MATHEMATICS OF OPERATIONS RESEARCH
Vol. 26, No. 2, May 2001, pp. 358–374
Printed in U.S.A.

BUFFER REQUIREMENTS AND SERVER ORDERING IN A TANDEM QUEUE WITH CORRELATED SERVICE TIMES

BENJAMIN AVI-ITZHAK AND HANOCH LEVY

We analyze the intermediate buffer requirements in a tandem queue where service times of each customer are deterministically correlated between the servers and arbitrarily distributed between customers. The major issue at hand is the determination of intermediate buffer sizes assuring no blocking when the arrivals pattern is arbitrary and unpredictable. The analysis shows that the worst arrival process is the Just-in-Time (JIT) process. Further, it shows that ordering of the servers with respect to service rates may be detrimental, and that the most vulnerable architectural design is that in which the servers have almost the same service rates. It is shown that the total buffer requirement in the system may be quite sensitive to the server ordering: A proper ordering requires just $O(M)$ (where M is the number of queues) buffer size, while an improper ordering may require $O(M^2)$.

1. Introduction. This work considers a tandem arrangement of $M+1$ servers, indexed $0, 1, \ldots, M$ and denoted for short by s_i, $i = 0, 1, \ldots, M$. Each server has a buffer, which includes one service position and an unlimited number of waiting positions. Customers arrive at s_0 according to some arbitrary arrival (input) process and are served in a FCFS order. When the jth arriving customer, denoted by c_j, completes its service at s_i, $i = 0, 1, \ldots, M-1$, it moves to the next queue. If s_{i+1} is idle at that time, the customer is admitted to service; otherwise the customer joins the end of the waiting line, which is processed in a FCFS order. The customer departs upon completion of service at s_M.

We are interested in the model in which service requirement may vary from one customer to another, but for a given customer it is the same at all servers. Nevertheless, because the service rates of the servers may differ, the service times of the given customer may differ accordingly. Let $S^{(j)}$ be the service requirement of c_j and let r_i denote the service rate of s_i; then the service time of c_j at s_i, denoted by $S_i^{(j)}$, is given by the ratio,

$$S_i^{(j)} = \frac{S^{(j)}}{r_i}, \tag{1}$$

and is deterministically correlated to the service times of the customer at all other servers of the tandem.

This model has applications in various areas. One is telecommunications, in which the customers represent messages, while the servers represent communications switches in a network. The assumption that the service requirement of a customer is fixed along the tandem is very natural in these applications, since the service requirement is typically proportional to the message size. Note that, in these applications, it is common that one message will differ from another in its size, but a single message will not change its size while moving along the network. It is also common to have switches that differ from each other in their actual speed, because of differences either in technology between the switches, or in the functionality they perform on the messages.

Received July 1, 1999; revised August 30, 2000, and January 19, 2001.
MSC 2000 subject classification. Primary: 60K25.
OR/MS subject classification. Primary: Queues/tandem.
Key words. Queuing, tandem queues, buffer.

0364-765X/01/2602/0358/$05.00
1526-5471 electronic ISSN,

Another area is that of manufacturing, where the assemblies go through a series of assembling stages. Here it is likely that the amount of work associated with an assembly may differ among models but will be fixed for a certain model when moving along the assembling stages. The speed of processing at the different stages may vary mainly because of the different functionalities of the stages.

In this study, we are mainly interested in analyzing the intermediate buffer sizes needed to avoid blocking and thus prevent loss in the case of communications, or to maximize throughput in the case of manufacturing. The intermediate buffer sizes needed to avoid blocking are calculated in this work for the manufacturing blocking scheme, where blocking may take place after service completion: If at the time that c_j completes service at s_i the buffer of s_{i+1} is fully occupied, c_j is blocked. For the communication blocking scheme, where c_j cannot start service at s_i unless the buffer at s_{i+1} is not fully occupied, one has to increase our results by 1. A more comprehensive discussion of blocking schemes is presented in Avi-Itzhak and Levy (1995).

The only generality-limiting assumption on the service requirements is

$$0 < a \leq S^{(j)} \leq b < \infty, \qquad j \geq 1, \tag{2}$$

where a and b are respectively the smallest and the largest possible service requirements of any customer.

To prevent the possibility of infinite accumulation of workload or customers in any of the intermediate buffers, we assume that $r_0 \leq r_i$, $i = 1, 2, \ldots, M$, i.e., s_0 is the slowest server.

In previous research, the throughput behavior (Kelly 1982, 1984, 1985) and the intermediate buffer size required to avoid blocking (Ziedins 1993) were studied under the assumptions that service times of a given customer are identical at all servers (equivalent to $r_0 = r_1 = \cdots = r_M$ in our study), and Just-in-Time (JIT) arrivals, where s_0 is busy at all times (equivalent to either saturation or coordination of arrivals at s_0). Under these assumptions, it was shown by Kelly (1982) and then by Ziedins (1993) that the intermediate buffer size required is given by $\lceil b/a \rceil$, where $\lceil x \rceil$ is the smallest integer greater than, or equal to, x. The particular situation where no intermediate buffers are available was studied by Avi-Itzhak and Halfin (1993), who showed that the case of $r_0 = r_1 = \cdots = r_M$ is the worst with respect to throughput rate. An extensive list of additional earlier related publications is provided in Ziedins (1993).

Our objective in this work is to relax the assumption of exact equality of service rates and study the impact of service rates variability on the buffer sizes required for blocking avoidance; of specific interest is how sensitive the buffer requirement is to the assumption of equal server rates.

Furthermore, we are interested in the effect of server ordering on the required buffer sizes and in widening the analysis to include non-JIT input.

The effect of server ordering on various measures of performance of tandem systems has been addressed by quite a number of authors, including, to mention a few, Kim and Avi-Itzhak (1995), Avi-Itzhak and Levy (1995), Yamazaki et al. (1992), Ding and Greenberg (1991), Huang and Weiss (1990), Whitt (1985), Pinedo (1982), Friedman (1965), and Avi-Itzhak (1965). A more extensive list is provided in Kim and Avi-Itzhak (1995). The magnitude of this effect ranges from zero to considerable, depending on the particular systems and measures of interest studied. It will be shown that, in the system studied in this work, the ordering of the servers may be detrimental.

The main results of this paper are four:

(1) The JIT arrival process is the worst process.

(2) If the servers are ordered in nondecreasing service rate, then the total buffer (sum of intermediate buffer sizes) required is proportional to $M\lceil b/a \rceil$.

(3) If the servers are ordered in decreasing service rate, except s_0, which is the slowest server by assumption, then the total buffer required is proportional to $(M^2)\lceil b/a \rceil$.

(4) In practical terms, the most vulnerable architectural design is that in which the servers have nearly the same service rates. Under such a design, we show that there exists a customers arrival pattern for which the buffer size requirement at s_i is approximately $\lceil i(b/a) \rceil$.

The last result is, perhaps, quite surprising since it implies that a system that is almost identical to the equal servers system may have radically worse performance. Theoretically, this means that the system performance is very sensitive at that setting. In practical terms, it implies that since it is very hard to guarantee precise equality of the servers, the selection of equal rate servers may be the worst of all. In fact, one may benefit from designing the servers to be slightly (but distinguishably) different from each other; while this may affect the throughput, presumably slightly, it can provide a factor of $O(M)$ reduction in the total buffer required to prevent blocking.

The structure of this paper is as follows: In §2, we introduce the notion of unprocessed work and use it to show stability conditions and to derive upper bounds on the work and number of customers in the system. Section 3 presents and analyzes the flow dynamics of the system and provides the necessary theorem and corollaries for establishing the bounds presented in §4, under various architectural designs. In §5 we present the "worst-case" scenarios and discuss the major conclusions of the paper.

The presentation in this work is in terms of deterministic sequences of arrivals and service requirements. The analysis should be considered as sample path analysis when stochastic arrivals and service requirements are assumed. Most of the results are extendable to the stochastic case. Applying in essence the same arguments used in this work to random variables defined on the same probability space is rather straightforward for many of our results. Nevertheless, seeking stochastic limit statements, as t or as j go to infinity, will involve additional elaborate treatment, along the ideas and lines offered by Kelly (1982), and may also require more restrictive assumptions on the random variables characterizing the service and arrivals.

