τ_{i+1} a *cycle*. Each cycle starts with an OP. After a certain time period, the system leaves the subspace $\{(q_1,q_2)\,|q_1>n_1\}$ and enters the subspace $\{(q_1,q_2)\,|q_1\leq n_1\}$; i.e., the system leaves the OP (from state (n_1+1,q_2), $q_2=0,\ldots,n_2+m$) and moves into the EP (at state (n_1,q_2), $q_2=0,\ldots,n_2+m$). Every cycle ends with the system leaving the subspace $\{(q_1,q_2)\,|q_1\leq n_1\}$ (entering the subspace $\{(q_1,q_2)\,|q_1>n_1\}$). Thus, each cycle is composed of an OP followed by an EP, and the expected length of any cycle is $E[L_{EP}]+E[L_{OP}]$. In contrast to Gandhi et al. (2013), the cycles we define may start at different states (with different q_2). Thus, while the lengths of these cycles are only dictated by the number of Station 1 deals and are i.i.d., the distribution of the number of Station 2 deals within these cycles depends on the starting state and is not necessarily i.i.d.

Note that κ_j is equivalent to the steady state proportion of time the system is at states $\{(q_1,q_2)\,|q_2=j,\,\forall q_1\}$. Let

$$\Phi_{\kappa_j}(t)=\begin{cases}1 \text{ if the system is at state } (q_1,q_2) \text{ s.t., } q_2=j,\forall q_1\\ 0 \text{ otherwise}\end{cases} \tag{OA.1}$$

be the rewards earned at time t. By the renewal reward theorem, the fraction of time the system spends at states $\{(q_1,q_2)\,|q_2=j,\,\forall q_1\}$ in steady state is the expected time spent at those states in a cycle divided by the average cycle length, i.e.,

$$\kappa_j=\frac{E\left[\int_{L_{OP}}\Phi_{\kappa_j}(t)\,dt\right]+E\left[\int_{L_{EP}}\Phi_{\kappa_j}(t)\,dt\right]}{E[L_{OP}]+E[L_{EP}]}. \tag{OA.2}$$

Continuing with the second step of QMCD, we now solve each of the subsystems. In Section A1.2, we derive $E[L_{EP}]$ and $E[L_{OP}]$. In Sections A1.3 and A1.4, we investigate the subspaces $\{(q_1,q_2)\,|q_1\leq n_1\}$ and $\{(q_1,q_2)\,|q_1>n_1\}$ separately for the expected reward in a cycle. Then, in Section A1.5, we express $E\left[\int_{L_{OP}}\Phi_{\kappa_j}(t)\,dt\right]+E\left[\int_{L_{EP}}\Phi_{\kappa_j}(t)\,dt\right]$ in Theorem A1.

A1.2. Expected Lengths of L_{EP} and L_{OP}

In what follows, we derive $E[L_{EP}]$ and $E[L_{OP}]$. Observation 1 states that from Station 1 deals' point of view, Station 1 operates as an $M/M/n_1+M$ system. As shown in Figure 2, the subspace $\{(q_1,q_2)\,|q_1\leq n_1\}$ is a finite MC. Let X_i be the expected first entrance time to level n_1+1, given that the system starts from level i, for $i=0,\ldots,n_1$. Clearly, from the definitions of $E[L_{EP}]$ and X_{n_1}, we have $E[L_{EP}]=X_{n_1}$.

PROPOSITION A1. *The expected length of L_{EP}, $E[L_{EP}] = X_{n_1}$, can be calculated from the following n_1+1 equations with n_1+1 unknowns:*

$$X_i = \begin{cases} \frac{1}{\lambda_1} + X_1 & \text{if } i = 0 \\ \frac{1}{\lambda_1+i(\mu_1+\theta_{1s})} + \frac{\lambda_1}{\lambda_1+i(\mu_1+\theta_{1s})} X_{i+1} + \frac{i(\mu_1+\theta_{1s})}{\lambda_1+i(\mu_1+\theta_{1s})} X_{i-1} & \text{if } i = 1,\ldots,n_1-1 \\ \frac{1}{\lambda_1+n_1(\mu_1+\theta_{1s})} + \frac{n_1(\mu_1+\theta_{1s})}{\lambda_1+n_1(\mu_1+\theta_{1s})} X_{n_1-1} & \text{if } i = n_1 \end{cases} . \tag{OA.3}$$

Note that (OA.3) is essentially a tridiagonal matrix equation, whose closed form solution can be derived by Gaussian elimination.

From the definition of $T^{(q_1)}$ in Section 3.1, it is clear that the L_{OP} is distributed identically to the $T^{(n_1+1)}$. Hence, $E[L_{OP}]$ is given in (7).

We next derive $E\left[\int_{L_{EP}} \Phi_{\kappa_j}(t)\,dt\right]$ and $E\left[\int_{L_{OP}} \Phi_{\kappa_j}(t)\,dt\right]$, and use (OA.2) to obtain κ_j. Note that $E[L_{EP}]$ and $E[L_{OP}]$ are independent of q_2; in contrast, both $E\left[\int_{L_{EP}} \Phi_{\kappa_j}(t)\,dt\right]$ and $E\left[\int_{L_{OP}} \Phi_{\kappa_j}(t)\,dt\right]$ depend on q_2 at the beginning of the EP and the OP. Therefore, deriving these quantities requires an analysis that conditions on q_2 at the beginning of a cycle, i.e., in a vector space that tracks q_2 at the beginning of cycles. This analysis in the vector space significantly extends Gandhi et al. (2013).

To demonstrate this challenge, we use κ_2 as an example but it can be replaced with other measures. Then, $\Phi_{\kappa_j}(t) = 1$ only at states $(q_1, 2)$, i.e., states with two Station 2 deals in the system. Now consider $E\left[\int_{L_{OP}} \Phi_{\kappa_j}(t)\,dt\right]$ in the MC in Figure 2 with a relatively large abandonment rate compared to the arrival and service rates; i.e., $\theta_{1s} = \theta_{1w} = 33 >> \lambda_1 = \mu_1 = \mu_2 = 1$. This means that after entering level n_1+1, with probability $\frac{2\mu_1+2\theta_{1s}+\theta_{1w}}{\lambda_1+2\mu_1+2\theta_{1s}+\theta_{1w}} > 0.99$, the OP will end after an $\exp(\lambda_1+2\mu_1+2\theta_{1s}+\theta_{1w})$ time period with an abandonment. Therefore, if an OP starts from state $(3,0)$, with probability greater than 0.99, no reward is collected in this OP, so that $E\left[\int_{L_{OP}} \Phi_{\kappa_j}(t)\,dt\right] \approx 0$. In contrast, if an OP starts from state $(3,2)$, the expected reward is at least $\frac{1}{\lambda_1+2\mu_1+2\mu_2+2\theta_{1s}+\theta_{1w}} = \frac{1}{104}$, i.e., $E\left[\int_{L_{OP}} \Phi_{\kappa_j}(t)\,dt\right] \geq \frac{1}{104}$. A similar discussion can be applied to the EP. To overcome this difficulty, we need to track the number of Station 2 deals in the system at the beginning of each EP and OP.

Let I be the identity matrix (of the required size). Let $\mathbf{r}^{(i)}$ be the vector of expected reward earned at level i and $\mathbf{r}^{(i)}_{q_2}$ represent the expected reward earned at state (i, q_2). Note from (OA.1) that the reward $\Phi_{\kappa_j}(t)$ is positive only at state (q_1, q_2) for $q_2 = j$. Using $v(i, q_2)$ in (3), we have

$$\mathbf{r}^{(i)}_{q_2} = \begin{cases} \frac{1}{v(i,q_2)} & \text{if } q_2 = j \\ 0 & \text{if } q_2 \neq j \end{cases} .$$

A1.3. Markov Chain's Transient Behavior during OP

We consider the OP, i.e., the subspace $\{(q_1, q_2) | q_1 > n_1\}$, with a focus on the expected *first passage reward* vector earned during an OP, α**,** based on the value of q_2 at the beginning of the OP; i.e., α_i represents the rewards earned during an OP, given that the OP starts with i Station 2 deals.

We apply the method of generating a recursive relation in Section 3 to the expected first passage reward vector, α.

