HEAVY TRAFFIC ANALYSIS OF POLLING SYSTEMS IN TANDEM

MARTIN I. REIMAN

Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974

LAWRENCE M. WEIN

Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142-1347, lwein@mit.edu

(Received September 1995; revisions received June 1996, July 1997, September 1997; accepted November 1997)

We analyze the performance of a tandem queueing network populated by two customer types. The interarrival times of each type and the service times of each type at each station are independent random variables with general distributions, but the load on each station is assumed to be identical. A setup time is incurred when a server switches from one customer type to the other, and each server employs an exhaustive polling scheme. We conjecture that a time scale decomposition, which is known to occur at the first station under heavy traffic conditions, holds for the entire tandem system, and we employ heavy traffic approximations to compute the sojourn time distribution for a customer that arrives to find the network in a particular state. When setup times are zero (except perhaps at the first station) and additional "product-form" type assumptions are imposed, we find the steady-state sojourn time distribution for each customer type.

We consider a tandem queueing system with K single-server stations. Customers of two types, denoted by A and B, arrive to station 1 according to independent renewal processes. Each customer is served once at each station, and customers exit after being served at station K. Although each customer type has its own general service time distribution at each station, we require the means of these distributions to be identical. A setup (or switchover) time is incurred at each station whenever a server switches from serving one customer type to the other. We assume that each server employs an exhaustive polling scheme: Serve all customers of the type that is currently set up; when there are no more of these customers in queue, switch to the other customer type and serve all its customers exhaustively.

Our goal is to derive the sojourn time distribution for each customer type in both the *transient* and *steady-state* cases. In the transient case, we attempt to find the sojourn time of a customer who arrives to the network and finds it in a particular state, which is taken to be the the $2K$-dimensional queue length process and the *location* (i.e., the customer type that is currently set up) of each server. Hence, in contrast to steady-state performance analysis, where unconditioned or conditioned sojourn times can be considered, the transient sojourn time distribution we compute is conditioned on the state, and hence is state-dependent.

The motivation for studying this problem comes from manufacturing systems. Although queueing network models for manufacturing typically do not include setup times, many factories incur significant setup times and/or setup costs (the recent trend is to reduce setup times at the expense of large material, equipment, and/or labor costs) when a machine switches from producing one type of product to another. To exploit these economies of scale, manufacturers are forced to produce their products in large batches, and the exhaustive policy considered here is the most natural way to reduce setups without incurring unnecessary idleness.

In manufacturing settings, sojourn time (also called manufacturing cycle time, lead time, or throughput time) distributions help managers to release and schedule work, quote delivery dates for customers, coordinate with downstream operations (such as distribution), price their products, and set performance metrics. Steady-state sojourn times are helpful for tactical and strategic level decisions and are the best available estimates for factories that cannot generate real time queue length information. For systems with real time information capability, transient sojourn times are preferable to steady-state estimates for operational decisions, and they can greatly enhance performance (see Wein 1991 for a simple example in due-date quotation).

Our queueing network model is essentially a set of polling systems in tandem. Although the performance analysis of polling systems has generated an enormous literature, there are no studies that consider a network of interacting polling systems. Karmarkar et al. (1985) develop a fixed batch size queueing network approximation that is useful for tactical level decisions, but does not capture the detailed system dynamics of the network. Because our queueing network model appears to defy exact analysis, we employ heavy traffic approximations to address the problem. Recently, Coffman et al. (1995, 1998), henceforth referred to as CPR I and CPR II, proved an averaging principle for single-server polling systems with and without switchover times. This principle is a result of a time scale decomposition that arises under the traditional heavy traffic normalizations: On the time scale giving rise to a

Subject classifications: Queues; heavy traffic approximations. Inventory/production: Multi-item, multi-stage systems with lot-sizing.
Area of review: Stochastic Models.

0030-364X/99/4704-0524 $05.00

diffusion process for the total workload, the individual workloads (for each class) move (asymptotically) infinitely fast. When viewed on the time scale that makes the rate of movement of the individual workloads positive and finite, the total workload remains constant, and the individual workloads move deterministically (i.e., behave as a fluid). Consequently, the dynamics of the individual workloads can be analyzed deterministically.

In this paper, we assume that the time scale decomposition uncovered by CPR I and CPR II holds for all stations in the tandem system, not just for station 1; heuristic arguments supporting this conjecture are given in §2. The primary contribution of this paper is the deterministic analysis of the individual fluid workloads for a tandem queueing system. For the system described at the beginning of this paper, our deterministic analysis yields a recursive procedure for the transient sojourn time distribution of each customer type at each station in isolation, as a function of the K-dimensional station workload process. To perform a steady-state sojourn time analysis of each customer type at each station in isolation, the stationary distribution of the vector describing the total workload at each station is required. We compute the steady-state sojourn time distribution by making the additional assumptions that setup times are zero and that the total workload vector has a product-form distribution (product-form conditions for an approximating Brownian network are slightly less restrictive than the corresponding conditions for a traditional queueing network). In addition, we derive a product-form solution in heavy traffic for the case where only the first station has setup times. We also briefly describe how our results can be used to compute a customer's transient sojourn time in the network conditioned on the entire $3K$-dimensional system state and the steady-state sojourn time in the network for each customer type.

The remainder of this paper is organized as follows. The model is formulated in §1, and the heavy traffic normalizations are introduced in §2. Section 3 contains the deterministic analysis that forms the basis of this paper. The deterministic results are used in §4 to obtain the transient and steady-state sojourn time distributions for the original queueing network. For several simple examples, we also compare our steady-state sojourn time estimates to those computed via simulation. In §5, we perform a steady-state analysis for the case where setup times are incurred at station 1. Concluding remarks are offered in §6.

1. THE MODEL

The queueing network has K single-server stations and two customer types, A and B. Customers of each type arrive to station 1 according to independent renewal processes with rates λ_A and λ_B, proceed through the tandem network from station 1 to station K, and exit after service is completed at station K. Following traditional terminology, we define a different class of customer for each stage of each type's route; these classes are denoted by Ai and Bi for $i = 1 \ldots, K$. Each class may have a different general service time distribution, but all service rates equal μ. Let $\rho_j = \lambda_j/\mu$ for $j = A, B$, and define the traffic intensity of this balanced system by $\rho = \rho_A + \rho_B$. Finally, let c_{j0}^2 denote the squared coefficient of variation of the interarrival times for type j and let c_{ji}^2 represent the squared coefficient of variation of the service times for class ji.

The server at each station serves each class to exhaustion, and then switches to the other class. We assume that a random setup time is incurred when a server switches from one class to the other. However, because of the scaling of time and the rarity of switchovers in heavy traffic, our transient sojourn time estimates are independent of the setup time; consequently, we do not introduce any setup time notation. Heavy traffic models possess an insensitivity property: In sufficiently heavy traffic, some fine details about the queueing system (such as the underlying probability distributions beyond the first two moments, the distinction between nonpreemptive and preemptive-resume in a priority queueing system, and the variance of setup times in steady-state analysis) are unimportant and do not appear in the results, and in this paper we have another example of this property. Nonetheless, it is important to keep in mind that setup times (and/or setup costs), while not appearing in the final results, *cause* the system to adopt an exhaustive polling scheme, and the results under an exhaustive polling scheme are fundamentally different than under a FCFS scheme; hence, the heavy traffic analysis captures the essence of system behavior of queueing networks with setup times.

2. HEAVY TRAFFIC PRELIMINARIES

In a typical heavy traffic analysis, one defines a sequence of queueing systems indexed by n that approaches heavy traffic as $n \to \infty$. Because a heavy traffic limit theorem will not be proved, we avoid unnecessary notation by considering a single large integer n satisfying $\sqrt{n}(\rho - 1) = c$, where c is negative and of moderate size. This condition requires each server to be busy the great majority of the time to satisfy demand. In §4, we see that our sojourn time estimates are independent of the system parameter n.

For class ji, $j = A, B$; $i = 1, \ldots, K$, let $\{L_{ji}(t), t \geq 0\}$ denote the *workload* process and $\{W_{ji}(t), t \geq 0\}$ be the *virtual waiting time* process. The quantity $L_{ji}(t)$ denotes the remaining work for the server at station i embodied in class ji customers at time t and is commonly referred to as the unfinished workload, and $W_{ji}(t)$ is the waiting time experienced at station i by a type j customer arriving to the network at time t. Using the standard heavy traffic scaling, we define the normalized processes $V_{ji}(t) = L_{ji}(nt)/\sqrt{n}$ and $Z_{ji}(t) = W_{ji}(nt)/\sqrt{n}$. As is typical in heavy traffic systems, if we let $\tilde{W}_{ji}(t)$ denote the actual waiting time at station i of the first type j customer to arrive to the system after time t, then $\tilde{W}_{ji}(nt)/\sqrt{n}$ converges together with the

normalized virtual waiting time process $W_{ji}(nt)/\sqrt{n}$. Finally, let $V_i(t) = V_{Ai}(t) + V_{Bi}(t)$ denote the normalized workload at station i at time t.