2. Work in the system. Let $P_i(t)$ be the total amount of service requirement processed by s_i up to time t; then the unprocessed work at s_i at time t, denoted by $u_i(t)$, is given by

$$u_i(t) = P_{i-1}(t) - P_i(t), \qquad i = 1, 2, \ldots, M,\ t \geq 0. \tag{3}$$

The unprocessed work in the system up to and including s_i is

$$U_i(t) = \sum_{j=1}^{i} u_j(t) = P_0(t) - P_i(t) \geq 0, \qquad i = 0, 1, \ldots, M,\ t \geq 0. \tag{4}$$

Note that by this definition, $U_0(t) = 0 \,\forall t \geq 0$. In this paper, we will assume that $U_i(0) = 0, i = 1, 2, \ldots, M$, i.e., the system is empty at time zero.

The function $U_i(t)$ is piecewise linear and continuous. Its slope $U_i'(t)$ is determined by the idle-or-busy states of s_0 and s_i at time t.

$$U_i'(t) = \begin{cases} 0, & \text{if } s_i \text{ and } s_0 \text{ are idle,} \\ r_0, & \text{if } s_i \text{ is idle and } s_0 \text{ is busy,} \\ -(r_i - r_0), & \text{if } s_i \text{ and } s_0 \text{ are busy,} \\ -r_i, & \text{if } s_i \text{ is busy and } s_0 \text{ is idle.} \end{cases} \tag{5}$$

We note that the assumption $r_0 \leq r_i$ implies that $U_i'(t) \leq 0$ if s_i is busy.

THEOREM 1. BOUND ON UNPROCESSED WORK. *Let the service requirements be bounded as in* (2). *Then*

$$U_i(t) \le ib, \quad i = 0, 1, \ldots, M, \ t \ge 0. \tag{6}$$

PROOF. From (4), we have $U_0(t) = 0, t \ge 0$. Thus the theorem is true for $i = 0$. Assume it is true up to some $i = k - 1 \ge 0$ and not true for $i = k$. Since $U_i(0) = 0 \ \forall i$, then there must exist a value of $t > 0$, say $t = \tau$, where $U_k(t)$ crosses from below kb to above kb. Hence $U_k(\tau) = kb$ and $\exists \delta > 0$ such that $U_k(\tau+\varepsilon) > kb \forall \varepsilon \in (0, \delta]$. By assumption, however, $U_{k-1}(\tau+\varepsilon) \le (k-1)b \ \Rightarrow \ U_k(\tau+\varepsilon) - U_{k-1}(\tau+\varepsilon) > b \Rightarrow$ there is at least one customer at s_k and therefore s_k is busy. But $r_0 \le r_k \ \Rightarrow \ U_k'(t) \le 0$ when s_k is busy $\Rightarrow U_k'(\tau+\varepsilon) \le 0 \forall \varepsilon \in (0, \delta] \Rightarrow U_k(\tau+\varepsilon) \le U_k(\tau) = kb$, which is a contradiction. □

An upper bound for the unfinished work at s_i is obtained from the relation

$$u_i(t) \le U_i(t) \le ib, \qquad i = 1, 2, \ldots, M, \ t \ge 0. \tag{7}$$

The two following corollaries are immediate results of Theorem 1.

COROLLARY 1. BOUND ON QUEUE SIZES. *Let $n_i(t)$ be the number of customers at s_i at time t and let*

$$N_i(t) = \sum_{k=1}^{i} n_k(t), \qquad i = 1, 2, \ldots, M, \ t \ge 0. \tag{8}$$

Then,

$$n_i(t) \le N_i(t) \le \left\lceil \frac{U_i(t)}{a} \right\rceil \le \left\lceil i\frac{b}{a} \right\rceil, \qquad t \ge 0. \tag{9}$$

COROLLARY 2. STABILITY. *If $r_0 \le r_i$, then $u_i(t), U_i(t), n_i(t)$, and $N_i(t)$ are bounded for all $t \ge 0, \ i = 1, 2, \ldots, M$.*

Theorem 1 applies to any subsequence of adjacent servers, $s_k, s_{k+1}, \ldots, s_i$, $0 \le k \le M-1$, $k < i \le M$, satisfying the condition $r_k \le r_m$, $k \le m \le i$.

For example, if $r_{i-1} \le r_i$ for some $0 < i \le M$, then the bound in (7) takes the form $u_i(t) \le b$ and the bound in (9) takes the form $n_i(t) \le \lceil b/a \rceil$. This agrees with Kelly (1982) and Ziedins (1993), who show this property for $r_0 = r_1 = \cdots = r_M$.

When the values of the service rates $r_0, r_1, \ldots, r_M$ are known, it is possible as is shown later, to obtain tighter bounds than those presented in this section.

3. Holding times. In this section, we investigate the holding times experienced by customers in the system. Our key results, provided in Theorem 2, classify the customers into two types: (a) *dominating customers*, whose service requirement is greater than or equal to that of all their predecessors, and (b) *dominated customers*, which are all other customers. We bound the total holding time of a dominated customer by that of the dominating customer preceding it, and provide the exact expression for the holding time of a dominating customer. Propositions 1 and 2 and Lemma 1 lay the ground for the proof of the Key Theorem.

We denote the departure time epoch of c_j from s_i by $d_i^{(j)}$ and its waiting and holding time at s_i, respectively, by $w_i^{(j)}$ and $h_i^{(j)}$. Then

$$d_i^{(j+1)} - d_i^{(j)} \ge S_i^{(j+1)} = \frac{S^{(j+1)}}{r_i}, \qquad i \ge 0, \ j \ge 1, \tag{10}$$

and, under our assumption of an unlimited number of intermediate waiting positions,

$$h_i^{(j)} = w_i^{(j)} + S_i^{(j)} = d_i^{(j)} - d_{i-1}^{(j)} \geq S_i^{(j)} = \frac{S^{(j)}}{r_i}, \quad i > 0,\ j \geq 1, \tag{11}$$

where (10) is an equality if $w_i^{(j+1)} > 0$ and an inequality otherwise, while the right side of (11) is an equality if $w_i^{(j)} = 0$ and is a strict inequality otherwise.

The total holding time of c_j in $s_1, s_2, \ldots, s_i$ denoted by $H_i^{(j)}$ is given as

$$H_i^{(j)} = d_i^{(j)} - d_0^{(j)} = \sum_{m=1}^{i} h_m^{(j)} = H_k^{(j)} + \sum_{m=k+1}^{i} h_m^{(j)}, \quad i, j \geq 1,\ k = 1, 2, \ldots, i. \tag{12}$$

The total holding time, $H_i^{(j)}$, is the time-in-the-system, where the system is defined as servers $1, 2, \ldots, i$.

Substitution of (11) in (12) yields

$$H_i^{(j)} \geq H_k^{(j)} + \sum_{m=k+1}^{i} S_i^{(j)}, \quad \forall i, j \geq 1 \quad \text{and} \quad k = 1, 2, \ldots, i. \tag{13}$$

From the dynamics of the system (Relations (11) and (12)), we have $\forall i \geq 0,\ j \geq 1$ the basic relation:

$$d_i^{(j+1)} = \begin{cases} d_{i-1}^{(j+1)} + S_i^{(j+1)}, & \text{iff } w_i^{(j+1)} = 0 \Leftrightarrow d_{i-1}^{(j+1)} \geq d_i^{(j)}, \\ d_i^{(j)} + S_i^{(j+1)}, & \text{iff } w_i^{(j+1)} > 0 \Leftrightarrow d_{i-1}^{(j+1)} < d_i^{(j)}, \end{cases} \tag{14}$$

where $\Leftrightarrow$ stands for an iff relation.

This basic relation can be rewritten as

$$d_i^{(j+1)} = S_i^{(j+1)} + \max(d_{i-1}^{(j+1)}, d_i^{(j)}), \quad \forall i,\ j \geq 1. \tag{15}$$

Note that the basic Equations (14) and (15) hold for arbitrary service times and are therefore true for the whole class of FCFS tandem models. By substituting $S^{(j)}/r_i$ for $S_i^{(j)}$ we obtain the forms for our particular model.

