Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Using Strategic Idleness to Improve Customer Service Experience in Service Networks

Opher Baron, Oded Berman, Dmitry Krass, Jianfu Wang

To cite this article:
Opher Baron, Oded Berman, Dmitry Krass, Jianfu Wang (2014) Using Strategic Idleness to Improve Customer Service Experience in Service Networks. Operations Research 62(1):123-140. http://dx.doi.org/10.1287/opre.2013.1236

OPERATIONS RESEARCH
Vol. 62, No. 1, January–February 2014, pp. 123–140
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.2013.1236

Using Strategic Idleness to Improve Customer Service Experience in Service Networks

Opher Baron, Oded Berman, Dmitry Krass
Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada
{opher.baron@rotman.utoronto.ca, berman@rotman.utoronto.ca, krass@rotman.utoronto.ca}

Jianfu Wang
Nanyang Business School, Nanyang Technological University, Singapore 639798, Republic of Singapore, jianfu.wang.ntu@gmail.com

The most common measure of waiting time is the overall expected waiting time for service. However, in service networks the perception of waiting may also depend on how it is distributed among different stations. Therefore, reducing the probability of a long wait at any station may be important in improving customers' perception of service quality. In a single-station queue it is known that the policy that minimizes the waiting time and the probability of long waits is nonidling. However, this is not necessarily the case for queueing networks with several stations. We present a family of threshold-based policies (TBPs) that strategically idle some stations. We demonstrate the advantage of strategically idling by applying TBP in a network with two single-server queues in tandem. We provide closed form results for the special case where the first station has infinite capacity and develop efficient algorithms when this is not the case. We compare TBPs with the nonidling and Kanban policies, and we discuss when a TBP is advantageous. Using simulation, we demonstrate that the analytical insights for the two-station case hold for a three-station serial queue as well.

Subject classifications: strategic idleness; threshold-based policy; customer service experience; service network.
Area of review: Stochastic Models.
History: Received March 2012; revisions received December 2012, June 2013; accepted September 2013.

1. Introduction

Multistage service networks, where customers must visit several stations during a single service encounter, abound in modern economy. Examples range from call centers, where a typical service path may include an automated response system, followed by a generalist call taker, and eventually (and if required) a specialist, to hospital emergency rooms, where the initial triage stage may be followed by any number of medical tests and procedures.

Although there are many determinants of service quality, the link between customer waiting times and the perceived service quality is well recognized (Friedman and Friedman 1997, Taylor 1994). Waiting times have long been the focus of much of the queueing literature. The most common measure of waiting time is the overall expected waiting time for service (see, e.g., the survey by Gans et al. 2003). A related measure is the probability that the total waiting time exceeds a certain predefined threshold. These measures take a macro view of the network, treating it as a one-stage system.

However, considering only such macrolevel measures might not be sufficient to measure service quality and may even be misleading. There is a strong body of evidence showing that it is also important to consider what happens within the network. A poor level of service received at a particular station may not be compensated by an exceptional service at another station, even if the overall measure appears to be acceptable. The adverse impact of a long waiting time at a particular station is further supported by marketing literature, e.g., Soman and Shi (2003), and by the psychology of queueing literature, e.g., Larson (1987). Baron et al. (2008), Baron and Milner (2009), de Vericourt and Jennings (2011), and references therein also focused on the probability of a long waiting time as a service-level measure.

Several other papers looked beyond the traditional mean waiting time measures. de Véricourt and Zhou (2005) analyzed a call-routing problem while considering both the call resolution probability and the average service time in the overall service-level measure. Mehrotra et al. (2012) considered a similar problem with heterogeneous servers. Saghafian et al. (2012) analyzed the service policy in emergency departments while considering the weighted average of the expected length of stay and the expected time to first treatment.

We recently encountered an explicit example of focusing on the probability of overly long waits at any single station at a company we call XYZ (name changed to protect confidentiality), one of the leaders in preventive healthcare services in North America. The company's primary clientele are executives and busy professionals, so its primary focus is on providing excellent customer service experience. XYZ operates a service network with 15–20 stations. In addition to closely tracking macrolevel measures, the company also records all instances where a customer waits longer than 20 minutes at a station. Any such incident results in a *red face* flashing on the manager's screen, who takes immediate steps to expedite the customer. All red face incidents

are regarded as service failures, irrespective of whether customer's overall waiting time in the system was acceptable or not. We note that this example is not unique, e.g., the proportion of customers waiting longer than a specified time at a station is a common key performance indicator in call centers. The focus on long waits implies that service quality is affected not only by the overall waiting time, but also by the distribution of waiting among stations.

The focus of this paper is to simultaneously consider two objectives in a service network, one based on some macrolevel measure and one based on the *probability of excessive wait* at any one station. The difference in managing these two objectives can be rather dramatic. Indeed, the macrolevel service measures are typically minimized by using work-conserving policies, where system resources are not idled as long as there is work in the system. Such policies are optimal with respect to minimizing overall service times and are the focus of most studies of queueing networks (see, e.g., Chen and Yao 2001 and reference therein). However, using a work-conserving policy is not necessarily a good idea when it comes to the second objective. Consider a situation where one station in the network accumulates a long queue, while the waiting times are low at the upstream stations. In such a case, continuing to operate upstream stations at the normal rate may increase the probability of excessive waits at downstream stations. A better idea may be to temporarily reduce the service rate or idle the upstream stations, allowing the downstream queue to dissipate. By intentionally idling some resources we are effectively redistributing the waiting times more evenly within the network. As long as such redistribution does not significantly increase the overall system times (i.e., the first objective), it may well improve the overall customer service experience.

Our objective is to propose and analyze a class of scheduling policies that intentionally idle some resources to reduce the probability of excessive waits at any one station. We refer to such intentional idling of resources as *strategic idleness* (*SI*). Note that the classical way of reducing waiting time and probabilities of long waits is to add resource capacity to the system (e.g., adding a doctor in the healthcare setting), which is often quite expensive. On the other hand, changing the scheduling rules to intentionally idle some resources can often be done at a negligible cost. Thus, a switch to an SI policy may be very cost-effective of improving customer service experience. Indeed, we establish that in contrast to the single station queue, where a nonidling (NI) scheduling policy minimizes both the sojourn time and the probability of long waits, for a multistage queueing network, policies with SI may significantly reduce the probability of long waits while only slightly increasing the overall time in the system. To the best of our knowledge, ours is the first paper to systematically study SI as a mechanism for reducing the probability of excessive waits and improving the customer service experience.

In service networks, long waits can be measured in a variety of ways. For example, consider a two-station tandem queue with station 1 as the upstream machine and station 2 as the downstream one. The specific measure we consider is $PW(t) = \frac{1}{2}\sum_{i=1}^{2} P\{W_i > t\}$, where W_i is the steady state customers' waiting time for station i, and t is the time threshold designating an "excessive wait." We interpret $PW(t)$ as the frequency with which customers experience excessive waits. We note that in place of $PW(t)$ one can use other related measures, e.g., $1-P\{W_1 < t, W_2 < t\}$, i.e., the probability that a customer experiences at least one excessive wait.

There are many possible policy classes that involve SI. Our primary focus is on a specific family of *threshold-based policies* (*TBPs*). The idea behind the TBP is simple, it compares the difference between queue lengths at different stations and idles some upstream stations if this difference is larger than a predetermined threshold. For example, consider the two-station tandem queue described above: let q_1, q_2 be the lengths of queues in front of the respective stations. A TBP, defined by the value of the threshold TH, idles station 1 whenever the difference $q_2 - q_1 \geqslant TH$ (we only consider $TH \geqslant 0$ as using $TH < 0$ is clearly counterproductive, e.g., with $TH = -1$ when $q_1 = 1$, $q_2 = 0$, station 1 would be idled).

We note that, assuming Poisson arrivals to station 1 and exponentially distributed and independent service times at both stations, the performance of the NI policy is easy to analyze (see, e.g., Ross 2000, Chapter 8). However, such an analysis for the system operating under the TBP is quite challenging for several reasons. First, the process is not reversible, so arrivals to station 2 do not follow a Poisson process. Second, as explained in §3, a customer's waiting time for station 1 depends on future arrivals, so Little's distributional law (see, e.g., Bertsimas and Nakazato 1995, Bertsimas and Mourtzinou 1996) does not hold.

We develop efficient algorithms to calculate the distribution of waiting time for each station and the system sojourn time under the TBP. These algorithms use a new analysis of the waiting time faced by specific customers. Using these results we present trade-off curves between the probability of long waits and the expected sojourn time. (Note that the distribution of the system sojourn time can provide other measures than the mean, but the trade-offs between $PW(t)$ and these measures are similar to the trade-off between $PW(t)$ and the mean sojourn time.) For the asymptotic case when $\mu_1 = \infty$, we derive closed form expressions for the performance measures. We derive interesting insights that also hold in the case of finite processing capacity for both stations.

Our results show that TBP can significantly reduce the probability of long waits (as expressed by $PW(t)$ or similar measures) versus the NI policy as long as the waits of length t are sufficiently rare in the system. If, on the other hand, the frequency of such "excessive" waits is high under the NI policy (indicating that they are not, in

fact, excessive), then the TBP is unlikely to provide an improvement—the only way to decrease such waits is by adding capacity.

We also consider the class of TBPs in a tandem queue network with three stations. By developing a simulation model, we show that a TBP can reduce the probability of long waits while only slightly increasing sojourn times. A comparison with Kanban policies indicates that the TBPs perform significantly better in this case.

We note that service systems, such as XYZ, do not always reach steady state before the end of a business day. Moreover, such systems often operate a nonserial queueing network. However, the results for the serial system under the steady state assumption still provide valuable insights for such systems. Specifically, polices with SI such as the TBP can improve customers' perception of the service level with little cost. In Baron et al. (2014), we tested a generalized TBP with a simulation model of the open-shop operation of XYZ; we indeed established that TBP can be effective in improving customers' perception of the service level.

The outline of this paper is as follows. In the next section, we provide a brief discussion of other policies with idling. After introducing the TBP for the two-station network in §3, we consider the asymptotic $\mu_1 = \infty$ case in §4. In §5, we analyze the case of finite processing rate for both stations. In §7, we discuss generalization of the TBPs, to n-station serial queues and list several open questions. All proofs are in Section EC.1 of the e-companion (available as supplemental material at http://dx.doi.org/10.1287/opre.2013.1236).

2. Literature Review—Other Policies with Idling

Note that the main idea behind the TBP—idling an upstream station when a downstream station is facing a large workload—can be achieved by other policy classes. We next briefly review classes of policies that are discussed in the literature of manufacturing systems.

Masin et al. (2005) developed a unified model that encompasses and compares a wide range of production control policies. We follow their exposition focusing on a serial manufacturing system with M stations, and each station i has an input pile, IP_i, and an output pile, OP_i, for $i = 1, \ldots, M$. Let OP_0 represent an ample pile of raw materials, i.e., $OP_0 = \infty$. Each part waits in IP_i before being processed at station i and then transferred to OP_i, and stays in OP_i until it can be transferred to IP_{i+1}.

There are four well-known static control policies (i.e., controls that are independent of the system state): the *fixed buffer* policy (see, e.g., Conway et al. 1988) places a finite buffer FB_{i+1} between stations i and $i+1$, i.e., $IP_1 < FB_1$ and $OP_i + IP_{i+1} \leqslant FB_{i+1}$ for $i = 1, \ldots, M-1$; the Kanban policy, implemented by Toyota (Sugimori et al. 1977), places an upper bound KB_i on the total number of parts associated with station i, i.e., $IP_i + OP_i \leqslant KB_i$ for $i = 1, \ldots, M$; the *constant work in process* (*CONWIP*) policy, first presented by Spearman et al. (1990), places an upper bound CW on the total number of parts in the system, i.e., $\sum_{j=1}^{M}(IP_j + OP_j) \leqslant CW$ (for a recursive calculation of several performance measure in a resulting closed queueing network, see Solberg 1977); the *base-stock* policy (see, e.g., van Ryzin et al. 1993) places an upper bound BS_i on the total number of parts at the downstream of station i, i.e., $\sum_{j=i}^{M}(IP_j + OP_j) \leqslant BS_i$ for $i = 1, \ldots, M$.

More sophisticated dynamic control policies where controls depend on the state of the system were also studied. Weber and Stidham (1987) considered a general model for control of service rates ($\mu_i \in [0, \bar{\mu}_i]$) in a serial or closed queueing network, where control policies depend on the entire state vector $q = (q_1, q_2, \ldots, q_M)$, where $q_i = OP_{i-1} + IP_i$. They considered the sum of total inventory holding cost and station operating cost as the objective function. They provided necessary conditions, called the "monotonicity result," for any control policy to be optimal: (1) the optimal service rate at station i does not decrease as a customer finishes service at another station; (2) the optimal service rate at station i does not increase as a customer finishes service at station i. They applied their monotonicity result to models where stations can only be turned on or off ($\mu_i = 0$ or $\bar{\mu}_i$) and showed that it is optimal to turn an off-station on as the numbers of customers at its downstream stations decrease, or as the numbers of customers at upstream stations increase. Note that the four control policies discussed above and TBP all satisfy this monotonicity result. Veatch and Wein (1994) considered the optimal control of a two-station tandem production/inventory system with a similar objective function. They compared these four policies, gave conditions under which certain simple controls are optimal, and computed the dynamic optimal controls using dynamic programming.