The first step in analyzing the processes $V_{ji}(t)$ is to analyze the processes $V_i(t)$. In the case where the setup times are zero and the service time distributions depend only on the station, the heavy traffic limit theorem in Reiman (1984) can be applied to obtain the limiting vector workload process $\{(V_1(t), \ldots, V_K(t)), t \geq 0\}$. This limit process is a reflected Brownian motion in the K-dimensional nonnegative orthant. The result of Reiman (1984), which is for a FCFS network, applies to any nonpreemptive work conserving discipline where the service time distribution depends only on the station, because the queue length processes (and hence workload processes) are equal in distribution.

With service time distributions that depend on customer class, no setup times, and FCFS discipline at each station, the heavy traffic limit theorem in Peterson (1991) also yields a reflected Brownian motion in the K-dimensional nonnegative orthant as the limiting vector workload process. The vector workload process under exhaustive polling may not be the same as with the FCFS discipline, so Peterson's results do not directly apply. However, it seems clear from heavy traffic scaling arguments that the heavy traffic limit of the vector total workload process (with zero setup times) is the same under exhaustive polling as in FCFS, as long as the mean service time depends only on the station. The heavy traffic scaling argument is as follows. The drift of the limit diffusion process depends only on arrival rates and mean service times, so it is the same in both cases. The variance of the limit diffusion process depends on variations in the centered and normalized processes over $O(n)$ time; on this scale, based on results in CPR I and CPR II, the fraction of each type served is the same as in FCFS, so the variances should match. We use this conjectured "extension" to Peterson's results in our analysis. When the mean service times at a station are different for the two types, the vector workload process is different for exhaustive polling and FCFS, as described in §5.1.

As mentioned in the introduction, the individual workload components $V_{ji}(t)$ move infinitely quickly in the limit, so no "standard" limit theorem is available for the $2K$-dimensional workload process $\{(V_{A1}(t), V_{B1}(t), \ldots, V_{AK}(t), V_{BK}(t), t \geq 0\}$. This difficulty is circumvented in CPR I and CPR II by considering smoothed versions of the individual workloads, where the smoothing is obtained by integrating over a time interval. The averaging principles of CPR I and CPR II imply that, in heavy traffic, for any bounded continuous function f and any $T > 0$,

$$\int_0^T f(V_{j1}(t))\, dt \text{ is well approximated by}$$

$$\int_0^T \left(\int_0^1 f(uV_1(t))\, du \right) dt, \quad j = A, B. \tag{1}$$

This result reveals a time scale decomposition of the total workload V_1 and the individual workloads V_{A1} and V_{B1}. The one-dimensional total workload V_1 varies as a diffusion process (a Bessel process if setup times are positive and a reflected Brownian motion if setup times are zero), whereas the two-dimensional process (V_{A1}, V_{B1}) moves infinitely quickly in the heavy traffic limit. If time is slowed down by a factor of $\sqrt{n}$ (we are considering $\bar{V}_{j1}(t) = V_{j1}(t/\sqrt{n}) = L_{j1}(\sqrt{n}t)/\sqrt{n}$), so that the two-dimensional workload moves at a finite and positive rate, then the total workload $\bar{V}_1$ remains constant and the movement of the two-dimensional workload process $(\bar{V}_{A1}, \bar{V}_{B1})$ is deterministic. Moreover, the setup times do not affect the deterministic movement of the normalized two-dimensional process. (It may be a bit confusing to see the claim that $\bar{V}_1(t) \equiv V_1(t/\sqrt{n})$ is constant while $V_1(t)$ is not, since $\bar{V}_1(\sqrt{n}) = V_1(1)$. The resolution is that t is meant to be $O(1)$: weak convergence results hold for $t \in [0, T]$, where T is fixed.) CPR I and CPR II also use (1) to derive an averaging principle for virtual waiting times in a single-server polling system, which implies that,

$$\int_0^T f(Z_{j1}(t))\, dt \text{ is well approximated by}$$

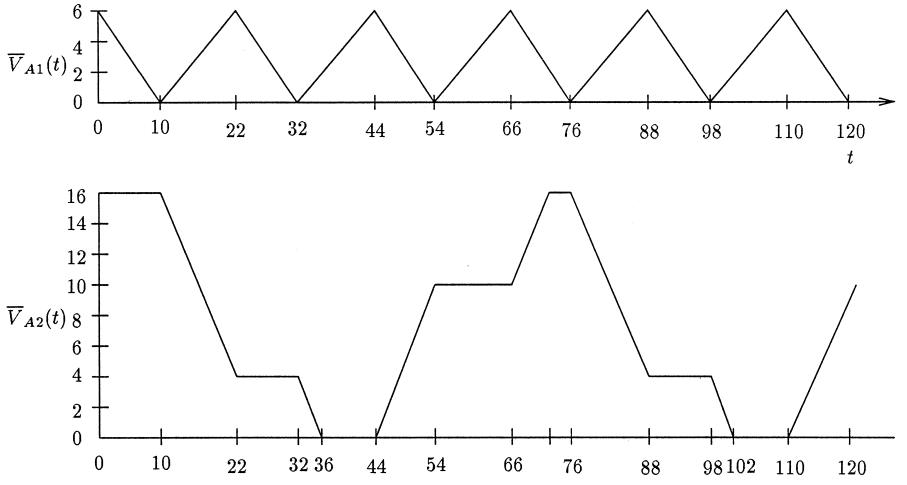
$$\int_0^T \left(\int_0^1 f\left(\frac{uV_1(t)}{\rho_j} \right) du \right) dt, \quad j = A, B. \tag{2}$$

In this paper, we assume that the time scale decomposition holds not just for station 1, but for each station $i = 1, \ldots, K$ in the tandem network. We also assume that properly modified versions of the averaging principles (1) and (2) hold for these downstream stations. We do not provide any proofs here, only a heuristic argument.

Examining the proof of (1) in CPR I and CPR II, we see that it follows roughly the intuition given under (1). Time is divided into intervals during which the total workload does not move much. Then over each of these intervals the individual workload processes are examined on the $\sqrt{n}$ time scale. A fluid-type (strong law of large numbers) analysis shows that the individual workloads move in a simple, deterministic, periodic manner: The individual workloads grow linearly from zero to the total workload, and then linearly shrink back to zero. This fluid limit yields the inner integral on the right hand side of (1) over each small interval; piecing the intervals together (actually piecing together upper and lower bounds) yields the outer integral.

The above method can be extended roughly as follows. We divide time into intervals over which the total vector workload does not move much. Over each of the intervals the individual workloads can be examined on the $\sqrt{n}$ time scale. Because of the tandem structure, this analysis can be carried out in a sequential manner, using the results from the analysis of station 1 to analyze station 2, etc. The structure of the limiting fluid that arises from this analysis is the subject of §3. This analysis shows that the fluid limits for stations $2, \ldots, K$ have a periodic structure, but it is not

Figure 1. The workload process dynamics for an example: $\lambda_A = 0.4$, $\lambda_B = 0.5$, $\mu = 1$, $\bar{V}_{A1}(0) = 6$, $\bar{V}_{B1}(0) = 0$, $\bar{V}_{A2}(0) = 16$, $\bar{V}_{B2}(0) = 0$, and $s_1(0) = s_2(0) = A$.



linear as it is for station 1. As a consequence of the more complicated movement, which includes having the individual workload stay constant for certain lengths of time (see Figure 1, bottom graph), the associated workload averaging principle is more involved than for station 1; because our interest lies more in sojourn times, we do not display it here. As shown in §3, this more complex structure is also mirrored in the fluid-level virtual waiting times, which take on a discrete distribution.

Our analysis also uses the *snapshot principle*, which we assume to hold for our system. This principle, which was first observed by Foschini (1980) and proved by Reiman (1984) in the context of a single-class queueing network, asserts that in the limiting heavy traffic time scale, the vector of total workloads $(V_1, \ldots, V_K)$ does not change during a customer's sojourn in the network. This is clearly a consequence of the time scale decomposition assumed above, and so does not represent another assumption. With $O(\sqrt{n})$ customers in the system, customers spend $O(\sqrt{n})$ time in the system. As mentioned above, the vector total workload process does not change over time intervals of this magnitude.