PROPOSITION 1. *If $w_i^{(j)} > 0$ for some $1 \leq i \leq M$ and $j > 1$, then*

(a) $H_i^{(j)} \leq H_i^{(j-1)} - S^{(j)}(1/r_0 - 1/r_i) \leq H_i^{(j-1)}$,

(b) $h_i^{(j)} \leq h_i^{(j-1)} - S^{(j)}(1/(r_{i-1}) - 1/r_i)$.

PROOF. From (14), we have that when $w_i^{(j)} > 0$,

$$d_i^{(j)} = d_i^{(j-1)} + \frac{S^{(j)}}{r_i}, \tag{16}$$

and from (10), we have

$$d_0^{(j)} \geq d_0^{(j-1)} + \frac{S^{(j)}}{r_0}. \tag{17}$$

Subtracting (17) from (16) results in

$$d_i^{(j)} - d_0^{(j)} \leq d_i^{(j-1)} - d_0^{(j-1)} - S^{(j)}\left(\frac{1}{r_0} - \frac{1}{r_i}\right). \tag{18}$$

Substituting $H_i^{(j)}$ as defined in (12) and noting that $r_0 \leq r_i$ yields the relation given in (a). Note that for JIT input (17) is an equality resulting in (18) being an equality too.

To prove (b) subtract $d_{i-1}^{(j)}$ from (16) to obtain

$$h_i^{(j)} = d_i^{(j)} - d_{i-1}^{(j)} = d_i^{(j-1)} - d_{i-1}^{(j)} + \frac{S^{(j)}}{r_i}. \tag{19}$$

But from (10), we have

$$d_{i-1}^{(j)} \geq d_{i-1}^{(j-1)} + \frac{S^{(j)}}{r_{i-1}}. \tag{20}$$

Substitution of (20) in (19) yields the result. □

COROLLARY 3.

(1) *If* $r_i \geq r_{i-1}$ *and* $w_i^{(j)} > 0$ *for some* $1 \leq i \leq M$ *and* $j > 1$, *then* $h_i^{(j)} \leq h_i^{(j-1)}$.

(2) *If* $r_i \geq r_{i-1}$ *and* $S^{(j-1)} \geq S^{(j)}$ *for some* $1 \leq i \leq M$ *and* $j > 1$, *then* $h_i^{(j)} \leq h_i^{(j-1)}$.

(3) *For* $j > n \geq 1$, *if* $S^{(n)} \geq S^{(m)}$ $\forall m = n, n+1, \ldots, j$ *and* $r_i \geq r_{i-1}$ *for some* $1 \leq i \leq M$, *then* $h_i^{(j)} \leq h_i^{(n)}$.

PROOF. The proof of (1) follows immediately from Part (b) of Proposition 1.

To prove (2), note that (1) asserts that this is true if $w_i^{(j)} > 0$. Suppose $w_i^{(j)} = 0$ then obviously we have $h_i^{(j)} = S^{(j)}/r_i$, but $h_i^{(j-1)} \geq S^{(j-1)}/(r_{i-1})$ by definition, and $S^{(j-1)}/(r_{i-1}) \geq S^{(j)}/r_i$ by assumption.

The proof of (3) follows by induction from (1) and (2). From (2), we have that $h_i^{(n+1)} \leq h_i^{(n)}$. Assume then that $h_i^{(m)} \leq h_i^{(n)}$ for $m = n+1, n+2, \ldots, j-1$. If $w_i^{(j)} > 0$ we have from (1) that $h_i^{(j)} \leq h_i^{(j-1)} \Rightarrow h_i^{(j)} \leq h_i^{(n)}$. If $w_i^{(j)} = 0$ then from (14) it follows that $h_i^{(j)} = S^{(j)}/r_i$, but $S^{(j)}/r_i \leq S^{(n)}/r_i \leq h_i^{(n)} \Rightarrow h_i^{(j)} \leq h_i^{(n)}$. □

PROPOSITION 2. *If for some* $j \geq 1$, $S^{(j)} \geq S^{(j+1)}$, *then* $H_i^{(j+1)} \leq H_i^{(j)}$ $\forall i \geq 1$.

This proposition states that if the service requirement of a customer does not exceed the service required by its predecessor then its total holding time, up to any server in the tandem, cannot exceed the corresponding total holding time of its predecessor.

PROOF. If c_{j+1} is not delayed at Servers $1, 2, \ldots, i$, i.e., $w_k^{(j+1)} = 0$, $k = 1, 2, \ldots, i$, then from (11) and (12) we have

$$H_i^{(j+1)} = S^{(j+1)} \sum_{k=1}^{i} \frac{1}{r_k},$$

while

$$H_i^{(j)} \geq S^{(j)} \sum_{k=1}^{i} \frac{1}{r_k},$$

and therefore $H_i^{(j)} \geq H_i^{(j+1)}$.

If c_{j+1} is delayed before exiting s_i let s_k, $k \geq 1$, be the last station among $s_1, s_2, \ldots, s_i$ at which it is delayed. Then from (11) and (12), we get

$$d_i^{(j+1)} = d_k^{(j+1)} + S^{(j+1)} \sum_{m=k+1}^{i} \frac{1}{r_m}. \tag{21}$$

Since $w_k^{(j+1)} > 0$ we get from (14) that

$$d_k^{(j+1)} = d_k^{(j)} + \frac{S^{(j+1)}}{r_k}. \tag{22}$$

Substitution in (21) yields

$$d_i^{(j+1)} = d_k^{(j)} + \frac{S^{(j+1)}}{r_k} + S^{(j+1)} \sum_{m=k+1}^{i} \frac{1}{r_m}. \tag{23}$$

For $i = 0$, we have from (10),

$$d_0^{(j+1)} \geq d_0^{(j)} + \frac{S^{(j+1)}}{r_0}. \tag{24}$$

(Note that for JIT input, the equality in (24) holds for all $j \geq 1$.)

Subtracting (24) from (23), and using the definition of $H_i^{(j)}$ given in (12) we obtain

$$\begin{aligned} H_i^{(j+1)} &= d_i^{(j+1)} - d_0^{(j+1)} \\ &\leq d_k^{(j)} - d_0^{(j)} + S^{(j+1)} \sum_{m=k+1}^{i} \frac{1}{r_m} - S^{(j+1)}\left(\frac{1}{r_0} - \frac{1}{r_k}\right) \\ &= H_k^{(j)} + S^{(j+1)} \sum_{m=k+1}^{i} \frac{1}{r_m} - S^{(j+1)}\left(\frac{1}{r_0} - \frac{1}{r_k}\right), \end{aligned} \tag{25}$$

where for JIT input the equality holds.

From (13) we can write

$$H_i^{(j)} \geq H_k^{(j)} + S^{(j)} \sum_{m=k+1}^{i} \frac{1}{r_m}, \tag{26}$$

and then subtracting (25) we obtain

$$H_i^{(j)} - H_i^{(j+1)} \geq \left(S^{(j)} - S^{(j+1)}\right) \sum_{m=k+1}^{i} \frac{1}{r_m} + S^{(j+1)}\left(\frac{1}{r_0} - \frac{1}{r_k}\right). \tag{27}$$

Since $S^{(j)} \geq S^{(j+1)}$ and $r_0 \leq r_k$ the right-hand side of (27) is nonnegative. □

Note that

$$H_i^{(j+1)} \leq H_i^{(j)} \Leftrightarrow d_i^{(j+1)} - d_i^{(j)} \leq d_0^{(j+1)} - d_0^{(j)}, \tag{28}$$

which means that the interdeparture time, between c_{j+1} and c_j, at s_i, is smaller than or equal to their interdeparture time at s_0, i.e., c_{j+1} is narrowing the time gap to its predecessor.