There are several conceptual differences between the control policies discussed above, tailored to manufacturing systems, and the TBP, tailored to service systems. First, the main motivation behind developing policies in manufacturing setting is the control of expected inventory costs. This motivation is different for service systems focusing on the effect of the distribution of waiting time on customers' experience. As we demonstrate below, this different motivation also leads to a different analysis. In fact, to the best of our knowledge, no analysis of the distribution of waiting times under the policies mentioned above is available; such an analysis appears to be subject to many of the challenges as in the analysis of the TBP. Second, another important modeling difference is that the control for manufacturing systems is often modeled as a make-to-stock system, whereas the control for service systems must be modeled as a make-to-order system. Third, from a modeling perspective, the supply and demand models are also different in a service system: the service at a first station is initiated by an exogenous arrival process, and customers

leave the system as they complete service at the last station, whereas in manufacturing the exogenous demand arrives to the last station. A final difference is with respect to admission control. In contrast to our model, where all customers are accepted, models for manufacturing systems often operate with admission control where not all arriving orders are fulfilled. (Note that IP_1 is bounded in the four policies above, so *not all* arriving customers are admitted. Still, if all customers need to be admitted, IP_1 can be removed from all constraints. For example, a CONWIP policy could place an upper bound CW on the total number of parts without considering IP_1, i.e., $OP_1 + \sum_{j=2}^{M}(IP_j + OP_j) \leqslant CW$.)

Despite these differences, the control policies developed for manufacturing systems can be applied in service systems (sometimes with a few modifications). When applied in a two-station tandem queue service system without admission control, the fixed buffer, Kanban, CONWIP, and base-stock policies can all be shown to be equivalent. To illustrate the equivalence of Kanban policy and fixed buffer policy, note that a Kanban policy with KB_1 and KB_2 is equivalent to a fixed buffer policy with buffer size $FB_2 = KB_1 + KB_2$ between the two stations; and a fixed Buffer policy with buffer size FB_2 is equivalent to a Kanban policy with $KB_1 = 1$ and $KB_2 = FB_2 - 1$. Thus, in the two-station tandem queue service system we consider in this paper, we focus on a Kanban policy that idles station 1 whenever $q_2 \geqslant BS$, where BS is the size of the buffer between the two stations.

In this paper, we compare our TBP with the Kanban policy. Note that in the two-station case, our TBP is a more sophisticated dynamic control policy, where the upper bound of q_2 is a linear function of q_1, i.e., station 1 is idled whenever $q_2 \geqslant q_1 + TH$; and a Kanban policy idles station 1 based only on $q_2 \geqslant BS$ irrespective of the value of q_1, and thus—intuitively—it provides less flexible control than a TBP. This intuition appears to be supported by our results. For the asymptotic case when station 1 has infinite processing capacity, we derive closed form expressions for the $PW(t)$ measure under a Kanban policy, allowing us to make analytical comparisons to a TBP. For the finite capacity case, we use Monte Carlo simulation to compare TBP and Kanban policies. Our results indicate that, similar to a TBP, the Kanban policy allows for the trade-off between the $PW(t)$ measure and expected service times. However, this policy appears to be less efficient than the TBP.

In closing this section we note that (i) the idea of intentionally idling a capacitated resource has also been considered by Afèche (2013). In the revenue management context, he showed how such delays can allow a seller to differentiate between customer types and thus improve the overall profit. His motivation and analysis are much different from ours. (ii) Recent polices for control of manufacturing systems often considered prioritization among several customer classes, but are focused on a single-stage system. Ha (1997a, b) was the first to discuss inventory rationing problems in a centralized make-to-stock system. He focused on base-stock-level production control. (iii) In the revenue management context, Caldentey and Wein (2006) developed a diffusion approximation for profit maximization with two classes of customers. They showed that a dynamic control policy based upon the inventory or backlog level is effective.

Finally, we are aware that there are other policies that consider the entire system state. This paper serves as a stepping stone motivating the analysis of such policies in service systems.

3. Two Queues in Tandem—Preliminary Analysis

Consider the two-station tandem queueing network with two sequential single server stations and infinite buffer space discussed before. We define a simple TBP for this network as follows: upon completing service, station 1 is idled and will not admit the next customer to service if

$$\delta(q_1, q_2) = q_2 - q_1 \geqslant TH.$$

Station 1 will resume work once $\delta(q_1, q_2) < TH$. When no ambiguity arises, we will use δ instead of $\delta(q_1, q_2)$. We denote $TBP(TH)$ as the TBP with threshold TH. We say that a customer is *stopped* (at station 1) if this customer is waiting at station 1 while this station is *idled*.

Three events can occur in this tandem queueing network:

1. *Arrival*—Arrival to the network decreases δ by 1. Arrivals occur with rate λ at any state.

2. *Completion* 1—Service completion at station 1 increases δ by 2. This happens with rate μ_1, if $q_1 \geqslant 1$ and $\delta < TH$ (when station 1 is not idled).

3. *Completion* 2—Service completion at station 2 decreases δ by 1. This event has rate μ_2, if $q_2 \geqslant 1$.

Note that since δ decreases when station 2 completes service or when a new customer arrives to station 1, either of these two events may cause station 1 to resume work.

From these three events, we conclude that there are two situations when station 1 is idled: $\delta = TH$ or $\delta = TH + 1$. When $\delta = TH + 1$, station 1 is idled, so only arrival or Completion 2 can happen in the network. After a time period, which is distributed $\sim \exp(\lambda + \mu_2)$, one of these events happen, reducing δ to TH. Note that station 1 remains idled. This sequence repeats and after another time period $\sim \exp(\lambda + \mu_2)$, δ is reduced to $TH - 1$, at which point station 1 resumes work, and its idle period ends. We define *stoppage* as the time period from the moment when the value of δ changes and station 1 becomes idled until the moment when either arrival or Completion 2 happens. With this definition, when $\delta = TH + 1$, customers in station 1 experience two stoppages before station 1 resumes work; when $\delta = TH$, they experience only one stoppage.

Let $Q_i(t)$, $i = 1, 2$ be the random variable (RV) denoting the total number of customers at station i (in queue and in service) at time t. Given TH, the process $(Q_1(t), Q_2(t))$ is a continuous time Markov chain (MC). Let π_{q_1, q_2}

denote the steady state probability of $MC(Q_1, Q_2)$. Let S be the sojourn time for any customer, i.e., $S =$ *total waiting time* + *total service time*. Let $\rho_2 = \frac{\lambda}{\mu_2}$.

To investigate the trade-off between $PW(t)$ and $E[S]$ under the TBP, we first characterize the distribution of three steady state service measures: the waiting time at station 1, W_1; the waiting time at station 2, W_2; and the sojourn time, S. We can calculate the distributions of these three measures by conditioning on the state (q_1, q_2) seen by a random arrival. Let X^{q_1,q_2} be any one of these three measures experienced by a *tagged customer* (TC) who arrives in state (q_1, q_2). Then, the steady state distribution of X can be calculated as

$$\begin{aligned} &P\{X > t\} \\ &= \sum_{q_1,q_2} P\{X^{q_1,q_2} > t \mid \text{TC sees } (q_1,q_2) \text{ at arrival}\} \\ &\qquad \cdot P\{\text{TC sees } (q_1,q_2) \text{ at arrival}\} \\ &= \sum_{q_1,q_2} P\{X^{q_1,q_2} > t \mid \text{TC sees } (q_1,q_2) \text{ at arrival}\}\pi_{q_1,q_2}, \quad (1) \end{aligned}$$

where the second equality follows by Poisson Arrivals See Time Average.

Similar to (1), the Laplace transform (LT) of X can be written as

$$L_X(h) = \sum_{q_1,q_2} L_{X^{q_1,q_2}}(h \mid \text{TC sees } (q_1,q_2) \text{ at arrival})\pi_{q_1,q_2}. \quad (2)$$

4. Asymptotic Case: Station 1 Has an Infinite Service Capacity

We next calculate the steady state performance measures under the TBP and compare them with the measures for the nonidling network and the Kanban policy when station 1 has infinite capacity. For convenience, we denote quantities related to this asymptotic case with a $\hat{}$, e.g., $\hat{W}_i$ is the waiting time at station i. A full list of notation can be found in Table EC.1 in Section EC.5 of the e-companion.

The MC for the $\mu_1 = \infty$ case is depicted in Figure 1. As described in §3, three events occur in this MC: arrival, Completion 1, and Completion 2. However, since Completion 1 happens instantaneously, only two events are shown on the figure: arrival (at rate λ) and Completion 2 (at rate μ_2). Consider the state $(0, TH)$, where station 1 is idled under the TBP. An arrival momentarily bring the MC to state $(1, TH)$, where $\delta = TH - 1$, and thus station 1 resumes work, instantaneously bringing the MC to state $(0, TH + 1)$ and idling station 1 again. At the next arrival the MC transitions to state $(1, TH + 1)$, where $\delta = TH$ and thus the newly arrived customer is stopped. This stoppage lasts until either a new arrival, which allows the system to process the first customer from station 1 and sends the system to state $(1, TH + 2)$, or Completion 2, which also releases the customer from station 1 and sends the system to $(0, TH)$. In general, whenever $q_1 > 0$, station 1 is idled and the system is either in state $(q_1, q_1 + TH)$ or $(q_1, q_1 + TH + 1)$.

The steady state distribution of this simple birth and death MC (similar to the solution of an $M/M/1$ queue), for $q_1 = 0, q_2 = 0, \ldots, TH + 1$ and for $q_1 > 0$, $q_2 = q_1 + TH$, $q_1 + TH + 1$, is

$$\pi_{q_1,q_2} = \rho_2^{q_1+q_2}(1 - \rho_2). \quad (3)$$

Remark 1. If we consider $q_1 + q_2$ as the total queue length, this network has the same steady state probability distribution as an $M/M/1$ queue with $\rho_2 = \lambda/\mu_2$. Because station 2 works as long as there are customers in the network, the sojourn time is the same as the sojourn time in the system with $\mu_1 = \infty$ operating under a nonidling policy. Thus, in the asymptotic case the TBP does not increase the sojourn times, and we can focus solely on the $PW(t)$ measure.

Remark 2. Suppose the system is in state $(q_1, q_1 + TH + 1)$ for $q_1 > 0$ (station 1 is idled). The next arrival (Completion 2) event sends the system to state $(q_1 + 1, q_1 + TH + 1)$ (state $(q_1, q_1 + TH)$), with $\delta = TH$, and station 1 is stopped again. Thus, the next event must be another arrival or Completion 2. Similarly, suppose the system is in state $(q_1, q_1 + TH)$ for $q_1 > 0$ (station 1 is idled). The next event must be arrival or Completion 2, which will trigger a Completion 1 event and send the system to $(q_1, q_1 + TH + 1)$ or $(q_1 - 1, q_1 + TH)$, respectively, with $\delta = TH + 1$ in both cases. Thus, as long as $q_1 > 0$, between any two Completion 1 events there are always two other events. This leads to the following proposition.

Proposition 1. *Let $\hat{M}^{q_1,q_2}$ be the number of stoppages a TC sees before entering station 2, given she arrives in state (q_1, q_2). Then either $q_2 < TH$ and $\hat{M}^{q_1,q_2} = 0$ or $q_2 \in \{q_1 + TH, q_1 + TH + 1\}$ and $\hat{M}^{q_1,q_2} = q_1 + q_2 - TH$.*

4.1. Distribution of $\hat{W}_1$, Waiting Time at Station 1

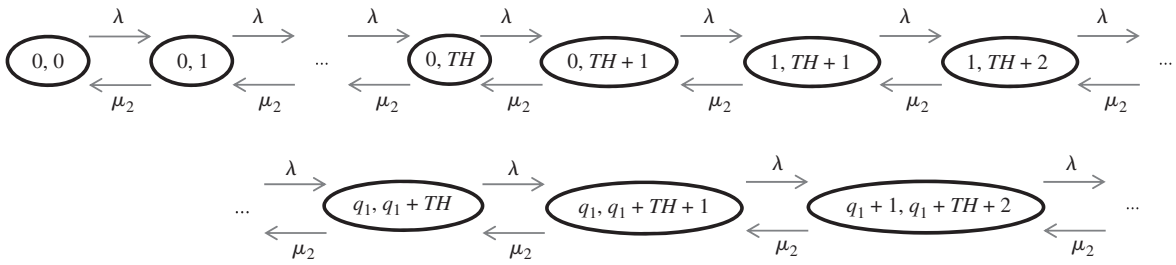
In general, the TC's waiting time for station 1 is composed of two parts: the service time of customers in front of her in station 1 and the stoppages of station 1. However, when $\mu_1 = \infty$, the service time of station 1 is zero, and thus $\hat{W}_1$ is only caused by stoppage.

