On the Saturation Principle for the Stability of Queues

F. Baccelli* and S. Foss†

August 20, 1993

Abstract

This paper focuses on the stability of open queueing systems under stationary ergodic assumptions. It provides a general framework ensuring that the stability region is given by the following saturation principle: 'saturate' the queues which are fed by the external arrival stream; look at the 'intensity' μ of the departure stream in this saturated system; then stability holds whenever the intensity of the arrival process, say λ satisfies the condition $\lambda < \mu$. This framework involves two main pathwise properties, external monotonicity and separability, which are satisfied by several classical queueing networks. The main tool for the proof of this principle is subadditive ergodic theory. Whenever this condition is satisfied, it is also shown that certain state variables associated with the network admit a finite stationary regime which is constructed pathwise using a Loynes type backward argument.

Keywords: Open queueing network, stationary point processes, monotonicity, separability, first and second-order ergodic properties, subadditive ergodic theorem.

*INRIA Centre Sophia Antipolis 06565 Valbonne France; baccelli@sophia.inria.fr; the work of this author was supported in part by a grant from the European Commission DG XIII, under the BRA Qmips contract

†NOVOSIBIRSK State University, 630090 Novosibirsk, Russia; foss@math.nsk.su; the work of this author was supported by a sabbatical grant from INRIA Sophia Antipolis

1 Introduction

A folk theorem of queueing theory states that the stability region of an open queueing systems can be obtained as follows: 'saturate' the queues which are fed by the external arrival stream with an infinite customer population; if μ denotes the 'intensity' of the departure stream in this saturated system, then the system is stable when the intensity of the arrival process, λ, satisfies $\lambda < \mu$. This engineering rule, which we will refer to as the saturation principle, was initially designed for Markovian queueing systems, where μ could be obtained by computing the steady state of a Markov chain of smaller dimension than that of the initial non-saturated system (see for instance [7]).

The aim of the present paper is to set a natural framework in which this principle can be rigorously proved when dropping the Markovian assumptions and replacing them by standard stationary ergodic assumptions, namely the arrival point process is a stationary ergodic marked point process of finite intensity. The conditions required by this framework consist of easy-to-check, pathwise properties of the system under consideration. The two main conditions are *external monotonicity* and *separability*, although two other mild conditions called *causality* and *translation linearity* are also assumed.

Whenever these properties holds, we prove that

- The time X_n of the last activity to take place in the system (this often coincides with the time to empty the system in the queueing context) whenever one starts with n customers, all arrived at time 0 (this is the adequate extension of the saturation idea to non-Markov systems), satisfies a strong law of large numbers:

$$\lim_n \frac{X_n}{n} = \gamma \quad \text{a.s.}, \tag{1}$$

  for some positive constant γ; this constant will also be finite if the input marked process satisfies natural integrability conditions to be specified in due time;

- Whenever the intensity of the input process satisfies the condition $\lambda < \gamma^{-1}$, then the system is stable; the stability here means that the time to

inactivity (usually to empty the system) when it is fed by the restriction of the point process on $(-\infty, t)$ admits a finite steady state regime. This usually implies most standard acceptations of stability. This stationary regime is constructed pathwise using a backward argument of the same nature as the argument of Loynes in the $G/G/1$ queue.

Several classes of queueing systems for which the stability is an open question (at least in the stationary ergodic case) fall within this *monotone separable framework*, like for instance Jackson type queueing networks or polling systems. It also contains other classes of discrete event systems like Petri nets.

The necessity of the main conditions of this framework for the saturation principle to hold is discussed in the final section.

2 The Monotone Separable Framework

Arrival Point Process Consider a marked point process with arrival points $\{t_n\}_{n\in\mathbb{Z}}$ and marks $\{\eta_n\}_{n\in\mathbb{Z}}$, where $\eta_n \in E$. Let $t = \{t_n\}$ (we have $t_n \leq t_{n+1},\ \forall n$) and $\eta = \{\eta_n\}$. In relation with this point process, we shall use the following notations:

$$u_n \stackrel{\text{def}}{=} t_n - t_0, \quad \tau_n \stackrel{\text{def}}{=} t_n - t_{n-1}, \quad \forall n. \tag{2}$$

Let $u = \{u_n\}$ and $\tau = \{\tau_n\}$. We shall stress the dependency on t by writing $u(t)$ in place of u or $\tau(t)$ in place of τ, with a similar meaning for $u_n(t)$ or $\tau_n(t)$. Finally, for c real, we shall denote $t + c$ the sequence $\{t_n + c\}$.

