Large Scale and Heavy Traffic Asymptotics for Systems with Unreliable Servers

Jean-François Dantzer, Isi Mitrani, Philippe Robert

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Large Scale and Heavy Traffic Asymptotics for Systems with Unreliable Servers

Jean-François Dantzer — Isi Mitrani — Philippe Robert

No 3807

Novembre 1999

THÈME 2

ISSN 0249-6399 ISRN INRIA/RR--3807--FR+ENG

Large Scale and Heavy Traffic Asymptotics for Systems with Unreliable Servers

Jean-François Dantzer , Isi Mitrani , Philippe Robert

Thème 2 — Génie logiciel
et calcul symbolique
Projets Algo

Rapport de recherche no 3807 — Novembre 1999 — 20 pages

Abstract: The asymptotic behaviour of the $M/M/n$ queue, with servers subject to independent breakdowns and repairs, is examined in the limit where the number of servers tends to infinity and the repair rate tends to 0, such that their product remains finite. It is shown that the limiting two-dimensional Markov process corresponds to a queue where the number of servers has the same stationary distribution as the number of jobs in an $M/M/\infty$ queue. Hence, the limiting model is referred to as the $M/M/[M/M/\infty]$ queue. Its numerical solution is discussed.

Next, the behaviour of the $M/M/[M/M/\infty]$ queue is analysed in heavy traffic. When the traffic intensity approaches 1, the distribution of the (suitably normalized) number of jobs in the system is approximately exponential. This result relies on two limiting processes—a diffusion and a normalized heavy traffic limit—being essentially the same.

Key-words: Queues with breakdowns. Scaled processes. Invariant measures. Diffusion approximations

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex (France)
Téléphone : +33 1 39 63 55 11 —Télécopie : +33 1 39 63 53 30

Étude asymptotique de systèmes de serveurs avec panne

Résumé : Le comportement asymptotique de la file $M/M/n$ avec des serveurs sujets à des pannes est étudié quand n tend vers l'infini et quand le taux de réparation η tend vers 0 de telle sorte que $n\eta$ soit constant. On montre que le système converge vers une file d'attente où le nombre de serveurs est le nombre de clients d'une file $M/M/\infty$. Ce modèle limite est noté $M/M/[M/M/\infty]$. Les aspects numériques de cette file d'attente sont discutés. Le comportement à la saturation de cette file d'attente est ensuite étudié quand l'intensité de trafic converge vers 1. On montre que le processus convenablement renormalisé converge vers une diffusion réfléchie.

Mots-clés : Serveurs avec panne. Mesures invariantes. Approximations aux diffusions.

1. Introduction

The multi-server queue subject to random breakdowns and repairs is a well known model, with applications in the fields of computing, communications and manufacturing. Its equilibrium distribution has been studied quite extensively (e.g., see [10, 14, 11] and references therein). In principle, the exact solution can be obtained numerically for any parameter setting. However, the computational complexity of all proposed solutions increases quite quickly with the number of servers (in some cases it may also depend on the offered load). Consequently, very large systems tend to be numerically intractable.

It is thus of interest, both from a theoretical point of view and for purposes of approximation, to examine the behaviour of this queue under various extreme parameter settings. Two different kinds of asymptotic regimes were analysed by Mitrani and Puhalskii in [12]: (a) the number of servers is fixed, the breakdown and repair rates are bounded away from 0 and the traffic intensity approaches 1; (b) the number of servers and the traffic intensity are fixed, while the breakdown and repair rates approach 0 in a fixed ratio. In case (a), the limiting normalized number of jobs in the system is distributed exponentially, while in case (b) it has a distribution with a rational Laplace transform with simple poles.

Here we are interested, first, in the behaviour of the $M/M/n$ queue when the number of servers, n, tends to infinity. The arrival, service and breakdown rates are kept fixed, while the repair rate, η, tends to 0 so that the product $n\eta$ approaches a finite limit, γ. In other words, the objective is to approximate very large systems where broken servers take relatively long time to repair. The main result of section 3 is to establish that the limiting two-dimensional Markov process corresponds to an unbounded FIFO queue served by servers which arrive into the system in a Poisson stream with rate γ, remain for an exponentially distributed period and then depart. In the steady-state, the number of servers present has the same distribution as the number of jobs in an $M/M/\infty$ queue. Hence, the limiting model is referred to as the $M/M/[M/M/\infty]$ queue.

The exact analytical solution of the $M/M/[M/M/\infty]$ queue is still an open problem. However, we show how one can compute an efficient numerical approximation without too much difficulty.

The second main result of this paper concerns the heavy traffic asymptotics of the $M/M/[M/M/\infty]$ queue. The random variable of interest is $J(1-\rho)$, where J is the steady-state number of jobs in the system and ρ is the traffic intensity, when the latter approaches 1. In fact, rather than consider the heavy traffic limit directly, we examine a diffusion limit which is intuitively equivalent. The normalized queue size is shown to be asymptotically exponentialy distributed, with mean depending on all system parameters. These developments are described in section 4.

2. The basic model

Our point of departure is an $M/M/n$ queue with independent random breakdowns and repairs. Jobs arrive in a Poisson stream at rate λ and join a single, unbounded queue. The required service times are i.i.d. random variables distributed exponentially with parameter μ. There are n identical parallel servers, each of which goes through alternating periods

of being operative and inoperative, independently of the others. The operative periods are i.i.d. random variables distributed exponentially with parameter ξ; similarly, the inoperative (repair) periods are i.i.d. random variables distributed exponentially with parameter η. Jobs are taken for service from the front of the queue, one at a time, by available operative servers. No operative server can be idle if there are jobs waiting to be served. If a service is interrupted by a breakdown, then the relevant job is returned to the front of the queue. When an operative server becomes again available for it, the service is resumed from the point of interruption; there are no switching overheads. This model is illustrated in figure 1.