Let $\hat{W}_1^{q_1,q_2}$ denote the TC's waiting time at station 1, given that she arrives at state (q_1, q_2). From Proposition 1, if $q_1 + q_2 \leqslant TH$, the TC sees no stoppage and $\hat{W}_1^{q_1,q_2} = 0$; similarly, if $q_1 + q_2 > TH$, then $\hat{M}^{q_1,q_2} = q_1 + q_2 - TH$, so that $\hat{W}_1^{q_1,q_2}$ is distributed as Erlang$(\lambda + \mu_2, q_1 + q_2 - TH)$. Thus, using (2) and (3), the LT of $\hat{W}_1$ is

$$\begin{aligned} &L_{\hat{W}_1}(h) \\ &= \sum_{i=0}^{TH} \rho_2^i(1 - \rho_2) + \sum_{i=TH+1}^{\infty} \rho_2^i(1 - \rho_2)\left(\frac{\lambda + \mu_2}{\lambda + \mu_2 + h}\right)^{i-TH} \\ &= (1 - \rho_2^{TH+1}) + \rho_2^{TH+1}\frac{\mu_2 - \lambda\rho_2}{(\mu_2 - \lambda\rho_2) + h}. \quad (4) \end{aligned}$$

From the transform of $\hat{W}_1$ we conclude that there is no waiting in station 1 with probability (w.p.) $1 - \rho_2^{TH+1}$, and

Figure 1. MC when $\mu_1 = \infty$ and $TH > 0$.



the waiting is distributed as an $\exp(\mu_2 - \lambda\rho_2)$ RV with probability ρ_2^{TH+1}. Hence,

$$P\{\hat{W}_1 > t\} = \rho_2^{TH+1} e^{-(\mu_2 - \lambda\rho_2)t}. \tag{5}$$

Note that given waiting (i.e., w.p. ρ_2^{TH+1}), $\hat{W}_1$ is distributed as the waiting time given waiting in an $M/M/1$ queue with arrival rate $\lambda\rho_2$ and service rate μ_2.

As intuition suggests, $P\{\hat{W}_1 > t\}$ is a decreasing function of TH. When TH decreases, customers see more stoppages, and thus wait more in station 1. When TH increases, the TBP's effect on the network is reduced, and customers' wait in station 1 is also reduced. The extreme case when $TH = \infty$ results in a nonidling network, so customers do not wait for station 1.

4.2. Distribution of the Waiting Time $\hat{W}_2$ and Service Measure $\widehat{PW}(t)$

We next derive $\hat{W}_2^{q_1, q_2}$, the TC's waiting time at station 2 given that she arrives at state (q_1, q_2), and then use (2) to calculate the LT of $\hat{W}_2$. Let K be the RV denoting (we omit the dependency on q_1, q_2) the number of customers in station 2 when the TC enters this station; thus $\hat{W}_2^{q_1, q_2}$ is distributed as Erlang(μ_2, K).

From Proposition 1, if $q_1 + q_2 \leqslant TH$, then $q_1 = 0$ and the TC gets into station 2 immediately, implying that $K = q_1 + q_2 = q_2$. Thus, for $q_1 + q_2 \leqslant TH$, the distribution of $\hat{W}_2^{q_1, q_2}$ is Erlang$(\mu_2, q_1 + q_2)$ with the LT given by

$$L_{\hat{W}_2^{q_1, q_2}}(h) = \left(\frac{\mu_2}{\mu_2 + h}\right)^{q_1 + q_2}. \tag{6}$$

Now suppose the TC arrives at state (q_1, q_2) with $q_1 + q_2 > TH$, implying that the number of stoppages is $\hat{M}^{q_1, q_2} = q_1 + q_2 - TH$. In this case, $\hat{M}^{q_1, q_2}$ arrival or Completion 2 events are required to end these stoppages, and $q_1 + q_2 - K$ of these are Completion 2 events, so $K \in [TH, q_1 + q_2]$. Since the probability that the next event is an arrival (Completions 2) is $\lambda/(\lambda + \mu_2)$ $(\mu_2/(\lambda + \mu_2))$, it follows that $q_1 + q_2 - K$ has the binomial distribution

$$P\{q_1 + q_2 - K = n\} = \binom{q_1 + q_2 - TH}{n} \left(\frac{\lambda}{\lambda + \mu_2}\right)^{q_1 + q_2 - TH - n} \left(\frac{\mu_2}{\lambda + \mu_2}\right)^n, \quad n = 0, \ldots, q_1 + q_2 - TH.$$

Thus

$$P\{K = k\} = \binom{q_1 + q_2 - TH}{q_1 + q_2 - k} \left(\frac{\lambda}{\lambda + \mu_2}\right)^{k - TH} \left(\frac{\mu_2}{\lambda + \mu_2}\right)^{q_1 + q_2 - k}, \quad k = TH, \ldots, q_1 + q_2.$$

Therefore, for $q_1 + q_2 > TH$, the LT of $\hat{W}_2^{q_1, q_2}$ is

$$\begin{aligned} L_{\hat{W}_2^{q_1, q_2}}(h) &= \sum_{k=TH}^{q_1 + q_2} \binom{q_1 + q_2 - TH}{q_1 + q_2 - k} \cdot \left(\frac{\lambda}{\lambda + \mu_2}\right)^{k - TH} \left(\frac{\mu_2}{\lambda + \mu_2}\right)^{q_1 + q_2 - k} \left(\frac{\mu_2}{\mu_2 + h}\right)^k \\ &= \left(\frac{\mu_2}{\mu_2 + h}\right)^{TH} \left(\frac{\mu_2}{\lambda + \mu_2} + \frac{\lambda}{\lambda + \mu_2} \frac{\mu_2}{\mu_2 + h}\right)^{q_1 + q_2 - TH} \\ &= \left(\frac{\mu_2}{\mu_2 + h}\right)^{q_1 + q_2} \left(\frac{\lambda + \mu_2 + h}{\lambda + \mu_2}\right)^{q_1 + q_2 - TH}. \end{aligned} \tag{7}$$

The second equality follows the Binomial Formula. The third equality follows because for $q_1 + q_2 > TH$,

$$\left(\frac{\mu_2}{\mu_2 + h}\right)^{TH} \left(\frac{\lambda\mu_2}{(\lambda + \mu_2)(\mu_2 + h)} + \frac{\mu_2}{\lambda + \mu_2}\right)^{q_1 + q_2 - TH} \cdot \left(\frac{\lambda + \mu_2}{\lambda + \mu_2 + h}\right)^{q_1 + q_2 - TH} = \left(\frac{\mu_2}{\mu_2 + h}\right)^{q_1 + q_2}. \tag{8}$$

We can now write the LT of $\hat{W}_2$ using (2), (3), (6), and (7):

$$L_{\hat{W}_2}(h) = \sum_{i=0}^{TH-1} \rho_2^i (1 - \rho_2) \left(\frac{\mu_2}{\mu_2 + h}\right)^i + \sum_{i=TH}^{\infty} \rho_2^{2i - TH} (1 - \rho_2^2) \left(\frac{\mu_2}{\mu_2 + h}\right)^i. \tag{9}$$

From the LT of $\hat{W}_2$ we know that $\hat{W}_2$ is distributed as an Erlang$(\mu_2, q_1 + q_2)$ RV with probability $\rho_2^{q_1 + q_2}(1 - \rho_2)$, for $0 \leqslant q_1 + q_2 < TH$ (i.e., when the TC experiences no stoppages), and as the sum of an Erlang$(\mu_2, TH - 1)$ RV and an

$\exp(\mu_2 - \lambda\rho_2)$ RV w.p. ρ_2^{TH}. Using (9), we can derive the tail distribution of $\hat{W}_2$ under the TBP with threshold TH:

$$P\{\hat{W}_2 > t\} = \begin{cases} \rho_2^{2-TH} e^{-(\mu_2-\lambda\rho_2)t} & \text{if } TH = 0, 1, \\ \rho_2^{2-TH} e^{-(\mu_2-\lambda\rho_2)t} + \rho_2 e^{-\mu_2 t} & \\ \quad \cdot \sum_{k=0}^{TH-2} \dfrac{(\mu_2 t)^k}{k!} \rho_2^k - \rho_2^2 e^{-\mu_2 t} \rho_2^{-TH} & \\ \quad \cdot \sum_{k=0}^{TH-2} \dfrac{(\mu_2 t)^k}{k!} \rho_2^{2k} & \text{if } TH \geqslant 2. \end{cases} \tag{10}$$

Using (5) and (10), the distribution of our main service-level measure under the TBP with threshold TH is

$$PW^{TBP(TH)}(t) = \frac{1}{2}(P\{\hat{W}_1 > t\} + P\{\hat{W}_2 > t\}) = \begin{cases} \frac{1}{2}\rho_2^2 e^{-(\mu_2-\lambda\rho_2)t} + \frac{1}{2}\rho_2 e^{-(\mu_2-\lambda\rho_2)t} & \text{if } TH = 0, 1, \\ \frac{1}{2}\rho_2^{TH+1} e^{-(\mu_2-\lambda\rho_2)t} + \frac{1}{2}\rho_2 e^{-\mu_2 t} & \\ \quad \cdot \sum_{k=0}^{TH-2} \dfrac{(\rho_2\mu_2 t)^k}{k!} + \frac{1}{2}\rho_2^{2-TH} e^{-(\mu_2-\lambda\rho_2)t} & \\ \quad - \frac{1}{2}\rho_2^{2-TH} e^{-\mu_2 t} \sum_{k=0}^{TH-2} \dfrac{(\mu_2 t)^k}{k!} \rho_2^{2k} & \text{if } TH \geqslant 2. \end{cases} \tag{11}$$

We observe that under the nonidling policy all waiting happens at station 2, and thus

$$PW^{NI}(t) = \tfrac{1}{2}\rho_2 e^{-(\mu_2-\lambda)t}, \quad t > 0. \tag{12}$$

Here, ρ_2 represents the probability of waiting, and $\exp(-(\mu_2 - \lambda)t)$ is the conditional probability of waiting more than t given an $M/M/1$ queue with parameters (λ, μ_2). In the expression for $PW^{TBP}(t)$ when $TH = 0, 1$ we see the same structure as in (12). The first term essentially has the probability of waiting reduced to ρ_2^2 from ρ_2 and the arrival rate reduced to $\rho_2\lambda$ from λ. The second term is just the probability of waiting longer than t in an $M/M/1$ queue with arrival rate $\rho_2\lambda$. Thus, the TBP effectively operates two $M/M/1$ stations with parameters $(\rho_2\lambda, \mu_2)$, where the probability of waiting at one of these stations is further reduced by ρ_2. The slower arrival rate (and the additional reduction in probability of waiting) brings the probability of wait longer than t at station 2 to below the level experienced at this station under the nonidling policy. However, the customer now has two chances to experience a long wait—once at each station.

4.3. Insight 1: Comparing the TBP and Nonidling Policy for the Asymptotic Case

Based on Remark 1 above, it suffices to compare $PW^{TBP}(t)$ with $PW^{NI}(t)$ since the expected service times are the same. From our earlier discussion, it is obvious that the number of stoppages is increased when TH is reduced. Thus, setting $TH = 0$ corresponds to the most aggressive redistribution of the waiting time from station 2 to station 1 achievable by a TBP (from (11)). On the other hand, $PW^{TBP(\infty)}(t) = PW^{NI}(t)$ since when $TH = \infty$, station 1 is never intentionally idled.

For any "excessive wait" value $t > 0$, let $TH^*(t) = \arg\min_{TH} PW^{TBP(TH)}(t)$ be the threshold value that minimizes $PW(t)$. This value is characterized in the following result.

Proposition 2. *For any t, the threshold $TH^*(t) \in \{0, \infty\}$. Specifically, let $t^* = \ln(1 + \rho_2)/(\lambda(1 - \rho_2))$ (note that $PW^{TBP(0)}(t^*) = (\rho_2/2)(\rho_2 + 1)^{-1/\rho_2}$). If $t \leqslant t^*$, then $TH^*(t) = \infty$, and if $t > t^*$, then $TH^*(t) = 0$.*

This proposition indicates that the optimal TBP is to idle station 1 as much as possible when t is sufficiently large (i.e., use $TH^* = 0$ when $t > t^*$), or to not idle it at all when t is small (i.e., $t \leqslant t^*$). The intuitive explanation behind this is that reducing the queue sizes at station 2 via the TBP reduces $P(\hat{W}_2 > t)$ but introduces $P(\hat{W}_1 > t) > 0$ (which is 0 under the NI policy). When t is large, the reduction in $P(\hat{W}_2 > t)$ is substantial, whereas the increase in $P(\hat{W}_1 > t)$ is small, and thus TBP outperforms the NI policy. However, if t is small, the waits longer than t are quite common at station 2 even if some customers are reallocated to station 1, whereas the increase in $P(\hat{W}_1 > t)$ may be substantial. Thus $TH^* = \infty$, and TBP is equivalent to the NI policy. In this case the re-allocation of waiting time will not solve the problem of excessive waits—the only solution is adding more capacity to the system.

