COMPARISONS OF SERVICE DISCIPLINES IN A TANDEM QUEUEING NETWORK WITH DELAY DEPENDENT CUSTOMER BEHAVIOR

D. Towsley and Francois Baccelli

COINS Technical Report 89-63

July 1, 1989

Comparisons of Service Disciplines in a Tandem Queueing Network with Delay Dependent Customer Behavior

Don Towsley* François Baccelli†

July 1, 1989

Abstract

In this paper we study the extremal properties of the stationary customer lag times in tandem G/GI/1 networks under different service disciplines in terms of convex and increasing convex orderings. Each customer carries a reference time with it and the lag time is defined to be the difference between the time that the customer departs from the system and its reference time. We show that among the class of work conserving nonpreemptive service disciplines that are service time indepedent, the service discipline that schedules customers with the smallest reference times (SR) and the service discipline that schedules the customer with the largest reference time (LR) provide the minima and maxima respectively. If we restrict ourselves to the subset of these disciplines that do not use reference times but do use arrival times in making scheduling decisions, then the FIFO and LIFO service disciplines provide the minima and maxima respectively. We also present similar results for G/M/1 queue when preemptions are allowed and for the class of service disciplines that are not work conserving.

*Department of Computer & Information Science, University of Massachusetts, Amherst, MA 01003 (U.S.A.)
The work of this author was supported by NSF under contract ASC 88-8802764, DCR 85-00332 and ONR under contract N00014-87-K-0796.
†INRIA-Sophia 06565 Valbonne (France)

1 Introduction and Summary

Numerous authors have studied the convex ordering properties of the FIFO and LIFO service disciplines in the context of the G/GI/1 queue [1,3,5] and the G/GI/c queue [1,6]. For the G/GI/1 queue, the following inequalities have been established,

$$T(FIFO) \leq_c T(\theta) \leq_c T(LIFO)$$

and for the G/GI/c queue,

$$W(FIFO) \leq_c W(\theta) \leq_c W(LIFO)$$

where $T(\theta)$ and $W(\theta)$ are the stationary sojourn time and wait time, respectively, under policy θ taken from the class of nonpreemptive work conserving disciplines that do not use service time information and where $X \leq_c Y$ means that the random variable X is smaller in the sense of convex ordering than the random variable Y, i.e., $E[f(X)] \leq E[f(Y)]$ for all convex functions f such that the mathemetical expectation exists. The interest of this ordering stems in particular from the fact that $X \leq_c Y$ implies $E[X] = E[Y]$, $E[X^r] \leq E[Y^r]$, r even (also odd values of r when X and Y are non-negative), and $\sigma_X^2 \leq \sigma_Y^2$.

In this paper we generalize these results to tandem queueing networks. Furthermore, we focus on a different random variable, *the customer lag time.* We assume that there is associated with each customer a *reference time* and define the lag time to be the difference betwen it and the customer departure time, i.e., a measure of how close the departure time is to the reference time. In the literature of real-time computing systems, the reference time is often referred to as a *soft real-time deadline.*

Specifically, we establish convex ordering results between the stationary lag times for service disciplines that are work conserving, nonpreemptive and that do not use service time information but may use reference time information. In this case we show

$$L(SR) \leq_c L(\Theta) \leq_c L(LR)$$

where Θ is a collection of non preemptive, work conserving disciplines at the M queues, none of which use service time information and SR and LR are such policies that schedule customers with the smallest reference times and largest reference times respectively. In addition, we

establish the following result for a subset of the above policies which are not allowed to use reference time information

$$L(FIFO) \leq_c L(\Theta) \leq_c L(LIFO).$$

This generalizes results in [2] which showed that SR minimizes the variance of the lag time in the case of a single G/D/1 queue.

The model and notation is presented in the next section. The proofs of the ordering properties are found in Section 3. Some remarks describing the analogous results for the class of preemptive policies and the class of non-work conserving policies are also found at the end of this section. Last, these results are applied to a system where customers have soft real-time deadlines. Under the assumption that customers must complete service, we show that out of the class of non-preemptive, work conserving, service independent, reference independent policies, LIFO maximizes the fraction of customers that complete by their deadlines provided that the deadlines satisfy certain concavity properties.