PROPOSITION A2. *The expected first passage reward vector earned during an OP is*

$$\alpha = \left(I - \mathbf{A}_1^{(n_1+1)} - \mathbf{A}_0^{(n_1+1)} - P\{\Theta_{1w} > L_{OP}\} \mathbf{A}_0^{(n_1+1)} \mathbf{G}^{(n_1+1)} \right)^{-1} \mathbf{r}^{(n_1+1)}. \tag{OA.4}$$

A1.4. Markov Chain's Transient Behavior during EP

We now consider the EP, i.e., the subspace $\{(q_1, q_2) \,|q_1 \leq n_1\}$, with a focus on two values:

1. $\mathbf{H}$, the *first passage probability matrix* in EPs; i.e., $\mathbf{H}_{ij}$ represents the probability that an EP ends at state (n_1+1, j), given that it starts from state (n_1, i). Note that, similar to the matrices $\mathbf{A}_0^{(i)}$, $\mathbf{A}_1^{(i)}$, and $\mathbf{A}_2^{(i)}$ in Section 2, $\mathbf{H}$ is of size $(n_1+m+1) \times (n_1+m+1)$.

2. β, the expected *first passage reward* vector earned during an EP based on the value of q_2 at the beginning of the EP; i.e., β_i represents the rewards earned during an EP, given that the EP starts with i Station 2 deals.

Because the EP (the subspace $\{(q_1, q_2) \,|q_1 \leq n_1\}$) has a finite number of states, using matrix analytic methods (see, e.g., Latouche and Ramaswami 1999) to derive the first passage probability matrix $\mathbf{H}$ in the EPs is straightforward. Let Y_i be the first passage probability matrix to level $i+1$, given the sample path starts from level i, for $i = 0, \ldots, n_1$. Clearly, from the definitions of Y_{n_1} and $\mathbf{H}$, we have $\mathbf{H} = Y_{n_1}$.

PROPOSITION A3. *The first passage probability matrix during an EP* $\mathbf{H} = Y_{n_1}$ *can be calculated from the following* n_1+1 *matrix equations with* n_1+1 *unknowns:*

$$Y_i = \begin{cases} \mathbf{A}_0^{(0)} + \mathbf{A}_1^{(0)} Y_0 & \text{if } i = 0 \\ \mathbf{A}_0^{(i)} + \mathbf{A}_1^{(i)} Y_i + \mathbf{A}_2^{(i)} Y_{i-1} Y_i & \text{if } i = 1, \ldots, n_1 - 1 \\ \mathbf{A}_0^{(n_1)} + \mathbf{A}_1^{(n_1)} Y_{n_1} + \mathbf{A}_2^{(n_1)} Y_{n_1-1} Y_{n_1} & \text{if } i = n_1 \end{cases} . \tag{OA.5}$$

Following the idea of Proposition A3, we can derive β, the expected first passage reward vector earned during an EP. Let $\mathbf{z}_i$ be the expected first passage reward vector to level $i+1$, given the sample path starts from level i, for $i = 0, \ldots, n_1$. Note that, by definition, we have $\beta = \mathbf{z}_{n_1}$.

PROPOSITION A4. *The expected reward earned during an EP,* $\beta = \mathbf{z}_{n_1}$*, can be derived by solving the following* n_1+1 *sets of linear equations with* n_1+1 *unknown vectors:*

$$\mathbf{z}_i = \begin{cases} \mathbf{r}^{(0)} + \mathbf{A}_1^{(0)} \mathbf{z}_0 & \text{if } i = 0 \\ \mathbf{r}^{(i)} + \mathbf{A}_1^{(i)} \mathbf{z}_i + \mathbf{A}_2^{(i)} \left(\mathbf{z}_{i-1} + Y_{i-1} \mathbf{z}_i \right) & \text{if } i = 1, \ldots, n_1 - 1 \\ \mathbf{r}^{(n_1)} + \mathbf{A}_1^{(n_1)} \mathbf{z}_{n_1} + \mathbf{A}_2^{(n_1)} \left(\mathbf{z}_{n_1-1} + Y_{n_1-1} \mathbf{z}_{n_1} \right) & \text{if } i = n_1 \end{cases} . \tag{OA.6}$$

A1.5. Expected Reward Earned in a Cycle

After developing the first passage probability matrices between the EP and OP and the expected first passage reward vectors in both time periods, we can derive the expected rewards earned in a cycle. This is the last step of QMCD: combining the two subsystems together and normalizing the solution.

THEOREM A1. *The expected first passage reward vector earned in one cycle is*

$$E\left[\int_{L_{OP}} \Phi_{\kappa_j}(t)\,dt\right] + E\left[\int_{L_{EP}} \Phi_{\kappa_j}(t)\,dt\right] = \omega\alpha + \omega\mathbf{G}^{(n_1+1)}\beta, \tag{OA.7}$$

where ω is the unique nonnegative solution of

$$\omega\mathbf{G}^{(n_1+1)}\mathbf{H} = \omega, \tag{OA.8}$$

$$\text{and } \omega\overrightarrow{1} = 1, \tag{OA.9}$$

and $\mathbf{G}^{(n_1+1)}$, α, $\mathbf{H}$, and β are given in Propositions 2, A2, A3, and A4, respectively.

Now, with the expected reward and average cycle length developed, by using the renewal reward theorem, we can easily derive κ_j, the probability of having j Station 2 deals in the system, by substituting (OA.7), (7), and $E[L_{EP}]$ obtained from Proposition A1 into (OA.2) for $j = 0, 1, \ldots$.

We stress that in our method, only a few matrices and vectors must be derived by solving matrix equations. The computation is much less complex than in the approach described at the end of Section 3.

A1.6. Other Service Level Measures

So far, we have focused on the distributions of Q_2 (the number of Station 2 deals) in a two-station tandem queueing network with abandonment as an example to illustrate our methodology. The same method can be used to derive other service level measures, and the selection of the reward function $\Phi(t)$ is quite flexible.

As illustrated in Figure 1, there are four streams of deals flowing out of the system: abandonment from Station 1's waiting room, and balking, abandonment, and departure from Station 2. The abandonment rate from Station 1 can be calculated as $\sum_{i=n_1+1}^{\infty} \eta_i (n_1\theta_{1s} + (i-n_1)\theta_{1w})$, where η_i is given in (2). Recall that balking from Station 2's waiting room occurs when a Station 1 deal completes service at Station 1, and Station 2's waiting room is full, while abandonment or departure from Station 2 takes place when there are Station 2 deals in Station 2. Thus, to calculate the balking, abandonment, and departure rates from Station 2, we set

$$\Phi_{B_2}(t) = \begin{cases} \frac{\mu_1 \min(q_1, n_1)}{v(q_1, n_2+m)} & \text{if the system is at state } (q_1, q_2) \text{ s.t., } q_2 = n_2 + m \\ 0 & \text{otherwise} \end{cases},$$

$$\Phi_{Ab_2}(t) = \frac{\theta_{2s}\min(q_2, n_2) + \theta_{2w}\max(q_2 - n_2, 0)}{v(q_1, q_2)} \text{ for } q_1 = 0, 1, 2, \ldots,$$

and

$$\Phi_{D_2}(t) = \begin{cases} \frac{\mu_2 \min(q_2, n_2)}{v(q_1, q_2)} & \text{if the system is at state } (q_1, q_2) \text{ s.t., } q_2 > 0 \\ 0 & \text{otherwise} \end{cases},$$

respectively, and apply Theorem A1. The sum of these four deal out-flows should equal the arrival rate, and the departure rate from Station 2 is the ST.

We note that the RRRE works well for service level measures in Station 2 where the numerator of the reward function $\Phi(t)$ is level-independent. For SLMs in Station 1, this approach can be complicated, as explained in the discussion of the departure time of deal A in the proof of Proposition 1. Fortunately, from Observation 1, the service level measures in Station 1 can be derived using fundamental queueing theory tools, as in Section 2.2.

A2. Proofs and Algorithms

A2.1. Proof of Corollary 2

Recall that $P\{\Theta_{1w} > L_{OP}\} = \tilde{L}_{OP}(\theta_{1w})$. Then, by letting $s \to 0$ in (6), we can write $P\{\Theta_{1w} > L_{OP}\}$ as a function of $E[L_{OP}]$:

$$\begin{aligned} P\{\Theta_{1w} > L_{OP}\} &= \lim_{s\to 0} \frac{(\theta_{1w} + n_1(\mu_1+\theta_{1s}) + s)\tilde{L}_{OP}(s) - \theta_{1w} - n_1(\mu_1+\theta_{1s})}{\lambda_1 \tilde{L}_{OP}(s) - \lambda_1} \\ &= \frac{\tilde{L}_{OP}(0) + (\theta_{1w} + n_1(\mu_1+\theta_{1s}))\tilde{L}'_{OP}(0)}{\lambda_1 \tilde{L}'_{OP}(0)} \text{ (L'Hospital's Rule)} \\ &= \frac{\theta_{1w} + n_1(\mu_1+\theta_{1s})}{\lambda_1} - \frac{1}{\lambda_1 E[L_{OP}]}. \end{aligned}$$