LEMMA 1. *For $j > n \geq 1$, if $S^{(n)} \geq S^{(m)}$ $\forall n \leq m \leq j$, then $H_i^{(j)} \leq H_i^{(n)}$.*

Lemma 1 states that the total holding time of c_j cannot exceed that of c_n, $n \leq j$, if the service requirements of $c_{n+1}, c_{n+2}, \ldots, c_j$ do not exceed the service requirements of c_n.

PROOF: BY INDUCTION. From Proposition 2, Lemma 1 is true for c_{n+1}. We assume that it is true for c_{j-1} and show first that $H_1^{(j)} \leq H_1^{(n)}$:

For $w_1^{(j)} > 0$, we have from Proposition 1, $H_1^{(j)} \leq H_1^{(j-1)}$, if $w_1^{(j)} > 0$. But $H_1^{(j-1)} \leq H_1^{(n)} \Rightarrow H_1^{(j)} \leq H_1^{(n)}$.

For $w_1^{(j)} = 0$, we have from (14), $H_1^{(j)} = S^{(j)}/r_0$, if $w_1^{(j)} = 0$. But $S^{(j)}/r_0 \leq S^{(n)}/r_0 \leq H_1^{(n)} \Rightarrow H_1^{(j)} \leq H_1^{(n)}$.

Assume now that $H_{i-1}^{(j)} \leq H_{i-1}^{(n)}$, then: If $w_i^{(j)} > 0$, it follows from Proposition 1 that $H_i^{(j)} \leq H_i^{(j-1)}$. But $H_i^{(j-1)} \leq H_i^{(n)}$ by the inductive assumption.

If $w_i^{(j)} = 0$ we have from (14), (12) and (11), $H_i^{(j)} = H_{i-1}^{(j)} + S^{(j)}/r_i$, but $H_{i-1}^{(j)} \leq H_{i-1}^{(n)}$, $S^{(j)} \leq S^{(n)}$ and $H_i^{(n)} \geq H_{i-1}^{(n)} + S^{(n)}/r_n \Rightarrow H_i^{(j)} \leq H_i^{(n)}$. □

Lemma 1 leads to the following definition of customers dominance: $c_j, j \geq 1$, is called *dominating* if $S^{(j)} \geq S^{(k)}, k = 1, 2, \ldots, j$. Otherwise it is called *dominated*.

By this definition, c_1 is dominating. Also, every customer $c_j, j > 1$, has a predominating customer. $c_n, 1 \leq n < j$, is the *predominating customer* of c_j if it is the highest indexed dominating customer among $c_1, c_2, \ldots, c_{j-1}$.

THEOREM 2. KEY. *Consider $c_j, j \geq 1$ whose predominating customer in the case $j > 1$, is c_n.*

(a) *If c_j is dominated, then*

$$H_i^{(j)} \leq H_i^{(n)},$$

and equivalently

$$d_i^{(j)} - d_i^{(n)} \leq d_0^{(j)} - d_0^{(n)}, \qquad i = 1, 2, \ldots, M, \tag{29}$$

where (29) *implies that for JIT arrivals,*

$$d_i^{(j)} - d_i^{(n)} \leq \frac{1}{r_0} \sum_{m=n+1}^{j} S^{(m)}. \tag{30}$$

(b) *If c_j is dominating, then $w_i^{(j)} = 0,\ i = 1, 2, \ldots, M$ and, consequently,*

$$h_i^{(j)} = \frac{S^{(j)}}{r_i},$$

and

$$H_i^{(j)} = S^{(j)} \sum_{m=1}^{i} \frac{1}{r_m}, \qquad i = 1, 2, \ldots, M. \tag{31}$$

(c) *If $r_i \geq r_{i-1}$ for some $1 \leq i \leq M$, then $h_i^{(j)} \leq \max(S^{(n)}, S^{(j)})/r_i$.*

PROOF. Relation (29) of Part (a) follows immediately from Lemma 1 and the definition of a predominating customer. Relation (30) follows from (29) and (10).

To prove (b) we note that, at time $t = 0$, all servers and their buffers are empty, except possibly for the buffer of s_0. Therefore (b) is true for c_1, which is the first dominating customer. From here we proceed by induction: Assume (b) is true for c_n, the predominant customer of c_j; then we have from Part (a) and from (31), respectively, that $H_i^{(j-1)} = d_i^{(j-1)} - d_0^{(j-1)} \leq H_i^{(n)}$ and $H_i^{(n)} = S^{(n)} \sum_{m=1}^{i} 1/r_m \Rightarrow$

$$d_i^{(j-1)} - d_0^{(j-1)} \leq S^{(n)} \sum_{m=1}^{i} \frac{1}{r_m}. \tag{32}$$

However, from the basic definitions (10), (11), and (12), we have $d_{i-1}^{(j)} - d_0^{(j)} \geq S^{(j)} \sum_{m=1}^{i-1} 1/r_m$ and $d_0^{(j)} - d_0^{(j-1)} \geq S^{(j)}/r_0 \Rightarrow$

$$d_{i-1}^{(j)} - d_0^{(j-1)} \geq S^{(j)} \sum_{m=0}^{i-1} \frac{1}{r_m}. \tag{33}$$

Subtracting (32) from (33) yields

$$d_{i-1}^{(j)} - d_i^{(j-1)} \geq (S^{(j)} - S^{(n)}) \sum_{m=1}^{i-1} \frac{1}{r_m} + \frac{S^{(j)}}{r_0} - \frac{S^{(n)}}{r_i}. \tag{34}$$

But c_j dominating $\Rightarrow S^{(j)} \geq S^{(n)}\ \forall n \leq j$. Also, $r_0 \leq r_j$ and therefore the right-hand side of (34) is nonnegative and $d_{i-1}^{(j)} \geq d_i^{(j-1)} \Rightarrow w_i^{(j)} = 0,\ \ i = 1, 2, \ldots, M$, which completes the proof of Part (b).

Part (c) follows from parts (a) and (b). If c_j is dominating, then from (b) we have $h_i^{(j)} = S^{(j)}/r_i \geq S^{(n)}/r_i$. For the case that c_j is dominated, we note that s_{i-1} and s_i can be viewed as a system with $M = 2$ servers, where $s_{i-1} \equiv s_0$ and $s_i \equiv s_1$ and $r_0 \geq r_1$. In this two-server system, $H_i^{(j)}$ is equal to $h_i^{(j)}$ of our original system. From (a) $\Rightarrow h_i^{(j)} \leq h_i^{(n)}$ and from (b) $\Rightarrow h_i^{(n)} = S^{(n)}/r_i > S^{(j)}/r_i$. Therefore,

$$h_i^{(j)} \leq \frac{\max(S^{(n)}, S^{(j)})}{r_i}. \quad \square \tag{35}$$

In essence, this theorem indicates the existence of a regeneration property. In the case where $S^{(1)},\ S^{(2)}, \ldots$ are i.i.d. and the input is JIT, time points at which customers requiring service of length b depart from s_0 are regeneration points for the queueing process of all later customers.

4. Upper bounds on holding times and queue sizes. The bounds provided in the following make use of the results of the previous section, and Theorem 2 in particular.

Corollary 4: Upper Bound on Total Holding Times. *Let $S^*(j)$ be the service requirement of the highest numbered dominating customer among $c_1, c_2, \ldots, c_j$, i.e., $S^*(j) = \max(S_1, S_2, \ldots, S_j),\ j \geq 1$, then*

$$H_i^{(j)} \leq S^*(j) \sum_{m=1}^{i} \frac{1}{r_m} \leq b \sum_{m=1}^{i} \frac{1}{r_m}, \qquad i \geq 1. \tag{36}$$

This bound follows from Parts (a) and (b) of Theorem 2.

Corollary 5: Upper Bound on Unprocessed Work at Departure Points. *Let $U_i^*(j) = \max_{k \leq j} U_i(d_i^{(k)})\},\ i \geq 1,\ j \geq 1$; then*

$$U_i^*(j) \leq r_0 \max_{k \leq j} \{H_i^{(k)}\} \leq S^*(j) \sum_{m=1}^{i} \frac{r_0}{r_m} \leq b \sum_{m=1}^{i} \frac{r_0}{r_m}, \tag{37}$$

where for JIT arrivals the first inequality is an equality.