Under these assumptions, we undertake in the next section a deterministic analysis of the $2K$-dimensional normalized workload process.

3. DETERMINISTIC ANALYSIS

Throughout this section, we slow down time so that the total workload vector $(\bar{V}_1(t), \ldots, \bar{V}_K(t)) = (V_1, \ldots, V_K)$ is fixed, and $\{\bar{V}_{ji}(t), t \geq 0\}$, $j = A, B$; $i = 1, \ldots, K$ moves at a finite rate in a deterministic fashion. Because it is extremely difficult to perform a transient analysis conditioned on the complete $3K$-dimensional state of the system at the time of a customer's arrival, our goal in this section is to derive the distribution of the normalized virtual waiting time Z_{ji}, $j = A, B$; $i = 1, \ldots, K$, given the constant workload vector $(\bar{V}_1(t), \ldots, \bar{V}_K(t)) = (V_1, \ldots, V_K)$. That is, we consider a customer that "randomly" arrives while the station 1 workload $\bar{V}_1(t)$ equals V, and suppress the finer state information. In Section 4, we describe how one can use our analysis to compute the virtual waiting time conditioned on the $3K$-dimensional system state. If we denote the arrival time of a randomly arriving customer as time 0, then the initial conditions are random, and consist of $\bar{V}_{ji}(0)$, $j = A, B$; $i = 1, \ldots, K$ and $s_i(0)$, $i = 1, \ldots, K$, where $\bar{V}_{Ai}(0) + \bar{V}_{Bi}(0) = V_i$ and $s_i(t)$ is the customer type being served by server i at time t. Hence, the deterministic path of $\{\bar{V}_{ji}(t), t \geq 0\}$ is dictated by $\bar{V}_{Al}(0)$, $s_l(0)$, $l = 1, \ldots, i$; as we will see shortly, it suffices to focus on $\bar{V}_{Ai}(t)$, $i = 1, \ldots, K$, because $\bar{V}_{Ai}(t) + \bar{V}_{Bi}(t)$ is equal to V_i for all t.

The derivation of our main result consists of three main steps: First, we show that, independent of initial conditions, the process $\{\bar{V}_{Ai}(t), t \geq 0\}$ enters a unique *limit cycle* within a finite amount of time; that is, the trajectory of the process keeps repeating the same cycle. Second, we identify the limit cycle for station i. The cycle consists of the service of class Ai customers for C_{Ai} time units followed by the service of class Bi customers for C_{Bi} time units, and the cycle repeats itself every $C_i = C_{Ai} + C_{Bi}$ time units; we refer to C_i as the *cycle length* for station i, and refer to C_{ji} as the cycle length for class ji. Finally, we derive the normalized virtual waiting time from the limit cycle.

Before stating the main result, we illustrate our method on the single-server polling model in CPR I and CPR II, and then informally describe the dynamics of $\bar{V}_{ji}(t)$, $j = A, B$; $i = 1, \ldots, K$ for a fixed total workload vector $(V_1, \ldots, V_K)$. Since arrivals to station 1 are exogenous, class $A1$ work arrives to station 1 at rate ρ_A, regardless of the system state. The server depletes work at rate one, and hence $\bar{V}_{A1}(t)$ decreases at rate $1 - \rho_A$ when class $A1$ is being served, and increases at rate ρ_A otherwise. Because $\rho = 1$ in the heavy traffic limit, we assume that $\bar{V}_{A1}(t)$ increases at rate $1 - \rho_B$, rather than ρ_A, when class $B1$ is being

served. Because setup times do not appear at this time scale in the heavy traffic limit, $\{\bar{V}_{A1}(t), t \geq 0\}$ follows the familiar "saw-tooth" path that arises in the economic production quantity model; see the top graph in Figure 1. Hence, regardless of the initial conditions $(\bar{V}_{A1}(0), s_1(0))$, the process $\{\bar{V}_{A1}(t), t \geq 0\}$ essentially enters a limit cycle at time zero. However, for convenience and without loss of generality, we specify the start of a cycle as the moment that $\bar{V}_{A1}(t)$ increases from zero. If $\bar{t}_1$ denotes the starting time of the first cycle at station 1 ($\bar{t}_1 = 10$ in Figure 1), then

$$\bar{V}_{A1}(t) = (1 - \rho_B)(t - \bar{t}_1) \quad \text{for } t \in \left[\bar{t}_1, \bar{t}_1 + \frac{V_1}{1 - \rho_B}\right], \tag{3}$$

and

$$\bar{V}_{A1}(t) = V_1 - (1 - \rho_A)\left(t - \left(\bar{t}_1 + \frac{V_1}{1 - \rho_B}\right)\right)$$
$$\text{for } t \in \left[\bar{t}_1 + \frac{V_1}{1 - \rho_B}, \bar{t}_1 + \frac{V_1}{1 - \rho_B} + \frac{V_1}{1 - \rho_A}\right]. \tag{4}$$

Class $B1$ customers are served during the interval described in (3) and class $A1$ customers are served in (4). It follows that

$$C_{A1} = \frac{V_1}{1 - \rho_A}, \quad C_{B1} = \frac{V_1}{1 - \rho_B} \text{ and}$$
$$C_1 = \frac{V_1}{1 - \rho_A} + \frac{V_1}{1 - \rho_B}. \tag{5}$$

Therefore, $\bar{V}_{A1}(t) = 0$ for $t = \bar{t}_1 + C_1 = 32$ and the cycle in (3) and (4) repeats itself at this time.

For station 1, we can use the limit cycle (3)–(4) to derive the transient normalized virtual waiting time Z_{A1} conditioned on the system state at station 1: $\bar{V}_{A1}(t)$, $\bar{V}_{B1}(t)$ and $s_1(t)$ (for reasons stated earlier, the corresponding quantity for downstream stations will not be derived). If a type A customer arrives at time $t \in [\bar{t}_1 + C_{B1}, \bar{t}_1 + C_1]$, then class $A1$ customers are being served, and Z_{A1} equals the class $A1$ workload at time t, or $V_1 - (1 - \rho_A)(t - \bar{t}_1 - C_{B1})$. A type A customer arriving at time $t \in [\bar{t}_1, \bar{t}_1 + C_{B1}]$ must wait $\bar{t}_1 + C_{B1} - t$ time units until all class $B1$ customers are served, and then wait an additional $\bar{V}_{A1}(t) = (1 - \rho_B)(t - \bar{t}_1)$ time units for the class $A1$ customers in front of him to be served.

We can also derive the normalized virtual waiting time Z_{A1} conditioned on V_1. As a consequence of the deterministic cycles ocurring infinitely quickly in the limit, a randomly arriving customer will arrive at a point in time that is uniformly distributed throughout the limit cycle (that is, uniformly distributed on $[\bar{t}_1, \bar{t}_1 + C_1]$). Hence, if we let U denote a uniform random variable on [0, 1] and again assume that $\rho_A + \rho_B = 1$, then

$$Z_{A1} = \begin{cases} UV_1 & \text{with probability } \rho_A, \\ UV_1 + \dfrac{(1 - U)V_1}{1 - \rho_B} & \text{with probability } \rho_B \end{cases}, \tag{6}$$

and hence

$$Z_{A1} \text{ is uniformly distributed on } \left[0, \frac{V_1}{\rho_A}\right]. \tag{7}$$

This result can be derived directly from the averaging principle of CPR I and CPR II for virtual waiting times by setting $V_1(t) = V_1$ for all t and letting $f(x) = I_{\{x \leq z\}}$ in (2); a similar analysis implies that

$$Z_{B1} \text{ is uniformly distributed on } \left[0, \frac{V_1}{\rho_B}\right]. \tag{8}$$

Notice that $E[Z_{j1}] = V_1$ if $\rho_A = \rho_B$. To develop an approximation scheme that is applicable when $\rho < 1$, we propose the following heuristic modification to (7)–(8):

$$Z_{j1} \text{ is uniformly distributed on } \left[0, \frac{V_{1\rho}}{\rho_j}\right] \quad \text{for } j = A, B. \tag{9}$$

This modification is chosen so that $E[Z_{j1}] = V_1$ if $\rho_A = \rho_B$ for all values of ρ. Because we assume perfect balance throughout the system (synchronous cycles are required to derive Theorem 1 below), similar refinements are not necessary for downstream stations.