From (11) and (12), the reduction in $PW(t)$ due to TBP for $t > t^*$ is

$$\frac{PW^{NI}(t) - PW^{TBP(0)}(t)}{PW^{NI}(t)} = 1 - (1 + \rho_2)e^{-\lambda(1-\rho_2)t}.$$

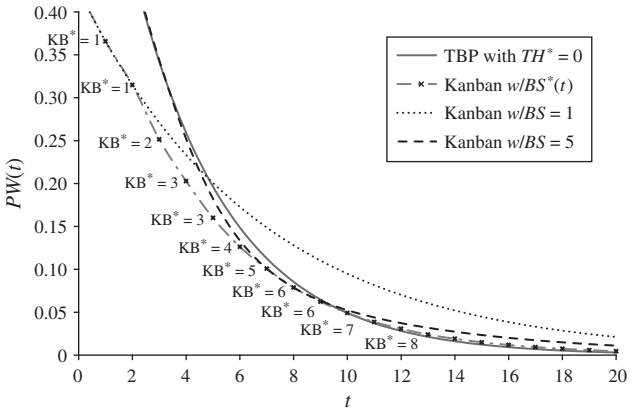
Thus, the relative improvement in $PW(t)$ increases with t, and approaches 100% as t increases. This shows that the TBP can dramatically reduce the incidence of excessive waits, but only if the designation of an "excessive" wait is used correctly, i.e., a wait is "excessive" if it is uncommon in the system.

The implications for the decision maker are clear: if waits of at least t adversely affect customer service experience, and $t > t^*$, the TBP can be used to improve $PW(t)$. If $t \leqslant t^*$, then the only way to improve $PW(t)$ is by adding capacity to the system (i.e., increasing μ_2). Most of the behaviors observed for the $\mu_1 = \infty$ case will also hold for the $\mu_1 < \infty$ case discussed in §5.

4.4. Insight 2: Comparing the TBP and the Kanban Policy for the Asymptotic Case

For the two-station tandem queue, a Kanban policy is defined by the buffer size ($BS \geqslant 1$) in front of station 2: station 1 is idled and will not admit the next customer to service whenever $q_2 \geqslant BS$.

Figure 2. $PW(t)$ as a function of t under TBP versus Kanban policy.



Observe that in the $\mu_1 = \infty$ case, under the Kanban(BS) policy station 2 operates as long as there are customers in the system for any $BS \geqslant 1$. Thus the expected sojourn time for any Kanban policy is the same as for an NI policy. Therefore, as in the TBP case, we focus only on $PW(t)$.

Using similar analysis as for the TBP, we have the following.

PROPOSITION 3. *For* $BS \geqslant 1$,

$$PW^{\text{Kanban}(BS)}(t) = \begin{cases} \dfrac{1}{2}\rho_2 e^{-(\mu_2-\lambda)t} & \text{if } BS = 1, \\ \dfrac{1}{2}\rho_2^{BS} e^{-(\mu_2-\lambda)t} + \dfrac{1}{2}e^{-\mu_2 t} & \\ \quad \cdot \displaystyle\sum_{k=0}^{BS-2} \frac{(\mu_2 t)^k}{k!}\rho_2^{k+1} & \text{if } BS \geqslant 2. \end{cases} \tag{13}$$

Note that $PW^{\text{Kanban}(1)}(t) = PW^{NI}(t) = PW^{\text{Kanban}(\infty)}(t)$. This is because when $BS = 1$, the Kanban policy shifts all waiting time to station 1 without changing the distribution of waiting times. This shows that by optimizing the buffer size, a Kanban policy can outperform the NI policy with respect to the $PW(t)$ measure. The second equation holds because when $BS = \infty$, station 1 is never idled.

In Figure 2 we compare $PW^{TBP(0)}(t)$ with $PW^{\text{Kanban}}(t)$ under different BS values, when $\lambda = 0.85$ and $\mu_2 = 1$. Note that $t^* = 4.82$ in this case, and thus TBP(0) outperforms the NI policy for $t > 4.82$. Recalling that the Kanban(1) policy is equivalent to the NI policy, we see that this is indeed the case in Figure 1, with the relative gap growing with t. Comparing TBP(0) with Kanban(5) we see that the TBP has a lower $PW(t)$ for $t > 9.28$, whereas the Kanban performs better for lower values of t.

Furthermore, using a similar analysis to the one in the proof of Proposition 2, we can obtain the buffer size $BS^*(t)$ that minimizes $PW^{\text{Kanban}(BS)}(t)$ for any t. (Specifically, the function $(\mu_2 t)^{BS-1}/(BS-1)! - (1-\rho_2)e^{\lambda t}$ has one or two zero points; if the function has two zero points, $BS^*(t)$ is the smaller zero point; otherwise $BS^*(t) = 1$.) The resulting Kanban$(BS^*(t))$ policy is plotted in Figure 2 along with the associated $BS^*(t)$ values. This policy achieves lower $PW(t)$ values than the TBP(0) for $t < 9.96$ and slightly higher values for $t > 9.96$.

For the asymptotic case the Kanban policies perform competitively with TBP(0), particularly when the buffer size is optimized for a given t value. We note that the TBP is more robust—because the same optimal threshold $TH^* = 0$ value applies over a wide range of t values, whereas the optimal buffer size $BS^*(t)$ is sensitive to t. More importantly, the performance of Kanban policies in the asymptotic case are somewhat misleading; we will see in the following sections that the performance in other cases may be significantly worse than that of the TBP.

5. Analysis of the Tandem Queue: General Case

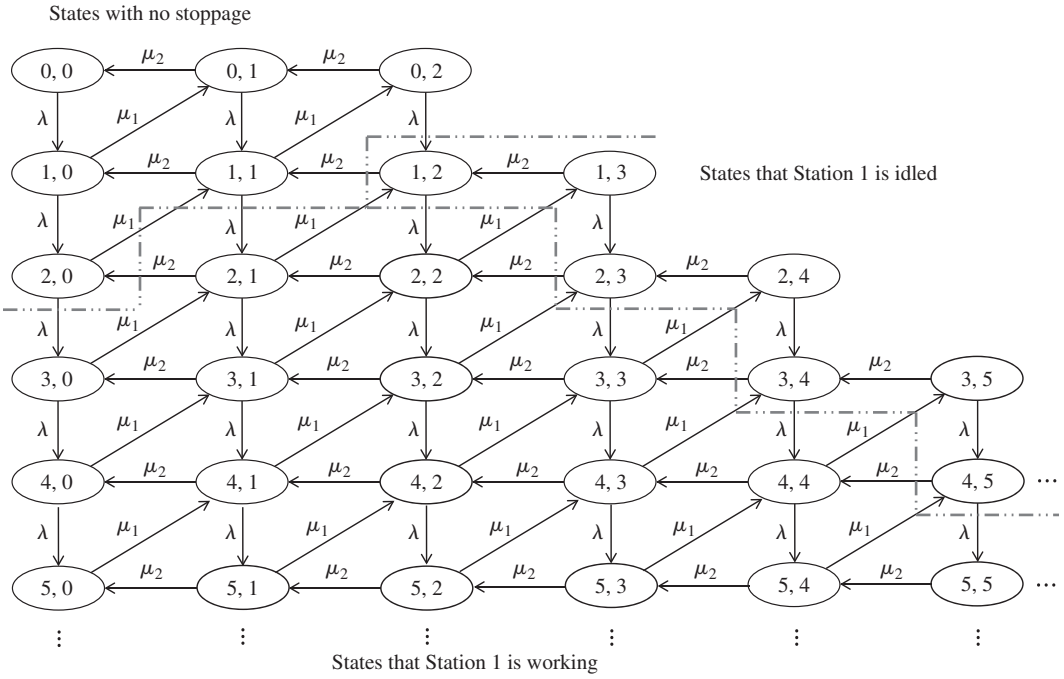
In this section, we begin by analyzing the TBP for the tandem queueing network when $\mu_1 < \infty$.

Figure 3 illustrates the MC of the tandem queueing network under the TBP with $TH = 1$. Recall that under the TBP it is not possible to reach a state (q_1, q_2) such that $q_2 - q_1 > TH + 1$. As illustrated in the figure, the states can be classified into three groups, depending on whether customers waiting for service at station 1 experience stoppage before they enter station 2. For example, if the system is currently in state $(2, 0)$, neither customer at station 1 can possibly experience any stoppages before entering station 2. The same is true for all the other states above the dashed line in the top left corner of Figure 3. On the other hand, in all states to the right of the dashed boundary line, station 1 is idled, and thus all customers at this station will experience one or more stoppage before entering station 2; state $(2, 3)$ is an example of this type.

Finally, customers at station 1 in all the states below and to the left of the dashed line may or may not experience a stoppage before entering station 2. Consider, for example, state $(3, 0)$. Whereas the first two customers at station 1 will not experience a stoppage, the situation is less clear for the last customer. We refer to this customer as the TC. If the next two events are both "Completion 1," the system will move to state $(1, 2)$ and the TC will be stopped. If, on the other hand, at least one of the next two events is arrival or Completion 2, the TC will not be stopped.

This discussion illustrates why the analysis of the TBPs is challenging. The number of stoppages experienced by the TC (and thus the distribution of her waiting time) depends on queue lengths at both stations and on customers arriving after the TC, i.e., this number depends on *future* events. The latter dependency prevents us from using distributional Little's law. Furthermore, the distribution of waiting time experienced by a customer depends not just on the state of the system, but also on the customer's position in the line at station 1. As discussed in the example above, the customer

Figure 3. The MC (Q_1, Q_2) for $TH = 1$.



immediately in front of the TC will not experience any stoppages, and thus his distribution of the waiting time is clearly different from that for the TC. This implies that the observed state (q_1, q_2) of the system is not sufficient to uniquely express the distribution of $W_1^{q_1, q_2}$.

To overcome this difficulty, we augment the state space with a position indicator for each customer. Specifically, for each TC, in addition to the queue length indicators, we also include the position of the TC in station 1; we denote this position s for $s \geqslant 1$. Note that each TC now generates a new MC upon arrival, which we name TCMC.

This TCMC has three dimensions. When the TC arrives in state (q_1, q_2), she joins station 1 and becomes the sth $= (q_1 + 1)$th customer, so that the first state of the TCMC is $(q_1 + 1, q_2, q_1 + 1)$. If we consider all states with the same s as one layer, each layer looks similar to the MC in Figure 3 except that there are no states with $q_1 > s$. The same three events discussed in Section 3 may occur in the TCMC as well. Their effect on state (q_1, q_2, s) is as follows.

1. *Arrival*—The TCMC transitions to state $(q_1 + 1, q_2, s)$. Arrivals occur with rate λ in any state.

2. *Completion* 1—The TCMC transitions to state $(q_1 - 1, q_2 + 1, s - 1)$. This happens with rate μ_1, if $q_1 > 0$ and $\delta < TH$ (when station 1 is not idled).

3. *Completion* 2—The TCMC transitions to state $(q_1, q_2 - 1, s)$. This event occurs with rate μ_2, if $q_2 \geqslant 1$.

When $s > 1$, the TC is waiting in station 1. When $s = 1$, the TC is either in service or is the first in line to enter service when the stoppage of station 1 ends. Since $\lambda < \mu_1$, the TCMC (q_1, q_2, s) will be absorbed in some state with $s = 0$, when the TC moves to station 2. Let $X^{q_1, q_2, s}$ represent the TC's performance measure, given the network is in state (q_1, q_2, s).

To obtain the performance measure using (2) we can keep track of the TCMC starting from the state $(q_1 + 1, q_2, q_1 + 1)$ and calculate the conditional performance measure according to all possible paths the TC may take until an absorbing state is reached. However, because the TCMC is three-dimensional, the required computational effort grows rapidly using this intuitive approach. We thus simplify the problem as shown below.

We first show that, similarly to the $\mu_1 = \infty$ case, the number of stoppages can be bounded.

Lemma 1. *If the TCMC is in state (q_1, q_2, s), then the maximum number of stoppages the TC may see, $M^{q_1, q_2, s}$, is*

$$M^{q_1, q_2, s} = \max\{2s - TH + \delta(q_1, q_2) - 1, 0\}.$$

Specially, if $\delta(q_1, q_2) \leqslant TH - 2s + 1$, there will be no stoppage for the TC. Thus, the performance measure experienced by a customer that reaches such states are independent of future arrivals.

It is easy to see that $\hat{M}^{q_1, q_2} = M^{q_1+1, q_2, q_1+1}$, i.e., Lemma 1 shows that, the number of stoppages the TC sees in the $\mu_1 = \infty$ case is the *maximum* number of stoppages the TC may see in $\mu_1 < \infty$ case. The reason is that when $\mu_1 = \infty$, the service time of station 1 is zero, so the set of sequential events Completion 1 $\Rightarrow$ Arrival (or Completion 2) $\Rightarrow$ Arrival (or Completion 2) repeats for sure; when $\mu_1 < \infty$, this set of sequential events repeats only in the worst case.

We define *a no-stoppage state* to be a state in the TCMC s.t. $\delta(q_1, q_2) \leqslant TH - 2s + 1$, i.e., $M^{q_1, q_2, s} = 0$. For example, consider again state $(3, 0)$ in the MC in Figure 3. As discussed previously, states $(3, 0, 1)$ and $(3, 0, 2)$ in the corresponding TCMC are no stoppage. On the other hand, by Lemma 1, $M^{3,0,3} = 1$, so stoppage may occur in state $(3, 0, 3)$.