A2.2. Proof of Proposition 2

In the beginning of any OP, the system is at level n_1+1 of the MC. Then, three types of transitions may happen: 1) The system moves to level n_1, with the one-step transition probability matrix $\mathbf{A}_2^{(n_1+1)}$, and the OP ends. In this case, the first passage probability matrix is $\mathbf{A}_2^{(n_1+1)}$. 2) The system moves to another state in the same level n_1+1, with one-step transition probability matrix $\mathbf{A}_1^{(n_1+1)}$. Because of the memoryless property, the system operates as if it starts from level n_1+1. In this case, the first passage probability matrix is $\mathbf{A}_1^{(n_1+1)}\mathbf{G}^{(n_1+1)}$. 3) The system moves to level n_1+2, with one-step transition probability matrix $\mathbf{A}_0^{(n_1+1)}$. Two Station 1 deals are now waiting for Station 1, and the repeating structure implies that the first passage probability matrix is $\mathbf{A}_0^{(n_1+1)}\left(P\{\Theta_{1w} \le L_{OP}\}\mathbf{G}^{(n_1+1)} + P\{\Theta_{1w} > L_{OP}\}\left(\mathbf{G}^{(n_1+1)}\right)^2\right)$. That is, in this case, if deal A has abandoned before the end of the OP initiated by deal B (w.p. $P\{\Theta_{1w} \le L_{OP}\}$), the first passage probability matrix is $\mathbf{G}^{(n_1+1)}$ (as observed by deal B); if deal A does not abandon during this period (w.p. $P\{\Theta_{1w} > L_{OP}\}$), the first passage probability matrix is $\left(\mathbf{G}^{(n_1+1)}\right)^2$ (the one observed by deal B followed by the one to be observed by deal A). Combining the above three points gives (10).

A2.3. Proof of Proposition 4

When $\lambda_1 < [\dot{n}_1]\mu_1$ and $\theta \to 0^+$, the BnchAR ensures that all arrivals can obtain service and then leave as ST.

When $\lambda_1 \ge [\dot{n}_1]\mu_1$, from Chapter 2.10.2 of Gross et al. (2008), we have Station 1's idling probability under BnchAR:

$$p_0 = \left(1 + \sum_{n=1}^{\infty}\prod_{i=1}^{n}\frac{\lambda_1}{\min([\dot{n}_1], i)\mu + i\theta}\right)^{-1},$$

which converges to zero when $\theta \to 0^+$. Thus, the utilization of the station with lower capacity converges to one; i.e., all its servers almost always work with no idleness, so the system's output converges to $Poisson(\min([\dot{n}_1]\mu_1, (N-[\dot{n}_1])\mu_2))$. Clearly, no other assignment rules can generate higher ST than the BnchAR.

A2.4. Algorithm for the First Passage Probability Matrix

Let ϵ be the error tolerance for the numerical algorithm.

Algorithm A1 *Deriving the first passage probability matrix* $\mathbf{G}^{(n_1+i)}$.

Step 1: Set $\mathbf{G}^{(n_1+i)} = \left(I - \left(\mathbf{A}_1^{(n_1+i)} + P\left\{\Theta_{1w} \leq L_{T^{(n_1+i)}}\right\}\mathbf{A}_0^{(n_1+i)}\right)\right)^{-1}\mathbf{A}_2^{(n_1+i)}$.

Step 2: Set $X = \mathbf{A}_1^{(n_1+i)} + P\left\{\Theta_{1w} \leq L_{T^{(n_1+i)}}\right\}\mathbf{A}_0^{(n_1+i)} + P\left\{\Theta_{1w} > L_{T^{(n_1+i)}}\right\}\mathbf{A}_0^{(n_1+i)}\mathbf{G}^{(n_1+i)}$.

Step 3: Set $\mathbf{G}^{(n_1+i)} = (I - \mathbf{X})^{-1}\mathbf{A}_2^{(n_1+i)}$.

Step 4: If $\max\left|\overrightarrow{1} - \mathbf{G}^{(n_1+i)}\overrightarrow{1}\right| > \epsilon$, *then go to Step 2; otherwise STOP.*

Clearly, the smaller the error tolerance ϵ the more accurate the result. The convergence of Algorithm A1 is guaranteed by Theorem 8.1.1 in Latouche and Ramaswami (1999).

Online Appendix to
"Staffing Tandem Queues with Impatient Customers – Application in Financial Service Operations"

OA1. Proofs and Algorithms

OA1.1. Proof of Proposition A1

The proof is based on the sample path method and the memoryless property of Markovian systems. Say the system is currently at level n_1. On average, the system stays at level n_1 for $\frac{1}{\lambda_1+n_1(\mu_1+\theta_{1s})}$ time units. Then

- with probability $\frac{\lambda_1}{\lambda_1+n_1(\mu_1+\theta_{1s})}$, the system moves to level n_1+1, and, in this case, X_{n_1} is zero;
- with probability $\frac{n_1(\mu_1+\theta_{1s})}{\lambda_1+n_1(\mu_1+\theta_{1s})}$, the system moves to level n_1-1. From the memoryless property, the system will operate as if it starts from level n_1-1. In this case, X_{n_1} is $\frac{n_1(\mu_1+\theta_{1s})}{\lambda_1+n_1(\mu_1+\theta_{1s})}X_{n_1-1}$.

Combining the above two points gives the first equation in (OA.3) for $i=0$ and a similar discussion gives the rest of (OA.3).

OA1.2. Proof of Proposition A2

We first use the MC in Figure 2 as an example to derive α, focusing on the probability of having two Station 2 deals in the system; i.e., $j=2$ in the definition of $\Phi_{\kappa_2}(t)$ in (OA.1). The proof for other rewards is similar.

We consider the sample path starting from state $(3,0)$. The system stays in state $(3,0)$ for an $\exp(v(3,0))$ time period, with no reward. If the next event is abandonment or completion 1 (w.p. $\frac{2\mu_1+2\theta_{1s}+\theta_{1w}}{v(3,0)}$), then OP ends with no reward. If the next event is an arrival at Station 1, following the same discussion as (9) and (10), the new arrival initiates an OP with the same distribution as L_{OP}. The number of Station 2 deals at the beginning of this OP (initiated by the new arrival) is zero, so the expected reward obtained in this time period is α_0. At the end of this time period, the distribution of the number of Station 2 deals is: 0 w.p. $\left[\mathbf{G}^{(n_1+1)}\right]_{00}$, 1 w.p. $\left[\mathbf{G}^{(n_1+1)}\right]_{01}$, and 2 w.p. $\left[\mathbf{G}^{(n_1+1)}\right]_{02}$. By using the memoryless property and conditioning on whether deal A is still waiting or not, we can write the expected reward earned in the OP initiated by deal A. From the above discussion, we get:

$$\begin{aligned}\alpha_0 = {} & \frac{2\mu_1+2\theta_{1s}+\theta_{1w}}{v(3,0)}\cdot 0+\frac{\lambda_2}{v(3,0)}\alpha_1+\frac{\lambda_1}{v(3,0)}\cdot(\alpha_0+P\{\Theta_{1w}\le L_{OP}\}\cdot 0\\ & +P\{\Theta_{1w}>L_{OP}\}\left(\left[\mathbf{G}^{(n_1+1)}\right]_{00}\alpha_0+\left[\mathbf{G}^{(n_1+1)}\right]_{01}\alpha_1+\left[\mathbf{G}^{(n_1+1)}\right]_{02}\alpha_2+\left[\mathbf{G}^{(n_1+1)}\right]_{03}\alpha_3\right)\big),\end{aligned} \tag{OB.1}$$

Following a similar discussion, for sample paths starting from states $(3,1)$, $(3,2)$, and $(3,3)$, we derive three other equations for α_1, α_2, and α_3, respectively:

$$\begin{aligned}\alpha_1 = {} & \frac{\mu_2+\theta_{2s}}{v(3,1)}\alpha_0+\frac{\lambda_2}{v(3,1)}\alpha_2+\frac{\lambda_1}{v(3,1)}\cdot(\alpha_1\\ & +P\{\Theta_{1w}>L_{OP}\}\left(\left[\mathbf{G}^{(n_1+1)}\right]_{10}\alpha_0+\left[\mathbf{G}^{(n_1+1)}\right]_{11}\alpha_1+\left[\mathbf{G}^{(n_1+1)}\right]_{12}\alpha_2+\left[\mathbf{G}^{(n_1+1)}\right]_{13}\alpha_3\right)\big),\end{aligned} \tag{OB.2}$$

$$\begin{aligned}\alpha_2 = {} & \frac{1}{v(3,2)}+\frac{2\mu_2+2\theta_{2s}}{v(3,2)}\alpha_1+\frac{\lambda_2}{v(3,2)}\alpha_3+\frac{\lambda_1}{v(3,2)}\cdot(\alpha_2\\ & +P\{\Theta_{1w}>L_{OP}\}\left(\left[\mathbf{G}^{(n_1+1)}\right]_{20}\alpha_0+\left[\mathbf{G}^{(n_1+1)}\right]_{21}\alpha_1+\left[\mathbf{G}^{(n_1+1)}\right]_{22}\alpha_2+\left[\mathbf{G}^{(n_1+1)}\right]_{23}\alpha_3\right)\big),\end{aligned} \tag{OB.3}$$

and

$$\alpha_3 = \frac{2\mu_2+\theta_{1w}}{v\,(3,3)}\alpha_2 + \frac{\lambda_1}{v\,(3,3)}\cdot(\alpha_3 + P\{\Theta_{1w} > L_{OP}\}\left(\left[\mathbf{G}^{(n_1+1)}\right]_{30}\alpha_0 + \left[\mathbf{G}^{(n_1+1)}\right]_{31}\alpha_1 + \left[\mathbf{G}^{(n_1+1)}\right]_{32}\alpha_2 + \left[\mathbf{G}^{(n_1+1)}\right]_{33}\alpha_3\right)) \quad \text{(OB.4)}$$

Solving these four equations with four unknowns gives α_0, α_1, α_2, and α_3.