Stations $2, \ldots, K$ behave in a fundamentally different way than station 1, and consequently the averaging principle for virtual waiting times (2) does not hold for the downstream stations. In particular, rather than receiving *steady* streams of *both* types of customers, downstream stations receive *alternating* streams of type A and type B customers. As mentioned earlier, the actual timing of the arrival streams of classes $A, i + 1$ and $B, i + 1$ to station $i + 1$ is dictated by the dynamic location (i.e., the class that is currently set up) of the server at station i.

Let us consider the behavior of $\{\bar{V}_{A2}(t), t \geq 0\}$, which is influenced by the locations of servers 1 and 2. If $s_1(t) = A$ then class $A2$'s queue is receiving work from station 1 at rate one, and if $s_2(t) = A$ then class $A2$'s work is being depleted at rate 1. Hence, if $s_1(t) = s_2(t) = A$ or $s_1(t) = s_2(t) = B$, then $\bar{V}_{A2}(t)$ remains constant; if $s_1(t) = A$ and $s_2(t) = B$ then $\bar{V}_{A2}(t)$ increases at rate one, and if $s_1(t) = B$ and $s_2(t) = A$ then $\bar{V}_{A2}(t)$ decreases at rate one. Moreover, $\bar{V}_{A2}(t) + \bar{V}_{B2}(t) = V_2$ for all $t \geq 0$.

We are now ready to state and prove our main result. Let us define

$$n_{Bi} = \left\lceil \frac{\bar{V}_{i+1}}{C_{Ai}} \right\rceil, \; n_{Ai} = \left\lceil \frac{\bar{V}_{i+1}}{C_{Bi}} \right\rceil; \text{ and } n_i = n_{Ai} + n_{Bi}$$
$$\text{for } i = 1, \ldots, K - 1. \tag{10}$$

THEOREM 1. *Fix* $(V_1, \ldots, V_N)$. *Then for* $i = 1, \ldots, K - 1$,

$$C_{A,i+1} = (n_i - 1)C_{Ai}, \quad C_{B,i+1} = (n_i - 1)C_{Bi}, \text{ and}$$
$$C_{i+1} = (n_i - 1)C_i, \tag{11}$$
$$Z_{A,i+1} = \begin{cases} V_{i+1} + jC_{Bi} \text{ with probability} \\ \quad \dfrac{1}{n_{i-1}} \text{ for } j = 0, \ldots, n_{Bi} - 1, \\ V_{i+1} - jC_{Bi} \text{ with probability} \\ \quad \dfrac{1}{n_{i-1}} \text{ for } j = 1, \ldots, n_{Ai} - 1, \end{cases} \tag{12}$$

and

$$Z_{B,i+1} = \begin{cases} V_{i+1} + jC_{Ai}, \textit{ with probability} \\ \quad \dfrac{1}{n_{i-1}} \text{ for } j = 0, \ldots, n_{Ai} - 1, \\ V_{i+1} - jC_{Ai}, \textit{ with probability} \\ \quad \dfrac{1}{n_{i-1}} \text{ for } j = 1, \ldots, n_{Bi} - 1. \end{cases} \quad (13)$$

PROOF. The proof is by induction on i. Let us first focus on deriving the workload (unfinished work) at station 2. Our first step is to show that the process $\{\bar{V}_{A2}(t), t \geq 0\}$ enters a limit cycle within a finite amount of time. Without loss of generality, we define the first cycle at station 2 to begin at time $\bar{t}_2 = \inf\{t \geq 0: \bar{V}_{A1}(t) = V_1, \bar{V}_{A2}(t) = 0\}$. At time $\bar{t}_2$, the server at station 1 switches from class $B1$ to class $A1$. Since $\bar{V}_{A2}(\bar{t}_2) = 0$, the server at station 2 will be serving class $B2$ just after time $\bar{t}_2$ (and perhaps just before time $\bar{t}_2$ also); hence, $\bar{V}_{A2}(t)$ increases at rate 1 starting at time $\bar{t}_2$.

Since V_1 and V_2 are fixed, the deterministic process $\{\bar{V}_{A2}(t), t \geq 0\}$ is fully specified by the initial conditions $(\bar{V}_{A1}(0), \bar{V}_{A2}(0), s_1(0), s_2(0))$. We now compute $\bar{t}_2$ for the four cases characterized by the values of $(s_1(0), s_2(0))$. If $s_1(0) = s_2(0) = A$, then $\bar{V}_{A2}(t)$ stays constant for the first $\bar{V}_{A1}(0)/(1 - \rho_A)$ time units. Thereafter, until $\bar{V}_{A2}(t)$ reaches zero, it alternates between decreasing at rate 1 for C_{B1} time units and staying constant for C_{A1} time units, and

$$\bar{t}_2 = \frac{\bar{V}_{A1}(0)}{1 - \rho_A} + \left\lceil \frac{\bar{V}_{A2}(0)}{C_{B1}} \right\rceil C_{B1} + \left(\left\lceil \frac{\bar{V}_{A2}(0)}{C_{B1}} \right\rceil - 1 \right) C_{A1}; \quad (14)$$

readers are referred to the bottom graph in Figure 1, where $\bar{t}_2 = 44$.

If $s_1(0) = s_2(0) = B$, then after remaining constant for $\bar{V}_{B1}(0)/(1 - \rho_B)$ time units, $\bar{V}_{A2}(t)$ rises and stays constant during $\lceil \bar{V}_{B2}(0)/C_{A1} \rceil$ periods of length C_{A1} and $\lceil \bar{V}_{B2}(0)/C_{A1} \rceil - 1$ periods of length C_{B1}, respectively. At this point $\bar{V}_{A2}(t) = V_2$, $s_1(t) = B$, and $s_2(t) = A$; then the process $\bar{V}_{A2}(t)$ alternates between decreasing and constant phases until $\bar{t}_2$, where

$$\bar{t}_2 = \frac{\bar{V}_{B1}(0)}{1 - \rho_B} + \left\lceil \frac{\bar{V}_{B2}(0)}{C_{A1}} \right\rceil C_{A1} + \left(\left\lceil \frac{\bar{V}_{B2}(0)}{C_{A1}} \right\rceil - 1 \right) C_{B1} + \left\lceil \frac{V_2}{C_{B1}} \right\rceil C_{B1} + \left(\left\lceil \frac{V_2}{C_{B1}} \right\rceil - 1 \right) C_{A1}. \quad (15)$$

If $s_1(0) = A$ and $s_2(0) = B$, then $\bar{V}_{A2}(t)$ grows initially. Before reaching $\bar{t}_2$, it must rise by $\bar{V}_{B2}(0)$ and then drop to zero. The length of the first growth period (which starts at time zero) for $\bar{V}_{A2}$ is $\bar{V}_{A1}(0)/(1 - \rho_A) \wedge \bar{V}_{B2}(0)$, and the length of any subsequent growth periods (there are none if $\bar{V}_{B2}(0) < \bar{V}_{A1}(0)/(1 - \rho_A)$) is C_{A1}. Once $\bar{V}_{A2}(t)$ starts decreasing, it proceeds as in the previous case. Hence,

$$\bar{t}_2 = \frac{\bar{V}_{A1}(0)}{1 - \rho_A} + \left\lceil \frac{\bar{V}_{B2}(0) - \dfrac{\bar{V}_{A1}(0)}{1 - \rho_A}}{C_{A1}} \right\rceil C_1 I\left\{ \bar{V}_{B2}(0) > \frac{\bar{V}_{A1}(0)}{1 - \rho_A} \right\} + \left\lceil \frac{V_2}{C_{B1}} \right\rceil C_{B1} + \left(\left\lceil \frac{V_2}{C_{B1}} \right\rceil - 1 \right) C_{A1}, \quad (16)$$

where $I_{\{x\}}$ is the indicator function for the event x. Similarly, if $s_1(0) = B$ and $s_2(0) = A$, then

$$\bar{t}_2 = \frac{\bar{V}_{B1}(0)}{1 - \rho_B} + \left\lceil \frac{\bar{V}_{A2}(0) - \dfrac{\bar{V}_{B1}(0)}{1 - \rho_B}}{C_{B1}} \right\rceil C_1 I\left\{ \frac{\bar{V}_{B1}(0)}{1 - \rho_B} < \bar{V}_{A2}(0) \right\}. \quad (17)$$