Note that, except for the case where $r_i = r_0$, the function $U_i(t)$ cannot reach a maximum at a point of departure from s_i. Nevertheless, both $N_i(t)$ and $n_i(t)$ have drops of 1 at such points and are nondecreasing elsewhere. Both must, therefore, attain maxima immediately before these epochs.

Proof. The unprocessed work by server i at time $d_i^{(k)}$ is the amount of work processed by s_0 during the interval $(d_0^{(k)}, d_i^{(k)})$, which is of length $H_i^{(k)}$. Hence,

$$U_i(d_i^{(k)}) \leq r_0 H_i^{(k)}, \qquad \forall i,\ k \geq 1, \tag{38}$$

where under JIT input the equality sign holds. From Corollary 6, it follows that

$$U_i(d_i^{(k)}) \leq S^*(j) \sum_{m=1}^{i} \frac{r_0}{r_m}, \tag{39}$$

which completes the proof. $\square$

COROLLARY 6: UPPER BOUND ON UNPROCESSED WORK. *Let* $\hat{S}(t) = \max\{S^{(j)}: d_0^{(j)} < t\}$, $t \geq 0$; *then*

$$U_i^*(t) = \max_{0 \leq x \leq t}(U_i(x)) \leq \hat{S}(t) \sum_{m=0}^{i-1} \frac{r_0}{r_m} \leq b \sum_{m=0}^{i-1} \frac{r_0}{r_m}. \tag{40}$$

PROOF. $U_i(x)$ can strictly increase only when s_i is idle (see relation (5)). The only points in $[0, t]$ where it can attain a maximum are the end of the interval, t, and ends of idle periods of s_i which must also be departure points from s_{i-1}. Hence,

$$U_i^*(t) \leq \max_{\{k: d_0^{(k)} < t\}} \left\{U_{i-1}(d_{i-1}^{(k)}) + S^{(k)}\right\}, \qquad t \geq 0. \tag{41}$$

From Part (b) of Theorem 2 and Corollary 5 we get relation (40). □

Defining

$$U_i^* = \max_{t \geq 0} U_i^*(t), \qquad i \geq 1, \tag{42}$$

Corollary 6 provides the bound

$$U_i^* \leq b \sum_{m=0}^{i-1} \frac{r_0}{r_m}, \qquad i \geq 1, \tag{43}$$

where for JIT arrivals the equality sign holds. This bound is tighter than the one obtained in Theorem 1, relation (6).

COROLLARY 7. UPPER BOUND ON NUMBER IN THE SYSTEM. *Let* $N_i^*(j) = \max_{k \leq j} \{N_i(d_i^{(k)-})\}$, $j \geq 1$; *then*

$$N_i^*(j) \leq \left\lceil \frac{U_i^*(j)}{a} \right\rceil \leq \left\lceil \frac{S^*(j)}{a} \sum_{m=1}^{i} \frac{r_0}{r_m} \right\rceil \leq \left\lceil \frac{b}{a} \sum_{m=1}^{i} \frac{r_0}{r_m} \right\rceil. \tag{44}$$

This result follows from Corollary 5 and from a and b being the shortest and the longest possible service requirements, respectively.

Defining

$$N_i^* = \max_j (N_i^*(j)), \tag{45}$$

we have

$$N_i^* \leq \left\lceil \frac{b}{a} \sum_{m=1}^{i} \frac{r_0}{r_m} \right\rceil. \tag{46}$$

This upper bound, which is tighter than the one obtained in Corollary 1, will be realized in the event of an arrival with service requirement of size b followed by a consecutive JIT sequence of arrivals with service requirement of size a.

COROLLARY 8: UPPER BOUND ON UNPROCESSED WORK AND QUEUE SIZE WHEN $r_i \geq r_{i-1}$. *Define*

$$\begin{aligned} u_i^*(j) &= \max_{k \leq j} \{u_i(d_i^{(k)})\}, \qquad i \geq 1, \\ n_i^*(j) &= \max_{k \leq j} \{n_i(d_i^{(k)-})\}, \qquad i \geq 1, \end{aligned} \tag{47}$$

and assume $r_i \geq r_{i-1}$. *Then,*

$$u_i^*(j) \leq \frac{S^*(j) r_{i-1}}{r_i} \leq b \frac{r_{i-1}}{r_i}, \qquad i \geq 1, \tag{48}$$

$$n_i^*(j) \leq \left\lceil \frac{S^*(j)}{a} \frac{r_{i-1}}{r_i} \right\rceil \leq \left\lceil \frac{b}{a} \frac{r_{i-1}}{r_i} \right\rceil, \qquad i \geq 1, \tag{49}$$

and there exist arrival patterns and values of j *for which*

$$\left\lceil \frac{S^*(j)}{a} \frac{r_0}{r_i} \right\rceil \leq n_i^*(j). \tag{50}$$

PROOF. Since in general $u_i(d_i^{(k)}) \leq r_{i-1} h_i^{(k)}$ the upper bounds in (48) and (49) follow immediately from Part (c) of Theorem 2. To prove the lower bound of (50), suppose c_n, $1 \leq n \leq j$ is dominating and is followed at s_0 by a JIT sequence of at least $\lceil (S^{(n)}/a)(r_0/r_i) \rceil$ customers whose service requirement is of size a. From Proposition 2, we know that the interdeparture times of these customers from s_{i-1} do not exceed their interdeparture time from s_0, which is by assumption equal to a/r_0. From Part (b) of Theorem 2 we have $h_i^{(n)} = S^{(n)}/r_i$; hence

$$n_i(d_i^{(n)-}) \geq \left\lceil \frac{S^{(n)}}{a} \frac{r_0}{r_i} \right\rceil \qquad \forall c_n \text{ dominating.} \tag{51}$$

Since $S^*(j)$ is the service requirement of the highest indexed dominating customer among $c_1, c_2, \ldots, c_j$, the lower bound of (50) follows. Note also that this lower bound does not require that $r_i \geq r_{i-1}$. □

Define

$$n_i^* = \max_{j \geq 1} n_i^*(j), \qquad i \geq 1; \tag{52}$$

then, if $r_i \geq r_{i-1}$, we have

$$\left\lceil \frac{b}{a} \frac{r_0}{r_i} \right\rceil \leq n_i^* \leq \left\lceil \frac{b}{a} \frac{r_{i-1}}{r_i} \right\rceil, \qquad i \geq 1, \tag{53}$$

where the second inequality holds for all arrival processes while the first inequality is true for some arrival processes.

n_i^* is the minimal buffer size needed for preventing blocking at s_i when counting the service position as part of the buffer. If all servers are arranged in increasing order of service rates, $r_0 \leq r_1 \leq \cdots \leq r_M$, the total buffer size, B, required in the worst case is

$$\sum_{i=1}^{M} \left\lceil \frac{b}{a} \frac{r_0}{r_i} \right\rceil \leq B \leq \sum_{i=1}^{M} \left\lceil \frac{b}{a} \frac{r_{i-1}}{r_i} \right\rceil \leq M \left\lceil \frac{b}{a} \right\rceil, \tag{54}$$

which is one of the main results listed in §1.

If $r_1, r_2, \ldots, r_m$ are close in value, the bounds on B, in such an ordering of the servers, increase approximately linearly with the number of servers M. Furthermore, if all rates are the same, we have from (54) that $n_i^* = \lceil b/a \rceil \Rightarrow B = M \lceil b/a \rceil$, which is in accord with Kelly (1982) and Ziedins (1993) results.

THEOREM 3: BOUND ON UNPROCESSED WORK AT s_i. *Define* $u_i^* = \max_{j \geq 1} u_i^*(j)$, $i \geq 1$. *Then, for* $i = 1, 2, \ldots, M$:

$$u_i^* \leq b + (b - a) \sum_{m=1}^{i-1} \frac{r_0}{r_m}. \tag{55}$$

PROOF. We first note that when $u_i(t)$ is strictly increasing, s_{i-1} must be busy and, furthermore, $u_i(t)$ will continue to increase until the next departure point from s_{i-1}. Therefore all maxima of $u_i(t)$ occur at points of departure from s_{i-1}. Similarly, $n_i(t)$ has jumps of $+1$ at points of departure from s_{i-1}, and is nonincreasing otherwise.