Observe that once the TCMC reaches a no-stoppage state, the network acts like a nonidling tandem queueing network for the TC, and the distributions of the three steady state service measures can be calculated directly (see below). In the following sections, we treat no-stoppage states as absorbing states and use a recursion method to develop all three performance measures as follows:

- If state (q_1, q_2, s) is a no-stoppage state, i.e., $\delta \leqslant TH - 2s + 1$, then the distribution of $X^{q_1, q_2, s}$ can be calculated from Propositions 4 and 5 below.
- If station 1 is stopped, i.e., $\delta = TH$ or $TH + 1$, both arrival (w.p. $\lambda/(\lambda+\mu_2)$) and Completion 2 (w.p. $\mu_2/(\lambda+\mu_2)$) can happen in the TCMC. Using conditional probability, the distribution of $X^{q_1, q_2, s}$ can be recursively calculated from the distributions of $X^{q_1+1, q_2, s}$ and $X^{q_1, q_2-1, s}$.
- For states (q_1, q_2, s) such that $TH - 2s + 1 < \delta \leqslant TH - 1$, arrival (w.p. $\lambda/(\lambda + \mu_1 + \mu_2)$), Completion 1 (w.p. $\mu_1/(\lambda + \mu_1 + \mu_2)$), and Completion 2 (w.p. $\mu_2/(\lambda + \mu_1 + \mu_2)$) can all happen in the TCMC. Using conditional probability, the distribution of $X^{q_1, q_2, s}$ can be calculated from the distributions of $X^{q_1+1, q_2, s}$, $X^{q_1-1, q_2+1, s-1}$, and $X^{q_1, q_2-1, s}$.

Calculating the LT of X, similarly to (2), requires the steady state probability vector of the MC (Q_1, Q_2). It is easily seen that this MC is irreducible and aperiodic, and has equilibrium probabilities, π_{q_1, q_2}. The balance equation for the (Q_1, Q_2) MC are as follows (these are easier to follow when looking at Figure 3):

(1) When $\delta < TH$ and $q_1 = q_2 = 0$, we have $\lambda\pi_{0,0} = \mu_2\pi_{0,1}$.

(2) When $\delta < TH$ and $q_1 > 0$, $q_2 = 0$, we have $(\lambda+\mu_1) \cdot \pi_{q_1,0} = \lambda\pi_{q_1-1,0} + \mu_2\pi_{q_1,1}$.

(3) When $\delta < TH$ and $q_1 > 0$, $q_2 > 0$, we have $(\lambda + \mu_1 + \mu_2)\pi_{q_1,q_2} = \lambda\pi_{q_1-1,q_2} + \mu_1\pi_{q_1+1,q_2-1} + \mu_2\pi_{q_1,q_2+1}$.

(4) When $\delta \leqslant TH$ and $q_1 = 0$, $0 < q_2 \leqslant TH$, we have $(\lambda+\mu_2)\pi_{0,q_2} = \mu_1\pi_{1,q_2-1} + \mu_2\pi_{0,q_2+1}$.

(5) When $\delta = TH$ and $q_1 > 0$ (then $q_2 = q_1 + TH$), we have $(\lambda+\mu_2)\pi_{q_1,q_2} = \lambda\pi_{q_1-1,q_2} + \mu_1\pi_{q_1+1,q_2-1} + \mu_2\pi_{q_1,q_2+1}$.

(6) When $\delta = TH + 1$ and $q_1 \geqslant 1$ (implying $q_2 = q_1 + TH + 1$), we have $(\lambda+\mu_2)\pi_{q_1,q_2} = \mu_1\pi_{q_1+1,q_2-1}$.

(7) We also require $\sum_{q_1,q_2}\pi_{q_1,q_2} = 1$.

To solve these balance equations, we approximate π_{q_1,q_2} by assuming that station 1 has a finite waiting room of size *Limit*. For any finite value of *Limit*, we can calculate an approximation of π_{q_1,q_2} by solving the balance equations numerically. When *Limit* goes to infinity, the approximation approaches π_{q_1,q_2}. In our numerical experiments we found that $P\{q_1 = 100\} < 10^{-5}$, so *Limit* $= 100$ appears to be an adequate value for our parameter choices.

5.1. Distribution of Waiting Time for Station 1: W_1

In this section, we consider the TC's waiting time for station 1, W_1. Note that there are two components of W_1: the time spent waiting for $s-1$ Completion 1 events and the time spent when station 1 is idled. The first component depends only on s, and the second one is determined by s and $\delta = q_2 - q_1$. Thus, given s and δ, W_1 does not depend on the values of q_1 and q_2. Indeed, from Lemma 1 we see that the maximum number of stoppage $M^{q_1,q_2,s}$ only depends on s and δ; thus, we will next use $M^{s,\delta}$ to denote the maximum number of stoppages for a customer that is at a position s in queue 1 when $q_2 - q_1 = \delta$. A revised TCMC, with the state description (s, δ), is illustrated in Figure 4 for the case $TH = 1$; this simplified TCMC will be used to compute W_1.

Arrival or Completion 2 events do not affect s; these events only decrease the value of δ by 1. Completion 1 decreases the value of s by 1 and increases the value of δ by 2. If $\delta = TH$ or $TH + 1$, station 1 is idled, so that the next event can only be arrival or Completion 2.

In Figure 4, the column on the right-hand side, starting from $(1, 0)$, represents the no-stoppage states established in Lemma 1. The states above the doted line are states where station 1 is idled, i.e., with $\delta \geqslant TH = 1$.

Let $W_1^{s,\delta}$ be the TC's waiting time for station 1 while the network is in state (s, δ), and denote its LT by $L_{W_1^{s,\delta}}(h)$. Note that $W_1^{s,\delta}$ is composed of two parts. The first part is the service time of the $s-1$ customers in front of the TC in station 1. This service time distribution is Erlang$(s-1, \mu_1)$. The second part consists of stoppages in station 1. As in the $\mu_1 = \infty$ case, the length of each stoppage is distributed as an $\exp(\lambda+\mu_2)$ RV.

Let $B_1^{s,\delta}$ denote the actual number of stoppages the TC will experience if she is in state (s, δ). Thus, the LT of $W_1^{s,\delta}$ is

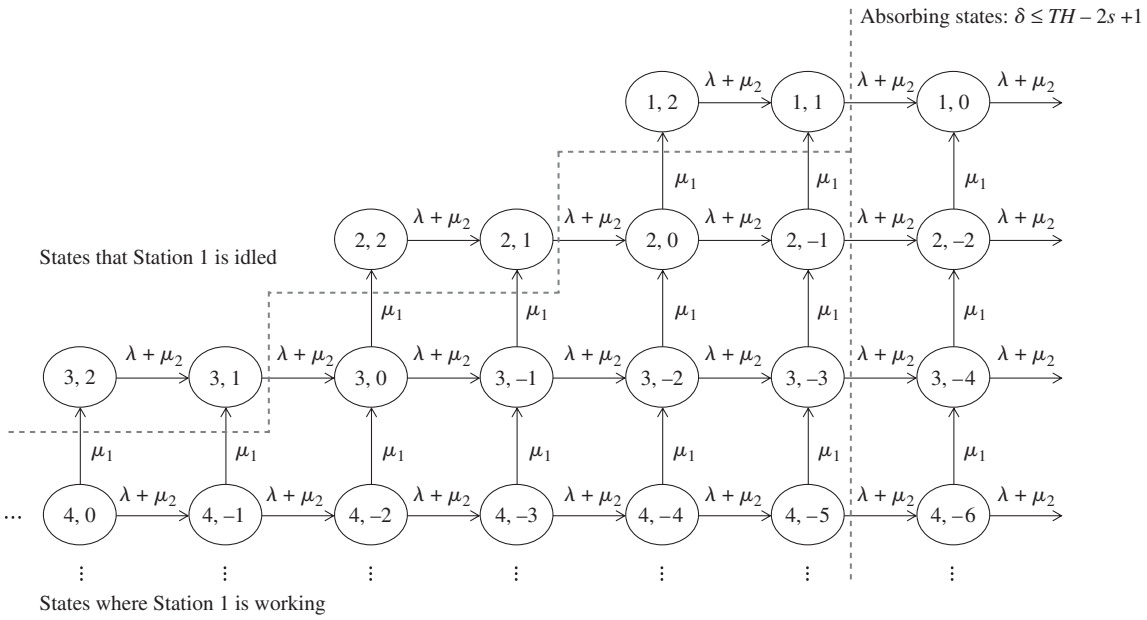
$$L_{W_1^{s,\delta}}(h) = \left(\frac{\mu_1}{\mu_1+h}\right)^{s-1} \sum_{i=1}^{M^{s,\delta}} P\{B_1^{s,\delta} = i\} \left(\frac{\lambda+\mu_2}{\lambda+\mu_2+h}\right)^{i}, \quad (14)$$

where $M^{s,\delta}$ can be found from Lemma 1, and $\sum_{i=1}^{M^{s,\delta}} P\{B_1^{s,\delta} = i\} = 1$.

Thus, finding $L_{W_1^{s,\delta}}(h)$ is equivalent to finding the distribution of $B_1^{s,\delta}$, for any $s \geqslant 1$, $\delta \leqslant TH + 1$. This can be done as follows:

- If (s, δ) is a no-stoppage state, i.e., $\delta \leqslant TH - 2s + 1$, then $B_1^{s,\delta} = 0$ from Lemma 1.
- If station 1 is stopped, i.e., for states with $\delta = TH$ or $TH + 1$, $B_1^{s,\delta}$ has the same distribution as $1 + B_1^{s,\delta-1}$.
- Otherwise, for states (s, δ) such that $TH - 2s + 1 < \delta \leqslant TH - 1$, the TCMC will go to state $(s, \delta - 1)$ (w.p. $(\lambda + \mu_2)/(\lambda + \mu_1 + \mu_2)$), or to state $(s-1, \delta+2)$ (w.p. $\mu_1/(\lambda + \mu_1 + \mu_2)$). Therefore, $B_1^{s,\delta}$ is distributed the same as $B_1^{s,\delta-1}$ or $B_1^{s-1,\delta+2}$, depending on which state the TCMC transitions to.

Figure 4. The revised TCMC (s, δ), when $TH = 1$.



Since $s \in \{1, \ldots, Limit\}$ and $\delta \in \{-Limit, \ldots, TH + 1\}$, the distribution of $B_1^{s,\delta}$ can now be computed iteratively; see Algorithm 1 in Section EC.2 of the e-companion for details.

5.2. Distribution of Waiting Time for Station 2: W_2

In this section we calculate $W_2^{q_1, q_2, s}$—the TC's waiting time for station 2, given that the network is at state (q_1, q_2, s). Let $K^{q_1, q_2, s}$ be the number of customers the TC sees when she enters station 2. Given $K^{q_1, q_2, s} = k$, we know that $W_2^{q_1, q_2, s} \sim \text{Erlang}(\mu_2, k)$. So once we know the distribution of $K^{q_1, q_2, s}$, the LT of $W_2^{q_1, q_2, s}$ can be expressed as

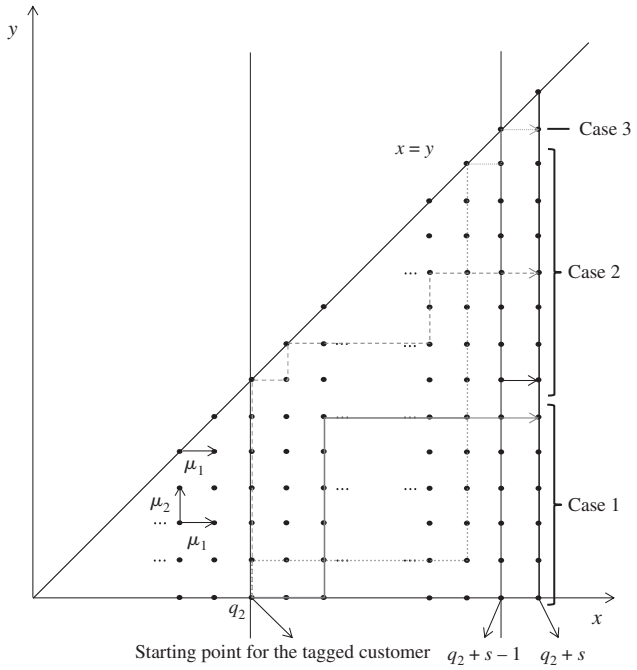
$$L_{W_2^{q_1, q_2, s}}(h) = \sum_{k=0}^{q_2+s-1} \left(\frac{\mu_2}{\mu_2 + h}\right)^k P\{K^{q_1, q_2, s} = k\}. \tag{15}$$

We next derive the distribution of $K^{q_1, q_2, s}$, first for no-stoppage states and then for states with stoppages.

5.2.1. Distribution of W_2 at No-Stoppage States. First, assume that the network is currently in a no-stoppage state, i.e., (q_1, q_2, s), and $\delta \leqslant TH - 2s + 1$. Given Lemma 1, station 1 will not be idled before the TC enters station 2. Thus, the arrival process does not affect the network, and we need to only consider the service processes of stations 1 and 2. Still, it is possible for station 2 to be *starved*, i.e., $q_2 = 0$, before the TC enters this station. We next discuss how to consider the starvation periods when calculating the distribution of $K^{q_1, q_2, s}$.

We represent the service operation of the TC by a Random Walk (RW) process in a two dimensional lattice graph, where the x and y axes represent the number of customers served by the first and second servers, respectively. Let the TC be the Nth arrival to the original tandem queue. Denote the total number of customers served by stations 1 and 2 before the TC's arrival by X_N and Y_N, respectively. Note that $X_N \in [0, \ldots, N-1]$, $Y_N \in [0, \ldots, X_N]$, and $q_2 = X_N - Y_N$. The RW process is depicted in Figure 5. Obviously, the RW cannot go above the line $x = y$ (service 1 must finish before service 2). When station 1 completes service the RW moves to the right, and when station 2 completes service the RW moves up. Because both service completions are exponentially distributed, when both stations are busy, $P\{RW\ moves\ right\} = \mu_1/(\mu_1 + \mu_2)$ and

Figure 5. Lattice graph of number of customers served by each station.