2 Model and Notation

We consider a tandem network of M queues, each with a single server being fed by an exogenous stream of arriving customers. This exogenous stream of customers has pattern $a_1 = 0 < a_2 < \cdots < a_n < \cdots \in IR^+$, where a_i is the arrival time of the i-th customer. Associated with customer i is a *reference time* y_i, $1 \leq i$ whose importance will be described later. We define $r_i = y_i - a_i$ to be the *relative reference time associated with customer* i. In the sequel, these relative reference times will be assumed to form a stationary and ergodic sequence of random variables. We associate with queue j, $1 \leq j \leq M$, a sequence $\{\sigma_n^j\}_1^\infty$, where $\sigma_n^j \in IR^+$ represents the service time requirement of the n-th customer to receive service at the j-th queue. Let θ_j denote the service discipline used at queue j. We consider several classes of disciplines from which θ_j can be chosen.

- Σ_0 - class of work conserving non preemptive policies that do not use information regarding either service time or reference times.
- Σ_1 - superset of Σ_0 that uses information regarding reference times.

We use the notation $\Theta = (\theta_1, \cdots, \theta_M)$ to denote the collection of scheduling policies at all nodes. We define $T(\Theta)$ to be the stationary customer sojourn time under policy Θ when it exists. We are interested in the *stationary lag time* $L(\Theta) = T(\Theta) - r$.

Remark. When $r_i = 0$ we get $L(\Theta) = T(\Theta)$, and the lag times boil down to classical sojourn times. In other contexts, y_i may be interpreted as the time a customer turns *bad*, [3,4].

We define the following two scheduling policies:

- SR - this is the policy in $(\Sigma_1)^M$ that schedules the customer in each queue with the smallest reference time.
- LR - this is the policy in $(\Sigma_1)^M$ that schedules the customer in each queue with the largest reference time.

The discussion will be conducted under the following set of assumptions

H:

1. *The service times, the interarrival times and the reference times are three mutually independent sequences of random variables.*
2. *The service times are i.i.d.*
3. *The interarrival times and the reference times sequences are stationary and ergodic.*

With these notations, the main results of the paper read: Under assumption H,

$$L(FIFO) \leq_c \quad L(\Theta) \quad \leq_c L(LIFO), \quad \Theta \in (\Sigma_0)^M, \tag{1}$$

$$L(SR) \leq_c \quad L(\Theta) \quad \leq_c L(LR), \quad \Theta \in (\Sigma_1)^M. \tag{2}$$

We introduce some additional notations before stating and proving relations (1) and (2).

Let d_n^i denote the time of the n-th departure from queue i. Let $I_n(\Theta)$ be the position in the departure stream of the last queue of the n-th customer to arrive to the network when the scheduling policy is Θ. We have

$$L_n(\Theta) = d^M_{I_n(\Theta)}(\Theta) - y_n \tag{3}$$

where

$$d_n^k = \begin{cases} a_n, & k = 0, \\ \max(d_{n-1}^k, d_n^{k-1}) + \sigma_n^k, & k = 1, \cdots, B. \end{cases} \tag{4}$$

3 The Convex Ordering Result

We define the following two convex programming problems. Consider a tandem system in which the n-th customer arrives at time a_n with relative reference time r_n, $1 \leq n \leq N$. The n-th customer to receive service at queue i is given service time σ_n^i, $1 \leq i \leq M$; $1 \leq n \leq N$. Let

$$F_N(\Theta) = \sum_{i=0}^{N} f(L_i(\Theta)) \tag{5}$$

where f is a convex function on the real numbers, $f : IR \rightarrow IR$. The problems are

(P1) minimize $E[F_N(\Theta)]$
subject to $\Theta \in (\Sigma_1)^M$.

(P2) maximize $E[F_N(\Theta)]$
subject to $\Theta \in (\Sigma_1)^M$.

We show that the policies that solve P1 and P2 are SR and LR respectively. In order to do so, we make use of the following result.

Lemma 1 *For all convex functions $f : IR^+ \rightarrow IR^+$, and all real valued vectors (x_1, y_1) and (x_2, y_2) with $x_1 < y_1$ and $x_2 < y_2$, if $(x_1, y_1) \leq (x_2, y_2)$ (componentwise) and $x_2 \leq y_1$, then*