Using the one-step transition matrices $\mathbf{A}_0^{(n_1+1)}$ and $\mathbf{A}_1^{(n_1+1)}$, we can write (OB.1-OB.4) in matrix form:

$$\alpha = \mathbf{r}^{(3)} + \left(\mathbf{A}_1^{(3)} + \mathbf{A}_0^{(3)}\left(I + P\{\Theta_{1w} > L_{OP}\}\,\mathbf{G}^{(n_1+1)}\right)\right)\alpha.$$

Recall that $r^{(3)} = \left[0, 0, \frac{1}{v(3,2)}, 0\right]^T$ is the expected reward vector earned at level 3.

Following the same thinking, for the general case, we can write a matrix equation:

$$\alpha = \mathbf{r}^{(n_1+1)} + \left(\mathbf{A}_1^{(n_1+1)} + \mathbf{A}_0^{(n_1+1)}\left(I + P\{\Theta_{1w} > L_{OP}\}\,\mathbf{G}^{(n_1+1)}\right)\right)\alpha.$$

Note from the discussion in the proof of Proposition A3, the matrix $I - \mathbf{A}_1^{(n_1+1)} - \mathbf{A}_0^{(n_1+1)} - \mathbf{A}_0^{(n_1+1)}P\{\Theta_{1w} > L_{OP}\}\,\mathbf{G}^{(n_1+1)}$ is invertible. Thus, α can be solved as (OA.4).

OA1.3. Proof of Proposition A3

As we did for Proposition A1, we prove Proposition A3 by discussing the sample path and using the memoryless property of Markovian systems.

If the system is at level n_1, three possible transitions may happen next: 1) The system moves to level n_1+1, with the one-step transition probability matrix $\mathbf{A}_0^{(n_1)}$. In this case, the EP ends, and the first passage probability matrix is $\mathbf{A}_0^{(n_1)}$. 2) The system moves to a different state at the same level n_1, with the one-step transition probability matrix $\mathbf{A}_1^{(n_1)}$. Using the memoryless property, the system will operate as if it had started from level n_1, yielding a first passage probability matrix of $\mathbf{A}_1^{(n_1)}Y_{n_1}$. 3) The system moves to level n_1-1, with the one-step transition probability matrix $\mathbf{A}_2^{(n_1)}$. Now, the sample path needs to return to level n_1 with a first passage probability matrix Y_{n_1-1}, before it enters level n_1+1. Using the memoryless property, the first passage probability matrix when it moves to level n_1+1 is once again Y_{n_1}. Therefore, in this case, the first passage probability matrix is $\mathbf{A}_2^{(n_1)}Y_{n_1-1}Y_{n_1}$. Combining the above three points gives the last equation in (OA.5) for $i = n_1$.

Note that if a matrix X has the property that $\lim_{i\to\infty} X^i = 0$, then $I - X$ is invertible and $(I-X)^{-1} = \sum_{i=0}^{\infty} X^i$. Clearly, $\lim_{i\to\infty}\left(\mathbf{A}_1^{(n_1)} + \mathbf{A}_2^{(n_1)}Y_{n_1-1}\right)^i = 0$, so (OA.5) can be written as $Y_{n_1} = \left(\mathbf{I} - \mathbf{A}_1^{(n_1)} - \mathbf{A}_2^{(n_1)}Y_{n_1-1}\right)^{-1}\mathbf{A}_0^{(n_1)}$.

In a similar fashion, we derive the other n_1 equations (OA.5) and solve these n_1+1 matrix equations with n_1+1 unknowns $Y_0, Y_1, \ldots, Y_{n_1}$ recursively from Y_0 to Y_{n_1}.

OA1.4. Proof of Proposition A4

As in the proofs of Propositions A1 and A3, for Proposition A4 we discuss the next moves of the sample path.

Say the system is at level n_1. A reward of $r^{(n_1)}$ will be collected before one of the next three possible transitions: 1) The system moves to level n_1+1, with the one-step transition probability matrix $\mathbf{A}_0^{(n_1)}$. In this case, the EP ends, and no more reward is collected. 2) The system stays at

level n_1 after a transition, with the one-step transition probability matrix $\mathbf{A}_1^{(n_1)}$. Then, from the memoryless property, the expected future reward is z_{n_1}. 3) The system moves to level n_1-1, with the one-step transition probability matrix $\mathbf{A}_2^{(n_1)}$. A reward z_{n_1-1} is collected before the sample path returns to level n_1, according to the first passage probability matrix Y_{n_1-1}, (derived in Section A1.4). Then, using the memoryless property, the expected future reward is the same as if the sample path started from level n_1, z_{n_1}. The above discussion gives the first equation in (OA.6) for $i=0$.

Following a similar process, we derive the other n_1 equations in (OA.6). Using a discussion similar to the one in the proof of Proposition A3, we can solve (OA.6) recursively for $z_0, z_1, \ldots, z_{n_1}$.

OA1.5. Proof of Theorem A1

First, let us review the process for each cycle. Every cycle starts with the sample path entering the subspace $\{(q_1,q_2)\,|q_1>n_1\}$, i.e., when the OP starts. During the OP, the expected reward vector, α, depends on the value of q_2 at the beginning of the OP. At the end of the OP, an EP starts; i.e., the sample path enters the subspace $\{(q_1,q_2)\,|q_1\le n_1\}$, according to the first passage probability matrix $\mathbf{G}^{(n_1+1)}$, and this gives the distribution of Q_2 at the beginning of the EP. Similarly, during the EP, we collect the expected reward, β, and exit according to the first passage probability matrix H. After this renewal epoch, another cycle starts, following the same procedure.

From this discussion, we observe that the first passage probability matrix for one cycle is $\mathbf{G}^{(n_1+1)}\mathbf{H}$. As the system reaches steady state when $t\to\infty$, the limit $\lim_{i\to\infty}\left(\mathbf{G}^{(n_1+1)}\mathbf{H}\right)^i$ exists. It has identical rows, and we denote each row by ω. From Theorem 4.1 of Ross 2007, we know ω is the unique nonnegative solution of (OA.8-OA.9).

Thus, in steady state, every cycle (or each OP) starts with i Station 2 deals with probability ω_i, for $i=0,\ldots,n_2+m$. Similarly, in steady state, the number of Station 2 deals in the beginning of each EP is distributed as $\omega\mathbf{G}^{(n_1+1)}$.

Given the steady state probability distribution of Q_2 at the beginning of each OP and EP, (OA.7) is straightforward.

OA2. No Abandonments during Service

Applications of tandem queueing system, where customers need to visit several stations in sequence, include call centers, where customers talk to general call-takers before being transferred to specialists, hospital emergency rooms, where patients are admitted by triage nurses and then diagnozed by a doctor. In these applications, customers rarely abandon during service. Moreover, due to the waiting cost already incurred, customers waiting for the downstream station may abandon less often than those waiting for the upstream station. To adapt our general model to these applications, we simply use $\theta_{1s}=\theta_{2s}=0$ and $\theta_{1w}\ge\theta_{2w}>0$. In this section, we carry out an initial numerical study in this direction with $\theta_{1w}=\theta_{2w}=\theta$, but no abandonments during service, directly departures from Station 1, or external arrivals to Station 2, i.e., $p=1$, $\lambda_2=0$ and $\theta_{1s}=\theta_{2s}=0$, and develop managerial insights into the operations of such systems.