$P\{RW\ moves\ up\} = \mu_2/(\mu_1+\mu_2)$. Any point on the line $x = y$ means that station 2 is starved, and the next possible move for the RW is only to the right. We call points on the line $x = y$ points with station 2 starved and other points in Figure 5 points with station 2 working.

For any TC, we can ignore Y_N, because these customers have already left the network. Therefore, upon arrival of the TC, we reset the starting point of the RW to $(X_N, Y_N) = (q_2, 0)$.

When the TC arrives to state (q_1, q_2, s) , there are q_2 customers in station 2, which corresponds to the point $(q_2, 0)$ on Figure 5. When the TC finishes service in station 1, this station has finished s customers, which represents the RW moving right s steps and reaching the line $x = q_2 + s$. By this time, station 2 has served n customers, where $0 \leqslant n \leqslant q_2 + s - 1$. Thus, the sojourn time of the TC at station 1 corresponds to the time the RW moves from point $(q_2, 0)$ to a point on the line $(q_2 + s, n)$, with $0 \leqslant n \leqslant q_2 + s - 1$.

Let $B_2^{q_1,q_2,s}$, the number of times station 2 is starved from when the TC arrives to the network and until she finishes service at station 1. The joint distribution of n and $B_2^{q_1,q_2,s}$ can be calculated using the result from Milch and Waggoner (1970). This gives us the marginal distribution of n. Since the number of customers the TC sees upon entering station 2 is $K^{q_1,q_2,s} = q_2 + s - 1 - n$, this also provides the distribution of $K^{q_1,q_2,s}$:

Proposition 4. *For any state* (q_1, q_2, s) *with* $\delta \leqslant TH - 2s + 1$, *the distribution of* $K^{q_1,q_2,s}$ *is*

$$
\begin{aligned}
&P\{K^{q_1,q_2,s} = k\}\\
&= \left[\binom{2s+q_2-3}{s-1} - \binom{2s+q_2-3}{s+q_2-1}\right]\\
&\quad\cdot\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{s-1}\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{q_2+s-1}\\
&\quad+\sum_{i=2}^{s}\left[\binom{2s+q_2-i-2}{s+q_2-2} - \binom{2s+q_2-i-2}{s+q_2-1}\right]\\
&\quad\cdot\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{s-i}\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{q_2+s-1} \quad \text{if } k = 0\\
&= \left[\binom{2s+q_2-k-2}{s-1} - \binom{2s+q_2-k-2}{s+q_2-1}\right]\\
&\quad\cdot\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{s}\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{q_2+s-k-1}\\
&\quad+\sum_{i=1}^{s-k}\left[\binom{2s+q_2-k-i-2}{s+q_2-2}\right.\\
&\qquad\left.-\binom{2s+q_2-k-i-2}{s+q_2-1}\right]\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{s-i}\\
&\qquad\cdot\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{q_2+s-k-1} \quad \text{if } 0 < k \leqslant s-1\\
&= \binom{q_2+2s-2-k}{q_2+s-1-k}\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{s}\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{q_2+s-1-k}\\
&\qquad\text{if } s-1 < k \leqslant q_2+s-1. \qquad (16)
\end{aligned}
$$

5.2.2. Distribution of W_2 for States with Stoppages. To calculate $K^{q_1,q_2,s}$ for states with stoppages, we observe the following:

- If station 1 is stopped, i.e., $\delta = TH$ or $TH + 1$, both arrival (w.p. $\lambda/(\lambda+\mu_2)$) and Completion 2 (w.p. $\mu_2/(\lambda+\mu_2)$) can happen in the TCMC. So $K^{q_1,q_2,s}$ will be distributed as $K^{q_1+1,q_2,s}$ or $K^{q_1,q_2-1,s}$, depending on which event happens.
- For states (q_1, q_2, s) such that $TH - 2s + 1 < \delta \leqslant TH - 1$, arrival (w.p. $\lambda/(\lambda+\mu_1+\mu_2)$), Completion 1 (w.p. $\mu_1/(\lambda+\mu_1+\mu_2)$), and Completion 2 (w.p. $\mu_2/(\lambda+\mu_1+\mu_2)$) can all happen in the TCMC. So $K^{q_1,q_2,s}$ will be distributed as $K^{q_1+1,q_2,s}$, $K^{q_1-1,q_2+1,s-1}$, and $K^{q_1,q_2-1,s}$, with these probabilities respectively.

Notice that the distribution of $K^{q_1,q_2,s}$ depends only on which no-stoppage state the process finally reaches, and is independent of the other details of the service process before that. Algorithm 2, given in Section EC.2 of the e-companion, uses these three conditions and Proposition 4 to express $K^{q_1,q_2,s}$ for any state (q_1, q_2, s). The distribution of $W_2^{q_1,q_2,s}$ can now be computed from (15).

Remark 3. It may be of interest to compute the distribution of the total wait in the system for the TC, $W^{q_1,q_2,s} = W_1^{q_1,q_2,s} + W_2^{q_1,q_2,s}$. First note that $W_1^{q_1,q_2,s}$ and $W_2^{q_1,q_2,s}$ are not independent: since station 2 is never intentionally idled, the longer the TC stays in station 1, the fewer customers she will see, on average, when she enters station 2. Still, in a similar way to Algorithms 1 and 2, one can calculate the distribution of $W^{q_1,q_2,s}$.

5.3. Distribution of Sojourn Time: S

In this section, we calculate the LT of sojourn time, which is the sum of waits and services in both stations, for the TC. This derivation allows us to express both $E[S]$ and $P\{S > t\}$.

We focus on station 2. The TC's sojourn time, $S^{q_1,q_2,s}$, is between her arrival to the network and her departure, i.e., the time when station 2 finishes serving $q_2 + s$ customers. We note that if there are customers in the network, station 2 always serves customers when station 1 is idled; and station 1 always serves customers (if there are any customers in station 1) when station 2 is starved. Thus, the TC's sojourn time is composed of two parts. The first part is the service time of the $q_2 + s$ customers at station 2, which is Erlang$(\mu_2, q_2 + s)$. The second part is the total time that station 2 starves until it serves the TC. This time may depend on the behavior of the network after the TC's arrival and is therefore more challenging to characterize.

We know that the number of times station 2 is starved $B_2^{q_1,q_2,s} \leqslant s$, because in the worst case Completion 2 happens q_2 times and then the {Completion 1, Completion 2} sequence repeats until the TC is served at station 2, so that $\sum_{i=0}^{s} P\{B_2^{q_1,q_2,s} = i\} = 1$.

Similar to (15), the common form of the LT of $S^{q_1,q_2,s}$ is

$$L_{S^{q_1,q_2,s}}(h)=\left(\frac{\mu_2}{\mu_2+h}\right)^{q_2+s}\cdot\sum_{i=0}^{s}P\{B_2^{q_1,q_2,s}=i\}\left(\frac{\lambda+\mu_2}{\lambda+\mu_2+h}\right)^{i}. \quad (17)$$

This transforms the problem to finding the distribution of $B_2^{q_1,q_2,s}$, for any state (q_1,q_2,s). We first consider no-stoppage states. As in the proof of Proposition 4, for the no-stoppage states we use the joint distribution of $q_2+s-1-K^{q_1,q_2,s}$ and $B_2^{q_1,q_2,s}$, $P\{q_2+s-1-K^{q_1,q_2,s}=n, B_2^{q_1}, q_2, s=i\}$. Using this distribution and the law of total probability, we get the following:

Proposition 5. *For any state (q_1,q_2,s) with $\delta\leqslant TH-2s+1$, the distribution of $B_2^{q_1,q_2,s}$ is*

$$\begin{aligned}
&P\{B_2^{q_1,q_2,s}=i\}\\
&=\sum_{n=0}^{q_2-1}\binom{n+s-1}{n}\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{s}\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{n}\\
&\quad+\sum_{n=q_2}^{q_2+s-2}\left[\binom{s+n-1}{s-1}-\binom{s+n-1}{s+q_2-1}\right]\\
&\quad\cdot\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{s}\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{n},\quad \text{if } i=0\\
&=\sum_{n=q_2}^{q_2+s-2}\left[\binom{s+n-2}{s+q_2-2}-\binom{s+n-2}{s+q_2-1}\right]\\
&\quad\cdot\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{s-1}\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{n}\\
&\quad+\left[\binom{2s+q_2-3}{s-1}-\binom{2s+q_2-3}{s+q_2-1}\right]\\
&\quad\cdot\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{s-1}\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{q_2+s-1},\quad \text{if } i=1\\
&=\sum_{n=q_2}^{q_2+s-2}\left[\binom{s+n-i-1}{s+q_2-2}-\binom{s+n-i-1}{s+q_2-1}\right]\\
&\quad\cdot\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{s-i}\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{n}\\
&\quad+\left[\binom{2s+q_2-i-2}{s+q_2-2}-\binom{2s+q_2-i-2}{s+q_2-1}\right]\\
&\quad\cdot\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{s-i}\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{q_2+s-1},\\
&\qquad\qquad \text{if } 2\leqslant i<n-q_2+1. \quad (18)
\end{aligned}$$

We can now calculate $B_2^{q_1,q_2,s}$ for any state (q_1,q_2,s) as follows:

- If the state (q_1,q_2,s) is in a no-stoppage state, i.e., $\delta\leqslant TH-2s+1$, the distribution is given by Proposition 5.
- If station 1 is idled, i.e., $\delta=TH$ or $TH+1$, both arrival (w.p. $\lambda/(\lambda+\mu_2)$) and Completion 2 (w.p. $\mu_2/(\lambda+\mu_2)$) can happen in the TCMC. So $B_2^{q_1,q_2,s}$ will be distributed as $B_2^{q_1+1,q_2,s}$ or $B_2^{q_1,q_2-1,s}$.
- For states (q_1,q_2,s) such that $TH-2s+1<\delta\leqslant TH-1$ and $q_2=0$, there is no customer in station 2. Arrival (w.p. $\lambda/(\lambda+\mu_1)$) and Completion 1 (w.p. $\mu_1/(\lambda+\mu_1)$) can happen in the TCMC. So $B_2^{q_1,0,s}$ is distributed as $B_2^{q_1+1,0,s}$ or $B_2^{q_1-1,1,s-1}+1$.
- For states (q_1,q_2,s) such that $TH-2s+1<\delta\leqslant TH-1$ and $q_2\neq 0$, arrival (w.p. $\lambda/(\lambda+\mu_1+\mu_2)$), Completion 1 (w.p. $\mu_1/(\lambda+\mu_1+\mu_2)$), and Completion 2 (w.p. $\mu_2/(\lambda+\mu_1+\mu_2)$) can all happen in the TCMC. So $B_2^{q_1,q_2,s}$ will be distributed as $B_2^{q_1+1,q_2,s}$, $B_2^{q_1-1,q_2+1,s-1}$ or $B_2^{q_1,q_2-1,s}$.

Algorithm 3 in Section EC.2 of the e-companion uses these four conditions to compute the distribution of $B_2^{q_1,q_2,s}$ for any state (q_1,q_2,s). The LT of the sojourn times $S^{q_1,q_2,s}$ can then be computed from (17).

6. Insights for the $\mu_1<\infty$ Case

In this section, we compare the performance of the TBP, the nonidling policy, and the Kanban policy with respect to the expected sojourn time, $E[S]$, and the probability of excessive waits, $PW(t)$.

6.1. Insight 3: Comparing the TBP and the Nonidling Policy

First, we compare the performance of TBP and the nonidling policy. The key questions are as follows: (1) What degree of improvement can be achieved by the TBP for the $PW(t)$ measure? And (2) by how much do sojourn times have to increase to achieve this improvement? We note that service measure $P\{S>t'\}$ could be used in place of $E[S]$. Numerical results show that the trade-off curves of $PW(t)$ and $P\{S>t'\}$ behave the same as the trade-off curves of $PW(t)$ and $E[S]$, so only $E[S]$ is considered in our numerical results.