From the definition of unprocessed work, we have

$$U_i(d_{i-1}^{(j)}) = U_{i-1}(d_{i-1}^{(j)}) + u_i(d_{i-1}^{(j)}), \qquad i, j \geq 1, \tag{56}$$

where the value of $u_i(d_{i-1}^{(j)})$ is not affected by the arrival times and service requirements of $c_{j+1}, c_{j+2}, \ldots$.

From Corollary 6, we have the bound,

$$U_{i-1}(d_{i-1}^{(j)}) + u_i(d_{i-1}^{(j)}) \leq b \sum_{m=0}^{i-1} \frac{r_0}{r_m}. \tag{57}$$

Suppose for the contradiction that for some $j \geq 1$, $u_i(d_{i-1}^{(j)}) > b + (b-a)\sum_{m=1}^{i-1} r_0/r_m$ and the arrival process after c_j is JIT. In this case,

$$U_{i-1}(d_{i-1}^{(j)}) = r_0 H_{i-1}^{(j)} \geq r_0 \sum_{m=1}^{i-1} \frac{a}{r_m}, \tag{58}$$

which violates (57). By way of a contradiction the claim follows. □

COROLLARY 9: BOUNDS ON QUEUE SIZE.

$$\left\lceil \frac{b}{a}\frac{r_0}{r_i} \right\rceil \leq n_i^* \leq \left\lceil \frac{b}{a} + \frac{(b-a)}{a} \sum_{m=1}^{i-1} \frac{r_0}{r_m} \right\rceil, \qquad i \geq 1. \tag{59}$$

The upper bound follows from Theorem 3. The lower bound is the same as in (53), meaning that there exist arrival patterns for which it is true.

COROLLARY 10: IMPROVED UPPER BOUND ON QUEUE SIZE. *For $i \geq 1$, let $s_{i'}$, $0 \leq i' \leq i$ be the highest numbered server such that $r_{i'} \leq r_i$. Then,*

$$u_i^* \leq b + (b-a) \sum_{m=i'+1}^{i-1} \frac{r_{i'}}{r_m}, \tag{60}$$

and

$$n_i^* \leq \left\lceil \frac{b}{a} + \frac{b-a}{a} \sum_{m=i'+1}^{i-1} \frac{r_{i'}}{r_m} \right\rceil. \tag{61}$$

$s_{i'}$ is called the dominating server of s_i.

PROOF. The subsystem $s_{i'}, s_{i'+1}, \ldots, s_i$ satisfies all the requirements of the tandem system under study, namely, the first server is the slowest, the service times are deterministically correlated, and the buffer spaces are unlimited. Hence, (60) and (61) follow from Theorem 3 and Corollary 9, respectively. □

The upper bounds provided by Theorem 3 and its corollaries are useful only if $r_{i-1} > r_i$, since, in the complementary case where $r_{i-1} \leq r_i$, the upper bounds provided by Corollary 8 are lower. If $r_{i-1} > r_i$, the upper bound on n_i^* is the smaller of the two given in Corollaries 9 and 10, relations (59) and (61). Suppose then that the servers, except for s_0, are ordered in

decreasing service rate: $r_1 > r_2 > \cdots > r_M \geq r_0$. In this case, s_0 is the dominating server of s_i, $i = 1, 2, \ldots, M$, and the upper bound on the total buffer required for preventing blocking is

$$B \leq \sum_{i=1}^{M} \left\lceil \frac{b}{a} + \frac{b-a}{a} \sum_{m=1}^{i-1} \frac{r_0}{r_m} \right\rceil < \sum_{i=1}^{M} \left\lceil i \frac{b}{a} \right\rceil - \frac{M(M-1)}{2}, \tag{62}$$

which is another main result listed in §1.

If $r_1, r_2, \ldots, r_M$ are close in value, the rate of growth of this bound is $O(M^2)$, as compared to $O(M)$ for the reverse ordering. From Corollaries 8, 9, and 10, we know that in such a case, deviating from the ordering $r_1 > r_2 > \cdots > r_M$, will result in a reduction in the value of the upper bound on B. The case where the service rates are close in value is of special interest. In designing tandem systems it is often attempted to "balance" the system, and thus improve its performance by equalizing the service rates. In the following section, we will discuss a possible drawback of this approach.

5. Worst-case arrival process and scenario. In this section we characterize arrival patterns that impose the highest buffer requirements on the system. We show that the family of JIT arrivals can be characterized as being "worst." We demonstrate situations in which the upper bound on the buffer size is materialized, that is, $O(M^2)$ is indeed required.

An arrival process $A \equiv \{(d_0^{(j)}, S^{(j)}),\ j \geq 1\}$ is characterized by a sequence of departure times from s_0 and a sequence of customer service requirements. The arrival process $\hat{A} \equiv \{(\hat{d}_0^{(j)}, \hat{S}^{(j)}), j \geq 1\}$, where $\hat{S}^{(j)} = S^{(j)}$, $j \geq 1$, is worse than A with respect to holding time if $\hat{h}_i^{(j)} \geq h_i^{(j)}\ \forall i, j \geq 1$. $\hat{A}$ is worse than A with respect to queue sizes if $\hat{n}_i(\hat{d}_i^{(j)}) \geq n_i(d_i^{(j)}), \forall i, j \geq 1$. If at least one of the inequalities are strict then $\hat{A}$ is strictly worse than A. The process A is worst if there does not exist a process that is strictly worse than A. We note that the arrival process, as defined here, contains the necessary information for calculating any of the associated queueing processes, e.g., if $Q \equiv \{d_0^{(j)}, d_1^{(j)}, \ldots, d_M^{(j)}, S^{(j)}, j \geq 1\}$ then $d_i^{(j)}, j \geq 1, i \geq 1$ can be calculated from A using basic relation (15) repeatedly.

A worst-case scenario is a situation where B is either equal to, or is arbitrarily close to, the upper bound given in (62).

We proceed to show that a JIT process is the worst and to discuss worst-case scenarios.

PROPOSITION 3. *For each arrivals process* $A \equiv \{(d_0^{(j)}, S^{(j)}),\ j \geq 1\}$, *we define a perturbed arrivals process* $\tilde{A} \equiv \{(\tilde{d}_0^{(j)}, \tilde{S}^{(j)}),\ j \geq 1\}$, *such that* $\tilde{S}^{(j)} = S^{(j)}, \forall j \geq 1$; $\tilde{d}_0^{(j)} = d_0^{(j)}, j = 1, 2 \ldots, k-1$; $\tilde{d}_0^{(j)} = d_0^{(j)} + \varepsilon_0, j \geq k$ *for some* k, $k \geq 1$, *and* $\varepsilon_0 > 0$. *Then,*

(a) *The interdeparture times under* $\tilde{A}$ *are at least as long as under* A:

$$\tilde{d}_i^{(j+1)} - \tilde{d}_i^{(j)} \geq d_i^{(j+1)} - d_i^{(j)}, \qquad i, j \geq 1. \tag{63}$$

(b) A is worse than $\tilde{A}$, with respect to holding times:

$$\tilde{h}_i^{(j)} \leq h_i^{(j)}, \qquad i, j \geq 1, \tag{64}$$

(c) A is worse than $\tilde{A}$ with respect to queue sizes

$$n_i(d_i^{(j)}) \geq \tilde{n}_i(\tilde{d}_i^{(j)}), \qquad i, j \geq 1. \tag{65}$$

Note that regular notation refers to the system under A and notation with an added $\sim$ indicates the system under $\tilde{A}$. Note also that the perturbed process is obtained by increasing by ε_0 the departure times from s_0 of c_k and all later customers. The purpose of doing so will become apparent in the next theorem.