We consider two questions:

1. *How can we assign $N\ge 2$ servers into a two-station tandem queueing network with abandonments to maximize throughput?*

2. *What is the minimum number of servers needed to achieve a throughput target in such a network?*

We start with the first question. In Section OA2.1, we discuss how the assignment rule affects the throughput and use enumeration to search for the optimal assignment rule. In Section OA2.2, we define an easily calculable benchmark assignment rule and then compare the optimal and the benchmark assignment rules to gain insight into refining the search method. In Section OA2.3, we answer the second question by generating a list of best performances of different total numbers of servers. At this point, the staffing problem in a tandem queue service system can be fully addressed by choosing the optimal staffing level for any throughput target.

OA2.1. Optimal Assignment Rule, Given N Servers

For a fixed total number of servers, $N \geq 2$, suppose n_1 $(1 \leq n_1 \leq N-1)$ servers are assigned to Station 1, and the rest, $n_2 = N - n_1$, are assigned to Station 2. When no confusion arises, we use n_1, instead of (n_1, n_2), to represent an assignment rule.

On the one hand, when n_1 is small (i.e., n_2 is large), Station 2 is able to accept most customers before they abandon, but many customers abandon Station 1's waiting room before reaching Station 1, making the input rate to Station 2 too low. In this case, we can assign some of Station 2's servers to Station 1 to increase the system's throughput. On the other hand, when n_1 is large (i.e., n_2 is small), Station 1 is able to capture most customers before they abandon, but Station 2 does not have enough capacity to handle all the input from Station 1, causing many customers to abandon Station 2's waiting room; consequently, Station 1's work on these customers is wasted. Therefore, it is better to move some servers from Station 1 to Station 2 to assure a higher throughput. From this discussion, we see that the throughput is an increasing function of n_1, when n_1 is small, and a decreasing function of n_1, when n_1 is large – i.e., close to N.

Let $TP(n_1)$ denote the throughput of this tandem queueing system under assignment rule n_1. Using the method described in Section 3 and Appendix A1, for a series of systems with $n_1 = 1, 2, \ldots, N-1$ servers at Station 1, we calculate $TP(n_1)$ and find the optimal assignment rule (OptAR) $n_1^* = \arg\max_{1 \leq n_1 \leq N-1} TP(n_1)$ through enumeration. Figure 8 records the throughput as functions of n_1 and the OptAR n_1^*, for $\lambda_1 = 40$, $N = 40$, $\theta = 1$, and $(\mu_1, \mu_2) \in \{(1,1), (2,2), (1,2), \text{and } (2,1)\}$. We see from Figure 8 that our intuition is valid: the throughput $TP(n_1)$ is, in fact, *concave* (initially increasing and then decreasing) in n_1 and has a global maximum. The concavity of the throughput holds for all other parameter settings we test. We summarize the observation below.

Observation OA1 *For any fixed λ_1, μ_1, μ_2, θ, and $N \geq 2$, the throughput $TP(n_1)$ is an initially increasing and then decreasing* ***concave*** *function of n_1, for $n_1 = 1, \ldots, N-1$.*

OA2.2. Optimal Assignment Rule vs. Benchmark Assignment Rule

When the total number of servers N is large, the search time for the OptAR n_1^*, using enumeration, can be long. In this section, we make observations that can help reduce the search space.

Given N and service rates in two stations (μ_1 & μ_2), we define $\{\dot{n}_1 | \dot{n}_1\mu_1 = \dot{n}_2\mu_2 \text{ and } \dot{n}_1 + \dot{n}_2 = N\}$ as the *benchmark assignment rule* (BnchAR), under which, Stations 1 and 2 have identical capacities. Note that $\dot{n}_1$ is independent of θ and may be a fraction; however, at this stage, we only use it for comparison, as rounding has little effect on our analysis. For now, we consider BnchAR as a *virtual* assignment rule and assume servers can be assigned in fractions; after making the comparison, we return to the rounding issue. We call $\dot{n}_1\mu_1$ the *benchmark capacity*. Specifically,

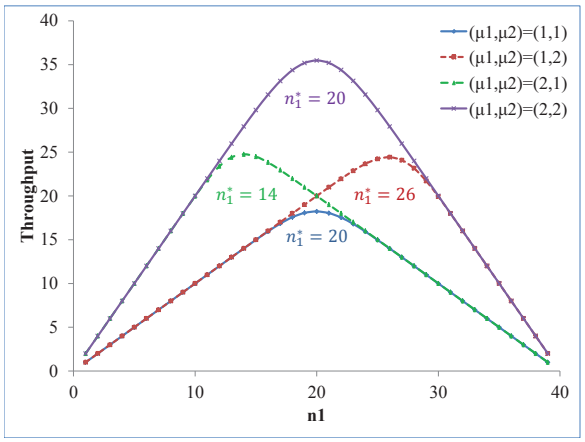

Figure 8 **Throughput as a function of** $n_1 \in \{1,\ldots,N-1\}$ **when** $\lambda = 40$, $N = 40$, $\theta = 1$, **and** $(\mu_1,\mu_2) \in \{(1,1),\ (2,2),\ (1,2),$ **and** $(2,1)\}$.

we compare the OptAR n_1^* with the BnchAR $\dot{n}_1$ to provide intuitions and guidelines for staffing in tandem queueing systems with impatient customers. These intuitions and guidelines were not previously available because of a lack of exact evaluation methods for such systems.

We start with the following intuitive proposition for the BnchAR when $\theta \to 0^+$.

PROPOSITION OA1. *In a tandem queueing system with abandonment rate* $\theta \to 0^+$*, the BnchAR is optimal in the domain of virtual assignment rules, and the maximum throughput is* $\min(\lambda_1, \dot{n}_1\mu_1)$.

Proof of Proposition OA1 When $\lambda_1 < \dot{n}_1\mu_1$ and $\theta \to 0^+$, the BnchAR ensures that all arrivals can obtain service and then leave as throughput. When $\lambda_1 \geq \dot{n}_1\mu_1$, from Chapter 2.10.2 of Gross et al. (2008), we have Station 1's idling probability under the BnchAR:

$$p_0 = \left(1 + \sum_{n=1}^{\infty}\prod_{i=1}^{n}\frac{\lambda_1}{\min(\dot{n}_1, i)\,\mu + i\theta}\right)^{-1},$$

which converges to zero when $\theta \to 0^+$. Thus, Station 1's utilization converges to one; i.e., all its $\dot{n}_1$ servers almost always work at rate μ_1 with no idleness, so Station 1's output, i.e., the arrival to Station 2, converges to $Poisson(\dot{n}_1\mu_1)$. The same discussion can be applied to Station 2 to show that the throughput converges to $\dot{n}_2\mu_2 = \dot{n}_1\mu_1$ when $\theta \to 0^+$. Clearly, no other assignment rules can generate higher throughput than the BnchAR. □

Similar to the discussion in Section OA2.1, when $\theta > 0$, there are two conflicting effects of Station 1's capacity on Station 2; both effects dictate how the OptAR changes from the BnchAR. For an arrival to contribute to the throughput of our system, Station 1 needs to capture her before abandonment and pass her on to Station 2 as input, and Station 2 needs to serve her before abandonment. On the one hand, for Station 1 to capture enough customers before they abandon and thus maintain an adequate input rate to Station 2, there is pressure to assign more servers to Station 1. We call this the *deal-capturing effect*. On the other hand, part of Station 1's capacity is wasted on customers who eventually abandon Station 2's waiting room. Thus, there is a tendency to move servers from Station 1 to Station 2 to increase Station 2's capacity and reduce this undesirable abandonment. We call this the *customer-loss effect*. When θ increases, the deal-capturing and

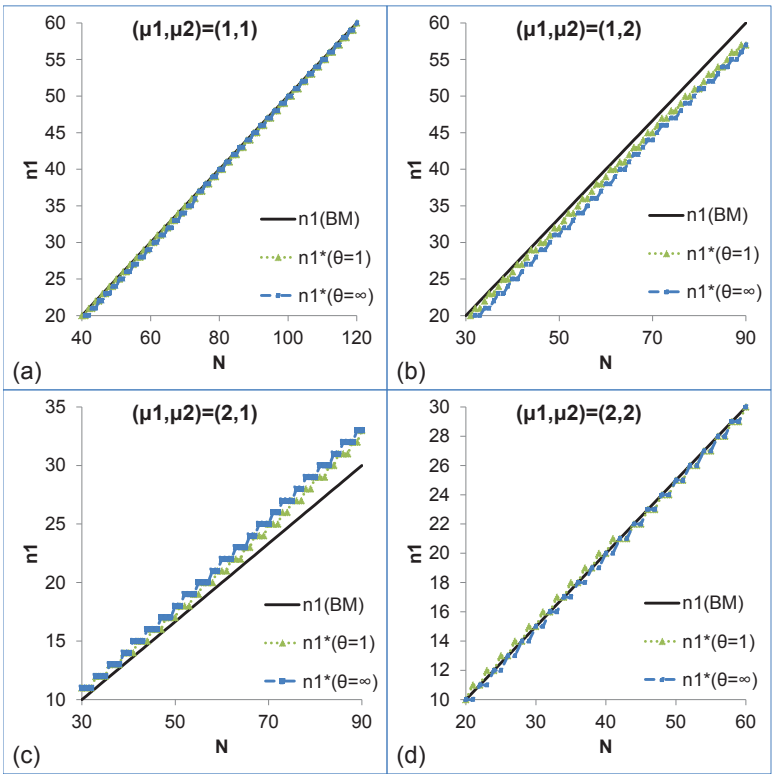

Figure 9 **OptAR n_1^* and BnchAR $\dot{n}_1$ as functions of $N \in \left\{ \left\lfloor \frac{1}{2}\left(\frac{\lambda}{\mu_1}+\frac{\lambda}{\mu_2}\right) \right\rfloor, \ldots, \left\lceil \frac{3}{2}\left(\frac{\lambda}{\mu_1}+\frac{\lambda}{\mu_2}\right) \right\rceil \right\}$ under $\lambda = 40$ and $\theta \in \{1, \infty\}$ for $(\mu_1, \mu_2) =$ (a) $(1,1)$; (b) $(1,2)$; (c) $(2,1)$; (d) $(2,2)$.**

