A stochastic analysis of a network with two levels of service

Vianney Bœuf1 · Philippe Robert2

Received: 22 December 2018 / Revised: 6 May 2019 / Published online: 25 May 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract
In this paper, a stochastic model of a call center with a two-level architecture is analyzed. A first-level pool of operators answers calls, identifies, and handles non-urgent calls. A call classified as urgent has to be transferred to specialized operators at the second level. When the operators of the second level are all busy, the operator of first-level handling the urgent call is blocked until an operator at the second level is available. Under a scaling assumption, the evolution of the number of urgent calls blocked at level 1 is investigated. It is shown that if the ratio of the number of operators at level 2 and 1 is greater than some threshold, then, essentially, the system operates without congestion, with probability close to 1 no urgent call is blocked after some finite time. Otherwise, we prove that a positive fraction of the operators of the first level is blocked due to the congestion of the second level. Stochastic calculus with Poisson processes, coupling arguments and formulations in terms of Skorokhod problems are the main mathematical tools to establish these convergence results.

Keywords Network with blocking · Stochastic models · Call centers

Mathematics Subject Classification 60K35 · 60K25

1 Introduction

The motivation of the model analyzed in this paper originates from a collaboration with “Préfecture de police de Paris”, the police department of Paris, and “Brigade de sapeurs-pompiers de Paris”, the fire department of Paris, to design an emergency
call center in charge of receiving calls for police and firemen in the Paris area. The previous organization had two independent call centers with a single level of operators for each organization. The new call center has an architecture with two levels of operators:

(a) A first-level pool of operators handles numerous incoming calls. It has to detect and transfer the calls, which will be then classified as “urgent”, to a second-level pool of more specialized operators, policemen or firemen, depending on the call. The other incoming calls may be processed by another pool of servers or put on hold.

(b) A second-level operator with technical expertise who, after analysis, may dispatch emergency means, if needed.

It should be stressed that in the flow of incoming calls, no indication is available to determine the nature of any call, urgent or not. A call is classified as urgent only by the first level of operators, not by users who call. Only an operator determines that. An important fraction of the incoming calls concerns general information on these organizations. The first-level pool operates therefore as a filter so that, optimally, the second-level pool should process only urgent calls.

An additional, natural, constraint is that if a first-level operator has detected an urgent call, this operator releases the call only when a second-level operator can handle it. It cannot be put on hold for security reasons, given that the user is potentially in danger. This is a strict rule of these two organizations. This is a standard practice also in most of call centers of emergency units in France, and probably elsewhere.

In a congested situation, when all servers of the second level are busy, this implies that when a first-level operator detects an urgent call, it will have to wait until an operator of the second level is available. In this situation, the processing capacity of the first level is reduced since a server is blocked and, worse, an urgent call experiences a delay. One of the key issues in the design of this new organization was of determining a minimal number of (expensive) second-level operators necessary so that this blocking phenomenon had a small probability even during a congested phase.

A mathematical model is introduced to study this important phenomenon which is, to the best of our knowledge, original in the vast literature of call centers.

The flows of urgent and non-urgent calls form two independent processes, and we assume that the system is in the following state:

(a) a large number of calls are already waiting to be processed by the first level. This will be assumed to be infinite in our mathematical model.

(b) no call reneges because of congestion.

This is a worst-case scenario, since usually only a bounded number of incoming calls are waiting for processing. Additionally, in a congested state, non-urgent calls will be more likely impatient and leave the system, and therefore release the pressure on urgent calls which are generally more patient.

The purpose of this paper is therefore to analyze the behavior of a key component of an architecture, the second level of operators, during a congestion episode. This was the main motivation of this study. The waiting times experienced by users are, in particular, not meaningful in our study. We are investigating only the robustness of
the architecture under stress, with the number of busy servers of the second level of operators as a congestion indicator.

Calls require random processing times whose distribution depends on the level and the class of the call (urgent or non-urgent). We now give a quick description of this system in terms of a queueing model.

1.1 A queueing description of the system

As input, there is an infinite queue of jobs waiting to enter the system, this is the saturation assumption mentioned above.

(a) The first level has $C_1$ servers. Every time a server of this level is idle, it immediately receives a job from the infinite queue. It is of class $i \in \{0, 1\}$ with probability $p \in [0, 1]$ and $1 - p$, respectively. A job of class $i$ requires an exponentially distributed service with rate $\mu_{i1}$ at this level. Class 0 jobs are defined as “urgent calls” and have to be processed by level 2.

