Stability Criteria For Yet Another Multidimensional Distributed System

Leonidas Georgiadis

Wojciech Szpankowski
Purdue University, spa@cs.purdue.edu

Report Number:
91-071
STABILITY CRITERIA FOR YET ANOTHER MULTIDIMENSIONAL DISTRIBUTED SYSTEM

Leonidas Georgiadis
Wojciech Szpankowski

CSD-TR-91-071
September 1991
We present sufficient and necessary conditions for stability of the time-limited token passing ring introduced recently by Lueng and Eisenberg. This was an open problem up to date. In general, establishing stability for multidimensional distributed systems is notoriously a difficult problem. We record here that the standard Lyapunov test function method often fails when applied to such systems (e.g., token passing rings, ALOHA-type systems, etc.). This particularly applies to time-limited and other token passing rings. We note also that our recent papers on this topic—including this one—establish a useful and alternative approach that turns out to be very successful for deriving stability conditions for several distributed systems.

1. INTRODUCTION

Distributed multiqueue systems which share a single scarce resource (i.e., server) such as a communication channel or a processor, have received a considerable amount of attention in the recent literature. Important examples of such distributed multiqueue systems are local area networks (e.g., ALOHA system, Ethernet, token passing ring, FDDI ring, etc.), multiprocessor systems, distributed computations, distributed data base, and so forth. Of special interest is the token passing ring (cf. [10], [25], [26]) due to a number of reasons. In particular, it appears that determination of sound measures of performance for such a system, under realistic assumptions such as asymmetric traffic, finite or infinite buffers, non-exhaustive service and general input are fairly difficult to obtain, as can be witnessed

*This research was supported by NSF Grant CCR-8900305, and in part by AFOSR Grant 90-0107, and by Grant R01 LM05118 from the National Library of Medicine.
from the literature ([2], [3], [9], [13], [26]). For example, it is known that obtaining the distribution of the number of messages queued in each station is a formidable open problem, as is the problem of obtaining the waiting time distribution. Surprisingly enough, the stability condition for the token passing ring was heuristically predicted by Kuehn [10] in 1979, and then reproduced with some minor changes in many other papers (e.g., [7]). But ... , Watson [29] observed that in the performance evaluation of token passing rings "it is convenient to derive stability conditions ... (without proof)". This particularly applies to time-limited token passing rings introduced recently by Leung and Eisenberg (cf. [11], [12]). In such a system each station transmits messages for at most an amount of time, $\tau$. If the transmission time exceeds $\tau$, the station completes the transmission of the message in progress and sends the token to the next station. In this paper we establish rigorously stability condition for such a system.

Despite a vigorous research in the area of stability over last twenty years (cf. [27], [23], [28]), very few computable stability criteria are known for multidimensional processes, in particular multidimensional Markov chains. The most popular approach through the Lyapunov test function (cf. [27]) did not succeed in the past to provide general computable criteria for multidimensional Markov chains. However, due to pioneering work of Malyshev [15], continued by Mensikov [17], and Malyshev and Mensikov [16] some progress has been made in obtaining stability condition for a class of two-dimensional and three-dimensional Markov chains. Recently, weaker stability criteria for two dimensional chains have been presented by Fayolle [5] and Rozenkrantz [20]. Unfortunately, these conditions are still difficult to apply in practice for higher dimensional processes (see [8] for an application of this to a multidimensional ALOHA system). A more practical approach to stability of multidimensional Markov chains arising in queueing applications was discussed in Szpankowski [22] (for more details see survey [23]), and Georgiadis and Szpankowski [6].

Our approach to the stability of token passing rings follows the idea suggested in Szpankowski [24], and differs significantly from the standard methodology of the test function (cf. [27]). It resembles, however, the general idea of stability criteria proposed by Malyshev and Mensikov [16]. Our approach is based on a simple idea of stochastic dominance technique, and application of Loynes [14] stability criteria for an isolated queue. We use the stochastic dominance to verify technical stationarity requirements in Loynes' criteria. We note that this approach is not restricted to time-limited token passing rings, and stability of several other distributed systems can be assessed by this methodology.

In the rest of this paper, we will consider the gated version of the $\tau$-limited policy, i.e., the customers that are allowed to be served at queue $i$ are only those that are present at the
instant of token arrival at that queue. This is done for ease of exposition. The methodology presented in this paper can be easily extended to cover the case when customers arriving at a queue while it is being served can also be transmitted.

We now summarize our main results. We shall analyze the token passing ring with Poisson arrivals with parameter \( \lambda_i \) for the \( i \)th station, general distribution of service times \( \{S^k_i\}_{k=1}^\infty \) and switchover times \( \{U^k_i\}_{k=1}^\infty \), and transmission time limit \( \tau_i \). Our first result (cf. Theorem 4) establishes a stochastic dominance among token passing rings, and it can be used to establish some bounds on the performance evaluation of the system (cf. [13]). We use this result to prove our main result regarding stability of the system (see also Theorem 6 and Theorem 8). To formulate it in a compact form, we define \( \bar{L}_i = \min\{k : \sum_{j=1}^k S^j_i \geq \tau_i\} \) and \( \ell_i = E\bar{L}_i \). Clearly, \( \ell_i \) is the average of the maximum number of customers served during \( \tau_i \).

**Proposition.** Consider a token passing ring consisting of \( M \) stations with \( \tau \)-limited service schedule for the \( i \)th station, and Poisson arrivals. Then the system is stable if and only if \( \sum_{j=1}^M \rho_j < 1 \) and

\[
\lambda_j \leq \frac{\ell_i}{u_0}(1 - \rho_0) \quad \text{for all} \quad j \in \mathcal{M} = \{1, \ldots, M\},
\]

where \( u_0 = \sum_{i=1}^M E U_i \) is the average total switchover time, and \( \rho_0 = \sum_{i=1}^M \rho_i \) with \( \rho_i = \lambda_i s_i \) and \( s_i = ES_i \) being the average service time at the \( i \)th station.