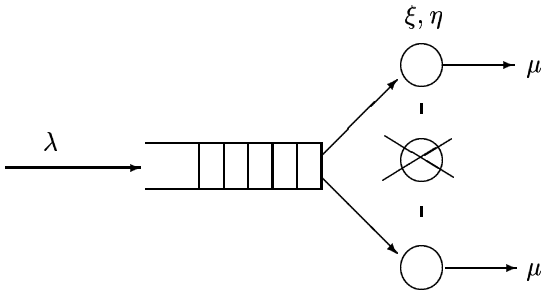

FIGURE 1. The $M/M/n$ queue with breakdowns and repairs

The evolution of the system state is represented by the irreducible Markov process $X_n = \{[I_n(t), J_n(t)] \ ; \ t \geq 0\}$, where $I_n(t)$ is the number of operative servers, out the n that are available, and $J_n(t)$ is the number of jobs (in the queue and/or in service), at time t. Denote the stationary distribution of X_n by

$$p_n(i,j) = \lim_{t \to \infty} \mathbb{P}(I_n(t) = i, J_n(t) = j) \ ; \ i = 0, 1, \ldots, n \ ; \ j = 0, 1, \ldots \ . \tag{1}$$

The probabilities $p_n(i,j)$ satisfy the following balance equations

$$\begin{aligned} \lambda + \mu_{i,j} + i\xi + (n-i)\eta]p_n(i,j) &= \lambda p_n(i,j-1) + \mu_{i,j+1} p_n(i,j+1) \\ &\quad + (i+1)\xi p_n(i+1,j) + (n-i+1)\eta p_n(i-1,j) \ , \end{aligned} \tag{2}$$

where $\mu_{i,j}$ is the state-dependent instantaneous service rate, $\mu_{i,j} = \min(i,j)\mu = (i \wedge j)\mu$, and $p_n(-1,j) = p_n(n+1,j) = p_n(i,-1) = 0$ by definition.

It is important to note that, since server breakdowns and repairs occur independently of the arrivals and services of jobs, the marginal distribution of the number of operative servers is binomial:

$$p_n(i,\cdot) = \binom{n}{i} \left(\frac{\eta}{\xi+\eta}\right)^i \left(\frac{\xi}{\xi+\eta}\right)^{n-i} \ ; \ i = 0, 1, \ldots, n \ . \tag{3}$$

Hence, the processing capacity of the system, which is defined as the average number of operative servers, is equal to

$$\mathbb{E}(I_n) = \frac{n\eta}{\xi + \eta} \ . \tag{4}$$

The process X_n is ergodic if, and only if, the offered load is less than the processing capacity:

$$\frac{\lambda}{\mu} < \frac{n\eta}{\xi + \eta} \ . \tag{5}$$

It is assumed that this condition is satisfied. In terms of the traffic intensity, ρ_n, the assumption is that

$$\rho_n = \frac{\lambda(\xi + \eta)}{n\eta\mu} < 1 \ . \tag{6}$$

The fact that the state space of the process X_n depends on n is a source of some inconvenience. To get around that, we shall formally extend the state space to all non-negative integer pairs $\{(i,j) \,|\, i \geq 0, j \geq 0\}$. The instantaneous transition rates out of state (i,j) are as before $(\lambda, \mu_{i,j}, i\xi)$, except that the transition rate to state $(i+1,j)$, which is equal to $(n-i)\eta$ when $i < n$, is defined as 0 for $i \geq n$. Thus, all states with more than n operative servers, $\{(i,j) \,|\, i > n\}$, are unreachable and their steady-state probabilities are 0. The extended process is no longer irreducible, but the solution of the balance and normalising equations is unaffected.

3. Large scale limit

Consider an infinite sequence, $\{X_n \ ; \ n = 1, 2, \dots\}$ of the above processes. The parameters λ, μ and ξ are kept fixed, but the average repair time is allowed to grow with the number of servers, in a ratio that approaches a finite limit:

$$\lim_{n \to \infty} n\eta = \gamma \ ,$$

for some $\gamma > 0$.

Every state, (i,j), is reachable by all X_n such that $n \geq i$. In the limit $n \to \infty$, the transitions out of state (i,j) are

to state $(i, j+1)$ with rate λ;
to state $(i+1, j)$ with rate γ (since $(n-i)\eta \to \gamma$);
to state $(i, j-1)$ with rate $(i \wedge j)\mu$;
to state $(i-1, j)$ with rate $i\xi$.

We observe that these are precisely the transition rates corresponding to a queueing system where jobs and servers arrive in independent Poisson processes, with rates λ and γ respectively; each server remains in the system for an independent, exponentially distributed period (with parameter ξ), during which it may serve jobs at rate μ, and then departs. Services that are interrupted due to a server departure are eventually resumed on another server. We refer to this system as the $M/M/[M/M/\infty]$ queue (see figure 2), because the number of servers present behaves like the number of jobs in an $M/M/\infty$ 'queue' with parameters γ and ξ.

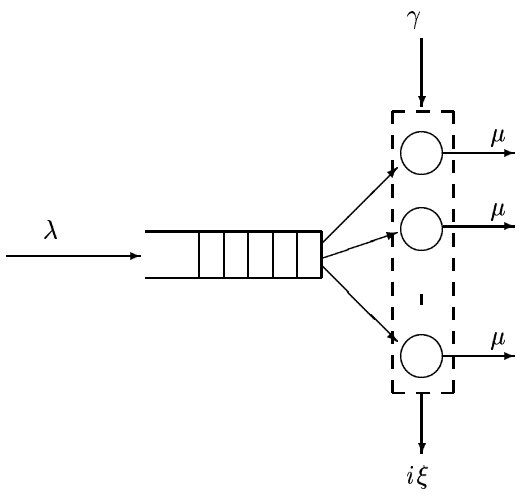

FIGURE 2. The $M/M/[M/M/\infty]$ queue

Let $X = \{[I(t), J(t)] \; ; \; t \geq 0\}$ (where $I(t)$ is the number of servers and $J(t)$ is the number of jobs at time t), be the Markov process modelling the $M/M/[M/M/\infty]$ queue. Denote its stationary distribution by

$$p(i,j) = \lim_{t \to \infty} \mathbb{P}(I(t) = i, J(t) = j) \; ; \; i = 0, 1, \ldots \; ; \; j = 0, 1, \ldots \; . \tag{7}$$

The marginal stationary distribution of the number of servers always exists and is given by

$$p(i,\cdot) = \frac{1}{i!} (\gamma/\xi)^i \, e^{-\gamma/\xi} \; ; \; i = 0, 1, \ldots \; . \tag{8}$$

The average number of available servers is $\mathbb{E}(I) = \gamma/\xi$. Hence, the condition for ergodicity of X is clearly

$$\frac{\lambda}{\mu} < \frac{\gamma}{\xi} \, , \tag{9}$$

or, in terms of the traffic intensity,

$$\rho = \frac{\lambda \xi}{\gamma \mu} < 1 \, . \tag{10}$$

So far, we have seen that the generator matrices of X_n converge, element by element, to the generator matrix of X. Now our aim is to demonstrate a much stronger result, namely the existence and similar convergence of stationary distributions:

Theorem 1. *If the condition (9) holds, then X, and X_n for all sufficiently large n, are ergodic, and*

$$\lim_{n\to\infty} p_n(i,j) = p(i,j) \; ; \; i = 0,1,\dots \; ; \; j = 0,1,\dots \; .$$

The proof relies on a rather general theorem by Malyshev and Menshikov, [5], which we restate here in a form suitable to the present context. Let $Y_n = \{Y_n(k) \; ; \; k = 0,1,\dots\}$, $n = 1,2,\dots$, be a sequence of Markov chains, and $Y = \{Y(k) \; ; \; k = 0,1,\dots\}$ be a Markov chain, all on the same state space, S. Then

Theorem 2. *If there exist numbers $\beta > 0$ and $\alpha \geq 2$, a positive real-valued function $f(\cdot)$ defined over S, and a finite subset, F of S, such that for all sufficiently large n,*

(i) $\mathbb{E}(f(Y_n(k+1)) - f(x) \,|\, Y_n(k) = x) \leq -\beta$, *for* $x \in S - F$,
(ii) $\mathbb{E}(f(Y(k+1)) - f(x) \,|\, Y(k) = x) \leq -\beta$, *for* $x \in S - F$,

then those Y_n and Y are ergodic; denote their stationary distributions by $\pi_n(x)$ and $\pi(x)$ respectively. Moreover, if

(iii) $\sup_{n,x} \mathbb{E}(|f(Y_n(k+1)) - f(x)|^\alpha \,|\, Y_n(k) = x) < \infty$,
(iv) *the transition matrices of Y_n converge, element by element, to that of Y,*

then

$$\lim_{n\to\infty} \pi_n(x) = \pi(x) \;\; ; \;\; x \in S \; .$$

The function $f(\cdot)$ is sometimes called a Lyapunov function. Conditions (i), (ii) and (iii) refer to conditional negative drift, and integrability of transitions, of Y_n and Y with respect to the Lyapunov function.

Proof. For our purposes, Y_n and Y are the Markov chains embedded at the jump epochs of processes X_n and X respectively. The state space, S, is the set of all non-negative integer pairs. Finding a Lyapunov function that satisfies (i), (ii) and (iii) (condition (iv) is already satisfied), will establish Theorem 1 as a corollary of Theorem 2.

We start by defining a sequence of numbers, a_l, according to the recurrences

$$a_0 = 0 \;\; ; \;\; a_{l+1} = \frac{1}{\gamma}(l\xi a_l - \lambda + l\mu - \vartheta) \;\; ; \;\; l = 1,2,\dots \; , \tag{11}$$

where ϑ is a positive constant chosen so that

$$\frac{\lambda + \vartheta}{\gamma} < \frac{\mu}{\xi} \; .$$

Such a choice is possible whenever the inequality (9) holds. Then (11) implies $\xi a_1 > -\mu$. Hence, and from

$$\gamma(a_{l+1} - a_1) = l(\mu + \xi a_l) \; ,$$

it is easy to deduce that

$$a_{l+1} > a_1 + \frac{l}{\gamma}(\mu + \xi a_1) \;\; ; \;\; l \geq 1 \; .$$

Thus, (9) allows the construction of $a_1, a_2, \ldots$, such that (11) is satisfied, and $a_l \to \infty$ when $l \to \infty$.

Now, let the finite subset F mentioned in Theorem 2 be of the form $F = \{(i,j) \,|\, i \leq N, j \leq N\}$, where N is is a suitably large integer. For any given N, define the following Lyapunov function:

$$f(i,j) = \begin{cases} \sum_{l=0}^{N} a_l + i - N + j + c & \text{if } i > N, \\ \sum_{l=0}^{i} a_l + j + c & \text{if } i \leq N, \end{cases} \tag{12}$$

where c is a constant chosen so that $f(i,j)$ is positive. The conditional drift of Y with respect to f is given by:

$$\begin{aligned} d(i,j) = \mathbb{E}(f(Y(k+1)) - f(i,j) \,|\, Y(k) = (i,j)) = \frac{\lambda}{r_{i,j}} f(i,j+1) \\ + \frac{\gamma}{r_{i,j}} f(i+1,j) + \frac{(i \wedge j)\mu}{r_{i,j}} f(i,j-1) + \frac{i\xi}{r_{i,j}} f(i-1,j) - f(i,j) \,, \end{aligned} \tag{13}$$

where $r_{i,j} = \lambda + \gamma + (i \wedge j)\mu + i\xi$ is the total transition rate out of state (i,j). We shall show that, when N is sufficiently large and (i,j) is outside the corresponding set F, the right-hand side of (13) is bounded above by a negative constant.

There are three cases to be considered (see figure 3).

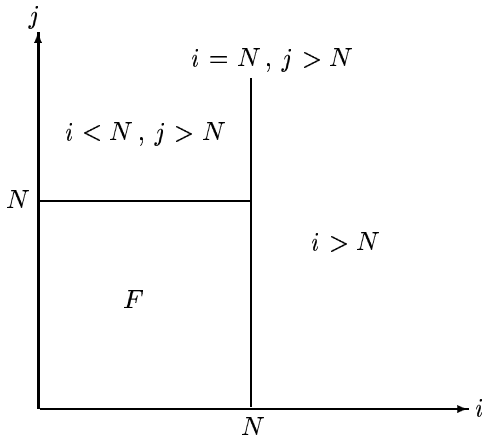

FIGURE 3. The sets of states F and $S - F$

Case 1, $i > N$. Substituting (12) into (13), we note that the sums of a_l cancel out:

$$d(i,j) = \frac{\lambda + \gamma - (i \wedge j)\mu - i\xi}{\lambda + \gamma + (i \wedge j)\mu + i\xi} < \frac{\lambda + \gamma - i\xi}{\lambda + \gamma + i\mu + i\xi} \,. \tag{14}$$

When $N \to \infty$, the right-hand side of (14) approaches $-\xi/(\mu+\xi)$. Hence, we can clearly find constants N_1 and $\beta_1 < \xi/(\mu+\xi)$, such that if N_1 is used to define F, $d(i,j) < -\beta_1$ for all $i > N_1$.