(i) When a job of class 0 completes its service at level 1, it goes to the second level if there is at least one idle server there.

(ii) If there is no idle server, then it remains at the first level and, consequently, blocks a server at this level. As soon as a job leaves the second level, a blocked job at the first level is sent to the second level and the first level can take a new job from the infinite queue.

When a job of class 1 completes its service, it leaves the system.

(b) The second level has $C_2$ servers and receives only class 0 jobs. A job at this level requires an exponentially distributed service with rate $\mu_{02}$.

See Fig. 1. A key feature of this network is that blocked jobs of class 0 at level 1 reduce the capacity of the system, since the corresponding servers at level 1 cannot process the calls waiting in the saturated queue.
1.2 Literature

1.2.1 Deterministic modeling

In this paper, the classes of calls and their processing times are assumed to be random. In a non-random setting, some aspects of this system have been investigated in Allamigeon et al. [1,2], where a performance analysis was carried out using deterministic Petri net modeling. A Petri net is a language describing systems in which resources circulate from place to place, incurring concurrency, synchronizations and bifurcations [5,19]. The dynamics of a Petri net can be translated into a dynamical system, whose stability and stable points can be analyzed; see Cohen et al. [9]. In Allamigeon et al. [1,2], a simplified model of the emergency call center is investigated. Computations on the stationary regimes of the dynamical system have shown a phase transition characterizing the different levels of congestion of the call center. The threshold is a critical ratio between the number of operators at level 1 and level 2.

The analysis of Petri net models may give general results for this class of systems in a deterministic framework. However, the dynamics investigated in the above articles do not take account of the random nature of the delays or the classes of calls in the call center, for example. In contrast, the queueing network analysis adopted in the present article focuses on a simpler system describing the transfer or the blocking of calls from level 1 to level 2. As it will be seen, it provides a deeper understanding of the behavior of this system in a random context.

It should be noted that the dynamical system obtained below in our results as the scaling limit of our model does correspond to the differential equations and the multi-phase behavior resulting from the continuous Petri net modeling of Allamigeon et al. [2], hence highlighting the consistency and strong relationship between both analyses.

1.2.2 Queueing models of blocking phenomena

A natural class of stochastic models related to the system described above is that of call centers. There is a huge literature dealing with the problem of staffing these systems. To the best of our knowledge, few seem to have considered jobs going through a series of call centers as in our case. The closest models of this literature seem to be multi-skill call centers where jobs can have different levels of quality of service depending on the call center chosen. They are nevertheless addressing quite different problems than the ones considered in this paper. See Pang and Perry [20], Bhulai and Koole [6], Dong and Perry [10] and Yang et al. [24], for example. See Koole and Mandelbaum [16] for a survey.

The model that we are studying can be described in terms of finite capacity queues with blocking in tandem. The blocking has the effect that, when a server at level 1 completes the service of a class 0 job, it cannot be used again until a server at level 2 is available. At level 1, a fraction of the servers, and consequently the corresponding calls, may be blocked. Related models have been investigated in the literature; see the survey Balsamo [4]. The papers study the corresponding finite Markovian models of these systems to express in particular the blocking probability at equilibrium. The corresponding equilibrium equations do not have, in general, a solution with a closed
form expression. When the values of the capacities (the numbers of operators) are not small, the dimension of the state space can be quite large, so that a numerical procedure can also be out of reach in practice. Some approximations have been proposed but, for the moment, without any convergence result which could give an idea of the accuracy of such estimations. Kelly [14] has investigated the problem of blocking of a series of queues, the analysis is concentrated on the estimation, via bounds, of the achievable throughput of such a system. To conclude, the literature of rigorous mathematical results for finite capacity queues with blocking is therefore somewhat scarce.

When the blocking is replaced by the following mechanism, defined as an exclusion process: a job blocked at some stage immediately repeats a service until the next stage can accommodate it, some of the mathematical models related to the asymmetric simple exclusion process can give some insights on the performances of these systems. Due to its relative mathematical tractability, the literature investigating these processes is also huge. See, for example, Liggett [18] for a general presentation of these important processes and Liggett [17] for a study of asymmetric simple exclusion process in finite dimension. These models are, however, quite different and do not seem to be usable, since the blocking phenomenon of interest is not really taken into account.