Note that the above stability criteria are represented in terms of a set of linear inequalities with respect to input rates \( \lambda_i \) for \( i \in \mathcal{M} \).

The paper is organized as follows. In the next section we present our preliminary results that are of their own interests for the performance evaluation of the token passing ring. In particular, we find Markovian representations of the system, prove a crucial stochastic dominance relationship, and establish some Wald's type formulas. Finally, in Section 3 we present our main construction that leads to the proof of the above Proposition.

2. PRELIMINARY RESULTS

In this section we present several results that are required to establish our main finding regarding the stability of the token passing ring. These results are of their own interests, and can be used to obtain some estimates for the performance evaluation of the system. In the sequel, we list our main assumptions, prove Markovian character of an imbedded queueing process, show two simple Wald's type identities, and finally establish a stochastic dominance relationship.
We start with a precise definition of our stochastic model. We shall adopt the following assumptions.

(A1) There are $M$ stations (queues) on a loop, each having infinite capacity buffer.

(A2) Maximum time customers are served during the token visit at a queue is limited to $\tau_i < \infty$ units of time. Only customers that are present at the instant of token arrival can be served. Moreover, we have nonpreemptive discipline, that is, the customer that is in the server when the time limit $\tau_i$ is reached, is served to completion before the token moves to the next queue.

(A3) Arrival process $A^i_t, t \in [0, \infty)$, to the $i$th queue is a Poisson process with parameter $\lambda_i > 0$. Here, $A^i_t$ is the number of arrivals at queue $i$ up to time $t$. The arrival process at a queue is independent of the arrival processes to other queues.

(A4) Service time process $\{S^k_i\}_{k=1}^{\infty}$ at queue $i$ is i.i.d. with $s_i = E S^1_i > 0$. The service time process at a queue is independent of the arrival processes at all queues and independent of the service time processes at other queues.

(A5) The switchover times between $i$ and $i+1 \mod M$ queue, $\{U^k_i\}_{k=1}^{\infty}$, are i.i.d., independent of the switchover times $\{U^k_j\}_{k=1}^{\infty}$ for $j \neq i$, and independent of the arrival and the service time processes. The average total switchover time is defined as $U_0 = \sum_{i=1}^{M} E U^1_i$.

To avoid unnecessary complications we assume that $P(U^i_1 > 0) = 1, i = 1, \ldots, M$.

Now we are ready to present a Markovian description of the system. We need a little bit of notation. By (A1), the token visits stations in a cyclic order. Let $n$ denote the $n$th visit of the token to any queue. Then, $k_n = \lfloor (n-1)/M \rfloor + 1$ denotes the cycle number in which the $n$th visits occurs (we start counting cycles from one and assume that the token starts from queue 1). Note that the queue visited at the $n$th visit is just $J_n = n - M(k_n - 1)$.

Let also $\{T_n\}_{n=1}^{\infty}$ be the time instant of the $n$th visit of the token to any queue. Define an $M$-dimensional process $\bar{N}^n = (\bar{N}_1^n, \ldots, \bar{N}_M^n)$, $n = 1, 2, \ldots$, where $\bar{N}_i^n$ as the number of customers in queue $i$ at time $T_n$. In addition, by $\bar{N}_i^n$ we mean the total number of customers served from queue $i$ up to time $T_n$. Theorem 1 below proves that $\bar{N}^n$ is a Markov chain.

**Theorem 1.** The process $\bar{N}^n$ is a nonhomogeneous Markov chain.

**Proof.** Let $\bar{L}^n_i = \min\{k: \sum_{j=1}^{k} S^j_i \geq \tau_i\}$ and $L^n_i = \min\{\bar{L}^n_i, N^n_i\}$. Note that $L^n_i$ is the number of customers served from queue $i$ at the $n$th visit of the token. The time $B_n$ that
elapses between $n$th and $n + 1$st visit of the token to any queue is

$$B_n = \sum_{j=1}^{L_n} S_j^N n + j + U_{j_n}^k,$$  \hspace{1cm} (1)

and the number of arrivals $X_i^n$ to queue $i$ between the $n$th and $n + 1$st visits are

$$X_i^n = A_i^{T_n} + B_n - A_i^T_n.$$  \hspace{1cm} (2)

Finally, the following recursions hold for the queue size in the $i$th station

$$\tilde{N}_i^n + 1 = \tilde{N}_i^n + X_i^n \hspace{1cm} if \hspace{1cm} i \neq J_n$$
$$\tilde{N}_i^n + 1 = [\tilde{N}_j^n - \tilde{L}_j^n] + X_i^n \hspace{1cm} if \hspace{1cm} i = J_n,$$  \hspace{1cm} (3)

where $[x]^+ = \max\{x, 0\}$. Since the transmission policy is nonpreemptive, no information is obtained from the history up to time $T_n$ about the service times of the customers that are in the queues at time $T_n$. Taking also into account assumptions (A3)-(A5), we conclude that the processes \(S_{i}^{N(n) + j} \to \infty \) \(A_i^{T_n} - A_i^T_n \), \(t \in [0, \infty)\), and the random variable $U_{j_n}^k$, are independent of $\tilde{N}^n$, $1 \leq m \leq n$. From the above discussion we conclude that $\tilde{N}^{n+1}$ is of the form $\tilde{N}^{n+1} = f(\tilde{N}^n, Y^n)$ for some (measurable) function $f(\cdot)$ where $Y^n$ is composed of the processes \(S_{i}^{N(n) + j} \to \infty \) \(A_i^{T_n} - A_i^T_n \), \(t \in [0, \infty)\) and $U_{j_n}^k$. Therefore, $\tilde{N}^n$ is a Markov chain (see, for example, page 34 of [19]).

Remark: It should be noted that the assumption that the service discipline is nonpreemptive (see assumption (A2)) is crucial for Theorem 1 to hold. Assume, for example, that preemptions were allowed so that a server could interrupt the service of the customer as soon as the limit $\tau$ was reached. Upon the next arrival of the token to the queue, the server could either complete the remaining service time or restart the service time of the interrupted customer. In both cases, the number of customers in queue $i$ at time $T_n$ will depend in general on the service time $S_i^{N(n) + 1}$ and the process of queue lengths will not be Markov.