Case 2, $i = N < j$. The expression for $d(N,j)$ is

$$d(N,j) = \frac{\lambda+\gamma-N\mu-N\xi a_N}{\lambda+\gamma+N\mu+N\xi} \; . \tag{15}$$

When $N \to \infty$, this approaches $-\infty$, since $a_N \to \infty$. Again, we can choose N_2 and β_2 such that $d(N_2,j) < -\beta_2$ for all $j > N_2$.

Case 3, $i < N < j$. The recurrences for a_l yield:

$$d(i,j) = \frac{\lambda+\gamma a_{i+1}-i\mu-i\xi a_i}{\lambda+\gamma+i\mu+i\xi} < \frac{-\vartheta}{\lambda+\gamma+N\mu+N\xi} \; . \tag{16}$$

Any β_3 smaller, in absolute value, than the right-hand side of (16), and any positive N_3, ensure that $d(i,j) < -\beta_3$ for all $i < N_3 < j$.

Thus, condition (ii) of Theorem 2 is satisfied by choosing

$$N = \max(N_1, N_2, N_3) \;\; ; \;\; \beta = \min(\beta_1, \beta_2, \beta_3) \; .$$

Almost exactly the same arguments show that condition (i) is satisfied, with the same values for N and β, for all sufficiently large n. Where terms of the form $(n-i)\eta$ appear, neglecting $i\eta$ enhances the inequalities, while $n\eta$ uniformly approaches γ.

It remains to verify condition (iii) for the value of N, and Lyapunov function, determined above. Take any $\alpha \geq 2$ and substitute (12) into the expressions for

$$D_n(i,j) = \mathbb{E}(|f(Y_n(k+1)) - f(i,j)|^\alpha \,|\, Y_n(k) = (i,j)) \; .$$

This yields, when $i > N$,

$$D_n(i,j) = 1 \; .$$

When $i = N$,

$$D_n(i,j) \leq \max(1, |a_N|^\alpha) \; .$$

Finally, when $i < N$,

$$D_n(i,j) \leq \max(1, |a_i|^\alpha, |a_{i+1}|^\alpha) \; .$$

Denoting $A = \max(1, |a_1|^\alpha, \ldots, |a_N|^\alpha)$, we see that $D_n(i,j) \leq A$ for all (i,j) and all n (including the chain Y).

This completes the verification of the conditions of Theorem 2, and hence the proof of Theorem 1. □

3.1. **Solution of the M/M/[M/M/∞] queue.** The stationary distribution of the Markov process associated to the $M/M/[M/M/\infty]$, X, satisfies the following balance equations:

$$\begin{aligned}[\lambda+\gamma+(i\wedge j)\mu+i\xi]\, p(i,j) \\ = \lambda\, p(i,j-1) + \gamma\, p(i-1,j) + (i\wedge j+1)\mu\, p(i,j+1) + (i+1)\xi\, p(i+1,j),\end{aligned} \tag{17}$$

for i, $j = 0, 1, \ldots$, where $p(-1, j) = p(i, -1) = 0$ by definition. These equations, together with the normalizing equation, determine, in principle, the probabilities $p(i, j)$.

Unfortunately, there is no analytical solution of (17). Methods that have been used successfully with other two-dimensional processes (eg, reduction to a boundary value problem) do not appear to work here. A numerical approximation can of course be obtained by truncating the state space to, say, $i \leq N$, $j \leq M$, for some integers N and M, and then solving the resulting finite set of equations. However, that approach can be very expensive, of uncertain accuracy, or both.

An efficient and accurate numerical solution can be obtained by exploiting the fact that the marginal stationary distribution of the number of servers is given by (8). Fix a tolerance level, ϵ, and find an integer, N_ϵ, such that

$$\sum_{i=N_\epsilon+1}^{\infty} \frac{1}{i!} (\gamma/\xi)^i \, e^{-\gamma/\xi} < \epsilon \; .$$

Truncate the state space by neglecting all states (i, j) for which $i > N_\epsilon$. The sum of the stationary probabilities of the discarded states is then known to be less than ϵ. The remaining state space is still infinite, but only in the j-dimension. Moreover, when $j > N_\epsilon$, the instantaneous transition rates do not depend on j. The resulting Markov process can be solved either by the matrix-geometric method [13] or by spectral expansion [11].

If this approach is used to compute a performance measure such as the average number of jobs in the system, then the result is not just an approximation but also an upper bound. This is because the truncation under-estimates the average number of servers in the system.

Figure 4 shows the effect of ϵ on the value of N_ϵ and on the computed performance measure $\mathbb{E}(J)$. The solution was obtained by spectral expansion.

The parameters used in this example were $\lambda = 3$, $\gamma = 0.3$, $\mu = 1$, $\xi = 0.05$. This is not a heavily loaded system; the average number of servers present is 6, and the traffic intensity is $\rho = 0.5$. The figure shows that a four-digit accuracy in $\mathbb{E}(J)$ is obtained for $\epsilon = 10^{-4}$. The corresponding level of truncation is $N = 18$; neither that, nor the truncation at $N = 21$, presents any computational problems.

It is worth pointing out that, while ϵ decreases exponentially, the truncation level, N, increases roughly linearly. This justifies the claim that the proposed solution is efficient (more examples of this type can be found in [4]). However, one can anticipate a deterioration in the accuracy of the truncated model, and hence intractably high values of N, when the offered load approaches the average number of servers. Then a small underestimate in $\mathbb{E}(I)$ can have a big effect on $\mathbb{E}(J)$.

The following section provides an analytic approximation which applies in the cases where the numerical approach breaks down.