PROOF. Let us define $\varepsilon_i^{(j)} \equiv \tilde{d}_i^{(j)} - d_i^{(j)}$, $j \geq 1$, $0 \leq i \leq M$. To carry out the proof of the proposition we first establish the following claim:

CLAIM. (i) $\varepsilon_i^{(j)} \geq \varepsilon_{i+1}^{(j)}$, $j \geq 1$, $0 \leq i < M$, and (ii) $\varepsilon_i^{(j+1)} \geq \varepsilon_i^{(j)}$, $j \geq 1$, $0 \leq i \leq M$.

To prove the claim, note that it holds trivially for $j \leq k-1$, since $\varepsilon_i^{(j)}$ is equal to 0 for this range. Further, Part (ii) holds trivially also for $i = 0$ and $j \geq k$, since $\varepsilon_0^{(j)} = \varepsilon_0$ for $j \geq k$. The proof of the claim for $j \geq k$ is now carried out by a double induction: Assuming (i) holds for j, namely $\varepsilon_0^{(j)} \geq \varepsilon_1^{(j)} \geq \cdots \geq \varepsilon_M^{(j)}$, and (ii) holds for i and j, namely $\varepsilon_i^{(j+1)} \geq \varepsilon_i^{(j)}$, we have to prove (i) $\varepsilon_i^{(j+1)} \geq \varepsilon_{i+1}^{(j+1)}$, and (ii) $\varepsilon_{i+1}^{(j+1)} \geq \varepsilon_{i+1}^{(j)}$. Using the basic relation (15) we get:

$$\begin{aligned} \varepsilon_{i+1}^{(j+1)} &= \tilde{d}_{i+1}^{(j+1)} - d_{i+1}^{(j+1)} = \max\{\tilde{d}_{i+1}^{(j)}, \tilde{d}_i^{(j+1)}\} - \max\{d_{i+1}^{(j)}, d_i^{(j+1)}\} \\ &= \max\{\varepsilon_{i+1}^{(j)} + d_{i+1}^{(j)}, \varepsilon_i^{(j+1)} + d_i^{(j+1)}\} - \max\{d_{i+1}^{(j)}, d_i^{(j+1)}\}. \end{aligned} \tag{66}$$

Now, by the inductive assumption $\varepsilon_i^{(j+1)} \geq \varepsilon_i^{(j)} \geq \varepsilon_{i+1}^{(j)}$, which yields

$$\begin{aligned} \varepsilon_i^{(j+1)} + \max\{d_{i+1}^{(j)}, d_i^{(j+1)}\} &\geq \max\{\varepsilon_{i+1}^{(j)} + d_{i+1}^{(j)}, \varepsilon_i^{(j+1)} + d_i^{(j+1)}\} \\ &\geq \varepsilon_{i+1}^{(j)} + \max\{d_{i+1}^{(j)}, d_i^{(j+1)}\}. \end{aligned} \tag{67}$$

Substituting into (66), we get

$$\varepsilon_i^{(j+1)} \geq \varepsilon_{i+1}^{(j+1)} \geq \varepsilon_{i+1}^{(j)}, \tag{68}$$

which completes the proof of the claim.

The proof of the proposition now immediately follows: First, (a) of the proposition follows from (ii) of the claim. Second, (b) of the proposition follows from:

$$\tilde{h}_i^{(j)} = \tilde{d}_i^{(j)} - \tilde{d}_{i-1}^{(j)} = d_i^{(j)} - d_{i-1}^{(j)} + \varepsilon_i^{(j)} - \varepsilon_{i-1}^{(j)} = h_i^{(j)} + \varepsilon_i^{(j)} - \varepsilon_{i-1}^{(j)} \leq h_i^{(j)}, \tag{69}$$

in which the inequality follows from (i) of the claim.

Lastly, Part (c) of the proposition now follows from Parts (a) and (b): From (b) we have that $h_i^{(j)} \geq \tilde{h}_i^{(j)}$ and from (a) we have $d_i^{(j+1)} - d_i^{(j)} \leq \tilde{d}_i^{(j+1)} - \tilde{d}_i^{(j)}$, $i, j \geq 1$. Since the queue size at s_i immediately before the departure of c_j is equal to the number of departures from s_{i-1} during the holding time of c_j at s_i, then

$$n_i(d_i^{(j)-}) \geq \tilde{n}_i(\tilde{d}_i^{(j)-}), \qquad i, j \geq 1. \quad \square$$

We mention that we are presenting here an inductive, rather than constructive, proof of Proposition 3 because of its brevity. A more constructive proof, providing better insight into the problem, is available but is considerably lengthier. The constructive proof builds on showing that after the perturbation, a customer cannot be delayed at a server at which it is not delayed before the perturbation.

THEOREM 4: WORST-CASE ARRIVALS. *The JIT arrivals process* $A \equiv \{(d_0^{(j)}, S^{(j)}), j \geq 1\}$ *is worst with respect to holding times and queue sizes.*

PROOF. Let $\hat{A} \equiv \{(\hat{d}_0^{(j)}, S^{(j)}), j \geq 1\}$ be any arrivals process with service requirements the same as in process A. Then $\hat{d}_0^{(j)} \geq d_0^{(j)}$, $j \geq 1$. Perturb A by increasing $d_0^{(j)}$, $j \geq 1$, by $\varepsilon_0 = \hat{d}_0^{(1)} - d_0^{(1)} = \hat{d}_0^{(1)} - S^{(1)}/r_0$ to produce $\tilde{A}_1$. From Proposition 3, A is worse than $\tilde{A}_1$. Perturb $\tilde{A}_1$ by increasing the departure times of c_j from s_0, $j \geq 2$, by $\varepsilon_0 = \hat{d}_0^{(2)} - \hat{d}_0^{(1)} - S^{(2)}/r_0$ to produce $\tilde{A}_2$. From Proposition 3, $\tilde{A}_1$ is worse than $\tilde{A}_2$. After n perturbations, the process $\tilde{A}_n$ is identical to A up to and including the nth arrival, since the "worse" relation is transitive, then A is worse than $\hat{A}$ up to any number n of arrivals. $\square$

We note that, using virtually the same proof, Theorem 4 holds for FCFS tandem systems with arbitrary service times.

Clearly, from Theorem 4 it follows that if a worst-case scenario exists, the arrival process involved must be JIT.

In the following, we show that a worst-case scenario is obtainable when servers are ordered in decreasing service rates and an arrival with service requirement b is followed at s_0 by a JIT string of arrivals with service requirement a.

THEOREM 5: WORST-CASE SCENARIO. *Assume JIT arrivals, and decreasing service rate ordering* $r_1 > r_2 > \cdots > r_M > r_0$, *and let* $S^{(1)} = b$ *and* $S^{(j)} = a \forall\, j > 1$; *then,*

(a) *There exist finite integers* $1 < n_0 \le n_1 \le n_2 \le \cdots \le n_M$ *such that* n_i *is the smallest integer, greater than* 1, for which $w_i^{(j)} = 0, \forall\, j \ge n_i$.

(b) The value of the total buffer required, B, can be made to be arbitrarily close to the upper bound given in (62).

PROOF. We assume for convenience that service of c_1 at s_0 starts at time $t = 0$, $w_0^{(j)} = 0, \forall\, j \ge 2$, and to avoid the trivial case where $n_i = 2$ for all or some values of $i > 0$; we also assume that $b/r_1 > a/r_0$. (Note that $n_0 = 2$.)