There are other Markovian descriptions of the system. For example, define $N_j^n(i)$ to be the number of customers at queue $j$ when the token visits queue $i$ for the $n$th time. Then, the process $N^n(i) = (N_j^n(i), \ldots, N_j^n(i))$ can be deduced from $\tilde{N}^n$ since $N^n(i) = N^{(n-1)M+i}$ and therefore it is a Markov chain. It is not difficult to verify also that,

Corollary 2. The process $N^n(i)$ of the queue lengths registered by the token when it visits (reference) queue $i$, is a homogeneous, irreducible and aperiodic Markov chain.
The fact that under assumption (A2) the service times of the customers at queue $i$ at instant $T_n$ are i.i.d. independent of the queue size at time $T_n$, permits us to consider a new model of the system which is stochastically equivalent to the original one and has the advantage of that under this new model, many of the arguments that follow become simpler. Specifically, in the new system assumptions (A1)-(A3) and (A5) are the same, while assumption (A4) is replaced with

(A4') Service times are assigned to the customers at queue $i$ upon beginning of service as follows. We consider a doubly infinite sequence of i.i.d random variables $\{S_i^{n,k}\}_{n,k=1}^{\infty}$ with $S_i^{n,k} = ES_i^{n,k} > 0$. The customers that are served during the $n$th arrival of the token to queue $i$ are assigned the service times $S_i^{n,1}, S_i^{n,2}, \ldots$. The sequence $\{S_i^{n,k}\}_{n,k=1}^{\infty}$ is independent of the sequence $\{S_j^{n,k}\}_{n,k=1}^{\infty}$ for $i \neq j$, and independent of the interarrival processes to the queues.

In the sequel we will study the system in which assumption (A4) is replaced by assumption (A4').

We will need some Wald's type relationships between the average number of customers served per token visit $L_i^n$ and the average cycle time $C_i^n$. The former quantity was defined in the proof of Theorem 1, and $C_i^n$ is the length of time between the $n$th and $n+1$st visits of the token to the reference queue $i$. By $EL_i$ and $EC$ we denote the limiting averages of $L_i^n$ and $C_i^n$. The following result is proved by an extension of the method used in [25, page 9]. For completeness, and since we will need some of the steps of the proof in later sections, we provide the proof here.

**Theorem 3.** Let the Markov chain $N^n(i)$ be positive recurrent (ergodic) for some $i \in M$. Then

1. $N^n(j)$ is ergodic for all $j \in M$,

2. $\rho_0 = \sum_{j=1}^{\infty} \rho_j < 1$, and

$$EL_j = \lambda_j EC, \quad j \in M$$

$$EC = \frac{u_0}{1 - \sum_{j=1}^{\infty} \rho_j},$$

where $u_0$ is the total average switchover time (cf. assumption (A5)) and $\rho_j = \lambda_j s_j$ is the utilization coefficient for the $i$th queue.

**Proof.** Without loss of generality, let $i = 1$. By the assumption, $N^n(1)$ is an ergodic Markov chain. Note that $N^n(1)$ has a natural regeneration structure, namely when all
queues are empty, that is, when the process returns to zero state \(0 = (0,0,\ldots,0)\). Assume
\[ N^1(1) = 0 \] and \(K^1 = 1\). Define
\[ K^{n+1} = \min\{m > K^n : N^m(1) = 0\} . \]
and \(R^n = K^{n+1} - K^n\). We shall also use \(R = R^1\) as the length of a regeneration cycle. Due
to the ergodicity of \(N^n(1)\) we have \(ER < \infty\). Observe that for \(j \in M\), \(N^n(j)\) is regenerative
with respect to \(R^n\). Since it is easily seen that \(R^n\) is aperiodic, it follows (see Asmussen [1, Chapter V]) that the process \(N^n(j)\) has a steady state distribution and therefore, is ergodic.

The sequences \(C^n_i, L^n_i\), \(n = 1, \ldots\) are regenerative with respect to \(R^n\). Therefore,
\[ \lim_{n \to \infty} \frac{\sum_{k=1}^n L^k_i}{n} = \frac{E \left( \sum_{k=1}^R L^k_i \right)}{ER} , \quad \lim_{n \to \infty} \frac{\sum_{k=1}^n C^k_i}{n} = \frac{E \left( \sum_{k=1}^R C^k_i \right)}{ER} \quad \text{a.s.} \quad (6) \]
Moreover, \(L^n_j\) and \(C^n_i\) converge in distribution to \(L_i\) and \(C_1\) such that
\[ EL_j = \frac{E \left( \sum_{k=1}^R L^k_j \right)}{ER} , \quad EC_1 = \frac{E \left( \sum_{k=1}^R C^k_1 \right)}{ER} . \quad (7) \]
Now we are in position to prove (4) and (5). Let now \(\hat{L}^n_i = \min\{k : \sum_{j=1}^k S^{n,j}_i \geq \tau_i\}\).
Clearly, \(\hat{L}^n_i\), \(n = 1, 2, \ldots\) are i.i.d. random variables and since \(s_1 > 0\), it is well known from
renewal theory that \(E\hat{L}^n_1 = \ell_1 < \infty\). Observe also that the event \(\{R \leq k\}\) is independent of
the sequence \(\hat{L}^n_1\), \(n > k\) and therefore, \(E \left( \sum_{k=1}^R \hat{L}^k_1 \right) = \ell_1 ER < \infty\). From the operation of
the policy we have that \(L^k_i \leq \hat{L}^k_i\) and therefore, \(E \left( \sum_{k=1}^R L^k_i \right) < \infty\). Observe next that in
the interval \([0, \sum_{k=1}^R C^k_1]\) all the arriving customers from all queues must be served. If \(A_j\) is
the number of arrivals to queue \(j\) in the interval \([0, \sum_{k=1}^R C^k_1]\), then \(E A_j = E \left( \sum_{k=1}^R L^k_j \right)\),
and due to the Poisson assumption (A3) we also have \(E A_j = \lambda_j E \left( \sum_{k=1}^R L^k_j \right)\). Therefore,
\[ E \left( \sum_{k=1}^R C^k_1 \right) = E \left( \sum_{k=1}^R L^k_1 \right) / \lambda_1 < \infty , \quad (8) \]
and
\[ E \left( \sum_{k=1}^R L^k_j \right) = \lambda_j E \left( \sum_{k=1}^R C^k_1 \right) , \quad j \in M . \quad (9) \]
The above and (7) lead to \(EL_j = \lambda_j EC_1\), which completes the proof of (4).