First Order State Variables We are given a set of functions $\{f_n\}$, $f_n : I\!R^n \times E^n \to I\!R$. For each m, n, t, define

$$X_{[m,n]}(t, \eta) = f_{n-m+1}\{(t_l, \eta_l),\ m \leq l \leq n\}. \tag{3}$$

When needed, we shall drop one of the parameters (either t or η), or both. We shall also assume that $t_0 = 0$, so that these variables can also be considered as functions of u or τ in place of t. We assume that the function f_n satisfy the following properties:

1. **(causality)** $\forall m \leq n,\ \forall t\ X_{[m,n]}(t) \geq t_n$;

2. **(external monotonicity)** $\forall m \leq n$, $X_{[m,n]}(t') \geq X_{[m,n]}(t)$, whenever $t' \geq t$ coordinatewise;

3. **(translation-linearity)** $\forall t$, $\forall c \in I\!R$, $\forall m \leq n$

$$X_{[m,n]}(t + c) = X_{[m,n]}(t) + c;$$

4. **(separability)** If for $m \leq l < n$, $X_{[m,l]}(t) \leq t_{l+1}$, then

$$X_{[m,n]}(t) = X_{[l+1,n]}(t).$$

We shall also use the notation

$$Z_{[m,n]}(t) = X_{[m,n]}(t) - t_n \equiv X_{[m,n]}(t - t_n). \tag{4}$$

By definition $Z_{[m,n]}(t)$ depends on $\{\tau_l\}_{m+1 \leq l \leq n}$ only. In particular,

$$Z_n(t) \stackrel{\text{def}}{=} Z_{[n,n]}(t) \tag{5}$$

does not depend on t.

Properties of the State Variables

Lemma 1 (internal monotonicity of X.) $\forall m \leq n, \quad \forall t$

$$X_{[m-1,n]}(t) \geq X_{[m,n]}(t) \tag{6}$$

Proof Consider the sequence $t' = \{t'_j\}$, where

$$t'_j = \begin{cases} t_j - Z_{m-1}(t) & \text{for } j \leq m-1; \\ t_j & \text{for } j \geq m\ . \end{cases} \tag{7}$$

Since the restriction of t and t' to $j \geq m$ coincide, $X_{[m,n]}(t) = X_{[m,n]}(t')$. The separability assumption implies that $X_{[m-1,n]}(t') = X_{[m,n]}(t')$. Finally, external monotonicity implies that $X_{[m-1,n]}(t') \leq X_{[m-1,n]}(t)$. □

An immediate consequence is:

Corollary 1 (internal monotonicity of Z.) $\forall m \leq n, \quad \forall t$

$$Z_{[m-1,n]}(t) \geq Z_{[m,n]}(t). \tag{8}$$

Lemma 2 (subadditive property of Z.) $\forall m \le l < n, \quad \forall t$

$$Z_{[m,l]}(t) + Z_{[l+1,n]}(t) \ge Z_{[m,n]}(t). \tag{9}$$

Proof Introduce two auxiliary sequences $t^1 = \{t_j^1\}$ and $t^2 = \{t_j^2\}$ defined by

$$t_j^1 = \begin{cases} t_j & \text{for } j \le l; \\ t_j + Z_{[m,l]}(t) & \text{for } j > l\ . \end{cases} \tag{10}$$

and

$$t_j^2 = \begin{cases} t_j - Z_{[m,l]}(t) & \text{for } j \le l; \\ t_j & \text{for } j > l\ . \end{cases} \tag{11}$$

So $t_j^2 = t_j^1 - Z_{[m,l]}(t)$ for all j. Then

$$\begin{aligned} X_{[m,n]}(t) &\overset{2}{\le} X_{[m,n]}(t^1) \overset{4}{=} X_{[l+1.n]}(t^1) \\ &\overset{3}{=} X_{[l+1.n]}(t^2) + Z_{[m,l]}(t) = X_{[l+1.n]}(t) + Z_{[m,l]}(t). \end{aligned}$$

Therefore

$$Z_{[m,n]}(t) = X_{[m,n]}(t) - t_n \le X_{[l+1,n]}(t) - t_n + Z_{[m,l]}(t) = Z_{[l+1,n]}(t) + Z_{[m,l]}(t).$$

