Stability Criteria For Queueing Networks with a Monotonicity Property

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

Report Number:
94-052
STABILITY CRITERIA FOR QUEUEING NETWORKS WITH A MONOTONICITY PROPERTY

Wojciech Szpankowski

Computer Sciences Department
Purdue University
West Lafayette, IN 47907

CSD TR-94-052
August 1994
Stability Criteria For Queueing Networks with a Monotonicity Property
September 24, 1993

W. Szpankowski*

A large class of queueing models possesses some special properties (most notably: monotonicity) that allows to derive stability criteria in a somewhat different and simpler manner (cf. [3, 4, 5, 7, 8, 9, 13, 15]). In this section, we first illustrate this new approach on the Jackson network, and then formulate general criteria for stability analysis of such queueing networks.

Informally, two properties are needed for our approach to work. One is the so called smaller copy property which requires that a queueing system with some queues saturated resembles the original network. The other postulate is the so called monotonicity property that refers to the increase of queue lengths (or waiting times) in all queues when some additional messages are introduced into the network (see below for formal definitions).

Consider now the Jackson network with $N$ queues, external arrival rate $\lambda_i$ and service rate $\mu_i$ for $i \in \mathcal{M} = \{1, 2, \ldots, N\}$. (The reader can assume at this moment of time that the arrival process to every queue is Poisson, and the service times are exponential, however, our analysis will work for a general arrival process with unbounded interarrival times which distribution is spread-out.) The routing is Markovian with the matrix $P = \{p_{ij}\}^{N}_{i,j}$. We denote such a system as $\mathcal{D}_N(\lambda, \mu, P)$ where $\lambda = (\lambda_1, \ldots, \lambda_N)$ and $\mu = (\mu_1, \ldots, \mu_N)$. Finally, let the effective input rate vector $\nu = (\nu_1, \ldots, \nu_N)$ satisfies the following system of linear equations

$$\nu = \lambda(I - P)^{-1} = \lambda M$$

where all elements $m_{ij}$ in the matrix $M = (I - P)^{-1}$ are nonnegative. This is due to the fact that $m_{ij}$ is the mean number of times a message arriving at queue $i$ visits queue $j$ before it leaves the network.

Our task is to show that the queue lengths vector $\xi(t) = (\xi_1(t), \ldots, \xi_N(t))$ is stable if and only if $\nu_i < \mu_i$ for all $i \in \mathcal{M}$. Since we plan to generalize our stability condition to general interarrival times and service times, we adopt here the following definition of stability. We say

*This research was primarily done while the author was visiting INRIA in Rocquencourt, France. The author wishes to thank INRIA (projects ALGO, MEVAL and REFLECS) for a generous support. Additional support was provided by NSF Grants NCR-9206315 and CCR-9201078, and in part by NATO Collaborative Grant 0057/89.
that a multidimensional processes \( \xi(t) = (\xi_1(t), \ldots, \xi_N(t)) \) (not necessarily Markovian) is stable if for \( x = (x_1, \ldots, x_N) \) the following holds

\[
\lim_{t \to \infty} \Pr\{\xi(t) < x\} = F(x) \quad \text{and} \quad \lim_{x \to \infty} F(x) = 1
\]

(2)

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

\[
\lim_{x \to \infty} \lim_{t \to \infty} \inf \Pr\{\xi(t) < x\} = 1 ,
\]

(3)

then the process is called either substable (cf. [12]) or tight or bounded in probability (cf. [14]). Otherwise, the system is unstable (cf. [12]).

We consider now a modified system in which the \( k \)-th queue always transmits maximum allowable number of messages, that is, one in our case, independently whether this queue is empty or not. In the case the queue is empty, it sends a dummy message. Note that the remaining queues form a smaller copy of the Jackson network with external arrival rate to queue \( i \in M - \{k\} := M_k \) equal to \( \lambda_i^{(k)} = \lambda_i + p_{ki} \mu_k \). We shall see that for the purpose of the stability analysis we can treat the \( k \)-th queue as a face \( \Lambda = \{k\} \), and the remaining queues as an induced Markov chain.