To prove (5) we note that the cycle length \(C^n_i\) is
\[ C^n_i = U^n + \sum_{j=1}^M \sum_{m=0}^{L^n_j} S_j^{m+\sum_{k=1}^{n-1} L^k_j} , \quad (10) \]
where \( U^n = \sum_{j=1}^{M} U_j^n \). Summing the above over first \( R \) visits of the token, taking the expectation of it, and using (9) one obtains the following

\[
E \left( \sum_{n=1}^{R} C^n_j \right) = u_0 \cdot ER + \sum_{j=1}^{M} \lambda_j s_j E \left( \sum_{n=1}^{R} C^n_j \right).
\]

Since \( ER > 1 \) and by (8) \( E \left( \sum_{n=1}^{R} C^n_1 \right) < \infty \), using (7) we obtain from the above that \( \sum_{j=1}^{M} \rho_j < 1 \) and \( EC_1 = EC = u_0/(1 - \sum_{i=1}^{M} \rho_i) \) as needed for (5).

The next result is our main finding in this section, and it is of prime importance for our stability analysis. Before we plunge into technical details, we first give a brief overview of our approach. In the process of estimating stability we need to build several dominant systems of the original token passing ring. For example, when we study stability of an isolated station, say the \( j \)th one, we partition all other stations into a class \( S \) of nonpersistent queues and a class \( \mathcal{U} \) of persistent queues. A nonpersistent queue serves customers in the normal way as in the original token passing ring. A persistent queue, however, always sends the maximum allowable number of customers, that is, \( \bar{L}_i \) for \( i \in \mathcal{U} \), by sending if necessary "dummy" customers. A question is whether such a new system dominates the original token passing ring in some sense. If the answer is yes, then by proving stability of the dominant system we establish stability of the original token passing ring.

We state the next result in a general form, since it can be useful in the performance evaluation of other service disciplines. Specifically, in the terminology of [13], we consider the class of "monotonic", "contractive" policies. This amounts to replacing assumption (A2) with the following more general one.

\[(A2') \text{ Let } \mathbf{A} \text{ denote a sequence of real numbers } \{a_1, a_2, \ldots \}. \text{ Let } f_i(m, \mathbf{A}) \text{ be the number of customers served from queue } i \text{ when there are } m \text{ queued messages at the instant of the } n \text{th token arrival at queue } i \text{ and } \{S^n_1, S^n_2, \ldots \} = \mathbf{A}. \text{ We assume that for fixed } \mathbf{A}, f_i(m, \mathbf{A}) \text{ is a nondecreasing function of } m. \text{ In addition, for a fixed } \mathbf{A}, \text{ the following relation holds}
\]

\[
f_i(m_1, \mathbf{A}) - f_i(m_2, \mathbf{A}) \leq m_1 - m_2 \quad \text{if } m_1 > m_2.
\]

Now we are ready to formulate our result. Consider two token passing rings, say \( \theta \) and \( \Theta \). Both satisfy assumptions (A1)-(A5) with (A2) replaced by the weaker assumption (A2'). The system \( \theta \) represents our original token passing ring. The system \( \Theta \) differs only in the switchover times, namely, we assume that the switchover time for \( \Theta \) is replaced by
We assume that for every $i \in \mathcal{M}$ and every $k \geq 0$ we have $\Delta_i^k \geq 0$. We make the following assumption for the process $\Delta_i^k$.

\[(A6)\] The random variable $\Delta_i^k$ is independent of the service times, switchover times and the Poisson increments of the arrival processes to all stations after time $T_M(k-1)+(i+1)-U_i^k$ (see Fig. 1).

**Theorem 4.** Let $\tilde{N}^n(\theta)$ and $\tilde{N}^n(\Theta)$ denote the queue lengths in both systems. Then, under the above assumptions, and under the condition that the token starts from the same queue, say queue number one, and with the same number of initial customers in both systems, the following holds

\[\tilde{N}^n(\theta) \leq_{st} \tilde{N}^n(\Theta).\] (13)

where $\leq_{st}$ means stochastically smaller.

**Proof.** To avoid cumbersome notation we present the proof only for $M = 2$ users. The proof can be easily extended to any number of users.

We define some new variables. For a system $\theta$ let $T_n^\theta$ and $D_n^\theta$ denote the instances of the $n$th visit and the $n$th token departure from any queue respectively. As before, $J_n^\theta$ denotes the queue number visited at the $n$th visit of the token. Finally, $L_i^\theta(\theta)$ as before denotes the number of customers served from queue $i$ at the $n$th visit of the token. Clearly, for our two station system $L_i^\theta(\theta) = 0$ for $n$ even, and $L_i^\theta(\theta) = 0$ for $n$ odd. In a similar manner we define respective quantities in the $\Theta$ system.

We will construct from the system $\theta$ a token passing ring $\tilde{\theta}$, which is stochastically equivalent to the system $\theta$ and for which we have that

\[\tilde{N}^n(\tilde{\theta}) \leq \tilde{N}^n(\Theta).\] (14)

Figure 1 should help to understand our construction. Assume $\tilde{N}^1_i(\tilde{\theta}) = \tilde{N}^1_i(\theta)$ for $i = 1,2$. The service times in system $\tilde{\theta}$ are assigned from the same sequences $S_i^{n,k}$ as in $\Theta$ (according to assumption (A4').) Also, the same functions $f_i(m,A)$, $i = 1,2$ are used in both systems. Therefore, the decision to switch to queue 2 will occur at the same time, namely $D_2^\theta = D_2^\Theta$. The switchover time for $\tilde{\theta}$ becomes now $U_1^\theta$, and of course $T_2^\theta \leq T_2^\Theta$ since $\Delta_1^1 \geq 0$ (see Fig. 1).