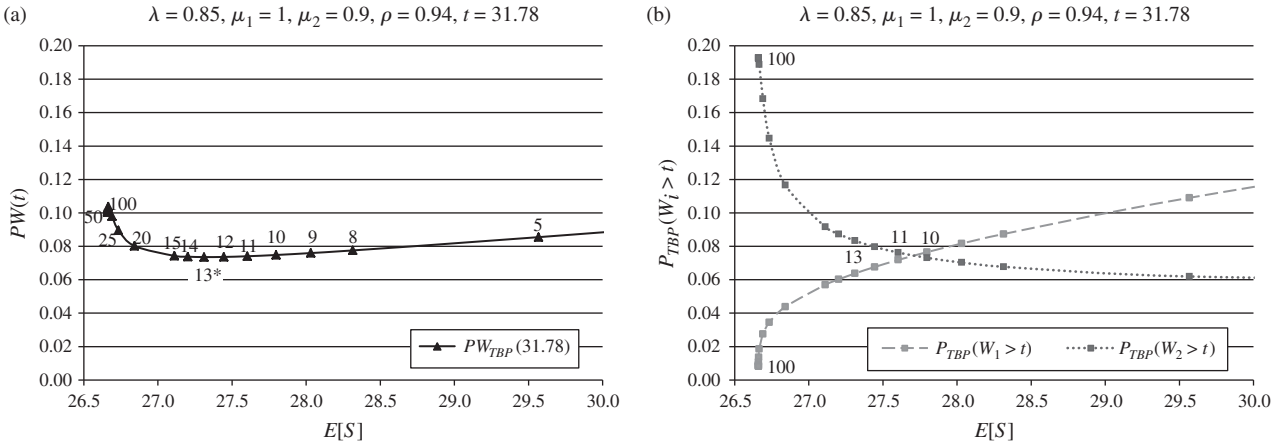
The expressions for the service measures for the nonidling policy are determined by λ, μ_1, μ_2, and t, and can be obtained from, e.g., using Burke's (1956) theorem:

$$E[S^{NI}]=\sum_{i=1}^{2}\frac{1}{\mu_i-\lambda};\quad PW^{NI}(t)=\frac{1}{2}\sum_{i=1}^{2}\frac{\lambda}{\mu_i}e^{-(\mu_i-\lambda)t}.$$

To illustrate the trade-off between $PW^{TBP}(t)$ and the expected sojourn time under the TBP, $E[S^{TBP}]$, we proceed as follows. We initially set $\lambda=0.85$, $\mu_1=1$, and $\mu_2=0.9$. Thus, station 2 is the bottleneck, and the system utilization ratio $\rho=\rho_2=0.85/0.9\approx 94\%$. Next we select t such that $PW^{NI}(t)=10\%$—from the expressions above, this value is $t=31.78$, and $E[S^{NI}]=26.67$.

We calculate the performance measures $E[S^{TBP}]$ and $PW^{TBP}(t)$ using $TH=100,99,\ldots,0$. For $TH=100$ the performance measures, $(E[S^{TBP}],PW^{TBP}(31.78))=(26.67,0.1)$ are identical to these measures for the nonidle

Figure 6. Trade-off curves corresponding to excessive wait probabilities of 10% (under the nonidling policy).



Note. The nonidling policy corresponds to the leftmost point on each curve.

system. The results in Figure 6(a) present the trade-off curve of the TBP for different thresholds. The points corresponding to selected *TH* values are labeled on the curve (they decrease from left to right).

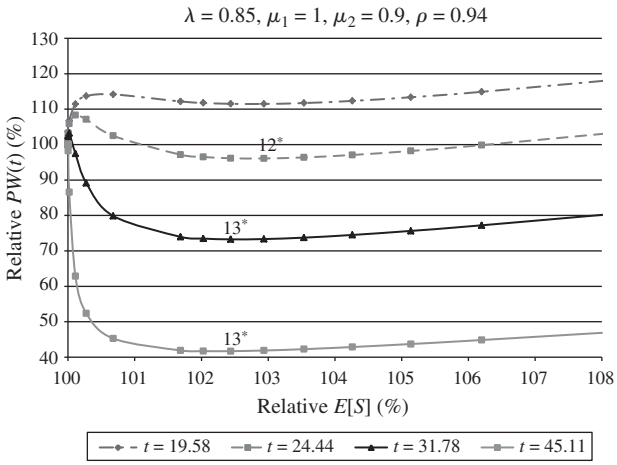
From the figure, we observe that the average sojourn times along the x-axis increase as *TH* values are decreased from 100: the lower the threshold, the more the TBP departs from the nonidling policy, with the incidents of idling of station 1 increasing. At $TH = 0$, $E[S^{TBP}] = 30.5$—a 14.4% increase over $E(S^{NI})$, the expected sojourn time under the nonidle policy. Initially, as *TH* is decreased from 100, the $PW(t)$ values are reduced, indicating that the TBP is achieving the desired trade-off between the two performance measures. The $PW^{TBP}(t)$ is minimized at just over 7%, corresponding to $TH^* = 13$ (labeled with a star). For this *TH* value, $E[S^{TBP}] = 27.31$. Thus, a TBP with $TH = 13$ achieves a nearly 30% improvement in the $PW(t)$ measure (7% versus 10%) at the cost of increasing the expected sojourn times by about 2% (from 26.67 to 27.31)—a trade-off that may be quite attractive. Reducing *TH* below 13 turns out to be counterproductive; thus, from the point of view of biobjective optimization, the *TH* values below 13 are Pareto inferior. However, all *TH* values greater than or equal to 13 are Pareto optimal.

To gain additional insight, in Figure 6(b), we plot $P(W_i > 31.78)$ for $i = 1, 2$ under the TBP. Since station 2 is the bottleneck in this case, the probability of waiting longer than t is much greater there under the nonidling policy. This is shown on the extreme left of the plot, where $TH = 100$ and the TBP is essentially identical to the NI policy. For very high *TH* values, most of the contribution to $PW(t)$ comes from station 2. As *TH* is reduced, $P(W_1 > t)$ increases and $P(W_2 > t)$ declines. Eventually, when *TH* decreases below 10, there is a much higher probability of long waits at station 1 than at station 2. It is interesting to note that $PW(t)$ is minimized at $TH^* = 13$ when the values of $P(W_1 > t)$ and $P(W_2 > t)$ are approximately equal. We have observed similar behavior with other parameter settings as well.

We have observed from numerical results with different parameter settings that $P(W_1 > t)$ is a concave increasing function and $P(W_2 > t)$ is a convex decreasing function of $E[S]$, as in Figure 6(b). However, for different values of t, the behavior of $PW(t) = \frac{1}{2}(P(W_1 > t) + P(W_2 > t))$ as a function of $E[S]$ varies, typically being convex in some regions and concave in others.

To illustrate the improvements that can be achieved with the TBP compared with the NI policy for different values of t, we plot the relative change in $PW(t)$ versus the relative change in $E[S]$ for four different values of t in Figure 7. Here 100% on both axes relates to corresponding values for the *NI* policy (or, equivalently, TBP(100) policy). Thus, on the x-axis the values increase from 100% since introducing SI can only hurt the expected service times, whereas on the y-axis we have values above and

Figure 7. Trade-off curves of the TBP corresponding to excessive wait probabilities of 20%, 15%, 10%, and 5% (under the nonidling policy) for different system parameters.



Note. The nonidling policy corresponds to the leftmost point on each curve.

below 100% since the TBP can improve or hurt the $PW(t)$ objective. The four values of $t = 45.11, 31.78, 24.44,$ and 19.58 were selected to correspond to "excessive wait" probabilities of 5%, 10%, 15%, and 20% under the NI policy, respectively.

For the case where excessive waits are rare ($t = 45.11$), the TBP provides very attractive trade-offs: decreasing $PW(t)$ by close to 60% at the cost of increasing $E[S]$ by just 2%. Moreover, most of the decrease in $PW(t)$ occurs for even smaller values of $E[S]$, corresponding to thresholds higher than the $PW(t)$-minimizing value of $TH^* = 13$. Thus, the value of TH that minimizes $PW(t)$ may not be the best choice. The reduction in $PW(t)$ provided by the TBP for the $t = 31.78$ case is a bit smaller, but is also quite substantial at nearly 30%, whereas the increase in $E[S]$ is just over 2%.

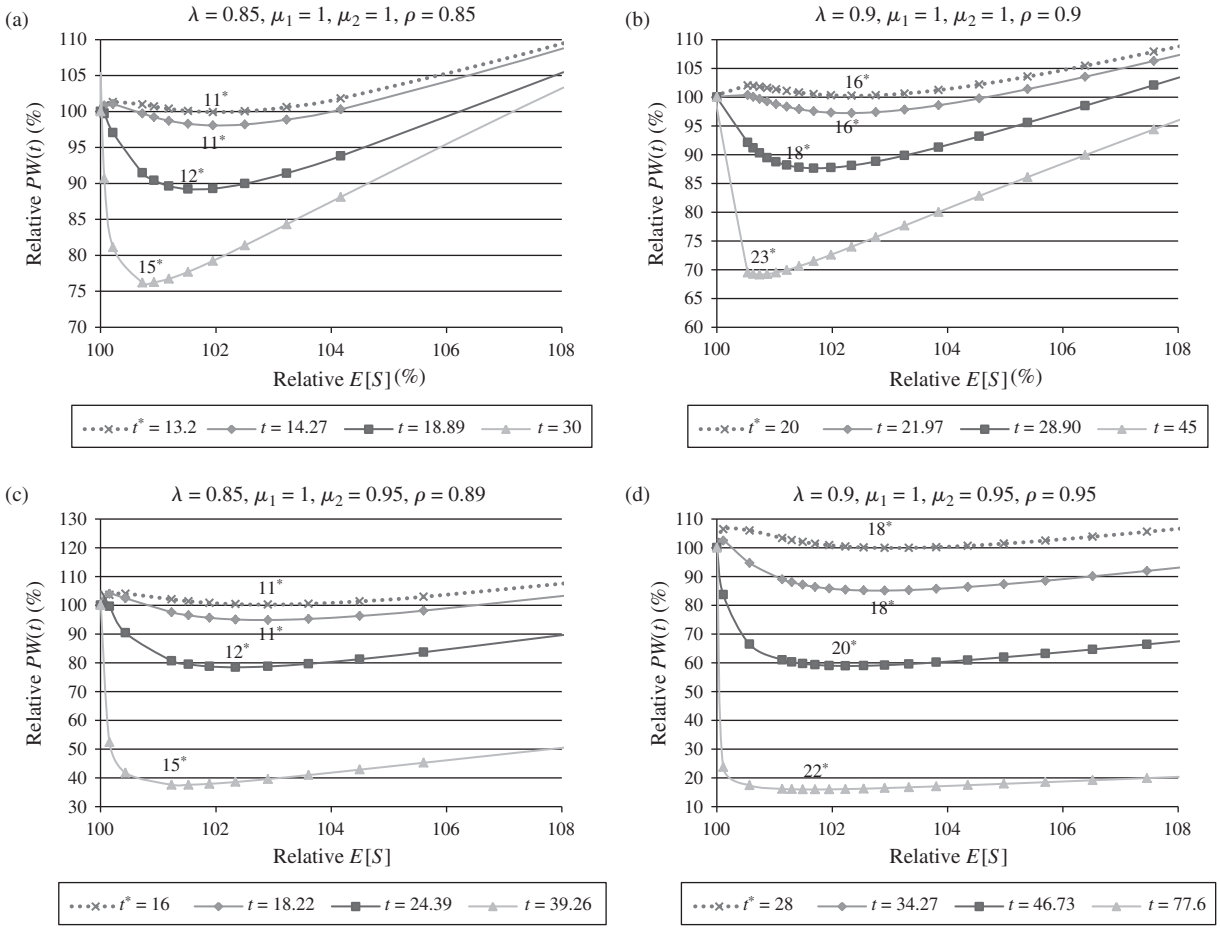
The TBP is much less successful for the $t = 24.44$ case where "excessive waits" occur 15% of the time under the NI policy. Here, as the threshold is decreased from 100, both objectives are initially hurt, with $PW(t)$ rising sharply. This is because the decrease in $P(W_2 > t)$ is very small, whereas $P(W_1 > t)$ increases rapidly. For lower TH values, the $PW(t)$ begins to fall, eventually falling about 5% below the value for the NI policy around $TH^* = 12$. The cost of this improvement is the 3% increase in $E[S]$. Thus, the trade-offs offered by the TBP are much less attractive in this case. We also observe that here $PW(t)$ is not a convex function of $E[S]$.

As the probability of excessive waits is increased to 20%, the Pareto-optimal trade-offs disappear: although the behavior of $PW(t)$ as TH values are increased is similar to the previous case (first an increase, then a slight decrease, followed by another increase), the level never gets below the value achieved for $TH = 100$, i.e., the value for the NI policy.

Thus, we observe similar patterns to the ones derived analytically for the asymptotic $\mu_1 = \infty$ case: the TBP reduces $PW(t)$ when the "excessive waits" are sufficiently rare in the system.

Since the TBP redistributes some waiting times from station 2 to station 1, intuitively it should be most effective when station 2 is the system's bottleneck. This intuition is supported by Figure 8. The four curves presented on four panels correspond to t^* (dashed line) and values of t such

Figure 8. Trade-off curves of the TBP corresponding to t^* and excessive wait probabilities of 10%, 5%, and 1% (under the nonidling policy) for different system parameters.



Note. The nonidling policy corresponds to the leftmost point on each curve.

that the probabilities of long waits are 1% (lower solid), 5% (middle solid), and 10% (top solid) under the NI policy. Figure 8 panels (a) and (b) present results for cases where the processing rates of stations 1 and 2 are identical. Figure 8 panels (c) and (d) present cases where the processing rate of station 2 is reduced to 0.95, making it more of a bottleneck. We see similar patterns to those described for the previous figure: the TBP reduces the $PW(t)$ in all cases at the cost of a small increase in $E[S]$; the relative improvement in $PW(t)$ is increasing in t. Moreover, we see that the improvements provided by the TBP are greater when station 2 is more of a bottleneck (Figure 8, (a) versus (c) and (b) versus (d)), even under similar utilization levels but different arrival rates (Figure 8, (b) versus (c)).