customer-loss effects are both strengthened, and the interplay of these two effects dictates the OptAR.

When $\mu_1 = \mu_2$, increasing θ strengthens the deal-capturing and customer-loss effects at similar scales, so that the OptAR remains close to the BnchAR. We next examine the $\mu_1 \neq \mu_2$ case. Consider two independent multi-server queues with identical arrival rates and capacities but different service rates. The system with a higher service rate has fewer servers and a longer expected waiting time. In the presence of abandonment, a longer waiting time leads to a higher probability of abandonment. Thus, in a tandem queueing system under the BnchAR, the station with the higher service rate will have more abandonments. If $\mu_1 > \mu_2$, the deal-capturing effect dominates the customer-loss effect and the OptAR has a tendency to assign more servers to Station 1 than the BnchAR; but if $\mu_1 < \mu_2$, the customer-loss effect dominates the deal-capturing effect, and the OptAR has a tendency to assign more servers to Station 2 than the BnchAR.

In Figure 9, we compare n_1^* with $\dot{n}_1$; when N increases from $\left\lfloor \frac{1}{2}\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_1}{\mu_2}\right) \right\rfloor$ to $\left\lceil \frac{3}{2}\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_1}{\mu_2}\right) \right\rceil$, the service capacity increases from scarcity to plenty for the following cases: (i) arrival rate $\lambda_1 = 40$; (ii) different service rates $(\mu_1, \mu_2) \in \{(1,1), (2,2), (1,2), \text{and } (2,1)\}$, representing the $\mu_1 = \mu_2$, $\mu_1 < \mu_2$, and $\mu_1 > \mu_2$ cases, respectively; (iii) abandonment rates $\theta \in \{1, \infty\}$, representing moderately and extremely impatient customers, respectively. (Note that the $\theta = \infty$ case corresponds to a tandem queueing system with no waiting room and can be solved without the derivation in this paper.)

We see from Figures 9(a, d) that when $\mu_1 = \mu_2$, the OptAR remains close to the BnchAR as expected. Further, the distance between n_1^* and $\dot{n}_1$ is, at most, one. Next, we observe from Figures

9(b, c) that when $(\mu_1,\mu_2)=(2,1)$, the OptAR assigns more servers to Station 1 than the BnchAR, and when $(\mu_1,\mu_2)=(1,2)$, the OptAR assigns more servers to Station 2 than the BnchAR. This fits our intuition discussed above.

These observations hold in other parameter settings. To substantiate this, for each instance in $\{(\mu_1,\mu_2)\,|\mu_1,\mu_2=1,\ldots,10\}$, we derive a set of $n_1^*-\dot{n}_1$ values $S(\mu_1,\mu_2)=\left\{n_1^*-\dot{n}_1\,\middle|N\in\left\{\left\lfloor\frac{1}{2}\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_1}{\mu_2}\right)\right\rfloor,\ldots,\left\lceil\frac{3}{2}\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_1}{\mu_2}\right)\right\rceil\right\}\right.$ and $\theta\in\{1,\ldots,10,\infty\}\Big\}$; we then record the mean, minimum, and maximum of set $S(\mu_1,\mu_2)$ in Table 1. Recall that n_1^* is an integer, while $\dot{n}_1$ may be a fraction, so $n_1^*-\dot{n}_1$ may be a fraction.

$\mu_1\backslash\mu_2$	1	2	3	4	5	6	7	8	9	10
1	-0.3^{0}_{-1}	$-1.8^{-0.3}_{-3.7}$	$-2.3^{-0.5}_{-4.5}$	$-2.5^{-0.6}_{-5.2}$	$-2.6^{-0.7}_{-5.3}$	$-2.6^{-0.6}_{-5.3}$	$-2.6^{-0.6}_{-5.5}$	$-2.6^{-0.6}_{-5.6}$	$-2.6^{-0.6}_{-5.5}$	$-2.5^{-0.6}_{-5.3}$
2	$1.4^{3.3}_{0.3}$	$-0.2^{0.5}_{-0.5}$	$-0.7^{0.2}_{-1.6}$	$-1.1^{0}_{-2.3}$	$-1.3^{0}_{-2.6}$	$-1.4^{-0.3}_{-3.0}$	$-1.4^{-0.1}_{-3.3}$	$-1.5^{-0.2}_{-3.4}$	$-1.5^{-0.3}_{-3.5}$	$-1.5^{-0.2}_{-3.3}$
3	$1.9^{4.5}_{0.5}$	$0.5^{1.4}_{-0.2}$	$-0.2^{0.5}_{-0.5}$	$-0.4^{0.3}_{-1.1}$	$-0.6^{0.3}_{-1.5}$	-0.8^{0}_{-2}	$-0.9^{0}_{-1.9}$	$-0.9^{0}_{-2.4}$	$-1^{0}_{-2.3}$	$-1^{0.1}_{-2.5}$
4	$2.1^{5.2}_{0.4}$	$0.9^{2.3}_{0}$	$0.3^{1}_{-0.3}$	$-0.2^{0.5}_{-0.5}$	$-0.3^{0.4}_{-1}$	$-0.4^{0.2}_{-1.2}$	$-0.6^{0.3}_{-1.4}$	$-0.6^{0}_{-1.7}$	$-0.7^{0.2}_{-1.5}$	$-0.7^{0.1}_{-2}$
5	$2.1^{5.2}_{0.3}$	$1.0^{2.6}_{0}$	$0.5^{1.4}_{-0.1}$	$0.1^{0.8}_{-0.4}$	$-0.2^{0.5}_{-0.5}$	$-0.2^{0.5}_{-0.8}$	$-0.3^{0.3}_{-1.1}$	$-0.4^{0.3}_{-1.3}$	$-0.5^{0.2}_{-1.3}$	$-0.5^{0}_{-1.3}$
6	$2.2^{5.3}_{0.3}$	$1.1^{3}_{0.3}$	$0.6^{1.7}_{0}$	$0.3^{1}_{-0.2}$	$0.1^{0.5}_{-0.5}$	$-0.1^{0.5}_{-0.5}$	$-0.2^{0.4}_{-0.8}$	$-0.2^{0.4}_{-0.9}$	$-0.3^{0.4}_{-1}$	$-0.4^{0.3}_{-1.1}$
7	$2.2^{5.5}_{0.3}$	$1.2^{2.8}_{0.1}$	$0.7^{1.9}_{0}$	$0.5^{1.3}_{-0.1}$	$0.2^{0.8}_{-0.3}$	$0.2^{0.6}_{-0.4}$	$-0.1^{0.5}_{-0.5}$	$-0.1^{0.3}_{-0.7}$	$-0.2^{0.4}_{-0.9}$	$-0.3^{0.3}_{-0.9}$
8	$2.2^{5.6}_{0.3}$	$1.3^{3.4}_{0.2}$	$0.8^{1.9}_{0}$	$0.5^{1.3}_{0}$	$0.3^{1.1}_{-0.3}$	$0.2^{0.7}_{-0.3}$	$0.1^{0.5}_{-0.3}$	$-0.1^{0.5}_{-0.5}$	$0^{0.4}_{-0.7}$	$-0.2^{0.3}_{-0.8}$
9	$2.1^{5.4}_{0.3}$	$1.3^{3.3}_{0.1}$	$0.8^{2.3}_{0}$	$0.6^{1.5}_{-0.2}$	$0.4^{1.2}_{-0.2}$	$0.2^{0.8}_{-0.2}$	$0.1^{0.6}_{-0.4}$	$0^{0.4}_{-0.4}$	$-0.1^{0.5}_{-0.5}$	$0^{0.4}_{-0.7}$
10	$2.1^{5.3}_{0.3}$	$1.3^{3.2}_{0.2}$	$0.9^{2.2}_{-0.1}$	$0.6^{1.6}_{-0.1}$	$0.5^{1.3}_{0}$	$0.3^{1}_{-0.3}$	$0.2^{0.8}_{-0.3}$	$0.1^{0.7}_{-0.3}$	$0^{0.3}_{-0.4}$	$-0.1^{0.5}_{-0.5}$

Table 1 Mean, minimum, and maximum, $Mean^{Max}_{Min}$ of $n_1^*-\dot{n}_1$ under any $\theta\in\{1,\ldots,10,\infty\}$ and $N\in\left\{\left\lfloor\frac{1}{2}\left(\frac{\lambda}{\mu_1}+\frac{\lambda}{\mu_2}\right)\right\rfloor,\ldots,\left\lceil\frac{3}{2}\left(\frac{\lambda}{\mu_1}+\frac{\lambda}{\mu_2}\right)\right\rceil\right\}$ for $\lambda=40$ and $(\mu_1,\mu_2)\in\{(\mu_1,\mu_2)\,|\mu_1,\mu_2=1,\ldots,10\}$.