The arrivals in the system $\tilde{\theta}$ in $[D_1^\theta, T_2^\theta]$ are now assumed to be identical to the arrivals in $[D_1^\Theta + \Delta_1^1, T_2^\Theta]$ in $\Theta$ system. Therefore clearly $\tilde{N}^2_i(\tilde{\theta}) \leq \tilde{N}^2_i(\Theta)$ for $i = 1,2$. The arrivals to system $\tilde{\theta}$ in $[T_2^\theta, T_2^\theta + S_2^1 + \cdots + S_2^{L_2^\theta}(\tilde{\theta})]$ are taken to be identical to the arrivals in
new arrival new arrival

\[ T^\varnothing_1 \rightarrow S^1_1 \rightarrow D^\varnothing_1 \rightarrow T^\varnothing_2 \rightarrow S^1_2 \rightarrow \Delta_1 \rightarrow U^1_1 \rightarrow S^1_3 \rightarrow \Delta_2 \rightarrow U^1_2 \rightarrow T^\varnothing_3 \]

identical arrivals

identical arrivals

identical arrivals

identical arrivals

Figure 1: Illustration to the proof of Theorem 4

\[ [T^\varnothing_2, T^\varnothing_2 + S^2_2 + \cdots + S^L_2(\hat{\theta})]. \] Note that this can be done since by (A2') \( L^1_2(\hat{\theta}) \leq L^1_2(\varnothing). \) Observe also that \( T^\varnothing_2 + S^1_2 + \cdots + S^L_2(\hat{\theta}) = D^{\varnothing}_2 \) (Fig. 1).

To complete the description of the system \( \hat{\theta} \) we have to specify the arrivals in \([D^\varnothing_2, D^\varnothing_2 + U^1_2]\). These are taken to be exactly the arrivals in \([D^\varnothing_2 + \Delta^1_2, D^\varnothing_2 + \Delta^1_2 + U^1_2]\) in the dominant system \( \varnothing \) (see Fig. 1). Note from the construction that

\[ \tilde{N}^3_2(\hat{\theta}) = \tilde{N}^3_1(\hat{\theta}) = \tilde{N}^3_1(\varnothing) + A_{[T^\varnothing_2, T^\varnothing_2]} \leq \tilde{N}^3_1(\varnothing) + A_{[T^\varnothing_2, T^\varnothing_2]} = \tilde{N}^3_1(\varnothing) \]

and also by (12),

\[ \tilde{N}_2^3(\hat{\theta}) = \tilde{N}_2^3(\hat{\theta}) - L^2_2(\hat{\theta}) \leq \tilde{N}_2^3(\varnothing) - L^2_2(\varnothing) = \tilde{N}_2^3(\varnothing). \]

We can now repeat exactly the same procedure to construct \( \hat{\theta} \) in the interval \([T^\varnothing_n, T^\varnothing_{n+1}]\), \( n \geq 3 \), in the same manner as it was constructed in the interval \([T_2, T_3]\). By construction the service times and switchover times of system \( \hat{\theta} \) are identically distributed to the corresponding variables of system \( \theta \) and are independent of the interarrival process. In addition, assumption (A6) and the fact that the servicing policy is nonanticipative assures that the times \( T^\varnothing_{n+1} - U^\varnothing_{j^n} \) are stopping times for the Poisson arrival processes to all stations. The independence of the increments of the Poisson process implies now, that the constructed interarrival process in system \( \hat{\theta} \) is Poisson with rate \( \lambda_i \) for queue \( i \). Moreover, by construction (14) holds. Since \( \hat{\theta} \) is stochastically equivalent to \( \theta \), we have that the distribution of \( \tilde{N}^n(\theta) \) is identical to the distribution of \( \tilde{N}^n(\hat{\theta}) \). This completes the proof of Theorem 4. \( \blacksquare \)
3. MAIN RESULTS

In this section we present a proof of our proposition from the Introduction. However, before we plunge into technical details an overview of our stability approach is discussed. We shall argue that our idea is novel (cf. [24]), and can be successfully used to establish stability of some other distributed systems (see Szpankowski [24], [23] for applications to the ALOHA system and coupled-processors system).

Our approach is based on three simple observations. At first, we note that a multidimensional process is stable if and only if its components are stable [23]. More precisely, if $N^n = (N^n_1, \ldots, N^n_M)$ is a stochastic process - not necessary a Markov chain - then say the process is stable if the distribution of $N^n$ as $n \to \infty$ exists and the distribution is honest. In other words, $N^n$ is stable if for $x \in \mathbb{R}^M$, where $\mathbb{R}$ is the set of real numbers, the following holds for all points of continuity of $F(x)$

$$
\lim_{n \to \infty} \Pr\{N^n \leq x\} = F(x) \quad \text{and} \quad \lim_{x \to \infty} F(x) = 1
$$

(15)

where $F(x)$ is the limiting distribution function, and by $x \to \infty$ we understand that $x_j \to \infty$ for all $j \in \mathcal{M} = \{1, \ldots, M\}$. If a weaker condition holds, namely,

$$
\lim_{x \to \infty} \lim_{n \to \infty} \inf \Pr\{N^n \leq x\} = 1,
$$

(16)

then the process is called \textit{substable} [14] or \textit{tight} or \textit{bounded in probability} sense. Otherwise, the system is unstable (for more details see Loynes [14]).