Now we look at the evolution of $\bar{V}_{A2}(t)$ during the first full cycle, which starts at time $\bar{t}_2$. Readers are referred to $\bar{V}_{A2}(t)$ for $t \in [44, 110]$ in Figure 1. The process $\bar{V}_{A2}(t)$ initially alternates between $n_{B1} - 1$ (possibly zero) growth periods of length C_{A1} and constant periods (that is, time intervals where $\bar{V}_{A2}(t)$ remains constant) of length C_{B1}, as described by

$$\bar{V}_{A2}(t) = (j - 1)C_{A1} + t - (\bar{t}_2 + (j - 1)C_1) \quad \text{for } t \in [\bar{t}_2 + (j - 1)C_1, \bar{t}_2 + (j - 1)C_1 + C_{A1}], \quad (18)$$

and

$$\bar{V}_{A2}(t) = jC_{A1} \quad \text{for } t \in [\bar{t}_2 + (j - 1)C_1 + C_{A1}, \bar{t}_2 + jC_1] \quad (19)$$

for $j = 1, \ldots, n_{B1} - 1$. The last growth period, which is given by

$$\bar{V}_{A2}(t) = (n_{B1} - 1)C_{A1} + t - (\bar{t}_2 + (n_{B1} - 1)C_1) \quad \text{for } t \in [\bar{t}_2 + (n_{B1} - 1)C_1, \bar{t}_2 + (n_{B1} - 1)C_1 + V_2 - (n_{B1} - 1)C_{A1}], \quad (20)$$

is truncated when $\bar{V}_{A2}(t)$ reaches its maximum value of V_2; at this point, the process remains constant until server 1 exhausts the class $A1$ customers in queue:

$$\bar{V}_{A2}(t) = V_2 \text{ for } t \in [\bar{t}_2 + (n_{B1} - 1)C_1 + V_2 - (n_{B1} - 1)C_{A1}, \bar{t}_2 + (n_{B1} - 1)C_1 + C_{A1}]. \quad (21)$$

The process then alternates between $n_{A1} - 1$ depletion intervals of length C_{B1} and constant intervals of length C_{A1}, given by

$$\bar{V}_{A2}(t) = V_2 - (j - 1)C_{B1} - [t - (\bar{t}_2 + (n_{B1} + j - 2)C_1 + C_{A1})] \quad \text{for } t \in [\bar{t}_2 + (n_{B1} + j - 2)C_1 + C_{A1}, \bar{t}_2 + (n_{B1} + j - 1)C_1], \quad (22)$$

and

$$\bar{V}_{A2}(t) = V_2 - jC_{B1} \text{ for } t \in [\bar{t}_2 + (n_{B1} + j - 1)C_1, \bar{t}_2 + (n_{B1} + j - 1)C_1 + C_{A1}] \quad (23)$$

for $j = 1, \ldots, n_{A1} - 1$. Once again, the last depletion cycle

$$\bar{V}_{A2}(t) = V_2 - (n_{A1} - 1)C_{B1} - [t - (\bar{t}_2 + (n_{A1} + n_{B1} - 2)C_1 + C_{A1})] \quad \text{for } t \in [\bar{t}_2 + (n_{B1} + n_{A1} - 2)C_1 + C_{A1}, \bar{t}_2 + (n_{A1} + n_{B1} - 2)C_1 + C_{A1} + V_2 - (n_{A1} - 1)C_{B1}] \quad (24)$$

is truncated when $\bar{V}_{A2}(t)$ reaches zero, and the limit cycle concludes when server 1 completes serving the class $B1$ customers in queue:

$$\bar{V}_{A2}(t) = 0 \text{ for } t \in [\bar{t}_2 + (n_{A1} + n_{B1} - 2)C_1 + C_{A1} + V_2 - (n_{A1} - 1)C_{B1}, \bar{t}_2 + (n_{A1} + n_{B1} - 1)C_1]. \quad (25)$$

Notice that at time $\bar{t}_2 + (n_{A1} + n_{B1} - 1)C_1$, the system state $(\bar{V}_{A1}(t), \bar{V}_{A2}(t), s_1(t), s_2(t))$ is the same as at time $\bar{t}_2$. Hence, the system repeats the same cycle of length $(n_1 - 1)C_1$ thereafter. Class $A2$ is served in (21)–(24) and class $B2$ is served in (18)–(20) and (25). Therefore, classes $A2$ and $B2$ are served in contiguous time blocks of length $(n_1 - 1)C_{A1}$ and $(n_1 - 1)C_{B1}$, respectively, and (11) holds for $i = 1$.

To calculate the virtual waiting time for class $A2$, notice that type A customers arrive to station 2 only when class $A1$ customers are being served. These time intervals correspond to the $n_{B1} - 1$ growth periods in (18), the truncated growth interval (20) and its subsequent constant period (21), and the $n_{A1} - 1$ constant intervals during $\bar{V}_{A2}(t)$'s descent in (23). If we view (20) and (21) together as one interval, then all $n_1 - 1$ intervals are of length C_{A1}; a randomly arriving class $A2$ customer is equally likely to arrive during each of these intervals.

A class $A2$ customer arriving during the constant periods in (23) finds the server serving type $A2$ customers, and hence the virtual waiting time there is given by $\bar{V}_{A2}(t)$. Therefore, $Z_{A2} = V_2 - jC_{B1}$ with probability $(n_1 - 1)^{-1}$ for $j = 1, \ldots, n_{A1} - 1$. A class $A2$ customer arriving at time t during (18) must wait until time $\bar{t}_2 + (n_{B1} - 1)C_1 + V_2 - (n_{B1} - 1)C_{A1}$, at which point class $A2$ begins exhaustive service, plus an additional $\bar{V}_{A2}(t)$ time units for the class $A2$ customers ahead of him to be served. Therefore, the virtual waiting time equals $\bar{V}_{A2}(t) + \bar{t}_2 + V_2 + (n_{B1} - 1)C_{B1} - t$. Substituting $\bar{V}_{A2}(t)$ from (18) into this expression gives $Z_{A2} = V_2 + [n_{B1} - j]C_{B1}$, which occurs with probability $(n_1 - 1)^{-1}$ for $j = 1, \ldots, n_{B1} - 1$. Similarly, class $A2$ customers arriving during (20) have virtual waiting time V_2. Finally, class $A2$ customers arriving during (21) find their class in service, and hence also have virtual waiting time V_2. Thus, (12) holds for $i = 1$.

A similar analysis of class $B2$ customers yields (13) for $i = 1$. Finally, notice that each station is directly affected only by its upstream station. Hence, if we define $\bar{t}_{i+1} = \inf\{t \geq 0: \bar{V}_{Ai}(t) = V_i, \bar{V}_{A,i+1}(t) = 0\}$ then our entire analysis holds for $i > 1$, except that we take $\rho_A = \rho_B = 0$ in Equations (14)–(17), because $\bar{V}_{ji}(t)$ decreases at rate one, not $1 - \rho_j$, for $i > 1$. □

4. PERFORMANCE ANALYSIS

In this section, we use the results of §3 to analyze the performance of the original tandem queueing system. Recall that two types of sojourn time analyses were identified in the introduction: transient and steady-state. Ideally, a transient analysis would estimate a customer's total sojourn time in the system conditioned on the complete three-dimensional system state, and a steady-state analysis would derive a customer's total sojourn time in the network. Although Theorem 1 and its proof provide a framework for such an analysis, these results are not explicitly derived here because they are extremely tedious to write out, and they add little to our understanding of the problem. Instead, we are content to derive the transient sojourn time for each station in isolation conditioned on the K-dimensional station workload process and derive the steady-state sojourn time for each station in isolation. At the end of this section, we briefly discuss how one would use our analysis to derive the more general quantities described above.