We see from Table 1 that

- When $\mu_1=\mu_2$ (i.e., all diagonal instances), all means are below zero, while all minimums and maximums are between -1 and 0.5, so every $n_1^*-\dot{n}_1$ in the set $S(\mu_1,\mu_2)$ is between -1 and 0.5. Thus, we have $\lfloor\dot{n}_1\rfloor-1\le n_1^*\le\lceil\dot{n}_1\rceil$ in this case.
- When $\mu_1>\mu_2$ (i.e., all lower triangular instances), all minimums are ≥-0.5, while all means and maximums are non-negative. This means that $n_1^*\ge\lfloor\dot{n}_1\rfloor$ in this case.
- When $\mu_1<\mu_2$ (i.e., all upper triangular instances), all maximums are ≤ 0.5, while all means and minimums are non-positive. Hence, in this case, $n_1^*\le\lceil\dot{n}_1\rceil$.

We therefore define the *rounded BnchAR* as

$$[\dot{n}_1]=\begin{cases}\lceil\dot{n}_1\rceil & \text{if } \mu_1<\mu_2\\ \\ \lfloor\dot{n}_1\rfloor & \text{if } \mu_1\ge\mu_2\end{cases}, \tag{OB.5}$$

and summarize the observations from Figure 9 and Table 1 as:

Observation OA2 *The relation between the rounded BnchAR and the OptAR is as follows:*

(i) When $\mu_1 = \mu_2$, the OptAR may deviate from the rounded BnchAR by adding or removing ***at most one*** *server to the upstream Station 1.*

(ii) When $\mu_1 \neq \mu_2$, the OptAR may deviate from the rounded BnchAR, and this deviation always favors the station with the higher service rate, independent of this station's position in the tandem queueing network.

It is interesting to note that the OptAR's tendency to move servers to the station with higher service rate is independent of this station's position - it does not matter if it is upstream or downstream.

Another interesting observation from Table 1 is that the OptAR's deviation from the BnchAR is not symmetrical. If we only look at means in Table 1, we see all diagonals entries are negative, and the absolute values of the upper triangular entries are greater (by up to 0.5) than those of the lower triangular ones. In other words, there is a small tendency for the OptAR to assign more servers to Station 2 than it would if the derivation were symmetrical. Providing a little more capacity to Station 2 is sensible because it reduces the loss of partially processed customers. However, the strength of this effect is so weak (less than a single server) that it is not instrumental in characterizing the OptAR.

We next look into the difference between the performances of the OptAR and the rounded BnchAR. Similar to our comparison of n_1^* and $\dot{n}_1$, we first describe $TP(n_1^*)$ and $TP([\dot{n}_1])$ as functions of N, for $\lambda_1 = 40$, $\theta \in \{1, \infty\}$, and $(\mu_1, \mu_2) \in \{(1,1), (2,2), (1,2), \text{and } (2,1)\}$, in Figure 10. We see that (i) when $\mu_1 = \mu_2$, $TP([\dot{n}_1])$ is almost identical to $TP(n_1^*)$; (ii) when $\mu_1 \neq \mu_2$, there is a visible difference between $TP(n_1^*)$ and $TP([\dot{n}_1])$. Somewhat surprisingly, the difference between $TP(n_1^*)$ and $TP([\dot{n}_1])$ is relatively stable over N.

In Table 2, we list the average difference between these two policies' performances under $N \in \left\{ \left\lfloor \frac{1}{2}\left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_1}{\mu_2}\right) \right\rfloor, \ldots, \left\lceil \frac{3}{2}\left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_1}{\mu_2}\right) \right\rceil \right\}$ and $\theta \in \{1, \ldots, 10, \infty\}$, i.e., $Average_{N \text{ and } \theta} TP(n_1^*) - TP([\dot{n}_1])$, in a broader parameter setting, for any instance in $\{(\mu_1, \mu_2) \,|\mu_1, \mu_2 = 1, \ldots, 10\}$. When $\mu_1 = \mu_2$, all average differences are less than 0.025; this means that if the rounded BnchAR is applied instead of the OptAR, the throughput will be reduced, but by less than 0.025, which is only 0.06% of λ_1. When $|\mu_1 - \mu_2|$ increases, the difference increases, and it can be substantial. For example, when $(\mu_1, \mu_2) = (1, 10)$, a throughput of 4.367 (i.e., 10.92% of λ_1) is saved by applying the OptAR instead of the rounded BnchAR.

These observations provide useful guidelines for optimally staffing a tandem queue service system with impatient customers:

- When $\mu_1 = \mu_2$, we can apply the rounded BnchAR $[\dot{n}_1]$ whose performance is almost identical to the OptAR, which is among $[\dot{n}_1] - 1$, $[\dot{n}_1]$, and $[\dot{n}_1] + 1$.
- When $\mu_1 \neq \mu_2$, an exhaustive search starting from the rounded BnchAR while moving servers to the station with the higher service rate will quickly identify the OptAR. Observation OA1 guarantees the convergence of this search method.

These guidelines significantly reduce the search space for the OptAR n_1^*, and our numerical method can provide n_1^* almost instantaneously for each parameter setting. Hence, for any practical purpose, our first managerial question has been fully addressed.

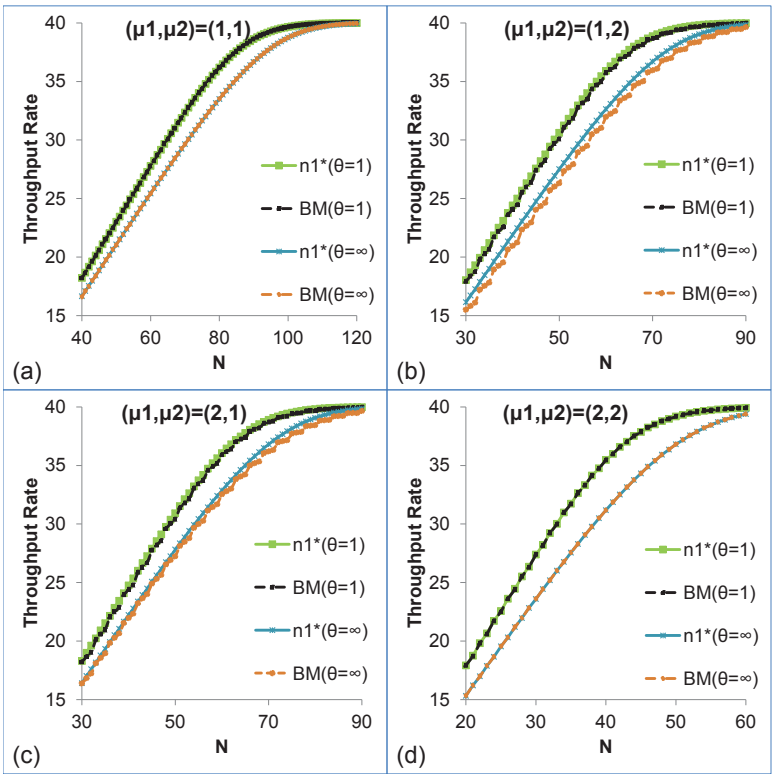

Figure 10 $TP(n_1^*)$ **and** $TP([\dot{n}_1])$ **as functions of** $N \in \left\{\left\lfloor \frac{1}{2}\left(\frac{\lambda}{\mu_1}+\frac{\lambda}{\mu_2}\right)\right\rfloor, \ldots, \left\lceil \frac{3}{2}\left(\frac{\lambda}{\mu_1}+\frac{\lambda}{\mu_2}\right)\right\rceil\right\}$ **under** $\lambda = 40$ **and** $\theta \in \{1, \infty\}$ **for** $(\mu_1, \mu_2) =$ **(a)** $(1,1)$**; (b)** $(1,2)$**; (c)** $(2,1)$**; (d)** $(2,2)$**.**

OA2.3. Optimal Staffing Scheme to Achieve a Required Throughput

In this section, we answer the second managerial question by generating the Optimal Staffing Scheme - a table with three columns: (i) the total number of servers, N; (ii) the OptAR n_1^* that maximizes throughput, given that N servers are available; (iii) the throughput under this OptAR, $TP(n_1^*)$.