4. The $M/M/[M/M/\infty]$ queue in heavy traffic

Suppose that the parameters of the $M/M/[M/M/\infty]$ queue are changed in such a way that the average number of servers remains constant (one could also handle the case where that number varies but remains bounded), while the offered load increases, causing the traffic

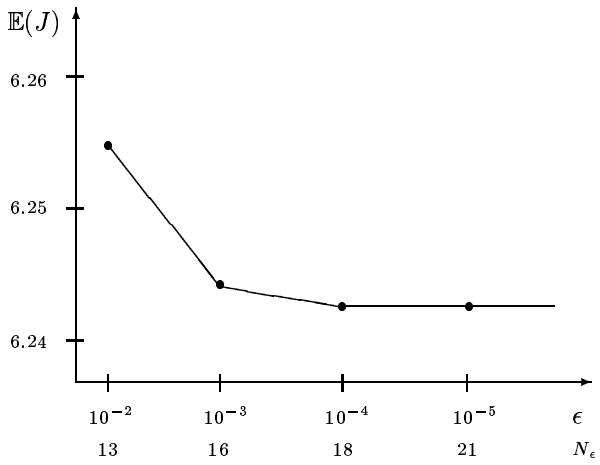

FIGURE 4. $\mathbb{E}(J)$ and N_ϵ for different values of ϵ

intensity to approach 1. To be precise, consider a sequence of parameter sets $\{\lambda_n, \mu_n, \gamma, \xi\}$, $n = 1, 2, \ldots$, such that

$$\lambda_n \to \lambda \ ; \ \mu_n \to \mu \ ; \ \rho_n = \frac{\lambda_n \xi}{\mu_n \gamma} \nearrow \frac{\lambda \xi}{\mu \gamma} = 1 \ , \tag{18}$$

for some positive constants λ and μ. Moreover, assume that $(1 - \rho_n)^{-1}$ approaches ∞ in the manner of $\sqrt{n}$, i.e.

$$(1 - \rho_n)\sqrt{n} \to 1 \ . \tag{19}$$

Denote the corresponding sequence of Markov processes by

$$X_n = ([I(t), J_n(t)] \ ; \ t \geq 0) \ \ ; \ \ n = 1, 2, \ldots \ ,$$

where $I(t)$ is the number of servers and $J_n(t)$ is the number of jobs in the system at time t.

Note that all the X_n's correspond to $M/M/[M/M/\infty]$ queues; they should not be confused with the sequence of $M/M/n$ processes introduced in the previous section.

Let $Q_n(t)$ be the following normalized queue length process:

$$Q_n(t) = \frac{J_n(nt)}{\sqrt{n}} \approx J_n(nt)(1 - \rho_n) \ . \tag{20}$$

The transformation (20) is known as 'diffusion scaling'.

We shall denote by $\xrightarrow{d}$ the convergence of distributions of processes in $D([0, +\infty[)$, the space of functions on $[0, +\infty[$ which are continuous on the right and have left limits. The space $D([0, +\infty[)$ is endowed with the Skorokhod topology (see Billingsley [2]).

The main result of this section is the following.

Theorem 3.

$$(Q_n(t)) \stackrel{d}{\to} \left(R\left[\sqrt{2\lambda(1+\frac{\mu}{\xi})}\, B(t) - \lambda t \right] \right) , \tag{21}$$

where $B(t)$ is the standard Brownian motion; the process in the square brackets is a diffusion with negative drift. $R[X(t)]$ is the Skorokhod reflection of $(X(t))$:

$$R[X(t)] = \sup_{u \le t} [X(t) - X(u)] .$$

(For a nice illustration of $(X(t))$ and $(R[X(t)])$, albeit in a slightly different context, see Feller [6], page 192).

Before proceeding with the proof, consider the implications of this theorem. First, let us assume that the stationary distribution of the limiting process in (21) is the same as the limiting stationary distribution of the normalized queue size:

$$\lim_{t\to\infty} \lim_{n\to\infty} (1-\rho_n) J_n(nt) = \lim_{n\to\infty} (1-\rho_n) \lim_{t\to\infty} J_n(nt) . \tag{22}$$

(The interchangeability of these two limits is normally taken for granted. There are very few instances where it has been proved, e.g. the $G/G/1$ queue [1, 9].)

Now, it is known that the stationary distribution of a Skorokhod reflection of a diffusion process with negative drift is given by

$$\lim_{t\to\infty} \mathbb{P}(R[aB(t) - bt] > x) = \mathbb{P}(\sup_{t\ge 0} [aB(t) - bt] > x) . \tag{23}$$

Also, it is known (see [1] or [15]) that

$$\mathbb{P}(\sup_{t\ge 0} [B(t) - bt] > x) = e^{-2bx} . \tag{24}$$

These last two results, together with the fact that $B(t)$ has the same distribution as $B(a^2t)/a$, imply that

$$\lim_{t\to\infty} \mathbb{P}(R[aB(t) - bt] > x) = e^{-2bx/a^2} . \tag{25}$$

The process in (21) has $b = \lambda$ and $a^2 = 2\lambda(1+\mu/\xi)$. Hence, Theorem 3 and the equivalence (22) imply that the limiting normalized stationary queue size is distributed exponentially:

$$\lim_{n\to\infty} \mathbb{P}((1-\rho_n) J_n > x) = e^{-\xi x/(\xi+\mu)} = e^{-\gamma x/(\gamma+\lambda)} \tag{26}$$

(the last equality follows from $\lambda\xi = \mu\gamma$).

Thus, the heavy traffic approximation for the average number of jobs in a system with traffic intensity ρ_n is

$$\mathbb{E}(J_n) \approx \frac{\gamma+\lambda}{\gamma(1-\rho_n)} . \tag{27}$$

Proof of Theorem 3. Let $N^x(t)$ be an independent Poisson process with parameter x. Different instances of such processes will be distinguished by indexing.

Assume, without affecting the long-term results, that $I(0) = J_n(0) = 0$. Since $I(t)$ is equal to the number of servers that arrive during the interval $(0,t)$, minus the number of servers that depart during that interval, we can write

$$I(t) = N^\gamma(t) - \sum_{k=1}^{\infty} \int_0^t 1_{\{I(u-)=k\}} \sum_{i=1}^{k} dN_i^\xi(u) \,, \tag{28}$$

where $1_{\{\cdot\}}$ is the indicator function and $u-$ is the instant 'just before' u.

Any Poisson process can be expressed in the form

$$N^x(t) = xt + M(t) \,, \tag{29}$$

where $M(t)$ is a locally square integrable martingale. Applying (29) to (28) and remembering that the sum of martingales is a martingale, we get

$$I(t) = \gamma t - \xi \sum_{k=1}^{\infty} k \int_0^t 1_{\{I(u)=k\}} \, du + M_I(t) \,,$$

or, exchanging the order of summation and integration,

$$I(t) = \gamma t - \xi \int_0^t I(u) \, du + M_I(t) \,, \tag{30}$$

where $M_I(t)$ is a locally square integrable martingale.

A similar argument shows that

$$J_n(t) = \lambda_n t - \mu_n \int_0^t [J_n(u) \wedge I(u)] \, du + M_{J,n}(t) \,, \tag{31}$$

where $M_{J,n}(t)$ is a locally square integrable martingale.