(a) c_{n_i} is the first customer, excluding c_1, whose waiting time at s_i is zero. Therefore, $n_i, i = 1, 2, \ldots, M$, must be the smallest integer j satisfying

$$d_{i-1}^{(j)} \ge d_i^{(j-1)}, \qquad j \ge 2. \tag{70}$$

For $i = 1$, we have

$$d_0^{(j)} = \frac{(b+(j-1)a)}{r_0} \ge d_1^{(j-1)} = b\left(\frac{1}{r_0} + \frac{1}{r_1}\right) + \frac{(j-2)a}{r_1}, \tag{71}$$

where the right-hand side is justified, since $c_2, \ldots, c_{j-1}$ are all delayed at s_1. This yields

$$n_1 = \left\lceil 1 + \frac{b-a}{a}\frac{r_0}{r_1 - r_0} \right\rceil \ge 2 = n_0. \tag{72}$$

Since $r_1 > r_0$, we have $w_1^{(j)} = 0, \forall\, j > n_1$. To prove for $i > 1$, assume (a) is true up to $i-1$. Then $c_j, j \ge n_{i-1}$ is not delayed in $s_1, s_2, \ldots, s_{i-1}$ and Condition (70) takes the form

$$\frac{1}{r_0}(b+(j-1)a) + a\sum_{m=1}^{i-1}\frac{1}{r_m} \ge b\sum_{m=0}^{i}\frac{1}{r_m} + \frac{(j-2)a}{r_i}, \tag{73}$$

which yields

$$n_i = 1 + \left\lceil \frac{b-a}{a}\frac{r_0 r_i}{r_i - r_0}\sum_{m=1}^{i}\frac{1}{r_m} \right\rceil. \tag{74}$$

For $n_i \ge n_{i-1}$, it is sufficient that

$$\frac{r_i}{r_i - r_0}\sum_{m=1}^{i}\frac{1}{r_m} \ge \frac{r_{i-1}}{r_{i-1} - r_0}\sum_{m=1}^{i-1}\frac{1}{r_m}. \tag{75}$$

This inequality can be rewritten as

$$\left(\frac{r_i}{r_i - r_0} - \frac{r_{i-1}}{r_{i-1} - r_0}\right)\sum_{m=1}^{i-2}\frac{1}{r_m} + \left(\frac{1}{r_i - r_0} - \frac{1}{r_{i-1} - r_0}\right) + \frac{r_i/r_{i-1}}{r_i - r_0} \ge 0. \tag{76}$$

Since $r_{i-1} > r_i > r_0$, each one of the three expressions in (76) is positive, which concludes the proof of (a).

We note that n_i increases at the rate of $1/(r_i - r_0)$, and is large when r_i is close to r_0.

(b) The function $u_i(t), i \geq 2$, equals zero in $[0, d_{i-1}^{(1)} - b/(r_{i-1})]$ and is increasing with slope $(r_{i-1} - r_i)$ in $[d_{i-1}^{(1)} - b/(r_{i-1}), d_{i-1}^{(n_{i-1}-1)}]$. It is easily shown that

$$u_i(t_i) = b + a(n_{i-1} - 2)\frac{r_{i-1} - r_i}{r_{i-1}}, \qquad i \geq 2, \tag{77}$$

where, for compactness of notation, $t_i = d_{i-1}^{(n_{i-1}-1)}$.

From (74), we have

$$n_{i-1} \geq 1 + \frac{b-a}{a}\frac{r_0 r_{i-1}}{r_{i-1} - r_0}\sum_{m=1}^{i-1}\frac{1}{r_m}, \qquad i \geq 2, \tag{78}$$

which when substituted in (77) yields

$$u_i(t_i) \geq b - a\frac{r_{i-1} - r_i}{r_{i-1}} + (b-a)\frac{r_{i-1} - r_i}{r_{i-1} - r_0}\sum_{m=1}^{i-1}\frac{r_0}{r_m}, \qquad i \geq 2. \tag{79}$$

Assuming r_1 close to r_0, the values of $n_i, i \geq 1$, are large and at time $t_i, i = 2, 3, \ldots, M$, the first customer is already departed from s_i. Hence,

$$n_i^* \geq \left\lceil \frac{b}{a} - \frac{r_{i-1} - r_i}{r_{i-1}} + \frac{(b-a)}{a}\frac{r_{i-1} - r_i}{r_{i-1} - r_0}\sum_{m=1}^{i-1}\frac{r_0}{r_m} \right\rceil, \qquad i \geq 2, \tag{80}$$

and

$$n_1^* \geq \left\lceil \frac{b}{a}\frac{r_0}{r_1} \right\rceil. \tag{81}$$

The right-hand side of (80) goes to the upper bound given in (59) as $r_i \to r_0$. Furthermore, the total buffer required in this scenario is

$$B \geq \left\lceil \frac{b}{a}\frac{r_0}{r_1} \right\rceil + \sum_{i=2}^{M}\left\lceil \frac{b}{a} - \frac{r_{i-1} - r_i}{r_{i-1}} + \frac{(b-a)}{a}\frac{r_{i-1} - r_i}{r_{i-1} - r_0}\sum_{m=1}^{i-1}\frac{r_0}{r_m} \right\rceil. \tag{82}$$

Defining $\delta_i = r_i - r_0, i \geq 1$, and letting $\delta_1 \to 0$ while $\delta_i/(\delta_{i-1}) \to 0, i = 2, 3, \ldots, M$ results in the right-hand side of (82) going to the right-most upper bound given in (62). □

6. Discussion. The major objective of this paper is to deal with situations where the arrivals pattern is hard to predict and the system must be designed to cope with worst case arrival patterns, or at least, not to be overly sensitive to such patterns. Two of the paper's results are quite meaningful for this purpose. The first, which may not be surprising, is that intermediate buffer size required will be relatively small, and not highly sensitive to worst case arrival pattern if slower servers are placed upstream and faster ones are placed downstream. Deviations from this type of ordering may be costly. The second major result, which applies to the common situation where the servers have close service rates, and is perhaps a rather surprising result, is that such systems, though effective from the servers utilization viewpoint, are highly unstable as far as buffer requirements are concerned. In many situations, the parameters $r_1, r_2, \ldots, r_M$ may undergo slight fluctuations in time and trying to design the system with parameters too close in value may result in stretches of time during which the worst case scenario is realized, for at least parts of the system. Since the parts of the systems affected may not be predictable and/or may be different at different stretches of time, the intermediate buffer required may be quite large. It might therefore be advantageous to have the servers slightly (but distinguishably) of increasing service rates by design.

References

Avi-Itzhak, B. 1965. A sequence of service stations with arbitrary input and regular service times. *Management Sci.* **11** 565–571.

Avi-Itzhak, B., S. Halfin. 1993. Servers in tandem with communication and manufacturing blocking. *J. Appl. Probab.* **31** 1061–1069.

Avi-Itzhak, B., H. Levy. 1995. A sequence of servers with arbitrary input and regular service times revisited. *Management Sci.* **41** 1039–1048.

Ding, J., B. S. Greenberg. 1991. Optimal order for servers in series with no queue capacity. *Probab. Engrg. Inform. Sci.* **5** 449–461.

Friedman, H. D. 1965. Reduction methods for tandem queueing systems. *Oper. Res.* **13** 121–131.

Huang, C. C., G. Weiss. 1988. Optimal order of *M* machines in tandem. *Oper. Res. Lett.* **9** 299–303.

Kelly, F. P. 1982. The throughput of a series of buffers. *Adv. Appl. Probab.* **14** 633–653.

———. 1984. Blocking, reordering, and the throughput of a series of queues. *Stochastic Proc. Appl.* **17** 327–336.

———. 1985. Segregating the input of a series of buffers. *Math. Oper. Res.* **10** 33–43.

Kim-Winch, J., B. Avi-Itzhak. 1995. Ordering of tandem constant-service stations to minimize in-process stock cost. *Probab. Engrg. Inform. Sci.* **9** 457–473.

Pinedo, M. 1982. On the optimal order of stations in tandem queues. *Appl. Probab.-Comput. Sci.: The Interface, Vol. II.* Birk Houser, Boston, MA, 307–325.

Whitt, W. 1985. The best order of queues in series. *Management Sci.* **31** 475–487.

Yamazaki, G., H. Sakasegawa, J. G. Shantikumar. 1992. On optimal arrangement of stations in a tandem queueing system with blocking. *Management Sci.* **38** 137–153.

Ziedins, I. 1993. Tandem queues with correlated service times and finite capacity. *Math. Oper. Res.* **18** 901–915.

B. Avi-Itzhak: RUTCOR and School of Business, Rutgers University, Piscataway, NJ 08854-8003; e-mail: aviitzha@rutcor.rutgers.edu

H. Levy: School of Computer Science, Tel Aviv University, Ramat Aviv, Tel Aviv 69978 Israel; e-mail: hanoch@cs.tau.ac.il