We first introduce an algorithm to generate the Optimal Staffing Scheme for any $N \in \left\{\left\lfloor \frac{1}{2}\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_1}{\mu_2}\right)\right\rfloor, \ldots, \left\lceil \frac{3}{2}\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_1}{\mu_2}\right)\right\rceil\right\}$.

Algorithm OA1 *Generate the Optimal Staffing Scheme Table for given* λ_1, μ_1, μ_2, *and* θ.

Step 1: Set $N = \left\lfloor \frac{1}{2}\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_1}{\mu_2}\right)\right\rfloor$.

Step 2: Set $n =$ *the rounded BnchAR given in (OB.5).*

Step 3: If $\mu_1 = \mu_2$, *go to Step 4; otherwise go to Step 6.*

Step 4: Derive $TP(i)$ *for* $i \in \{n-1, n, n+1\}$, *using the methods from Sections 3 and A1.6.*

Step 5: Set $n_1^* = \arg\max_{i \in \{n-1,n,n+1\}} TP(i)$ *and go to Step 9.*

Step 6: Set $\Delta = \begin{cases} -1 \text{ if } \mu_1 < \mu_2 \\ \\ 1 \text{ if } \mu_1 > \mu_2 \end{cases}$.

Step 7: Derive $TP(n)$ *and* $TP(n+\Delta)$, *using the methods from Sections 3 and A1.6.*

Step 8: If $TP(n) < TP(n+\Delta)$, *set* $n = n+\Delta$ *and go to Step 7; otherwise set* $n_1^* = n$.

Step 9: Record $[N, n_1^*, TP(n_1^*)]$ *as a new row of the Optimal Staffing Scheme Table.*

Step 10: Set $N = N+1$.

$\mu_1 \backslash \mu_2$	1	2	3	4	5	6	7	8	9	10
1	0.001	0.466	1.057	1.607	2.136	2.677	3.140	3.610	4.015	4.367
2	0.281	0.001	0.369	0.653	1.161	1.493	2.072	2.322	2.852	3.014
3	0.762	0.234	0.003	0.370	0.713	0.848	1.416	1.731	1.820	2.531
4	1.243	0.474	0.260	0.005	0.399	0.571	1.053	1.027	1.560	1.768
5	1.717	0.935	0.572	0.307	0.008	0.409	0.815	0.949	1.311	1.053
6	2.219	1.235	0.686	0.466	0.330	0.011	0.555	0.663	0.846	1.371
7	2.665	1.798	1.245	0.928	0.711	0.481	0.015	0.566	0.786	1.262
8	3.124	2.029	1.547	0.893	0.839	0.591	0.514	0.019	0.967	0.964
9	3.514	2.553	1.601	1.399	1.213	0.771	0.724	0.918	0.018	1.237
10	3.843	2.677	2.331	1.626	0.956	1.303	1.195	0.913	1.194	0.021

Table 2 **Average** $TP(n_1^*) - TP([\hat{n}_1])$ **under any** $\theta \in \{1, \ldots, 10, \infty\}$ **and** $N \in \left\{ \left\lfloor \frac{1}{2}\left(\frac{\lambda}{\mu_1} + \frac{\lambda}{\mu_2}\right) \right\rfloor, \ldots, \left\lceil \frac{3}{2}\left(\frac{\lambda}{\mu_1} + \frac{\lambda}{\mu_2}\right) \right\rceil \right\}$ **for** $\lambda = 40$ **and** $(\mu_1, \mu_2) \in \{(\mu_1, \mu_2) \,|\mu_1, \mu_2 = 1, \ldots, 10\}$.

Step 11: If $N \leq \left\lceil \frac{3}{2}\left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_1}{\mu_2}\right) \right\rceil$, *go to Step 2; otherwise STOP.*

We run Algorithm OA1 on a 64-bit desktop with an Intel Hexa-Core E5-1650 @ 3.5GHz processor. The run time of this algorithm depends on the range of N. For $N \leq 150$, each instance completes within 15 seconds, while for N close to 200, each instance takes up to 30 seconds. For example, for $\lambda_1 = 100$, $(\mu_1, \mu_2) = (1, 2)$, and $\theta = 1$, where $N \in \{75, \ldots, 224\}$, we use Algorithm OA1 to generate the Optimal Staffing Scheme. The algorithm completes in 45 minutes for 150 instances.

The Optimal Staffing Scheme generated by Algorithm OA1 is listed in Table 3 (due to page limits, we only list $N \in \{76, \ldots, 200\}$). Note that the total number of servers N covers a wide range: from scarce capacity, i.e., $N = 76$ where more than 50% of customers abandon, to plenty of capacity, i.e., $N = 200$ where less than 0.02% of customers abandon.

Using this table, answering the second managerial question is straightforward. For example, if the throughput target is 50, i.e., at least 50% of the customers finish service without abandonment, then at least 80 servers are required, and 52 of them should be assigned to Station 1. If the throughput target is 99, then at least 172 servers are needed, with 112 assigned to Station 1. Clearly, for any other parameter settings, a similar Optimal Staffing Scheme can easily be produced.

N	n1*	TP	N	n1*	TP	N	n1*	TP	N	n1*	TP	N	n1*	TP
76	50	47.5398	101	66	63.7577	126	83	79.9157	151	99	93.9103	176	115	99.4380
77	50	48.2017	102	67	64.4007	127	83	80.5615	152	100	94.3047	177	115	99.5054
78	51	48.8606	103	67	65.0277	128	84	81.2061	153	100	94.7047	178	116	99.5648
79	52	49.4764	104	68	65.7053	129	85	81.8149	154	101	95.0836	179	116	99.6154
80	52	50.1434	105	69	66.3465	130	85	82.4565	155	102	95.4249	180	117	99.6651
81	53	50.8000	106	69	66.9778	131	86	83.0890	156	102	95.7806	181	118	99.7057
82	54	51.4142	107	70	67.6532	132	87	83.6849	157	103	96.1035	182	118	99.7438
83	54	52.0860	108	71	68.2927	133	87	84.3189	158	103	96.3996	183	119	99.7772
84	55	52.7403	109	71	68.9279	134	88	84.9348	159	104	96.7016	184	119	99.8051
85	56	53.3530	110	72	69.6011	135	89	85.5134	160	105	96.9697	185	120	99.8324
86	56	54.0294	111	73	70.2387	136	89	86.1350	161	105	97.2245	186	121	99.8542
87	57	54.6816	112	73	70.8776	137	90	86.7292	162	106	97.4712	187	121	99.8746
88	58	55.2928	113	74	71.5484	138	91	87.2851	163	107	97.6877	188	122	99.8922
89	58	55.9736	114	75	72.1839	139	91	87.8888	164	107	97.9028	189	122	99.9067
90	59	56.6238	115	75	72.8260	140	92	88.4551	165	108	98.0988	190	123	99.9207
91	60	57.2336	116	76	73.4940	141	92	88.9906	166	108	98.2711	191	124	99.9317
92	60	57.9186	117	77	74.1269	142	93	89.5620	167	109	98.4467	192	124	99.9420
93	61	58.5669	118	77	74.7718	143	94	90.0940	168	110	98.5981	193	125	99.9506
94	61	59.1802	119	78	75.4363	144	94	90.6040	169	110	98.7387	194	125	99.9578
95	62	59.8643	120	79	76.0657	145	95	91.1353	170	111	98.8720	195	126	99.9645
96	63	60.5107	121	79	76.7124	146	96	91.6267	171	112	98.9855	196	127	99.9696
97	63	61.1289	122	80	77.3722	147	96	92.1051	172	112	99.0981	197	127	99.9746
98	64	61.8107	123	81	77.9969	148	97	92.5902	173	113	99.1965	198	128	99.9786
99	65	62.4554	124	81	78.6441	149	98	93.0353	174	113	99.2833	199	128	99.9819
100	65	63.0781	125	82	79.2976	150	98	93.4766	175	114	99.3676	200	129	99.9849

Table 3 **Optimal staffing scheme for $\lambda = 100$, $(\mu_1, \mu_2) = (1, 2)$, and $\theta = 1$.**