□

3 Proof of the Saturation Principle

Stationarity Assumptions Assume the variables $\{\tau_n, \eta_n\}$ are random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbf{P}, \theta)$, where θ is measure-preserving shift transformation, such that $(\tau_n, \eta_n) \circ \theta = (\tau_{n+1}, \eta_{n+1})$. The following integrability assumptions are also assumed to hold:

$$0 \le \mathbf{E}\tau_n \overset{\text{def}}{=} \lambda^{-1} < \infty, \tag{12}$$

and

$$\mathbf{E}Z_1 < \infty. \tag{13}$$

First-Order Ergodic Property

Lemma 3 *There exists a constant γ such that the a.s. limits*

$$\lim \frac{Z_{[-n,-1]}(\tau)}{n} = \lim \frac{\mathbf{E}Z_{[-n,-1]}(\tau)}{n} = \lim \frac{Z_{[1,n]}(\tau)}{n} = \lim \frac{\mathbf{E}Z_{[1,n]}(\tau)}{n} = \gamma \text{ a.s.} \tag{14}$$

holds.

Proof The proof follows directly from the Kingman's subadditive ergodic theorem. □

Corollary 2 *Under the foregoing assumptions*

$$\lim X_{[1,n]}(t) = \gamma + \lambda^{-1}. \tag{15}$$

0-1 Law Let A be the event

$$A = \{\lim Z_{[-n,0]}(\tau) = \infty\}. \tag{16}$$

Theorem 1 *Under the foregoing ergodic assumption,* $\mathbf{P}(A) \in \{0, 1\}$.

Proof Note that $A \circ \theta^{-1} = \{\lim Z_{[-n,-1]} = \infty\}$. But owing to the subadditive property,

$$Z_{[-n,-1]}(u) \geq Z_{[-n,0]}(u) - Z_0. \tag{17}$$

This and the integrability of Z_0 imply that $A \circ \theta^{-1} \supseteq A$. Since θ is ergodic, the proof is concluded. □

Scaling Factor For all $0 \leq C < \infty$, the sequences

$$X_{[m,n]}(Cu) \stackrel{\text{def}}{=} f_{n+1-m}\{(C \cdot u_l, \eta_l); m \leq l \leq n\} \tag{18}$$

and

$$Z_{[m,n]}(Cu) = X_{[m,n]}(Cu) - C \cdot u_n \tag{19}$$

enjoy all the monotonicity and subadditive properties mentioned above. In addition, for all n

(a) $Z_{[-n,-1]}(Cu)$ is decreasing in C;

(b) $X_{[1,n]}(Cu)$ is increasing in C $\forall n$.

So for all $C \geq 0$ we can define $\gamma(C)$; we have

(a) $\gamma(C)$ is decreasing in C;

(b) $\gamma(C) + C \cdot \lambda^{-1}$ is increasing in C.

Second-Order Ergodic Property

Theorem 2 *If* $\lim Z_{[-n,-1]} \to \infty$ *a.s., then* $\lambda \cdot \gamma(0) \geq 1$. *If* $\lambda \cdot \gamma(0) > 1$, *then* $\lim Z_{[-n,-1]} \to \infty$ *a.s.*

Proof We first prove the second assertion. Let 0 be the sequence with all its elements equal to 0: $0 = \{0\}$; for all n let t^n be the sequence with elements

$$t^n_j = u_{-n}, \quad \forall j. \tag{20}$$

Then

$$\begin{aligned} Z_{[-n,0]}(u) &= X_{[-n,0]}(u) \overset{2}{\geq} X_{[-n,0]}(t^n) \\ &\overset{3}{=} X_{[-n,0]}(0) - u_{-n} = Z_{[-n,0]}(0) - u_{-n} \end{aligned}$$

and

$$\liminf Z_{[-n,0]}(u)n \geq \gamma(0) + \lambda^{-1} > 0,$$

which concludes the proof of the second assertion.

As for the second one, for each integer $l \geq 1$, let N_l be the random variable

$$N_l = \min\{n \geq 1 : \quad Z_{[-n,0]}(u) \geq u_l\}. \tag{21}$$

Owing to the subadditive property, for all $n, l \geq 0$

$$Z_{[-n,l]}(u) \leq Z_{[-n,0]}(u) + Z_{[1,l]}(u) \leq Z_{[-n,0]}(u) + \sum_{i=1}^{l} Z_i, \quad \text{a.s.}, \tag{22}$$

where the random variables $Z_i = Z_i(u) = Z_0 \circ \theta^i$ do not depend on u and are integrable. For all $n \geq 1$, let t^n be the sequence with elements