Rather than attack directly the scaled queueing process, $Q_n(t)$, defined in (20), consider the scaled difference between the number of jobs and the number of servers:

$$Y_n(t) = \frac{J_n(nt) - I(nt)}{\sqrt{n}} \,. \tag{32}$$

First, using the fact that $J_n(u) \wedge I(u) = I(u) - [I(u) - J_n(u)]^+$, where $x^+ = x \vee 0$, rewrite (31) as

$$J_n(t) = \lambda_n t - \mu_n \int_0^t I(u) \, du + \mu_n \int_0^t [I(u) - J_n(u)]^+ \, du + M_{J,n}(t) \,. \tag{33}$$

Eliminating the first integral in the right-hand side of (33) with the aid of (30), setting $t = nt$ and dividing by $\sqrt{n}$, and bearing in mind that

$$J_n(t) - I(t) = [J_n(t) - I(t)]^+ - [I(t) - J_n(t)]^+ \,,$$

we obtain the following equation:

$$[Y_n(t)]^+ = Z_n(t) + V_n(t) \,, \tag{34}$$

where

$$Z_n(t) = -\frac{\lambda_n}{\rho_n}\sqrt{n}(1-\rho_n)t - \frac{\mu_n}{\xi}\frac{M_I(nt)}{\sqrt{n}} + \frac{M_{J,n}(nt)}{\sqrt{n}} + (\frac{\mu_n}{\xi} - 1)\frac{I(nt)}{\sqrt{n}} + \frac{[I(nt) - J_n(nt)]^+}{\sqrt{n}} , \tag{35}$$

and

$$V_n(t) = \frac{\mu_n}{\sqrt{n}} \int_0^{nt} [I(u) - J_n(u)]^+ \, du . \tag{36}$$

From the above definitions it is immediately clear that, for every n, the pair of processes $([Y_n(t)]^+)$ and $(V_n(t))$ have the following properties:

(a) $[Y_n(t)]^+ \geq 0$;
(b) $V_n(t)$ is non-decreasing and $V_n(0) = 0$;
(c) if $V_n'(t) > 0$ then $[Y_n(t)]^+ = 0$.

That pair is therefore the unique solution of the Skorokhod reflection problem for the process $(Z_n(t))$. Consequently, for $t \geq 0$,

$$[Y_n(t)]^+ = \sup_{u \leq t}[Z_n(t) - Z_n(u)] = R[Z_n(t)] . \tag{37}$$

Moreover, if $(Z_n(t))$ converges in distribution to some process $(Z(t))$ when $n \to \infty$, then $([Y_n(t)]^+)$ converges in distribution to the Skorokhod reflection $(R[Z(t)])$ (see, for example [8, 15, 16] and references therein).

The next step is to show that the sequence (35) converges in distribution to the diffusion process in square brackets in (21). The first term in the right-hand side of (35) approaches $-\lambda t$, due to (19).

The convergence of the two martingales in (35). To prove that they converge to Brownian motions, it is sufficient (see for example Theorem 1.4, page 339 of Ethier et Kurtz [3]) to show that their quadratic characteristics converge in probability to the quadratic characteristics of Brownian motions. If $\langle X \rangle$ denotes the increasing process of the process X (see Rogers and Williams [15]), we have

$$\left\langle \frac{M_I(nt)}{\sqrt{n}} \right\rangle = \frac{1}{n}\langle M_I(nt)\rangle = \gamma t + \xi\frac{1}{n}\int_0^{nt} I(u)\, du \tag{38}$$

(the two terms in the right-hand side are the continuously increasing parts of the server arrival and server departure processes, respectively; they have to be added because the jumps of those two processes are disjoint). For $t \geq 0$ the variables

$$\frac{1}{n}\int_0^{nt} I(u)\, du = t\frac{1}{nt}\int_0^{nt} I(u)\, du$$

converge in probability to $t\gamma/\xi$ by the ergodic theorem applied to $I(t)$. Hence,

$$\left\langle \frac{M_I(nt)}{\sqrt{n}} \right\rangle \to 2\gamma t \tag{39}$$

for the convergence in probability.

Similarly,

$$\left\langle \frac{M_{J,n}(nt)}{\sqrt{n}} \right\rangle = \lambda_n t + \mu_n \frac{1}{n} \int_0^{nt} [J_n(u) \wedge I(u)]\, du$$

$$= \lambda_n t + \mu_n \frac{1}{n} \int_0^{nt} I(u)\, du - \mu_n \frac{1}{n} \int_0^{nt} [I(u) - J_n(u)]^+ \, du \; . \tag{40}$$

By the same argument as above, and remembering that $\mu\gamma/\xi = \lambda$, the first two terms in the right-hand side of (40) approach $2\lambda t$. The third term vanishes, as shown by the first assertion of the following:

Proposition 1. *The variable*

$$\frac{1}{n} \int_0^{nt} [I(u) - J_n(u)]^+ \, du$$

converges in probability to 0 *and*

$$\left(\frac{I(nt)}{\sqrt{n}} \right) \xrightarrow{d} 0 \; .$$

The proof of this proposition is in the Appendix.

Since the jumps of $(M_I(nt)/\sqrt{n})$ and $(M_{J,n}(nt)/\sqrt{n})$ occur according to independent Poisson processes, the quadratic covariation of these martingales is zero,

$$\left\langle \frac{M_I(nt)}{\sqrt{n}}, \frac{M_{J,n}(nt)}{\sqrt{n}} \right\rangle = 0.$$

The above developments imply the convergence

$$\left(-\frac{\mu_n}{\xi} \frac{M_I(nt)}{\sqrt{n}} + \frac{M_{J,n}(nt)}{\sqrt{n}} \right) \xrightarrow{d} \left(\frac{\mu}{\xi} \sqrt{2\gamma} B(t) + \sqrt{2\lambda} B_1(t) \right) , \tag{41}$$

where $(B(t))$, $(B_1(t))$ are standard independent Brownian motions. The term on the right hand side of (41) has the same distribution as

$$\left(\sqrt{2\gamma \frac{\mu^2}{\xi^2} + 2\lambda}\, B(t) \right) = \left(\sqrt{2\lambda(1 + \frac{\mu}{\xi})}\, B(t) \right) . \tag{42}$$

The last two terms of (35). The second part of proposition 1 shows that they converge in distribution to 0 since

$$\frac{[I(nt) - J_n(nt)]^+}{\sqrt{n}} \leq \frac{I(nt)}{\sqrt{n}} \; .$$

Therefore,

$$(Z_n(t)) \xrightarrow{d} \left(-\lambda t + \sqrt{2\lambda(1 + \frac{\mu}{\xi})}\, B(t) \right) \; . \tag{43}$$

Thus, $([Y_n(t)]^+)$ converges in distribution to the Skorokhod reflection in the right-hand side of (21).