Secondly, to obtain stability conditions for a single isolated station in the token passing ring, we apply the technique of Loynes [14] who proved that a single $G|G|1$ queue is stable if the input rate is smaller than the average service time provided that service times and interarrival times are jointly stationary and ergodic. To verify a technical stationarity condition in Loynes' criteria we apply the stochastic dominance result of Theorem 4. More precisely, we partition the set of queues, $\mathcal{M}$, into a set $\mathcal{S}$ of nonpersistent queues and into a set $\mathcal{U}$ of persistent queues as was described in section 2. By Theorem 4 the new system stochastically dominates the original one, and by proving stability of it, we clearly establish stability conditions for the original token passing ring. We use induction to establish stability conditions for the nonpersistent queues in the new system, while the stability condition for a persistent queue is established by using Loyne's criteria.

To fulfill the above plan, we start by showing a result that will be useful in proving the condition for a persistent queue in the dominant system. More formally, as in Section 2 we consider a doubly infinite sequence of i.i.d random variables $\{S_{n,k}\}_{n,k=1}^{\infty}$ and define
\( \hat{L}^n = \min\{k : \sum_{j=1}^{k} S^{n,k} \geq \tau\}, \quad 0 \leq \tau < \infty. \) We consider further a queue with vacations such that upon the \( n\)th arrival of the server to the queue, \( \hat{L}^n \) customers (dummy if necessary) are served and then the server goes for a vacation. The service times of the customers served during the \( n\)th visit of the server are the random variables \( S^{n,k}, \quad k = 1, \ldots, \hat{L}^n. \) Let \( \{C^n\}_{n=1}^{\infty} \) be the process of cycle times (time intervals between successive visits to the queue). It is assumed that the processes \( \{C^n, \hat{L}^n\}_{n=1}^{\infty} \) are jointly stationary and ergodic (no independence is required). The arrival process \( A^t \) to this queue is a Poisson process with parameter \( \lambda, \) independent of the processes \( \{C^n, \hat{L}^n\}_{n=1}^{\infty} \). Let \( N^n \) represent the queue length at the beginning of the \( n\)th cycle. By \( X^n \) we denote the number of customers arrived during the \( n\)th cycle. Since \( A^t \) is Poisson and independent of the processes \( \{C^n, \hat{L}^n\}_{n=1}^{\infty}, \) the processes \( \{X^n, \hat{L}^n\}_{n=1}^{\infty} \) are jointly stationary and ergodic, and \( E \lambda = \lambda \mathcal{E} \) where \( \mathcal{E} = \mathcal{E}^1. \) Clearly, the process of queue lengths at the instants of the visits of the server to the queue satisfies the following recurrence

\[
N^{n+1} = \max\{N^n + X^n - \hat{L}^n, \quad X^n\}, \quad n = 1, 2, \cdots \tag{17}
\]

Let \( \ell = E \hat{L}^n. \) We prove the following stability result.

**Lemma 5.** Consider the queueing system just described. If \( \lambda \mathcal{E} < \ell, \) then the queue is stable in the sense of definition (15).

**Proof.** We apply Loynes' scheme to prove the lemma. We may assume without loss of generality that \( X^n, \hat{L}^n \) is a two-sided stationary process, that is, it is defined for \( -\infty < n < \infty. \) Note next, that the recursion (17) is such that the RHS of it represents a nondecreasing and left continuous (in \( N^n \)) function. Therefore, by Lemma 1 in Loynes [14] we conclude that there exists a stationary sequence \( \mathcal{N}^k \) satisfying recursion (17), such that \( N^n \) converges in distribution to \( \mathcal{N}^1 \) provided that \( N^1 = 0. \) Now, we need to find out when \( \mathcal{N}^k \) is honest. Recursion (17) is not quite the same as the one treated by Loynes, however we can use similar arguments as follows.

By telescoping the recurrence (17) we immediately obtain for \( n \geq 2 \)

\[
N^n = \max_{1 \leq r \leq n-1} \left\{ X^r + \sum_{k=r+1}^{n-1} \overline{X}^k \right\}, \tag{18}
\]

where \( \overline{X}^k = X^k - \hat{L}^k, \) provided \( N^1 = 0. \) Arguing as in Loynes [14] we have that \( \mathcal{N}^k \) is honest if and only if

\[
\limsup_{r \to \infty} \left\{ X^{-r} + \sum_{k=1}^{r} \overline{X}^{-k} \right\} < \infty. \tag{19}
\]
Observe now that
\[ X^{-r} + \sum_{k=1}^{r-1} X^{-k} = r \left( \frac{\sum_{k=1}^{r} X^{-k}}{r} - \frac{r-1}{r-1} \sum_{k=1}^{r} \tilde{L}^{-k} \right), \quad r = 2, 3, \ldots \]
Since by the ergodicity of the sequences \( X^n \), \( \tilde{L}^n \) we have that \( \lim_{r \to \infty} \sum_{k=1}^{r} X^{-k}/r = \lambda EC \) and \( \lim_{r \to \infty} \sum_{k=1}^{r} \tilde{L}^{-k}/r = \ell \), the condition \( \lambda EC < \ell \) assures the validity of (19). The assumption that \( N^1 = 0 \) can be removed as in [14].

Now we are ready to prove our main result described already in our Proposition of the Introduction. In the next theorem we show that the conditions of the Proposition are sufficient. The proof uses the idea presented in the overview above, however due to technical reasons we carry it out formally through the mathematical induction.

**Theorem 6.** The Markov chain \( N^n(i) \) representing the queue lengths in the token passing ring when it visits queue \( i \in M \) is ergodic if
\[ \lambda_j < \frac{\ell_j}{u_0}(1 - \rho_0) \quad \text{for all} \quad j \in M \] where \( \rho_0 = \sum_{j=1}^M \rho_j \) and \( \ell_i = E\tilde{L}_i \).

**Proof.** We use mathematical induction. For \( M = 1 \) the proof is simple. In this case, since the switchover times are independent of the service times and the interarrival times, the one dimensional process \( N^n(1) \) is a Markov chain satisfying (17) (here we do not make any assumptions regarding the stationarity of the process of cycle times). By the Lyapunov test function method (cf. [23], [27]) one easily notes that \( \lambda < \ell(1 - \rho_0)/u_0 \) is sufficient for stability, as needed.