We first formally show that the remaining queues form an \( N - 1 \)-dimensional Jackson network. Indeed, define for all \( i \neq k \) the following

\[
\mu_i^{(k)} = \mu_i \quad \text{(4)}
\]

\[
\lambda_i^{(k)} = \lambda_i + p_{ki} \mu_k \quad \text{(5)}
\]

and \( P^{(k)} \) is the routing matrix \( P \) with the \( k \)-th column and row deleted. Then, clearly the system consisting of \( N - 1 \) queues (except the \( k \)-th one) form a smaller copy of the original Jackson network, namely \( D_{N-1}^{(k)} = D_{N-1}(\lambda^{(k)}, \mu^{(k)}, P^{(k)}) \). We also note that the effective input rates \( \nu_i^{(k)} \) for \( i \neq k \) in such a system become

\[
\nu^{(k)} = \lambda^{(k)} (I^{(k)} - P^{(k)})^{-1}
\]

(6)

where \( I^{(k)} \) is the identity matrix with the \( k \)-th column and row deleted.

Finally, we consider an \( N \)-dimensional Jackson network, say \( \tilde{D}_N^{(k)} \), that consists of the smaller copy system \( D_{N-1}^{(k)} \), and the queue \( k \). The queue lengths vector in such a system we denote as \( \tilde{\xi}^{(k)}(t) = (\tilde{\xi}_i^{(k)}(t), \ldots, \tilde{\xi}_N^{(k)}(t)) \). The arrival rates in \( \tilde{D}_N^{(k)} \) are defined as:

\[
\tilde{\lambda}_i^{(k)} = \begin{cases} 
\lambda_k & i = k \\
\lambda_i^{(k)} = \lambda_i + \mu_i p_{ki} & i \neq k
\end{cases}
\]

(7)
Furthermore, the effective input rate vector $\vec{v}^{(k)}$ in the system $\tilde{D}^{(k)}_N$ becomes

$$\vec{v}^{(k)}_i = \begin{cases} \lambda_k + \sum_{j=1}^{N} p_{j,k} \vec{v}_{j}^{(k)} & i = k \\ \nu^{(k)}_{i} & i \neq k \end{cases}$$

where $\nu^{(k)}_{i}$ are defined in (6) for all $i \neq k$.

The last system of equations can be re-written in the matrix form as follows. Let $\vec{P}^{(k)}$ be $N \times N$ matrix as $\vec{P}$ except that the $k$th column of $\vec{P}$ is replaced by $(0, \ldots, p_{kk}, \ldots, 0)^T$ where $T$ stands for transpose. Then, the above becomes

$$\vec{v}^{(k)}(I - \vec{P}^{(k)}) = \vec{\lambda}^{(k)}.$$  

(8)

Now, we are ready to formulate our results. It is convenient, however, to start with an informal description of the method. We first note that $\vec{\xi}(t) \leq_{st} \vec{\xi}^{(k)}(t)$, hence it suffices to establish stability of the dominant system $\tilde{D}^{(k)}_N$. For this we use an induction argument and Loynes result [12]. More precisely, we observe that the dominant system is composed of $N - 1$ dimensional system $D^{(k)}_{N-1}$ and the $k$th single queue that is practically decoupled from the system $D^{(k)}_{N-1}$. Using induction arguments we shall conclude that the system $D^{(k)}_{N-1}$ is stable in the region $R^{(k)} = \{(\lambda_1, \ldots, \lambda_{k-1}, \lambda_{k+1}, \ldots, \lambda_N) : \vec{v}_{i}^{(k)} < \mu_i$ for all $i \neq k\}$. By Loynes result we shall argue that the $k$th queue is stable when $\vec{v}_{k}^{(k)} < \mu_k$, thus the original system is stable in the region $\tilde{R}^{(k)} = \{\lambda : \vec{v}_{k}^{(k)} < \mu_k$, and $\vec{v}_{i}^{(k)} < \mu_i$ for all $i \neq k\}$, where $\vec{v}_{i}^{(k)}$ are defined in (8). Finally, stability region $R$ of the original system $D_N$ is a union of stability regions $\tilde{R}^{(k)}$ since if for some $k$ the dominant system $\tilde{D}^{(k)}_N$ is stable, then the original system is stable, too.

Having in mind the above plan, we now prove rigorously some of the above statements in the following lemmas.

**Lemma 1.** For any Jackson network we have

$$\vec{\xi}(t) \leq_{st} \vec{\xi}(t)$$

provided $\vec{\xi}(0) = \vec{\xi}(0)$.