In our heavy traffic analysis, we estimate the normalized virtual waiting time Z_{ji} given the vector workload $(V_1, \ldots, V_i)$. Since a customer's service times are not known at the time of arrival to the system, the workload process V_i is not actually observable. However, if we let $Q_{ji}(t)$ denote the number of class ji customers in the system at time t, then we can define the observable process

$$L_i(t) = \frac{Q_{Ai}(t) + Q_{Bi}(t)}{\mu} \quad \text{for } t \geq 0. \quad (26)$$

Although L_i is not equal to the sum of L_{Ai} and L_{Bi}, which were defined in Section 2, $L_i(nt)/\sqrt{n}$ and $(L_{Ai}(nt) + L_{Bi}(nt))/\sqrt{n}$ converge together in the heavy traffic limit to the workload process V_i. Suppose a customer arrives to the system at time t and finds $L_i(t) = L_i$ for $i = 1, \ldots, K$. If we make the substitutions $W_{ji}/\sqrt{n}$ for Z_{ji} and $L_i/\sqrt{n}$ for V_i in Equations (5) and (9)–(13), then the heavy traffic parameter n cancels out of these expressions and we get

$$W_{j1} \text{ is uniformly distributed on } \left[0, \frac{\rho L_1}{\rho_j}\right] \quad \text{for } j = A, B, \quad (27)$$

and, for $i = 1, \ldots, K - 1$,

$$W_{A,i+1} = \begin{cases} L_{i+1} + jC_{Bi}, \text{ with probability } \dfrac{1}{n_i - 1} \\ \qquad \text{for } j = 0, \ldots, n_{Bi} - 1, \\ L_{i+1} - jC_{Bi}, \text{ with probability } \dfrac{1}{n_i - 1} \\ \qquad \text{for } j = 1, \ldots, n_{Ai} - 1, \end{cases} \quad (28)$$

and

$$W_{B,i+1} = \begin{cases} L_{i+1} + jC_{Ai}, \text{ with probability } \dfrac{1}{n_i - 1} \\ \qquad \text{for } j = 0, \ldots, n_{Ai} - 1, \\ L_{i+1} - jC_{Ai}, \text{ with probability } \dfrac{1}{n_i - 1} \\ \qquad \text{for } j = 1, \ldots, n_{Bi} - 1, \end{cases} \quad (29)$$

where

$$C_{A1} = \frac{L_1}{1 - \rho_A}, \quad C_{B1} = \frac{L_1}{1 - \rho_B} \text{ and}$$

$$C_1 = \frac{L_1}{1 - \rho_A} + \frac{L_1}{1 - \rho_B}, \quad (30)$$

and, for $i = 1, \ldots, K - 1$,

$$n_{Bi} = \left\lceil \frac{L_{i+1}}{C_{Ai}} \right\rceil, \quad n_{Ai} = \left\lceil \frac{L_{i+1}}{C_{Bi}} \right\rceil, \quad n_i = n_{Ai} + n_{Bi}, \tag{31}$$

$$C_{A,i+1} = (n_i - 1)C_{Ai}, \quad C_{B,i+1} = (n_i - 1)C_{Bi} \text{ and}$$
$$C_{i+1} = (n_i - 1)C_i. \tag{32}$$

If we let T_{ji} denote a random service time for class ji customers and let $S_{ji}(t)$ be the sojourn time of a type j customer at station i who arrives to the system at time t, then we have the heavy traffic estimate $S_{ji}(t) = T_{ji} + W_{ji}$, where W_{ji} is given in (27)–(29).

Now we estimate the steady-state sojourn time S_{ji} for class ji under some additional assumptions. Most importantly, we assume that setup times are zero. Peterson (1991) has shown (for the case where service time distributions depend only on the station, see the discussion in Section 2) that $(V_1, \ldots, V_K)$ has product-form stationary density $\pi(v_1, \ldots, v_K) = \Pi_{i=1}^K \hat{\theta}_i e^{-\hat{\theta}_i v_i}$, where $\hat{\theta}_i = 2\sqrt{n}(1-\rho)/\Sigma_{j=A,B} \lambda_j \mu^{-2}(c_{j0}^2 + c_{ji}^2)$, as long as the network data satisfy a certain skew-symmetry condition. The skew-symmetry condition, given by Peterson (1991, eq. (62)), reduces in our case to $\lambda_A c_{A0}^2 + \lambda_B c_{B0}^2 = \lambda_A c_{Ai}^2 + \lambda_B c_{Bi}^2$ for all $i = 1, \ldots, K$. A special case of this condition is $c_{ji}^2 = c^2$ for $j = A, B$ and $i = 0, \ldots, K-1$, which is slightly less restrictive than the standard product form conditions for Jackson networks. Of course, this condition also allows for other interesting solutions, such as $c_{ji}^2 = c_j^2$, $j = A, B$, $i = 0, \ldots, K-1$, where $c_A^2 \neq c_B^2$ is possible. Harrison and Williams (1986) show that the skew-symmetry condition, together with the stability condition $\rho_i < 1$ for all i, are necessary and sufficient for the exponential product-form solution. If skew-symmetry is not satisfied, then the numerical procedure developed by Dai and Harrison (1992) for the stationary distribution of reflected Brownian motion on the orthant can be employed in conjunction with our deterministic analysis to estimate steady-state sojourn times. For the remainder of this section, we assume that the skew-symmetry condition holds.

Since the random variable $Y = aX$ is exponential with parameter νa^{-1} if X is exponential with parameter ν, heavy traffic analysis predicts that $(L_1, \ldots, L_K)$ possesses an exponential product-form density with parameters

$$\theta_i = \frac{\hat{\theta}_i}{\sqrt{n}} = \frac{2(1-\rho)}{\sum_{j=A,B} \lambda_j \mu^{-2}(c_{j0}^2 + c_{ji}^2)}. \tag{33}$$

Hence, Equations (27)–(32) can be combined with (33) to characterize our estimate of the stationary virtual waiting time W_{ji}, which is independent of the heavy traffic parameter n.

Unfortunately, the recursive nature of (27)–(32) prevents us from explicitly writing the probability distribution for the virtual waiting time W_{ji}. However, performance measures of interest for W_{ji}, such as tail probabilities or moments, can be computed by integrating with respect to the stationary distribution of $(L_1, \ldots, L_i)$. For example, Equations (28)–(29) imply that

$$E[W_{A,i+1} | L_1, \ldots, L_{i+1}] = L_{i+1} + \frac{C_{Bi}(n_{Bi} - n_{Ai})}{2}, \tag{34}$$

and

$$E[W_{B,i+1} | L_1, \ldots, L_{i+1}] = L_{i+1} + \frac{C_{Ai}(n_{Ai} - n_{Bi})}{2}, \tag{35}$$

for $i = 1, \ldots, K-1$. Hence, the steady-state expected virtual waiting time for class $A2$ is

$$E[W_{A2}] = \theta_2^{-1} + \frac{1}{2} \int_0^\infty \int_0^\infty \left(\frac{x_1}{1-\rho_B} \right) \cdot \left(\left\lceil \frac{x_2(1-\rho_A)}{x_1} \right\rceil - \left\lceil \frac{x_2(1-\rho_B)}{x_1} \right\rceil \right) \cdot \theta_1 \theta_2 e^{-\theta_1 x_1} e^{-\theta_2 x_2} \, dx_1 \, dx_2. \tag{36}$$

Our results reveal several insights about system behavior. Expressions (31)–(32) quantify exactly how the service cycles at a station are forced to synchronize with the service cycles at its upstream neighbor, and how this effect ripples through the tandem network. Since $n_i \geq 2$ for all i by (31), it follows by (32) that C_{ji} is nondecreasing in i for $j = A, B$; that is, cycle lengths, and hence batch sizes, tend to be larger at downstream stations. Moreover, C_{Ai}/C_{Bi} is the same for all i, so that the cycle lengths of each customer type grow in the same proportions as one moves downstream. Also, the virtual workload is uniformly distributed at each station: it has a continuous uniform distribution at station 1 and a discrete uniform distribution at all downstream stations.

Finally, notice that since C_{ji} increases with i, the quantities n_{Ai} and n_{Bi} will equal one with greater frequency as i increases. Hence, Equations (34)–(35) suggest that as one moves downstream (i.e., as i increases), the values of $E[W_{Ai}]$ and $E[W_{Bi}]$ will both become closer in value to (but not necessarily converge to) the quantity θ_i^{-1}; consequently, we conjecture that the *imbalance in mean waiting time at a station between the customer types dissipates as one moves downstream*. Similarly, by (28)–(29), we conjecture that the *variability of the stationary sojourn times at each station decreases as one moves downstream*. Although we have not been able to prove these two conjectures, they have been borne out by Monte Carlo simulations (that is, by generating random samples for L_i) of the recursive Equations (31)–(32).

Let us now focus on the symmetric case where $\lambda_A = \lambda_B$. Then $n_{Ai} = n_{Bi}$ and $C_{Ai} = C_{Bi}$ for all i, and C_{i+1}/C_i is an odd positive integer. Hence, W_{Ai} and W_{Bi} are identically distributed for any given station i and $E[W_{ji}] = \theta_i^{-1}$, which is equal to the corresponding quantity in the heavy traffic analysis of a tandem system under the FCFS discipline; of course, the latter quantity coincides with exact results when one further restricts all interarrival and service time distributions to be exponential. Furthermore, by (4.3)–(4.4), the conditional variance of $W_{j,i+1}$ is given by

$$\text{Var}\,[W_{j,i+1} | L_1, \ldots, L_{i+1}] = \frac{C_{ji}^2 n_{ji}(n_{ji} - 1)}{3}, \tag{37}$$

Table 1. Simulation results for a symmetric network.