Figures 7 and 8 provide some intuitions on identifying t^*s and TH^*s for different parameter settings. We notice that TH^* is relatively stable for similar arrival rates, and that $PW^{NI}(t^*)$ is relatively stable for similar utilization level at station 2. Specifically, comparing Figure 7 with Figure 8, (a) and (c), $TH^* \in [11, 15]$ is stable for the same arrival rate, $\lambda = 0.85$. This is also supported by comparing Figure 8, (b) and (d), where $TH^* \in [16, 23]$. Similarly, $PW^{NI}(t^*)$ is stable under similar utilization levels. For example, when $\rho = 0.95$ (Figure 8(d) and Figure 7), $PW^{NI}(t^*)$ is about 15%, and when $\rho = .9$ (Figure 8, (b) and (c)), $PW^{NI}(t^*)$ is about 12%.

6.2. Insight 4: Comparing the TBP and Kanban Policies

Because the analytical derivation of the waiting time at each station is not available and is beyond the scope of this paper, to compare the performance of the Kanban policy and the TBP, we constructed a simulation model using MATLAB. We simulated one million customers under the Kanban policies with $BS = 100, 99, \ldots, 1$ and the TBPs with $TH = 100, 99, \ldots, 1$. (Despite having analytic results for the TBP, we use simulation so that we compare both policies under the same sample path.) The results are presented in Figure 9 for a system with $\mu_2 = 0.9$ in the left panel and $\mu_2 = 0.95$ in the right. In both cases, the value of t was chosen to correspond to 10% probability of long wait under the NI policy. With $BS = 100$, the Kanban policy performs identically to the NI one, which gives us the starting point on each panel. We then decrease the value of the buffer size BS in steps of 1 and plot the values of $PW(t)$ and $E[S]$ for each BS. We plot the TBP curve in a similar fashion.

First consider Figure 9(a). Although the TBP generally outperforms the Kanban policy (recall that the Pareto-optimal points are the ones on the southwestern frontier), when relative $E[S] \geqslant 101.54$, the Kanban policy outperforms the TBP, achieving lower $PW(t)$ values for the same sojourn times. We note that selecting the right BS value is very important—values that are too high or too low may lead to performance worse than that of the NI policy. In fact, our numerical experiments show that the BS^* that minimizes $PW(t)$ appears to be very sensitive to t, whereas the TBP is much more robust in this respect (see Table 1). This lack of robustness presents a challenge for implementing Kanban policies, because the exact value of t may differ among customers.

Now consider Figure 9(b), where $\mu_2 = 0.95$. Here the TBP clearly dominates the Kanban policy (which produces very few Pareto-optimal values). The intuition behind the poor performance of the Kanban policy in this case is that the Kanban policy ignores the queue size in front of station 1. Although this is not a major issue when station 2 is the main bottleneck in the system (as in the left panel), when the processing rates of stations 1 and 2 are similar (as in the right panel) and station 1 is idled even when facing a long queue, long wait times occur. Thus, whereas the Kanban policy performed very well for the asymptotic $\mu_1 = \infty$ case, the performance under more realistic

Figure 9. Trade-off curves corresponding to $t = 31.78$ under the TBP and the Kanban policy.

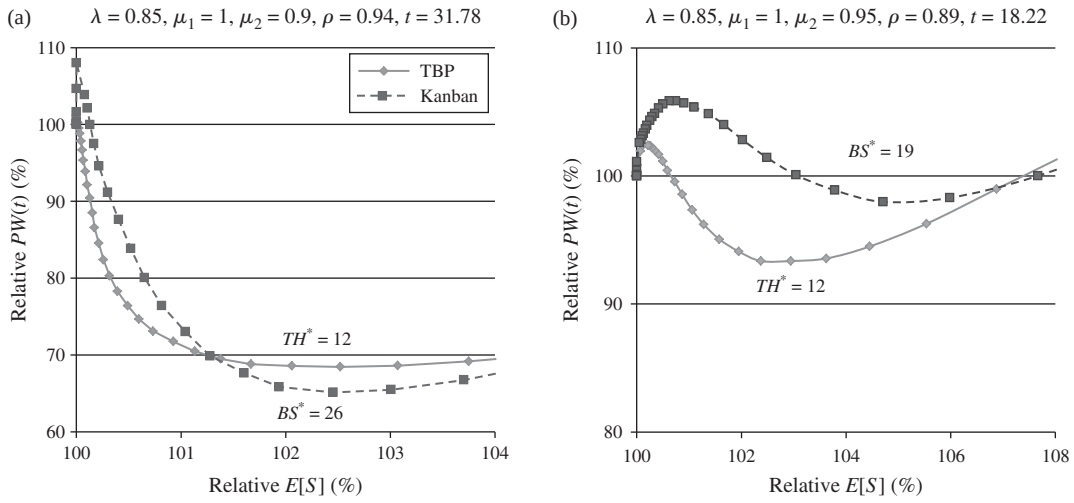


Table 1. Performance of TBP and Kanban policies for different values of t.

t	TH^*	$PW^{TBP(TH^*)}(t)$	$E^{TBP(TH^*)}[S]$	BS^*	$PW^{\text{Kanban}(BS^*)}(t)$	$E^{\text{Kanban}(BS^*)}[S]$	$PW^{NI}(t)$	$E^{NI}[S]$
15	100	0.2663	26.20	100	0.2664	26.20	0.2662	26.20
20	100	0.1930	26.20	100	0.1931	26.20	0.1930	26.20
25	12	0.1329	26.86	22	0.1243	27.76	0.1437	26.20
30	12	0.0810	26.86	25	0.0764	26.99	0.1082	26.20
35	12	0.0485	26.86	27	0.0470	26.71	0.0828	26.20
40	13	0.0284	26.74	31	0.0293	26.41	0.0630	26.20
45	12	0.0158	26.86	34	0.0180	26.30	0.0482	26.20
50	11	0.0079	27.00	37	0.0109	26.24	0.0371	26.20

conditions appears to be significantly worse. The additional flexibility afforded by the TBP, which takes both q_1 and q_2 into account, is apparently important in the case of a more balanced system.

7. Summary and Open Questions

In this paper, we studied strategic idling—i.e., purposefully idling some upstream stations when the downstream stations become too busy—in a two-station tandem queue network. The purpose of SI is to reduce the incidence of excessive waits and thus improve customer service experience in queueing networks. Numerical results indicate that TBP can be quite effective in reducing the incidence of excessive waits, without significantly increasing system sojourn times. Thus, the TBP makes it possible to improve the service experience of customers without adding any capacity to the system (by, instead, idling some of the existing capacity). A comparison with Kanban policies indicates that the TBP is more efficient.

We demonstrated that these insights hold in more general settings. Specifically, in Section EC.3 of the e-companion, we present a simple example that illustrates possible TBPs and Kanban policies for a three-station serial queueing network with exponential service time at each station and Poisson arrivals; and in our working paper, Baron et al. (2014), we consider an open-shop queueing network that does not reach steady state. Both studies used simulation. The results indicate that the managerial insights listed earlier for the two-station system likely hold in other more general settings as well. A generalization of the TBP to n-station tandem queue system is presented in the e-companion, Section EC.4.

Clearly, this paper undertakes only an initial study of the TBPs and SI, and much work remains to be done. It would be interesting to inspect the effect of the TBP in an emergency department setting and compare the result with that of Saghafian et al. (2012). It would be very beneficial to extend our analytical results to more general settings (n-station networks, nonstationary arrival rate, general service time, etc.), though this appears to be quite difficult. In particular, the structure of the optimal TBPs (i.e., the specification of δ functions and the TH values) needs to be investigated.

There are several other possible directions for future research. An analysis of waiting time distributions under either of the control policies developed for manufacturing settings is an obvious one. It would also be interesting to further investigate the application of the TBP and other policies with SI in additional settings such as open-shop queueing networks. Also, the trade-offs between other service-level measures can be explored. Finally, in practice there is value to adequately defining excessive wait and acceptable average sojourn times. Both measures should be related to customers patience and may be evaluated using customer surveys.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/opre.2013.1236.

Acknowledgments

This research was supported by NSERC grants to the first three authors. The authors are grateful to the associate editor and the anonymous referees for their invaluable comments that significantly improved this paper.

References

Afèche P (2013) Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manufacturing Service Oper. Management* 15(3):423–443.

Baron O, Milner J (2009) Staffing to maximize profit for call centers with alternate service level agreements. *Oper. Res.* 57(3):685–700.

Baron O, Berman O, Krass D (2008) Facility location with stochastic demand and constraints on waiting time. *Manufacturing Service Oper. Management* 10(3):484–505.

Baron O, Berman O, Krass D, Wang J (2014) Dynamic scheduling and strategic idling in an open-shop queueing network: Case study and analysis. Working paper, Rotman School of Management, Toronto, Ontario, Canada.

Bertsimas D, Mourtzinou G (1996) A unified method to analyze overtake free queueing systems. *Adv. Appl. Probab.* 28(2):588–625.

Bertsimas D, Nakazato D (1995) The distributional Little's law and its applications. *Oper. Res.* 43(2):298–310.

Burke P (1956) The output of a queueing system. *Oper. Res.* 4(6): 699–704.

Caldentey R, Wein L (2006) Revenue management of a make-to-stock queue. *Oper. Res.* 54(5):859–875.

Chen H, Yao DD (2001) *Fundamentals of Queueing Networks, Performance, Asymptotics, and Optimization* (Springer, New York).

Conway R, Maxwell W, McClain JO, Thomas LJ (1988) The role of work-in-process inventory in serial production lines. *Oper. Res.* 36(2): 229–241.

de Véricourt F, Jennings O (2011) Review on nurse staffing in medical units: A queueing perspective. *Oper. Res.* 59(6):1320–1331.

de Véricourt F, Zhou YP (2005) A routing problem for call centers with customer callbacks after service failure. *Oper. Res.* 53(6):968–981.

Friedman HH, Friedman LW (1997) Reducing the "wait" in waiting-line systems: Waiting line segmentation. *Bus. Horizons* 40(4):54–58.

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

Ha A (1997a) Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Sci.* 43(8):1093–1103.

Ha A (1997b) Stock-rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Res. Logist.* 44(5):457–472.

Larson RC (1987) Perspectives on queues: Social justice and the psychology of queueing. *Oper. Res.* 35(6):895–905.

Masin M, Herer Y, Dar-El EM (2005) Design of self-regulating production control systems by tradeoffs programming. *ITE Trans.* 37(3): 217–232.

Mehrotra V, Ross K, Ryder G, Zhou YP (2012) Routing to manage resolution and waiting time in call centers with heterogeneous servers. *Manufacturing Service Oper. Management* 14(1):66–81.

Milch PR, Waggoner MH (1970) A random walk approach to a shutdown queueing system. *Appl. Math.* 19(1):103–115.

Ross SM (2000) *Introduction to Probability Models*, 7th ed. (Academic Press).

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2012) Patient streaming as a mechanism for improving responsiveness in emergency departments. *Oper. Res.* 60(5):1080–1097.

Solberg JJ (1977) A mathematical model of computerized manufacturing systems. *4th Internat. Conf. Production Res., Tokyo*, 1265–1275.

Soman D, Shi M (2003) Virtual progress: The effect of path characteristics on perceptions of progress and choice. *Management Sci.* 49(9): 1229–1250.

Spearman M, Woodruff DL, Hopp WJ (1990) CONWIP: A pull alternative to Kanban. *Internat. J. Production Res.* 28(5):879–894.

Sugimori Y, Kusunoki K, Cho F, Uchikawa S (1977) Toyota production system and Kanban system materialization of just-in-time and respect-for-human system. *Internat. J. Production Res.* 15(6): 553–564.

Taylor S (1994) Waiting for service: The relationship between delays and evaluation of service. *J. Marketing* 58(2):56–69.

van Ryzin G, Lou SXC, Gershwin SB (1993) Production control for a tandem two-machine system. *IIE Trans.* 25(5):5–20.

Veatch MH, Wein LM (1994) Optimal control of a two-station tandem production/inventory system. *Oper. Res.* 42(2):337–350.

Weber RR, Stidham S Jr (1987) Optimal control of service rates in networks of queues. *Adv. Appl. Probab.* 19(1):202–218.

Opher Baron is an associate professor of operations management at the Joseph L. Rotman School of Management, University of Toronto. His research interests include applied probability and its application to facility location, service operations (such as healthcare), inventory planning, and revenue management.

Oded Berman is the endowed Sydney Cooper Chair in Business and Technology and a former associate dean of programs at the Joseph L. Rotman School of Management, the University of Toronto. He is a Fellow of the Institute for Operations Research and the Management Sciences, a cowinner of the Lifetime Achievement Award of the INFORMS Section on Location Analysis, and the 2012 recipient of the Canadian Operational Research Society Award of Merit. His main research interests include operations management in the service industry, location theory, network models, and stochastic inventory control.

Dmitry Krass is a professor of operations management and statistics at the Joseph L. Rotman School of Management, University of Toronto. His research interests include service operations, logistics, facility location and management issues, distributive service systems, and analytical modeling in marketing.

Jianfu Wang is an assistant professor of operations management at the Nanyang Business School, Nanyang Technological University, Singapore. His research interests include queueing theory and its application in service operations, healthcare operations, and revenue management.