**Proof.** This is a direct consequence of Foss lemma [5]. Indeed, consider two networks $D$ and $D'$ such that (sample path) arrival process epochs and service times are $\{T_i(n)\}$, $\{T'_i(n)\}$ and $\{S_i(n)\}$, $\{S'_i(n)\}$ ($i \in \mathcal{M}$ and $n = 1, \ldots$), respectively. Foss' lemma states: if $T_i(n) \leq T'_i(n)$ and $S_i(n) \leq S'_i(n)$ for all $i \in \mathcal{M}$ and $n = 1, \ldots$, then $D_i(n) \leq D'_i(n)$, where $D_i(n)$ and $D'_i(n)$ are departure epochs from $D$ and $D'$, respectively, provided $\sum_{n=1}^{\infty} S_i(n) = \infty$ (a.s.). In order to apply the above lemma to prove (9), one proceeds as follows: whenever a queue becomes empty in the original model $D$, then one immediately schedules the service time for the next customer.
(from the service times sequence \{S_i(n)\} in the system \mathcal{D}'. In other words, we shift forward the arrival times for some customers, and therefore \(T_i(n) \geq \tilde{T}_i(n)\). Since interarrival times and service times are i.i.d., an easily application of the Foss lemma gives the required dominance. \(\blacksquare\)

**Lemma 2.** Let

\[
\mathcal{R} = \{ \lambda : \nu_i < \mu_i \text{ for all } i \in \mathcal{M} \} \tag{10}
\]

\[
\tilde{\mathcal{R}}^{(k)} = \{ \lambda : \tilde{\nu}_k^{(k)} < \mu_k, \text{ and } \tilde{\nu}_i^{(k)} < \mu_i \text{ for all } i \in \mathcal{M}_k = \mathcal{M} - \{k\} \} . \tag{11}
\]

Then,

\[
\mathcal{R} = \bigcup_{k=1}^N \tilde{\mathcal{R}}^{(k)} . \tag{12}
\]

**Proof.** In the proof we shall use the following identity that follows from subtracting (8) from (1) (this is a correct version of Eq.(17) of [2])

\[
z(I - P) = v^{(k)} \tag{13}
\]

where

\[
v_i^{(k)} = \begin{cases} 
1 - p_{kk} (\tilde{\nu}_k^{(k)} - \mu_k) & i = k \\
0 & i \neq k 
\end{cases} \tag{14}
\]

and

\[
z_i = \begin{cases} 
\nu_k - \mu_k & i = k \\
\nu_i - \tilde{\nu}_i^{(k)} & i \neq k 
\end{cases} . \tag{15}
\]

Enriched in (13) we proceed as follows. We first prove that for every \(k\) we have \(\tilde{\mathcal{R}}^{(k)} \subset \mathcal{R}\). From (1), (13) and (14) we observe that \(z_i = v_i^{(k)} m_{ki}\) for all \(i \in \mathcal{M}\) where \(m_{ki} > 0\) are elements of the matrix \(\mathcal{M}\). Since \(\lambda \in \tilde{\mathcal{R}}^{(k)}\), hence \(\tilde{\nu}_k^{(k)} < \mu_k\), and therefore \(v_k^{(k)} < 0\) which implies \(z_k < 0\) and hence \(\nu_k < \nu_k\). Now, consider \(i \neq k\). From the above we have \(z_i < 0\), so \(v_i < \tilde{\nu}_i^{(k)}\). But, for \(\lambda \in \tilde{\mathcal{R}}^{(k)}\) the following holds \(\tilde{\nu}_i^{(k)} < \mu_i\), hence \(v_i < \mu_i\). This completes the proof of the inclusion \(\bigcup_{k=1}^N \tilde{\mathcal{R}}^{(k)} \subset \mathcal{R}\).

The proof of \(\mathcal{R} \subset \bigcup_{k=1}^N \tilde{\mathcal{R}}^{(k)}\) is only a little more intricate. As before, for some \(k\) we have \(z_i^{(k)} = v_i^{(k)} m_{ki}\). Since \(v_i < \mu_i\) for \(\lambda \in \mathcal{R}\) we conclude that \(z_k < 0\) and hence \(v_k^{(k)} < 0\) which finally implies that \(\tilde{\nu}_k^{(k)} < \mu_k\). We now must prove that this inequality holds for all \(i \neq k\) and some \(k \in \mathcal{M}\).