$$f(y_1 - x_1) + f(y_2 - x_2) \leq f(y_1 - x_2) + f(y_2 - x_1). \tag{6}$$

Proof. Observe that

$$y_1 - x_2 \leq y_1 - x_1 \leq y_2 - x_1$$

$$y_1 - x_2 \leq y_2 - x_2 \leq y_2 - x_1.$$

Inequality (6) follows from the convexity of f. □

Theorem 1 *Under the assumptions H, the policy SR solves problem P1.*

Proof. Assume that Θ_0 solves problem P1 and that it is not SR. If there is no unique optimum policy, then we choose Θ_0 so that it maximizes the index i corresponding to the first queue at which Θ_0 differs from SR. If there is more than one such policy, then we choose Θ_0 so that it differs from SR at queue i at the latest possible moment, say at the j-th customer scheduling. Now there may be many different values of arrival times, service times, and reference times for which Θ_0 exhibits this behavior. Choose one such sample path $\{a_n\}_{n=1}^{\infty}$, $\{\sigma_n^j : j = 1, \cdots M\}_{n=1}^{\infty}$, and $\{y_n\}_{n=1}^{\infty}$. Let $s_j^i(\Theta)$ denote the index of the customer (in the order of arrival) served in the j-th position at the i-th queue, $i = 1, \cdots, M$, $j = 1, \cdots, N$, for policy Θ. Suppose that SR schedules customer $s_j^i(SR) = l$ whereas Θ_0 schedules customer $s_j^i(\Theta_0) = l_0$ at queue i. Let $s_{j'}^i(\Theta_0) = l$ for some $j < j' \leq N$. We define a new policy Θ_1 that behaves exactly like Θ_0 at the first i queues except that $s_j^i(\Theta_1) = l$ and $s_{j'}^i(\Theta_1) = l_0$. In addition, we consider the version of this policy where the service times of these two customers are interchanged. This version of the policy is equivalent in law to the version without interchange due to our assumption that the service times are mutually independent i.i.d. sequences and are independent of the arrival and the reference times sequences. Let $m_{i+1}, \cdots, m_M$ and $n_{i+1}, \cdots, n_M$ be two sequences of indices such that $s_{m_p}^p(\Theta_0) = l$ and $s_{n_p}^p(\Theta_0) = l_0$, $p = i+1, \cdots, M$. We define Θ_1 to behave the same as Θ_0 at queues $i+1, \cdots, M$ with the exception that $s_{m_p}^p(\Theta_1) = l_0$ and $s_{n_p}^p(\Theta_1) = l$ for all $i+1 \leq p \leq M$ such that $m_p < n_p$. As for queue i, for each such permutation, we perform an additional interchange of the service times of the concerned customers. These interchanges do not change the law of $F(\Theta_1)$ for the same reasons as above.

Let us consider the effects of these scheduling and service changes on $F(\Theta_1)$. If we consider the departure times of all customers under Θ_0 and Θ_1 we observe that they are unchanged, with the possible exception of customers l and l_0. If these are unchanged, then $F(\Theta_0) = F(\Theta_1)$. If the departure times of l and l_0 are changed, then it follows from Lemma 1 that $F(\Theta_1) \leq F(\Theta_0)$. This construction can be applied to all sample paths such that Θ_0 deviates from SR for the first time at position j at queue i to yield a policy Θ_1 for which either $E[F(\Theta_0)] > E[F(\Theta_1)]$ or $E[F(\Theta_0)] = E[F(\Theta_1)]$ but where Θ_1 differs from SR for the first time either at some queue $i' > i$ or at position $j' > j$ at queue i. both of these contradict our initial hypothesis and we conclude that SR solves P1. □

Theorem 2 *Under the assumptions II, the policy LR solves the problem P2.*

Proof. The proof is similar to that of theorem 1. □

Theorem 3 *Under the assumptions H, for every* $\Theta \in (\Sigma_1)^M$ *such that the sequences* $L_i(SR)$, $L_i(LR)$, *and* $L_i(\Theta)$ *converge weakly towards finite random variables that will be denoted by* $L(SR)$, $L(LR)$, *and* $L(\Theta)$ *respectively, we have*

$$L(SR) \leq_c L(\Theta) \leq_c L(LR)$$

Proof. As a consequence of the previous theorems,

$$\frac{\sum_{i=0}^{N} E[f(L_i(SR))]}{N+1} \leq \frac{\sum_{i=0}^{N} E[f(L_i(\Theta))]}{N+1} \leq \frac{\sum_{i=0}^{N} E[f(L_i(LR))]}{N+1}, \quad \forall\, \Theta \in (\Sigma_1)^M.$$

the assumption that both $L_i(SR)$, $L_i(\Theta)$ and $L_i(LR)$ converge weakly allows one to conclude that

$$\begin{aligned}
\lim_{N\to\infty} \frac{\sum_{i=0}^{N} E[f(L_i(SR))]}{N+1} &= E[f(L(SR))], \\
\lim_{N\to\infty} \frac{\sum_{i=0}^{N} E[f(L_i(LR))]}{N+1} &= E[f(L(LR))], \\
\lim_{N\to\infty} \frac{\sum_{i=0}^{N} E[f(LW_i(\Theta))]}{N+1} &= E[f(L(\Theta))].
\end{aligned}$$