Class	Sojourn Time Mean	Sojourn Time Standard Deviation	Cycle Length Mean
$A1$	10.0 (±0.2)	12.1 (±0.3)	4.73 (±0.05)
$A2$	10.1 (±0.2)	11.6 (±0.3)	6.06 (±0.07)
$A3$	10.0 (±0.1)	11.2 (±0.3)	6.80 (±0.08)
$A4$	10.0 (±0.1)	11.0 (±0.2)	7.25 (±0.09)
$A5$	10.1 (±0.2)	10.9 (±0.3)	7.55 (±0.09)

for $i = 1, \ldots, K-1$ and $j = A, B$. To assess the accuracy of our estimates and to corroborate the insights described above, we display simulation results in Table 1 for a symmetric tandem system with $K = 5$ stations. Setup times are zero, all interarrival times and service times are exponential, and the system parameters are $\lambda_A = \lambda_B = 0.45$ and $\mu = 1$. Hence, $\theta_i = 1/9$ and $E[S_{ji}] = 10$, which agrees with the simulation results (this result can be derived directly from Little's formula in this special case). Equation (27) implies that $\mathrm{Var}[W_{A1}] = 135$, and hence the estimated standard deviation of class $A1$'s sojourn time is $\sqrt{136} = 11.66$, which is slightly less than the simulated value of 12.1 (numbers in parenthesis in Tables 1 and 2 correspond to 95% confidence intervals). Using (37), we estimate the stationary sojourn time standard deviation for class $A2$ to be 10.93, compared to the simulated value of 11.6. In contrast, under FCFS, the variance of the waiting time in queue is

$$\frac{2\rho - \rho^2}{(\mu - \lambda_A - \lambda_B)^2}, \tag{38}$$

and so the standard deviation of the sojourn time for each class is 10. As predicted, Table 1 shows that cycle lengths grow and sojourn time standard deviations decrease as one moves downstream.

Now let us assume that $\lambda_A \geq \lambda_B$. Then (30)–(32) imply that $C_{Ai} \geq C_{Bi}$ and $n_{Ai} \geq n_{Bi}$ for all i. Consequently, by (27)–(29), we have $E[Z_{Ai}] \leq E[Z_{Bi}]$ and $Var[Z_{Ai}] \leq Var[Z_{Bi}]$ for all i; that is, the customer type with the lower arrival rate incurs both a higher expected waiting time and a more variable waiting time. This is in contrast to a FCFS Markovian tandem network, where steady-state waiting times for each type are identically distributed. We simulated a five-station example for this case, where $\lambda_A = 0.6$, $\lambda_B = 0.3$ and $\mu = 1$. The predicted steady-state expected sojourn times are 7.75 for class $A1$, 14.5 for class $B1$, 8.56 for class $A2$, and 12.51 for class $B2$; the predicted sojourn time standard deviations are 8.77 for class $A1$ and 17.46 for class $B1$. All six of these quantities are reasonably close to the simulated values found in Table 2. Notice that the results in Tables 1 and 2 are quite different, whereas the corresponding results under FCFS would be identical. Table 2 confirms the qualitative insights predicted by heavy traffic theory: Type B customers have higher sojourn time means and variances, the difference in expected sojourn time between classes Ai and Bi decreases in i, and standard deviations for type B's sojourn times decrease as one moves downstream.

Table 2. Simulation results for an asymmetric network.

Station	Sojourn Time Mean		Sojourn Time Standard Deviation	
	Type A	Type B	Type A	Type B
1	8.10 (±0.07)	14.0 (±0.1)	9.03 (±0.12)	18.1 (±0.2)
2	8.58 (±0.07)	13.0 (±0.1)	9.05 (±0.12)	16.7 (±0.2)
3	8.83 (±0.08)	12.3 (±0.1)	9.04 (±0.12)	15.8 (±0.2)
4	9.02 (±0.07)	12.0 (±0.1)	9.04 (±0.12)	15.3 (±0.3)
5	9.12 (±0.07)	11.7 (±0.1)	9.02 (±0.11)	14.9 (±0.2)

We conclude this section by briefly describing how our analysis can be used to derive more general sojourn time results. A customer's total sojourn time in the system conditioned on the $3K$-dimensional system state can be derived by analyzing each station in a sequential manner. In the text above Equation (6), we define a customer's sojourn time at station 1 as a function of the three-dimensional state of station 1 at the time of the customer's arrival. Hence, we know what time the customer exits station 1 and arrives at station 2. Using the proof of Theorem 1, we can then find the three-dimensional state of station 2 at this point in time and derive the customer's sojourn time at this station. This procedure can be repeated for stations $3, \ldots, K$ to obtain the total sojourn time conditioned on the full system state. Notice that the fluid analysis leads to a deterministic transient sojourn time estimate. If one assumes that the fluid system state at the time of a customer's arrival to station 1 is consistent with the limit cycle (there is a unique limit cycle for each realization of the vector workload process $(\bar{V}_1, \ldots, \bar{V}_k)$), then our derivation of Theorem 1 can be used directly to perform the calculations in this recursive procedure; since $\bar{t}_i < C_i$, one suspects that in heavy traffic the system state spends most of its time in (or very close to) a limit cycle. To illustrate these calculations for the simple example in Figure 1, let us suppose that a type A customer arrives at time t to find the system in the state $\bar{V}_{A1}(t) = 2$, $\bar{V}_{B1}(t) = 4$, $\bar{V}_{A2}(t) = 10$, $\bar{V}_{B2}(t) = 6$, $s_1(t) = B$, $s_2(t) = B$; this state corresponds to time 58 in Figure 1. Hence, the customer exits station 1 and arrives at station 2 at time 68. The quantity $\bar{V}_{A2}$ is in station 2's truncated growth interval (Equation (20)) at time 68, and so the customer has a sojourn time of length $\bar{V}_2 = 16$ at station 2, giving a total system sojourn time of 26. If the initial fluid state is not consistent with the limit cycle then the recursive procedure still holds, although the analysis is considerably more difficult. One would have to track the transient dynamics as we did at the beginning of the proof of Theorem 1.

By Theorem 1, the limit cycle of a station is embedded in the limit cycle of its downstream station; hence, the limit cycle for station K is the limit cycle for the entire network. To obtain the steady-state distribution of the total system sojourn time for a customer type, we assume that customers arrive to station 1 uniformly over the network's limit cycle (recall that there is a different limit cycle for each

value of the K-dimensional workload vector), and calculate the deterministic total sojourn time conditioned on the arrival time to station 1, as described in the last paragraph. Then we uncondition on the arrival times by integrating the conditional total sojourn time with respect to the uniform arrivals. This integration yields a sojourn time distribution for each realization of the K-dimensional station workload vector. Finally, this conditional sojourn time distribution needs to be integrated with respect to the product-form stationary distribution of the workload vector to obtain the desired result. It would appear that a discretization (of the vector workload process and perhaps the uniform arrival process) procedure would be required to carry out this tedious agenda.

5. SETUP TIMES AT STATION 1

In this section we perform a steady-state analysis when setup times are incurred at the first station. The presence of setup times alters the vector station-level workload process; indeed, with only the first station present, it is shown in CPR II that the workload process converges to a Bessel process in heavy traffic. There is no analogous heavy traffic limit theorem for the tandem network with setup times in the first station. A heuristic derivation of the infinitesimal drift vector, along the lines presented in CPR II, yields $\mu_1(\mathbf{x}) = (\rho_A\rho_B s/x_1) - c$, $\mu_2(\mathbf{x}) = -\rho_A\rho_B s/x_1$, and $\mu_k(\mathbf{x}) = 0$, $2 < k \leqslant K$, where $c = \sqrt{n}(1-\rho)$, s is the mean setup time over a cycle (switch from A to B and from B to A), and $\mathbf{x}$ is the workload vector. Since our goal is to obtain steady-state performance measures, we assume that the covariance matrix satisfies the skew-symmetry condition described in §4, so that $\sigma_{ii}^2 = 2\alpha^{-1}$, $1 \leqslant i \leqslant K$, and $\sigma_{ij}^2 = -\alpha^{-1}1_{\{|i-j|=1\}}$, $1 \leqslant i \neq j \leqslant K$, where $\alpha = \mu^2/[\lambda_A c_{A_0}^2 + \lambda_B c_{B_0}^2]$. It was shown by Yamada (1986) that for $K = 2$ a process fitting an appropriately completed version of the above description (including boundary behavior) exists and is unique in distribution. (It was also remarked there that the extension to $K > 2$ is straightforward.) Yamada (1986) also proves a heavy traffic limit theorem for tandem networks with state-dependent service rate that gives some backing to our assumption here of convergence (especially in light of the discussion in CPR II linking their result to an earlier result of Yamada 1984).