Finally, note that

$$\frac{J_n(nt)}{\sqrt{n}} = [Y_n(t)]^+ + \frac{I(nt)}{\sqrt{n}} - \frac{[I(nt) - J_n(nt)]^+}{\sqrt{n}} \ .$$

Since the last two terms converge in distribution to 0, $(Q_n(t))$ and $([Y_n(t)]^+)$ converge in distribution to the same process. This completes the proof of Theorem 3. □

4.1. **Empirical evaluation of the approximation.** We have compared the approximation provided by (27), with estimates of $\mathbb{E}(J)$ obtained from simulations. Each estimate is obtained from a simulation run generating approximately one million job arrivals; this is divided into 21 portions of roughly equal size (of which the first is discarded), for the purpose of computing a 90% confidence interval.

Two sets of experiments were carried out. In the first, the parameters μ, γ and ξ are fixed as in the example of section 3.1: $\mu = 1$, $\gamma = 0.3$, $\xi = 0.05$; the average number of servers present is $\mathbb{E}(I) = 6$. The job arrival rate, λ, is increased from 5.2 to 5.8, with a corresponding increase in the traffic intensity from $\rho = 0.867$ to $\rho = 0.967$. The saturation point $\rho = 1$ is reached when $\lambda = 6$.

The results are illustrated in figure 5. It appears that the approximations are always pessimistic: they over-estimate the average number of jobs in the system. Moreover, the absolute difference between an observation and the corresponding approximation does not change much; of course, the relative difference decreases with the traffic intensity, as predicted by theorem 3. It should be pointed out that when the traffic intensity approaches 1, the simulation estimates become less and less reliable; the big variability in the observed queue sizes causes the confidence intervals to become very large.

In the second set of experiments, the breakdown rate is reduced by a factor of 10, to $\xi = 0.005$. Consequently, the average number of servers present is 60. The service rate is also reduced by a factor of 10, to $\mu = 0.1$, so that the saturation point is again reached when $\lambda = 6$.

These changes do not affect the heavy traffic approximation, but they do affect the queueing process (see figure 6). The same traffic intensity produces a lower average queue size in the second system than in the first (intuitively, this may be explained by the fact that the variance of the number of servers has increased by a factor of 'only' 10, rather than by a factor of 100). Thus, the absolute differences between the approximations and the simulated queue sizes tend to be larger. However, they still appear to be roughly constant as ρ increases.

These experiments suggest that the accuracy of the heavy traffic approximation (27) can be improved by simulating the system for one (moderately large) value of ρ, and using the observed difference between the approximated and simulated $\mathbb{E}(J)$, in order to calibrate the approximations for larger values of ρ.

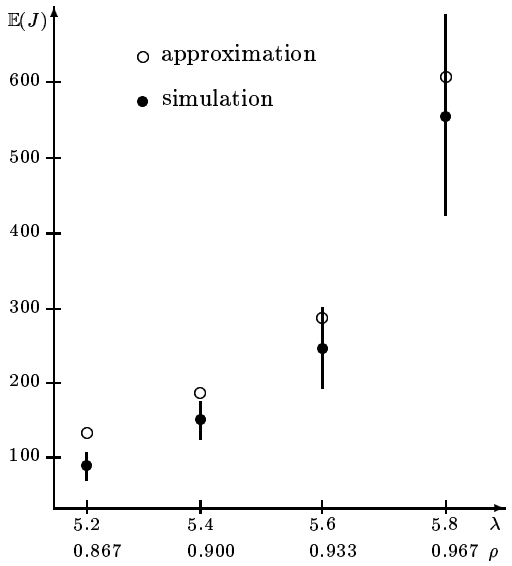

FIGURE 5. Approximation and simulation; set 1 — 90% confidence intervals

APPENDIX

Proof of Proposition 1. To have the first assertion, it is enough to show that the limit is 0 in expectation, i.e. that

$$\mathbb{E}\left(\frac{1}{n}\int_0^{nt}[I(u)-J_n(u)]^+\,du\right) \tag{44}$$

tends to 0 as n gets large.

We first prove that $(J_n(t))$ grows sufficiently as n tends to infinity. With the same notations as bove, the process $(J_n(t))$ satisfies the following stochastic differential equation,

$$dJ_n(t) = dN^{\lambda_n}(t) - \sum_{i=1}^{I(t-)} 1_{\{J_n(t-)\geq i\}}\, dN_i^{\mu_n}(t)\,.$$

One defines $(J_n'(t))$ by $J_n'(0)=0$ and

$$dJ_n'(t) = dN^{\lambda_n}(t) - 1_{\{J_n'(t-)>0\}} \sum_{i=1}^{I(t-)} dN_i^{\mu_n}(t)\,,$$

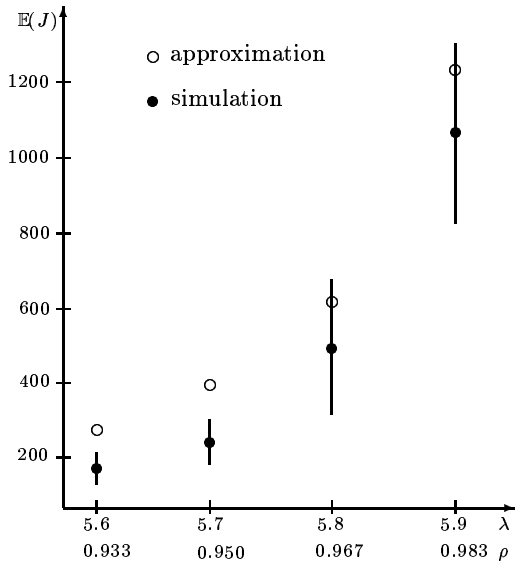

FIGURE 6. Approximation and simulation; set 2 — 90% confidence intervals

it is easily seen that $J'_n(t)$ is non negative and $J'_n(t) \leq J_n(t)$ for all $t \geq 0$. The process $(J'_n(t))$ can be also written as

$$J'_n(t) = Z'_n(t) + \mu_n \int_0^t 1_{\{J'_n(u-)=0\}}\, du \tag{45}$$

with

$$Z'_n(t) = \lambda_n t - \mu_n \int_0^t I(u)\, du + (N^{\lambda_n}(t) - \lambda_n t) - \sum_{i=1}^{+\infty} \left(\int_0^t 1_{\{i \leq I(u-)\}} (dN_i^{\mu_n}(u) - \mu_n\, du) \right) .$$

The identity (45) shows that the process $(J'_n(t))$ is the first component of the Skorokhod reflection problem for $(Z'_n(t))$. In this case, since J'_n does not play a rôle in the expression of Z'_n, it is easy to prove that $(Z'_n(nt)/\sqrt{n})$ converges in distribution to the Brownian motion with drift defined by (42). Consequently $(J'_n(nt)/\sqrt{n})$ converges in distribution to the corresponding reflected process.