Hence $E[f(L(SR))] \leq_c E[f(L(\Theta))] \leq_c E[f(L(LR))]$ and the theorem holds. □

The following result is a direct application of the theorem by setting $y_i = a_i$, $i = 0, 1, 2, \cdots$.

Corollary 1 *Under the assumptions H, we have*

$$T(FIFO) \leq_c T(\Theta) \leq_c T(LIFO), \quad \forall \Theta \in (\Sigma_1)^M$$

The following result can be proven in a manner similar to that of theorem 3.

Theorem 4 *Under the assumptions H, we have*

$$L(FIFO) \leq_c L(\Theta) \leq_c L(LIFO), \quad \forall \Theta \in (\Sigma_0)^M$$

Define the following new classes of policies,

- Σ_3 - class of non preemptive policies that do not use information regarding either service time or reference times, $\Sigma_0 \subset \Sigma_3$.
- Σ_4 - superset of Σ_3 that use information regarding reference times, $\Sigma_1 \subset \Sigma_4$. .

then we can show the following

$$\begin{aligned} L(FIFO) &\leq_{c\uparrow} L(\Theta), \quad \Theta \in (\Sigma_3)^M, \\ -L(LIFO) &\leq_{c\uparrow} -L(\Theta), \quad \Theta \in (\Sigma_3)^M, \\ L(SR) &\leq_{c\uparrow} L(\Theta), \quad \Theta \in (\Sigma_4)^M, \\ -L(LR) &\leq_{c\uparrow} -L(\Theta), \quad \Theta \in (\Sigma_4)^M. \end{aligned}$$

Remarks.

1. We have similar results for both preemptive work conserving and preemptive non-work conserving policies provided that the service times are exponential random variables.

2. We have the following interesting application to systems in which customers have soft real-time deadlines. Consider as a metric, the probability that a customer completes service by its reference time (deadline), $\Pr[L \geq 0] = \Pr[r \geq T]$. If the relative reference time, r, has a concave distribution function, then the following relations hold between FIFO, LIFO, and any policy $\Theta \in (\Sigma_0)^M$,

$$\Pr[L(FIFO) \geq 0] \leq \Pr[L(\Theta) \geq 0] \leq \Pr[L(LIFO) \geq 0], \quad \Theta \in (\Sigma_0)^M.$$

This inequality was first derived in the context of a single server queue in [3]. In addition, we have the following relation between LIFO and $\Theta \in (\Sigma_3)^M$,

$$\Pr[L(\Theta) \geq 0] \leq \Pr[L(LIFO) \geq 0], \quad \Theta \in (\Sigma_3)^M.$$

References

[1] M. Berg, M.J.M. Posner, "On the Regulation of Queues," *Operations Research Letters*, 4, 5, pp. 221-224, February 1986.

[2] T.M. Chen, J. Walrand, D.G. Messerschmitt, "Dynamic Priority Protocols for Packet Voice," *IEEE J. Sel. Areas of Commun.*, **7**, 5, pp. 632-643, June 1989.

[3] B.T. Doshi, E.H. Lipper, "Comparisons of Service Disciplines in a Queueing System with Delay Dependent Customer Behavior," *Applied Probability - Computer Science: The Interface*, Vol. II, R. L. Disney, T.J. Ott, Eds., Cambridge, MA:Birkhauser, pp. 269-301, 1982.

[4] B.T. Doshi, H. Heffes, "Overload Performance of Several Processor Queueing Disciplines for the M/M/1 Queue," *IEEE Trans. Communications*, **34**, 6, pp. 538-546, June 1986.

[5] J. G. Shantikumar, U. Sumita, "Convex Ordering of Sojourn Times in Single-Server Queues: Extremal Properties of FIFO and LIFO Service Disciplines," *J. Appl. Prob.*, **24**, 737-748 (1987).

[6] O.A. Vasicek, "An Inequality for the Variance of Waiting Time Under a General Queueing Discipline," *Operations Research*, **25**, pp. 879-884, 1977.