1.3 Contributions

With the above notation for our system, one of the main results of the paper, Theorem 2, shows that, under appropriate scaling conditions, if \( r \) is the ratio of the capacities of the two levels, \( r = \frac{C_2}{C_1} \), then the condition

\[
 r \left( \frac{p}{\mu_{01}} + \frac{1-p}{\mu_{11}} \right) > \frac{p}{\mu_{02}}
\]  

implies that there exists some fixed instant independent of the initial state such that after that time, with high probability, there are no blocked customers at level 1 on any finite time interval. See Theorem 2 and Corollary 2.

Otherwise, if the opposite (strict) inequality

\[
 r \left( \frac{p}{\mu_{01}} + \frac{1-p}{\mu_{11}} \right) < \frac{p}{\mu_{02}}
\]  

holds, then Theorem 1 shows that, under appropriate scaling conditions, the fraction of blocked customers at level 1 is positive after some time almost surely and it converges to

\[
1 - \frac{\mu_{02}}{\mu_{01}} \frac{C_2}{C_1} \left( \frac{(1-p)\mu_{01}}{p\mu_{11}} + 1 \right).
\]

See also Corollary 1.

Consequently, as the intuition suggests, if the ratio \( C_2/C_1 \) of the number of servers is larger than some parameter then the phenomenon of blocking will not occur with
high probability. Relation (1) gives therefore a rule for a convenient design of such a system.

As it will be seen in the next section, it is not difficult to figure out the conditions (1) and (2). It turns out that a formal proof requires important technical work which is the main contribution of the paper. It should be noted that, in a deterministic context, this is still an open problem.

1.3.1 A heuristic picture

Assume that there is no blocking at level 1 of class 0 jobs. Level 1 can be seen as a simple birth and death process described by the number of jobs \(Q(t)\) of class 0. A birth (resp. death) occurs when a job of class 1 (resp. 0) completing its service is replaced by a job of class 0 (resp. 1). Therefore, in state \(x \in \{0, \ldots, C_1\}\), the birth rate is \(p(C_1 - x)\mu_{11}\) and the death rate is \((1 - p)\mu_{01} x\). At equilibrium, these two rates should be of the same order and therefore, for a large \(C_1\), the number \(Q_0\) of class 0 jobs is of the order of

\[
Q_0 \sim C_1 \frac{p\mu_{11}}{(1 - p)\mu_{01} + p\mu_{11}}.
\]

To avoid congestion, the rate \(\mu_{01} Q_0\) at which class 0 jobs enter level 2 must be smaller than the maximal output rate of the second level, that is, \(C_2\mu_{02}\). This gives exactly Condition (1).

1.3.2 Mathematical aspects

Proving rigorously these intuitive results turns out to be, quite surprisingly, challenging. The Markov process associated with the queueing system has a finite state space of \(N^3\). It bears some similarity to classical loss networks of the literature but with a routing mechanism as in Jackson networks. See Kelly [13]. As such, little can be said for this process, in particular its invariant probability distribution does not seem to have a simple closed-form expression.

To get quantitative results on this system, a scaling approach is used. It is assumed that the capacities \(C_1\) and \(C_2\) are both large, so that \(C_2/C_1\) is close to some fixed constant \(r > 0\). In this framework, one investigates convergence results concerning stochastic processes indexed by the scaling parameter \(C_1\). The main technical difficulties lie in the behavior of the processes at the boundaries of the state space, when there are no blocked customers at level 1 or when there are no idle servers at level 2. As always with processes behaving locally as random walks, getting convergence results of scaled process in this context with two boundaries may be difficult. This situation has some similarities with the reflected random walks associated with classical queueing networks where the convergence results can be, sometimes, obtained by using a Skorohod problem formulation. See Harrison and Reiman [12], Chen and Mandelbaum [8] or Section 9.4 of Robert [21], for example. There is no such global Skorohod problem formulation for our model. An additional difficulty is the dependence on the
scaling parameter of the location of time intervals where blocking (or no blocking) occurs eventually.

To handle this complicated setting, we introduce two auxiliary processes which are first separately investigated in Sect. 3. For each of them, only one of the boundary conditions is involved. A generalized Skorohod problem formulation is used in both cases. Section 4 establishes the main convergence results. Stochastic calculus with Poisson processes, coupling arguments and the results obtained for auxiliary processes are the main ingredients of the proofs. See the proof of Proposition 6 for example.