We come back to the integral (44), interchange expectation and integration and write

$$\frac{1}{n}\int_0^{nt} \mathbb{E}([I(u)-J_n(u)]^+)\,du \le \frac{1}{n}\int_0^{nt} \mathbb{E}\left(I(u)1_{\{J_n(u)\le I(u)\}}\right)\,du$$
$$= \int_0^t \mathbb{E}\left(I(nu)1_{\{J_n(nu)\le I(nu)\}}\right)\,du\ .$$

The above integrand can be bounded by applying Cauchy-Schwartz's inequality:

$$\mathbb{E}\left(I(nu)1_{\{J_n(nu)\le I(nu)\}}\right) \le \sqrt{\mathbb{E}([I(nu)]^2)}\sqrt{\mathbb{P}(J_n(nu)\le I(nu))}\ . \tag{46}$$

Now, $\mathbb{E}([I(nu)]^2)$ approaches the second moment of the stationary number of jobs in the $M/M/\infty$ queue, which is finite (in fact this quantity can be calculated explicitly). For a fixed $u \ge 0$, using again the convergence in distribution of $I(nu)$, for $\varepsilon > 0$, one can find a $C > 0$ such that for n sufficiently large the inequality $\mathbb{P}(I(nu) > C) \le \varepsilon$ holds. Since

$$\mathbb{P}(J_n(nu)\le I(nu)) \le \mathbb{P}(J_n'(nu)\le I(nu))$$
$$\le \mathbb{P}(I(nu)>C) + \mathbb{P}(J_n'(nu)/\sqrt{n} \le C/\sqrt{n})$$

the convergence in distribution of $(J_n'(nu)/\sqrt{n})$ shows that the last term is arbitrarily small as n gets large. Hence the integrand (46) converges to 0, we conclude by using Lebesgue's theorem.

To prove the second part of Proposition 1, it is sufficient to show that for all $t \ge 0$, the variable $\sup_{0\le u\le t} I(nu)/\sqrt{n}$ converges to 0 in distribution. For $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{0\le u\le t}\frac{I(nu)}{\sqrt{n}} \ge \varepsilon\right) \le \mathbb{P}\left(H_{\varepsilon\sqrt{n}} \le nt\right),$$

where H_x is the hitting time of x by the process $(I(t))$. A classical result, see [7] for example, shows that the distribution function of $(\gamma/\xi)^x H_x/(x-1)!$ converges, as x tends to infinity, to a function continuous at 0. Consequently, the right hand side of the previous inequality converges to 0. This establishes the proposition. □

Acknowledgments. The authors wish to thank Tolya Puhalskii of the University of Denver for a number of helpful comments, in particular the ideas behind the proof of Theorem 3 were contributed by him.

References

[1] Søren Asmussen, *Applied probability and queues*, John Wiley & Sons Ltd., Chichester, 1987.

[2] Patrick Billingsley, *Convergence of probability measures*, Wiley series in probability and mathematical statistics, John Wiley & Sons Ltd, New York, 1968.

[3] Stewart N. Ethier and Thomas G. Kurtz, *Markov processes*, John Wiley & Sons Inc., New York, 1986, Characterization and convergence.

[4] M. Ettl and I. Mitrani, *Applying spectral expansion in evaluating the performance of multiprocessor systems*, Performance Evaluation of Parallel and Distributed Systems, Part 1 (O.J. Boxma and G.M. Koole, eds.), CWI Tract, vol. 105, 1991, pp. 45–58.

[5] G. Fayolle, V. A. Malyshev, and M. V. Men'shikov, *Topics in the constructive theory of countable Markov chains*, Cambridge University Press, Cambridge, 1995.

[6] W. Feller, *An introduction to probability theory and its applications*, 2nd ed., vol. II, John Wiley & Sons Ltd, New York, 1971.
[7] Fabrice Guillemin and Alain Simonian, *Transient characteristics of an $M/M/\infty$ system*, Advances in Applied Probability **27** (1995), no. 3, 862–888.
[8] J.M. Harrison and M.I. Reiman, *Reflected brownian motion on an orthant*, Annals of Probability **9** (1981), no. 2, 302–308.
[9] J.F.C. Kingman, *The heavy traffic approximation in the theory of queues*, Proc. Symp. on Congestion theory (Chapel Hill), Univ. of North Carolina Press, 1965, pp. 137–169.
[10] I. Mitrani and B. Avi-Itzhak, *A many-server queue with service interruptions*, Operations Research **16** (1968), 628–638.
[11] I. Mitrani and R. Chakka, *Spectral expansion solution for a class of markov models: Application and comparison with the matrix-geometric method*, Performance Evaluation **23** (1995), 241–260.
[12] I. Mitrani and A. Puhalskii, *Limiting results for multiprocessor systems with breakdowns and repairs*, Queueing Systems Theory Appl. **14** (1993), no. 3-4, 293–311.
[13] Marcel F. Neuts, *Matrix-geometric solutions in stochastic models*, Dover Publications Inc., New York, 1994, An algorithmic approach, Corrected reprint of the 1981 original.
[14] Marcel F. Neuts and David M. Lucantoni, *A Markovian queue with N servers subject to breakdowns and repairs*, Management Sci. **25** (1979), no. 9, 849–861 (1980).
[15] L. C. G. Rogers and David Williams, *Diffusions, Markov processes, and martingales. Vol. 2: Itô calculus*, John Wiley & Sons Inc., New York, 1987.
[16] R. J. Williams, *Some recent developments for queueing networks*, Probability towards 2000 (New York, 1995), Springer, New York, 1998, pp. 340–356.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399