Now we assume that the theorem is true for \( M = 1 \) and prove that it can be extended to the \( M \) queue case. Let \( (\mathcal{U}, S) \), \( \mathcal{U} \neq \emptyset \), be a partition of the set \( M \) of \( M \) queues into persistent and nonpersistent queues. Assume for a moment that \( S \neq \emptyset \). Note that the cardinality \( |S| \) of \( S \) is not larger than \( M - 1 \). Let \( \overline{N^n(i)} = \{ \overline{N^1(i)}, \ldots, \overline{N_M(i)} \} \) be the queue lengths when the token visits the \( i \)th queue for the \( n \)th time in the \( (\mathcal{U}, S) \) system in which persistent queues \( \mathcal{U} \) send dummy packets as discussed above. Observe that the modified system differs from the original token ring system only in the switchover time from a persistent queue to the successor of that queue in the ring. Specifically, if \( i \in \mathcal{U} \), then the switchover times become,
\[ \overline{U}^k_i = \Delta^k_i + U^k_i, \]
where \( \Delta^k_i \) is the time needed to service the dummy messages at node \( i \) (if any). Since the queue length at the \( n \)th visit of the token to queue \( i \) is independent of future arrivals or
future service times, and the service times of the dummy messages are independent of the rest of the processes in the system, it is clear that $\Delta_i^k$ satisfies condition (A6) of Section 2. Therefore, according to Theorem 4, if $N^1(1) = \bar{N}^1(1)$, then

$$\bar{N}^n(j) \leq_{st} \bar{N}^n(j), \quad \text{for all } n, \quad j \in M. \quad (21)$$

Note now that the queues in $S$ constitute a token passing ring with $|S|$ stations satisfying conditions (A1)-(A5) of Section 2, whose operation is independent of the interarrival processes in the persistent queues. The total average switchover time $\bar{u}_0$ to this ring is equal to

$$\bar{u}_0 = u_0 + \sum_{i \in U} \ell_i s_i.$$  

Let the queue lengths in such a system be denoted as $\{\bar{N}_S^n(i)\}_{i \in S}$. Clearly, $\bar{N}^n_S$ is a Markov chain, and since $|S| \leq M - 1$ we can apply the induction hypothesis. Hence, for $i \in S$, $\bar{N}^n_S$ is ergodic if

$$\lambda_i < \frac{\ell_i}{\bar{u}_0 + \sum_{i \in U} \ell_i s_i} \left(1 - \sum_{i \in S} \rho_i\right), \quad i \in S. \quad (22)$$

Assume now that (22) holds, and consider a queue in $S$, say queue 1 and let $C^n_S(1)$ be the process of cycle lengths (successive visits to queue 1). The processes $N^n_S(1), C^n_S(1)$ are regenerative with respect to the renewal process of the successive visits of the process $N^n_S(1)$ to state 0 (see the proof of Theorem 3). Since $N^n_S(1)$ is ergodic, the renewal process has finite mean, and therefore, we can construct a strictly stationary version of $(N^n_S(1), C^n_S(1))$ [18]. Using now Lemma 5 (and some technical arguments which are omitted here but can be found in [6]) we have that the process $\bar{N}^n(i), \quad i \in U$ is stable provided that

$$\lambda_i < \frac{\ell_i}{EC^n_S(1)} = \frac{\ell_i}{\bar{u}_0 + \sum_{i \in U} \ell_i s_i} \left(1 - \sum_{i \in S} \rho_i\right), \quad i \in U, \quad (23)$$

where the equality in (23) follows from the fact that by Theorem 3,

$$EC^n_S(1) = \frac{\bar{u}_0}{1 - \sum_{i \in S} \rho_i} \quad (24)$$

Since the process $\bar{N}^n(i), \quad i \in S$ is stable by construction, it follows from (21) that the irreducible, aperiodic Markov chain $\bar{N}^n(1)$ is substable and therefore, ergodic. The fact that $\bar{N}^n(j)$ is ergodic for all $j \in M$ follows from Theorem 3.

Putting everything together, from (22) and (23) we finally have that the Markov chain $\bar{N}^n(j)$ is ergodic for every $j \in M$ if

$$\lambda_i < \frac{\ell_i}{u_0 + \sum_{i \in U} \ell_i s_i} \left(1 - \sum_{i \in S} \rho_i\right), \quad i \in M. \quad (25)$$
It is easy to see that (25) holds also in the case \( S = \emptyset \), since in this case the cycles are i.i.d. random variables. Since (25) holds for every partition \( P = (S, U) \) of the set \( M \) such that \( S \neq M \), we conclude that the sufficient condition for stability of the system is

\[
\mathcal{R} = \bigcup_{S \subseteq M} \mathcal{R}_S ,
\]

(26)

where

\[
\mathcal{R}_S = \{ \lambda = (\lambda_1, \ldots, \lambda_M) : \text{condition (25) holds} \} .
\]

(27)

Finally, to complete the proof we need to show that

\[
\bigcup_{S \subseteq M} \mathcal{R}_S = \{ \lambda = (\lambda_1, \ldots, \lambda_M) : \lambda_i < \frac{\ell_i}{u_0} (1 - \sum_{i=1}^{M} \rho_i) \quad i \in M \} .
\]

(28)

This requires only algebraic manipulations which are almost identical to the ones in [6], therefore we omit them here. The interesting reader should be able to reproduce this algebra. \( \blacksquare \)

We can use Theorem 6 to establish some other stability results. For example, it can be extended to the process of queue lengths at arbitrary time instants, that is the process \( \tilde{N}(t) = (\tilde{N}_1(t), \ldots, \tilde{N}_M(t)) \), where \( \tilde{N}_i(t) \) is the queue length at queue \( i \) at time \( t \). Assume that \( N^n(1) \) is ergodic. Using the notation of the proof of Theorem 3, we have from (9) that \( \mathbb{E} \left( \sum_{k=1}^{R} C_k^1 \right) < \infty \), i.e., the renewal process \( \tilde{C}^n \) of the length of time between two successive returns to state 0 of the process \( N^n(1) \) has finite expectation. Since the interarrival times are exponential, this renewal process is non-lattice. Since \( \tilde{N}(t) \) is regenerative with respect to \( \tilde{C}^n \), we conclude that

**Corollary 7.** The process \( \tilde{N}(t) \) is stable if (20) holds. \( \blacksquare \)

Finally, we show in the next theorem that the conditions of Theorem 6 are also necessary for the ergodicity of the Markov chain \( N^n(i), \ i \in M \). This will establish necessary condition for stability of the \( \tau \)-limited token passing ring, and therefore it completes the proof of our Proposition from the Introduction.