First of all, observe that from the definition of \(z_i\) and (13)-(15) for any \(i \in \mathcal{M}\) we have

\[
\tilde{\nu}_i^{(k)} = \nu_i + m_{ki} (1 - p_{kk}) (\mu_k - \tilde{\nu}_k^{(k)}) . \tag{16}
\]
It is also easy to see that $\nu_i$ as well as $\tilde{\nu}_i^{(k)}$ are increasing functions of $\lambda$ in sense that $\tilde{\nu}_i^{(k)}(\lambda_1, \ldots, \lambda_j + \delta, \ldots, \lambda_N) \geq \tilde{\nu}_i^{(k)}(\lambda_1, \ldots, \lambda_j, \ldots, \lambda_N)$ for $\delta > 0$ (and the same with $\nu_i$). Furthermore, $\tilde{\nu}_i^{(k)}$ is an increasing function of $\nu$ at least for $\tilde{\nu}_k^{(k)} < \mu_k$ (cf. (16)). Clearly, $\nu_k = \mu_k$ implies $\tilde{\nu}_k^{(k)} = \mu_k$. All together, (16) and the above lead to the following for some $k$

$$\tilde{\nu}_i^{(k)}(\nu_1, \ldots, \nu_k, \ldots, \nu_N) \leq \tilde{\nu}_i^{(k)}(\nu_1, \ldots, \mu_k, \ldots, \nu_N) = \mu_i$$

and this proves the inclusion $\mathcal{R} \subset \bigcup_{k=1}^N \hat{\mathcal{R}}^{(k)}$, as needed. 

Finally, we are in the position to prove our main result. We formulate it in the most general form to visualize the power of our method of proof.

**Theorem 3.** The Jackson network with general service times distribution and general unbounded and spread-out interarrival times distribution is stable if and only if $\lambda \in \mathcal{R}$ where $\mathcal{R}$ is defined as in (10), that is, iff $\nu_i < \mu_i$ for all $i \in \mathcal{M}$.

**Proof.** We first use induction arguments to establish the sufficient part of Theorem 3. More precisely: we first prove sub-stability (tightness) of the queue lengths vector $\xi(t)$.

* Sufficiency. For $N = 1$ the theorem holds. Consider now an $N$-dimensional Jackson network. By Lemma 1, it suffices to prove the stability of the dominant system $\tilde{D}^{(k)}$ represented by $\tilde{\xi}^{(k)}(t)$. Observe that the dominant system can be decomposed into two subsystems: the $N - 1$-dimensional copy $\mathcal{D}^{(k)}(\lambda^{(k)}, \mu, P^{(k)})$ of the original network, and a single queue, namely, the $k$th queue. By the induction, for

$$(\lambda_1, \ldots, \lambda_{k-1}, \lambda_{k+1}, \ldots, \lambda_N) \in \mathcal{R}^{(k)} = \{\nu_i^{(k)} < \mu_i \text{ for all } i \neq k\}$$

the subsystem $\mathcal{D}^{(k)}$ is stable. So, we only need to prove a stability criterion for the $k$th queue.

Consider now the single queue $k$. Observe that the total arrival process to this queue is generated by the subsystem $\mathcal{D}_{N-1}^{(k)}$. Since this subsystem is stable in $\mathcal{R}^{(k)}$ – as proved above – we can make it stationary. Then, the total arrival process to the $k$th queue is a superposition of stationary and ergodic processes, hence it is stationary and ergodic too (cf. [1]). Thus, the non-Markovian queue $k$ satisfies the assumption of the Loynes theorem (cf. [12]), so it is stable for $\nu_k^{(k)} < \mu_k$. Using the isolation lemma of [7, 15] we conclude that the aggregate queue lengths $\tilde{\xi}(t)$ for the system $\tilde{D}^{(k)}$ is substable in the region $\hat{\mathcal{R}}^{(k)}$ defined in (11). Finally, observe that if for some $k$ the dominant system $\tilde{D}^{(k)}$ is substable, then the original system $\mathcal{D}$ is substable. Thus, the original system $\mathcal{D}$ is substable in the region $\bigcup_{k=1}^N \hat{\mathcal{R}}^{(k)}$. But, by Lemma 2 this region is equal to $\mathcal{R} = \{\lambda : \nu_i < \mu_i, \text{ for all } i \in \mathcal{M}\}$, as desired.

For Jackson network with Poisson-exponential assumptions, the sub-stability of $\tilde{\xi}(t)$ automatically implies stability of the process since $\tilde{\xi}(t)$ is a Markov chain defined on a countable
state space (cf. [15]). In general case, however, we need to appeal to recent result of Meyn and Tweedie [14]. In particular, using the same arguments as in [3, 13] we conclude that under spread-out and unboundness assumptions of the interarrival time, the substability of $\tilde{\xi}(t)$ leads to its stability.