Although the existence of our process is covered by Yamada (1986), there are no results that characterize the stationary distribution of the process. We conjecture that the stationary distribution has a product form, with the first marginal a gamma distribution (obtained as the stationary distribution of a one-dimensional Bessel process), and the other marginals exponential. In particular, we conjecture that

$$\pi(x_1, \ldots, x_K) = \frac{(\alpha c)^K (\alpha c x_1)^\beta \prod_{i=1}^{K} e^{-\alpha c x_i}}{\Gamma(\beta+1)}, \quad x_i \geqslant 0,\ 1 \leqslant i \leqslant K, \tag{39}$$

where $\beta = \lambda\rho_A\rho_B s$. A brief discussion of the basis for these conjectures is in order. Note that it is clear that the marginal distribution for the first station is a gamma distribution. This follows because the first station is not affected by downstream stations, and the stationary distribution of the Bessel process (when $c > 0$) is gamma. If we could find a sequence of product-form networks whose (conjectured) heavy traffic limit diffusion is the same as the above diffusion, it seems reasonable that the steady-state distribution of the diffusion would be the limit of the steady-state distributions of the product-form networks (this constitutes an "invariance principle"). Consider a network with $K + 1$ stations, labeled 0, 1, ..., K, and two customer types. Type 1 is open, with a Poisson arrival process and exponential service times at each station. These customers arrive at station 1 and proceed sequentially to stations 2, ..., K, and then exit. Type 2 is closed, with $N \geqslant 1$ such customers in the system. These customers alternate between station 0, where there are N servers with rate γ, and station 1, where they are served in a FCFS manner along with type 1 and have the same service rate. This is a product-form network. We can write down the stationary distribution and take a limit as the network is driven into heavy traffic with N held fixed. The limit that emerges has the form (39). Because this network only makes sense for integer N, we cannot actually match $\mu_1(\mathbf{x})$ for all possible values of $\rho_A\rho_B$s, only a finite number of such values. Nonetheless we conjecture that (39) holds for all values of $\rho_A\rho_B$s.

The current theory of multidimensional diffusion processes does not provide a means of verifying the above conjecture. One way to lend it plausibility is to derive (without proof) a "Basic Adjoint Relationship" (BAR) for this process, analogous to the one proven to hold for reflected Brownian motion in an orthant by Harrison and Williams (1987), and show that our proposed solution satisfies the BAR. Rather than taking this lengthy detour, we merely point out that this program has been carried out, and it works.

We have restricted our attention to setup times at the first station. With setup times at other stations the state-dependent drift is extremely complicated because the kth component of the drift depends on the cycle time at station k, which depends on the workload at station k relative to station $k - 1$. For example, with $K = 2$, we have $\mu_2(x) = s/4x_1$ for $x_2 < 2x_1$; $\mu_2(\mathbf{x}) = s/12x_1$ for $2x_1 < x_2 < 4x_1$; $\mu_2(\mathbf{x}) = s/20x_1$ for $4x_1 < x_2 < 6x_1$; and so on, with the wedges continuing forever, getting progressively smaller.

6. CONCLUDING REMARKS

From the viewpoint of manufacturing systems applications, perhaps the biggest shortcoming in multiclass queueing network theory is the failure to incorporate non-FCFS queueing disciplines that are driven by large setup times and/or setup costs when a server changes class. In this

paper we make a first attempt at addressing this shortcoming by analyzing the performance of a set of perfectly balanced polling systems in tandem. Because even this idealized network is unlikely to yield to an exact analysis, we resort to heavy traffic approximations. With strong support from existing limit theorems (Coffman et al. 1995, 1998), we conjecture that a time scale decomposition holds in the heavy traffic limit, and derive sojourn time estimates for both the transient and steady-state cases. The simple form of Theorem 1 allows us to gain some understanding of the behavior of this complex system. In particular, as one moves from upstream stations to downstream stations, the batch sizes tend to get larger, the imbalance between customer types becomes less pronounced and the variability of sojourn times is reduced. Simulation results in Section 4 confirm that these three qualitative features occur in the actual queueing network.

Many manufacturing facilities evaluate their employees using performance metrics that encourage large batch sizes (Goldratt and Cox 1984). The exhaustive queueing discipline considered in this paper represents a situation where each workstation is trying to maximize its batch sizes subject to avoiding unnecessary idleness. The first of our three qualitative observations has important implications for these evaluation mechanisms: In a balanced tandem system, downstream stations are inherently better able to take advantage of the economies of scale than upstream stations.

In Reiman and Wein (1999 §5), we investigate four generalizations of the model: service rates that differ by customer type at each station, closed networks, networks with three customer types, and the incorporation of FCFS nodes. In §5 of the present paper we construct a product-form steady-state distribution (gamma marginal at station 1 and exponential marginals at the downstream stations) for the case where setup times are incurred only at station 1, even though the original Markovian version of the model does not possess a product-form solution. However, these calculations suggest that transient results for the general multitype, multistation tandem network and steady-state results for networks with setup times at downstream stations will not come easily.

ONLINE COMPANION

An online companion to this paper, where we study four natural generalizations of our basic model, can be found in the Online Collection:

http://grace.wharton.upenn.edu/~harker/opsresearch.html

at the Operations Research Home Page.

ACKNOWLEDGMENT

The second author was supported by a grant from the Leaders for Manufacturing Program at MIT, NSF grant DDM-9057297, and EPSRC grant GR/J71786. He thanks the Statistical Laboratory at the University of Cambridge for its hospitality while some of this work was carried out. The authors are grateful to Chalee Asavathiratham and Beril Toktay for performing the numerical computations reported in Section 4, and to the referees for their helpful comments.

REFERENCES

Boxma, O. J., H. Takagi. (Eds.) 1992. Special Issue on Polling Systems. *Queueing Systems* **11** (1, 2).

Coffman, E. G., Jr., A. A. Puhalskii, M. I. Reiman. 1995. Polling systems with zero switchover times: a heavy-traffic averaging principle. *Ann. Appl. Probab.* **5** 681–719.

——, ——, ——. 1998. Polling systems in heavy-traffic: a Bessel process limit. Math. Oper. Res. **23** 257–304.

Dai, J. G., J. M. Harrison. 1992. Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. *Ann. Appl. Probab.* **2** 65–86.

Foschini, G. J. 1980. Personal communication.

Goldratt, E. M., J. Cox. 1984. *The Goal*. North River Press, Inc., Croton-on-Hudson, NY.

Harrison, J. M. 1985. *Brownian Motion and Stochastic Flow Systems*. John Wiley, New York.

——, R. J. Williams. 1986. Multidimensional reflected Brownian motions having exponential stationary densities. *Ann. Probab.* **15** 115–137.

——, ——. 1987. Brownian models of open queueing networks with homogeneous customer populations. *Stochastics* **22** 77–115.

Hofri, M., K. W. Ross. 1987. On the optimal control of two queues with server set-up times and its analysis. *SIAM J. Comput.* **16** 399–420.

Karmarkar, U. S., S. Kekre, S. Kekre. 1985. Lotsizing in multi-item multi-machine job shops. *IIE Trans.* **17** 290–298.

Peterson, W. P. 1991. A heavy traffic limit theorem for networks of queues with multiple customer types. *Math. Oper. Res.* **16** 90–118.

Reiman, M. I. 1984. Open queueing networks in heavy traffic. *Math. Oper. Res.* **9** 441–458.

——, ——. 1999. Heavy traffic analysis of polling systems in tandem. Online version of present paper: http://grace.wharton.upenn.edu/~harker/opsresearch.html.

Takagi, H. 1986. *Analysis of Polling Systems*. MIT Press, Cambridge, MA.

Wein, L. M. 1991. Due-date setting and priority sequencing in a multiclass *M*/*G*/1 queue. *Management Sci.* **37** 834–850.

Yamada, K. 1984. Diffusion approximations for storage processes with general release rules. *Math. Oper. Res.* **9** 459–470.

——. 1986. Multi-dimensional Bessel processes as heavy traffic limits of certain tandem queues. *Stochastic Processes and Their Appl.* **23** 35–56.