2 The stochastic model

To analyze the stability properties of this network, it will be assumed that the capacities of the two levels of service are large, proportional to a scaling parameter $N$. Qualitative and quantitative properties of the system when $N$ gets large will be obtained. In particular, we will determine the conditions on the parameters for which the blocking probability is negligible or not. We begin with a brief reminder on Poisson processes and some notation used in this domain.

2.1 Notations for Poisson processes

Throughout the paper, for $\xi > 0$, one denotes by $\mathcal{N}_\xi = (t_n)$ a Poisson point process on $\mathbb{R}_+$ with rate $\xi$, and $(\mathcal{N}_\xi, i)$ denotes a sequence of such i.i.d. Poisson processes. In the following, we will use at some occasions the following coupling of Poisson processes: for $0 < \alpha \leq \beta$, one can construct a version of $\mathcal{N}_\alpha$ and $\mathcal{N}_\beta$ such that, for all $0 \leq s \leq t$,

$$\mathcal{N}_\alpha([s, t]) \overset{\text{def.}}{=} \int_s^t \mathcal{N}_\alpha(ds) \leq \mathcal{N}_\beta([s, t]).$$

This can be done in the following way: If $\mathcal{P}$ is a Poisson process on $\mathbb{R}_2^+$ whose intensity measure is Lebesgue on this space, then, for $\xi \in \{\alpha, \beta\}$, the order relation will hold if we take

$$\mathcal{N}_\xi(dt) = \mathcal{P}([0, \xi] \times dt).$$

The notation $\mathcal{N}_\xi = (t_n, B_n)$ is for a marked point Poisson process on $\mathbb{R}_+ \times \{0, 1\}$, where $(t_n)$ is a Poisson process with rate $\xi$ on $\mathbb{R}_+$ and $(B_n)$ is an i.i.d. sequence of Bernoulli random variables with parameter $p$. If $f$ is some positive Borelian function on $\mathbb{R}_+ \times \{0, 1\}$, we will use the (usual) notation

$$\int f(t, b), \mathcal{N}_\xi(ds, db) = \sum_{n \geq 1} f(t_n, B_n),$$

$(\mathcal{N}_\xi, i)$ denotes a sequence of such i.i.d. marked point Poisson processes. Concerning marked point Poisson processes, see Kingman [15] for example. They can be
interpreted as follows in our case: if $\xi \in \{\mu_{01}, \mu_{11}, \mu_{02}\}$, $u \in \{0, 1\}$ and the quantity $\overline{N}_{\xi,i}(dt, \{u\})$ is not 0, then a completion of a service occurs at time $t$, and if a new job enters the first level at this occasion, $u$ is the class of this job. Clearly the point process $\overline{N}_{\xi}(dt)$ has the same distribution as $\overline{N}_{\xi}(dt, \{0, 1\})$.

2.2 Scaling

The capacities $C_1$ and $C_2$ of levels 1 and 2 are assumed to depend on a scaling parameter $N$, $C_1 = C_1^N = N$ and $C_2 = C_2^N$ such that the convergence

$$\lim_{N \to +\infty} \frac{C_2^N}{C_1^N} = r$$

holds for some $r > 0$.

The evolution of the state of this system can described by the stochastic process $(X^N(t))_{t \geq 0} = (Y^N_1(t), Y^N_2(t), Z^N(t))$ with, for $t \geq 0$,

- $(Y^N_1(t))$ being the number of class 0 jobs blocked at level 1 at time $t$,
- $(Y^N_2(t))$, the number of class 0 jobs being served at level 1,
- $(Z^N(t))$, the number of idle servers at level 2.

For $t \geq 0$, we remark that at least one of the variables $Y^N_1(t)$ or $Z^N(t)$ is null. It is not difficult to see that $(X^N(t))$ is an irreducible Markov process on the state space

$$\mathcal{S}_N \defeq \{ x = (y_*, y, z) \in \mathbb{N}^3 : y + y_* \leq C_1^N, z \leq C_2^N, y_* z = 0 \}.$$

It will be assumed that the sequence of the initial states satisfies the relation

$$\lim_{N \to +\infty} \frac{1}{N}(Y^N_1(0), Y^N_2(0), Z^N(0)) = x_0 = (y_{*0}, y_0, z_0) \in [0, 1]^2 \times [0, r].$$