**Necessity.** The necessary part of the proof can be done in many ways. By the isolation lemma of [7, 15], to prove instability we need to show that at least one queue is unstable. Surprisingly simple proof of this fact can be obtained by an application of the Little's formula. Let the Jackson network be stable and in a stationary regime. We prove that if, say the $j$th queue is stable, then $\nu_j < \mu_j$. Consider a subsystem $S_j$ that consists only of the $j$th queue server. To such a subsystem we apply the Little's formula. Note that the waiting time in $S_j$ is equal to the service time, and the arrival process to $S_j$ has intensity $\nu_j$. Clearly, the queue length in $S_j$ is bounded by one. Thus, by Little's formula $\nu_j/\mu_j = \text{queue length in } S_j \leq 1$. We can eliminate the weak inequality in the above as in [7].

Finally, we offer some concluding remarks. By no means the method presented here is restricted to Jackson-type networks. However, it works only for queueing systems since the real engine behind it is the Loynes result. Before we discuss some further examples, we propose below a set of assumptions that are necessary for our method to work. More details can be found in [9, 15].

Let $D(\lambda, \mu, P) = D$ be a generic queueing model where $\lambda, \mu$ are the input rate vector and the service rate vector, respectively, while $P$ is a matrix of system parameters. By $\xi(t) = (\xi_1(t), \ldots, \xi_N(t))$ we denote the queue lengths vector in $D$. We partition the set of all users $M = \{1, \ldots, N\}$ into $P = (\Lambda, \Lambda)$ such that $M = \Lambda \cup \Lambda$. Users in $\Lambda$ can be treated as never empty since they are can transmit maximum allowable number of messages. Users in $\Lambda$ work as in the original system. In short, $\Lambda$ can be interpreted as a face while a process defined on $\Lambda$ as the induced Markov chain provided there is underlying Markovian property of the system (that is actually not required for the method to work!). For example, for the Jackson network analyzed above we assumed $\Lambda = \{k\}$ and $\Lambda = M_k = M - \{k\}$. Finally, the queue lengths vector $\xi^P(t)$ under partition $P$ can be partitioned as $\xi^P(t) = (\xi^{\Lambda}(t), \xi^{\Lambda}(t))$ where $\xi^{\Lambda}(t)$ (resp. $\xi^{\Lambda}(t)$) represents the queue lengths in the set $\Lambda$ (resp. $\Lambda$).

We postulate the following properties to hold:

(P1) **Smaller Copy Assumption.** Consider a modified system in which all queues in $\Lambda$ are treated as saturated, that is, they always send maximum allowable number of (dummy, if necessary) messages. We assume that such a system restricted to users in $\Lambda$, denoted as $D^\Lambda$, is a copy of the original system with (possibly) new parameters $\lambda^\Lambda, \mu^\Lambda$ and $P^\Lambda,$
that is, $D^\Lambda = D(\lambda^\Lambda, \mu^\Lambda, P^\Lambda)$.

(P2) **Weak Monotonicity.** Consider the modified system $\tilde{D}^\Lambda$ that consists of the system $D^\Lambda$ (defined in (P1)) and users in $\Lambda$. Then, for every $P = (\Lambda, \overline{\Lambda})$

$$\xi(t) \leq_{st} \tilde{\xi}^P(t).$$

(P3) **Stationarity.** Consider a queue $k \in \Lambda$. Let $\tilde{\xi}^k(t)$ be in a stationary regime, that is, $\tilde{\xi}^k(t)$ is a strictly stationary and ergodic process. Then, the (effective) input process and/or the (effective) service process of the $k$th queue are jointly stationary and ergodic (as needed for the Loynes result).

Generalizing our analysis of Jackson network, we can establish sufficient and necessary condition for stability of a class of queueing networks satisfying our postulates (P1)-(P3). This method was already successfully applied to token passing rings [7, 8], ALOHA system [15], rings with spatial reuse and finite quota [10]. We should mentioned, however, that the crucial monotonicity condition (P2) does not hold for some important queueing systems such as rings with spatial reuse and without quota, coupled processors, and ETHERNET (i.e., exponential back-off algorithm) [11].

Finally, we stress that if a stronger monotonicity property holds (that is, the queue length vector increases with the decrease of the service rates at a subset of users), then even simpler approach to stability criteria is possible (cf. [9]). For example, Jackson network of queues and token passing rings satisfy this stronger monotonicity property, while ALOHA system does not. For details we refer the reader to [9].

**References**