$$t_j^n = \begin{cases} u_j & \text{for } j \leq 0; \\ Z_{[-n,0]}(u) & \text{for } j \geq 1 \end{cases} \tag{23}$$

and let $\tilde{t}^n$ be the sequence

$$\tilde{t}_j^n = Z_{[-n,0]}(u), \quad \text{for all } j. \tag{24}$$

Then

$$\begin{aligned} X_{[-n,l]}(u)I(n \geq N_l) &\overset{2}{\leq} X_{[-n,l]}(t^n)I(n \geq N_l) \\ &\overset{4}{=} X_{[1,l]}(t^n)I(n \geq N_l) = X_{[1,l]}(\tilde{t}^n)I(n \geq N_l) \\ &\overset{3}{=} \left(Z_{[-n,0]}(u) + X_{[1,l]}(0)\right) I(n \geq N_l) \\ &= \left(Z_{[-n,0]}(u) + Z_{[1,l]}(0)\right) I(n \geq N_l). \end{aligned}$$

Therefore

$$\begin{aligned} Z_{[-n,l]}(u)I(n \geq N_l) &= \left(X_{[-n,l]}(u) - u_l\right) I(n \geq N_l) \\ &\leq (Z_{[-n,0]}(u) + Z_{[1,l]}(0) - u_l)I(n \geq N_l). \end{aligned}$$

Finally,

$$\begin{aligned} & Z_{[-n,l]}(u) - Z_{[-n,0]}(u) \\ &= \left(Z_{[-n,l]}(u) - Z_{[-n,0]}(u)\right) [I(n < N_l) + I(n \geq N_l)] \\ &\leq \left(\sum\nolimits_{i=1}^{l} Z_i\right) I(n < N_l) + \left(Z_{[1,l]}(0) - u_l\right) I(n \geq N_l) \\ &\leq \left(\sum_{i=1}^{l} Z_i + Z_{[1,l]}(0) + u_l\right) I(n < N_l) + Z_{[1,l]}(0) - u_l. \end{aligned} \tag{25}$$

Let

$$\psi_l \overset{\text{def}}{=} \sum_{i=1}^{l} Z_i + Z_{[1,l]}(0) + u_l. \tag{26}$$

Note that $\mathbf{E}\psi_l < \infty$. By making use of the relations $Z_{[-n,l]}(u) = Z_{[-n-l,0]}(u) \circ \theta^l$, $Z_{[-n-l,0]}(u) \geq Z_{[-n,0]}(u)$ and $\mathbf{E}Z_{[-n-l,0]}(u) < \infty$, we obtain from (25) that

$$\begin{aligned} 0 &\leq \mathbf{E}Z_{[-n,l]}(u) - \mathbf{E}Z_{[-n,0]}(u) \\ &\leq \mathbf{E}\{\psi_l \cdot I(n < N_l)\} + \mathbf{E}Z_{[1,l]}(0) - l\lambda^{-1}. \end{aligned} \tag{27}$$

If $\lim Z_{[-n,-1]} \to \infty$ a.s., then N_l is a.s. finite for all l. In this case, the right-hand side of the last equation tends to

$$\mathbf{E}Z_{[1,l]}(0) - l \cdot \lambda^{-1}$$

as $n \to \infty$. Therefore

$$\frac{\mathbf{E}Z_{[1,l]}(0)}{l} \geq \lambda^{-1}, \tag{28}$$

for all l. Finally, when letting l go to infinity and when making use of Lemma 3, we obtain

$$\gamma(0) = \lim_l \frac{\mathbf{E}Z_{[1,l]}(0)}{l} \geq \lambda^{-1}. \tag{29}$$

□

4 Examples and Counter-Example

The $G/G/1$ Queue. Here $X_{[m,n]}(t)$ is the departure epoch of customer n, when there are $n+1-m$ customers with arrival epochs t_l and service times σ_l, $m \leq l \leq n$; $Z_{[m,n]}(t)$ is then the sojourn time of customer n. The computation of $\gamma(0)$ is trivial, by the strong law of large numbers.