**Theorem 8.** If for some \( i \in M \) the Markov Chain \( N^n(i) \) is ergodic, then \( N^n(j) \) is ergodic for every \( i \in M \). Moreover, \( \sum_{j=1}^{M} \rho_j < 1 \), and

\[
\lambda_j < \frac{\ell_j}{u_0} (1 - \rho_0) , \quad j \in M .
\]
**Proof.** The first assertion follows from Theorem 3 and the remark following that theorem. All cycles in the following will refer to queue 1. For simplicity of notation we omit the queue index from the various variables. Let us define.

- $C^n$: length of the $n$th cycle.
- $C^n(r)$: length of the $n$th cycle during which $r$ customers from queue 1 were served.
- $M^n(r)$: number of cycles in regeneration cycle $R^n$, (see proof of Theorem 3 for the definition of $R^n$) during which $r$ customers were served. Clearly,

$$R = \sum_{r=0}^{\infty} M(r),$$

where $M(r) = M^1(r)$ and $R = R^1$.

Since (by the ergodicity of the chain $N^n(1)$) $ER < \infty$, we have the following formulas for the long run averages:

- average length of a cycle during which $r$ customers were served,

$$EC(r) = \lim_{n \to \infty} \frac{\sum_{k=1}^{n} C^k(r)}{n} = \frac{E(\sum_{k=1}^{M(r)} C^k(r))}{EM(r)};$$

- probability (proportion) of cycles during which $r$ customers were transmitted,

$$P(r) = \lim_{n \to \infty} \frac{\sum_{k=1}^{n} M^k(r)}{n} = \frac{EM(r)}{ER}.$$

Consider now the following system.

**System S.** Upon arrival of the token to queue 1, the number of customers (from queue 1) that will be served in the next cycle enters system S. These customers stay in S until the token visits queue 1 for the next time, at which time all customers depart.

Clearly, the number of customers that enter system S in the $n$th cycle is $L^n_1$. Let $A^n_S$ be the number of customers that arrived in system S by time $t$. Recall the definition of the renewal process $\hat{C}^n$ in the paragraph before Corollary 7. $A^n_S$ is regenerative with respect to $\hat{C}^n$. and the ergodicity of $N^n(1)$ implies by Theorem 3 that $E\hat{C}^n < \infty$. Hence we have that

$$\lambda_S = \lim_{y \to \infty} \frac{A^n_S}{t} = \frac{\sum_{k=1}^{R} L^k_1}{\sum_{k=1}^{R} C^k} = \lambda_1,$$
where the last equality follows from (9). Similarly, taking into account that \(E(\sum_{r=1}^{\infty} r M(r)) = E(\sum_{k=1}^{R} L_k^i) < \infty\), we have the following formulas for the long-run average queue size, \(E_{NS}\), and long-run average waiting time, \(E_{WS}\), in system \(S\).

\[
E_{NS} = \frac{E \left( \sum_{r=1}^{\infty} r \sum_{k=1}^{M(r)} C_k(r) \right)}{E \left( \sum_{k=1}^{R} C_k \right)} = \frac{\sum_{r=1}^{\infty} r E \left( \sum_{k=1}^{M(r)} C_k(r) \right)}{ECER} \tag{33}
\]

\[
E_{WS} = \frac{E \left( \sum_{r=1}^{\infty} r \sum_{k=1}^{M(r)} C_k(r) \right)}{E(\sum_{r=1}^{\infty} r M(r))} = \frac{\sum_{r=1}^{\infty} r E \left( \sum_{k=1}^{M(r)} C_k(r) \right)}{\sum_{r=1}^{\infty} r E(M(r))} \tag{34}
\]

Using (30), (31), we derive from (33), (34),

\[
E_{NS} = \frac{\sum_{r=1}^{\infty} r P(r) EC(r)}{EC} \tag{35}
\]

\[
E_{WS} = \frac{\sum_{r=1}^{\infty} r P(r) EC(r)}{\sum_{r=1}^{\infty} r P(r)}. \tag{36}
\]

As in Theorem 1, let \(L_1\) be a random variable distributed as the steady state distribution of the process \(\{L^n_i\}_{n=1}^{\infty}\). Then \(\sum_{r=1}^{\infty} r P(r) = EL_1\). Since no more than \(L_1^n\) customers from queue \(i\) are served during the \(n\)th cycle, it is easy to see that \(L_1 \leq_{st} L_1^n\). If \(EL_1 = EL_1^n\), then the stochastic dominance relation implies that \(P(L_1 = 0) = P(L_1^n = 0) = 0\) and from (31) it follows that \(EM(0) = 0\). But then, \(P(L_1^n \geq 1, n = 1, 2, \ldots) = 1\) and since \(L_1^n \geq 1\) if and only if \(N^n(1) \geq 1\), we have that \(P(N^n(1) = 0, n = 1, 2, \ldots) = 0\) which contradict the ergodicity of the chain \(N^n(1)\). Therefore, \(\sum_{r=1}^{\infty} r P(r) < E\tilde{L}_1 = \ell_1\) and

\[
E_{WS} > \frac{\sum_{r=1}^{\infty} r P(r) EC(r)}{\ell_1}. \tag{37}
\]

Using Little's law (cf. [21]), (32) (35) and (37) we have

\[
E_{NS} = \lambda S E_{WS} > \lambda_1 \frac{EN_{NS} EC}{\ell_1},
\]

and therefore,

\[
\lambda_1 EC < \ell_1, \tag{38}
\]

and this proves Theorem 8, and completes the proof of our Proposition. □

**References**