The $G/G/s$ Queue. $X_{[m,n]}(t)$ is the last departure epoch from the queue with $n+1-m$ customers with arrival epochs t_l and service times σ_l, $m \leq l \leq n$.

$$Z_{[m,n]}(t) = \max_{1 \leq i \leq s} W_{i,[m,n]}(t) = W_{s,[m,n]}(t), \tag{30}$$

where $W_{[m,n]} = (W_{1,[m,n]}(t), \ldots, W_{s,[m,n]}(t)$ is Kiefer-Wolfowitz's vector.
The constant $\gamma(0)$ is obtained as follows: there exists a stationary sequence ζ_n such that

$$W_{s,[m,n]}(0) - W_{1,[m,n]}(0) \leq \zeta_n \quad \text{a.s.} \tag{31}$$

for all $m \leq n$ (see for instance Borovkov,1984). Therefore

$$Z_{[1,n]} \leq W_{s,[1,\frac{n}{2}]}(0) + Z_{[\frac{n}{2}+1,n]} \quad \text{a.s.} \tag{32}$$

and

$$\frac{W_{s,[1,\frac{n}{2}]}(0)}{n} \to \frac{\mathbf{E}\sigma_1}{2} s. \tag{33}$$

FIFO Kelly-type Networks Customers are given a route through the network upon arrival (namely the route of customer n is a mark associated with point t_n). Then the external monotonicity property is not satisfied. Note that the conditions stating that the traffic intensity be less than one at each queue is not sufficient for monotonicity, even in the FIFO case.

Kelly-type Networks with Synchronization Constraints These networks were defined in [2]. The routing is mechanism is the same as above, but the service discipline is such that, on a given queue, service requirements brought by the n-th external arrival can in no case be served before all the service requirements brought by the $n-1$-st external arrival have been completed. There, $X_{[m,n]}(t)$ is also the last departure epoch from the system with restricted arrivals as above. The constant $\gamma(0)$ is the $(\max,+)$ Lyapunov exponent of a sequence of stationary random matrices defined from the routing and the service times (see [1] for the notions pertaining to $(\max,+)$-Lyapunov exponents).

Jackson-type Networks Whenever the routes and the service times are stationary ergodic sequences associated with stations (namely, the n-th customer to reach station j requires the service time σ_n^j, and is routed to station q_n^j, where $\{\sigma_n^j\}_n$ and $\{q_n^j\}_n$ are stationary ergodic sequences, then the network falls in the monotone separable framework. The stability region is given by the condition $\rho_j < 1$, for all queues, where ρ_j is the traffic intensity at queue j (see [3]).

Polling Systems

Stochastic Petri Nets Open FIFO stochastic event graphs fall in our framework, namely they satisfy the four conditions of § 2 (see [1]). The constant $\gamma(0)$ is also the $(\max,+)$ Lyapunov exponent of a sequence of random matrices defined from the topology of the network and the service times.
For general open stochastic Petri nets with switching, also defined in [1], take $X_{[m,n]}$ to be the epoch of the last firing time of whole net for the restricted arrival point process $\{t_l\}$, $m \leq l \leq n$. Then all the conditions of the framework are satisfied (see [4]).

A Non-separable Network for which the Principle Does Not Hold Consider an assembly queue with two independent Poisson arrival streams with the same intensity $\lambda/2$. The system starts empty. Whenever there are customers of both classes in the queue, service is provided at rate μ. The completion of a service consumes one customer of each class. Whenever the queue has no customers of either class, it is blocked. Let consider as input stream the superposition of the two Poisson processes properly marked. If one saturates the system with an infinite customer population, the (Markov) departure rate is μ. Similarly, if one takes that viewpoint of letting n customers of this input stream arrive at time 0, the last activity of the system takes place at time $\mu^{-1}n/2 + o(n)$. A naive application of the saturation principle would suggest that if $\lambda < \mu$, the system is stable. However, such queues are always unstable (see [1]), whatever the values of λ and μ. Note that this system does not satisfy the separability property.

A Non-monotone Network for which the Principle Does Not Hold

References

[1] F. Baccelli, G. Cohen, G. Olsder and J.P. Quadrat. Linearity and synchronization, Wiley, 1992.

[2] F. Baccelli and Z. Liu. On a class of stochastic evolution equation s. Ann. Appl. Prob. 20-2: 350-374, 1992.

[3] F. Baccelli and S. Foss. Stability of Jackson-type queueing networks. INRIA Report No. 1945, June 1993, QUESTA, to appear.

[4] F. Baccelli and S. Foss. Stability of stochastic Petri nets with general topology. In preparation.

[5] A.A. Borovkov. Asymptotic methods in queueing theory. Wiley, 1984.

[6] Foss, Chernova Polling systems?

[7] S. Lavenberg, Maximum Departure Rate of certain Open Queueing Networks Having Finite Capacity Constraints, RAIRO série bleue, 1978, 12-4, Paris