The vector $x_0$ will be referred to as the initial fluid state in the following. The transition rates are defined as follows, for $x = (y_*, y, z) \in \mathcal{S}$:

$$x \mapsto \begin{cases} 
(y_* + 1, y - 1, 0) & \text{at rate } \mu_{01}y\mathbb{1}_{\{z=0\}}, \\
(0, y - 1, z - 1) & \text{" } \mu_{01}y(1 - p)\mathbb{1}_{\{z>0\}}, \\
(0, y, z - 1) & \text{" } \mu_{01}yp\mathbb{1}_{\{z>0\}}, \\
(y_*, y + 1, z) & \text{" } \mu_{11}p(N - y_* - y), \\
(y_* - 1, y, z) & \text{" } (1 - p)\mu_{02}C_2^N\mathbb{1}_{\{y_* > 0\}}, \\
(y_* - 1, y + 1, z) & \text{" } p\mu_{02}C_2^N\mathbb{1}_{\{y_* > 0\}}, \\
(0, y, z + 1) & \text{" } \mu_{02}(C_2^N - z)\mathbb{1}_{\{y_* = 0\}}.
\end{cases}$$

Due to the constraints on the coordinates $y_*$ and $z$ of $x$ (at least one of them is 0), the Markov process $(X^N(t))$ can be seen as a two-dimensional process as depicted in Fig. 2.
2.3 Representation by stochastic differential equations

From the transition rates (5), the process $(X^N(t))$ can also be seen as the unique solution of the following stochastic differential equations:

$$
\begin{align*}
&dY^N_* (t) = \sum_{i=1}^{+\infty} \mathbb{1}_{\{i \leq Y^N(t-), Z^N(t-) = 0\}} \mathcal{N} \mu_{01,i} (dt, \{0, 1\}) \\
&\quad - \mathbb{1}_{\{Y^N_*(t-) > 0\}} \sum_{i=1}^{C^N_2} \mathcal{N} \mu_{02,i} (dt, \{0, 1\}), \quad (6) \\
&dY^N (t) = -\sum_{i=1}^{+\infty} \mathbb{1}_{\{i \leq Y^N(t-), Z^N(t-) = 0\}} \mathcal{N} \mu_{01,i} (dt, \{0, 1\}) \\
&\quad - \mathbb{1}_{\{Z^N(t-) > 0\}} \mathcal{N} \mu_{01,i} (dt, \{1\}) + \mathbb{1}_{\{Y^N_*(t-) > 0\}} \sum_{i=1}^{C^N_2} \mathcal{N} \mu_{02,i} (dt, \{0\}) \\
&\quad + \sum_{i=1}^{+\infty} \mathbb{1}_{\{i \leq C^N_2 - Z^N(t-), Y^N_*(t-) = 0\}} \mathcal{N} \mu_{11,i} (dt, \{0\}), \quad (7) \\
&dZ^N (t) = -\sum_{i=1}^{+\infty} \mathbb{1}_{\{i \leq Y^N(t-), Z^N(t-) > 0\}} \mathcal{N} \mu_{01,i} (dt, \{0, 1\}) \\
&\quad + \sum_{i=1}^{+\infty} \mathbb{1}_{\{i \leq C^N_2, Z^N(t-) = 0\}} \mathcal{N} \mu_{02,i} (dt, \{0, 1\}), \quad (8)
\end{align*}
$$
starting from some fixed initial state. The notation $f(t-) \equiv \lim_{s \to t-} f(s)$ stands for the left limit of $f$ at $t$.

### 2.4 Filtration

The $\sigma$-field $\mathcal{F}_t$ of the events up to time $t$ is classically defined as the $\sigma$-field generated by the random variables

$$\mathcal{N}_i([0, s] \times u), \text{ where } \xi \in \{\mu_0, \mu_{11}, \mu_{02}\}, s \in [0, t], u \in \{\{0\}, \{1\}\} \text{ and } i \in \mathbb{N}.$$  

With this definition, the process $(Y^N_*(t), Y^N(t), Z^N(t))$ is clearly $(\mathcal{F}_t)$-adapted. The martingale properties mentioned in the following are understood to be with respect to this filtration.

### 2.5 Evolution equations

The rescaled process is denoted by

$$\overline{(X^N(t))} \equiv (\overline{Y^N_*(t), Y^N(t), Z^N(t)}) \equiv \frac{1}{N} \left( Y^N_*(t), Y^N(t), Z^N(t) \right), \quad (9)$$  

the integration of the above SDEs and classical stochastic calculus give the relations

$$\overline{Y^N_*(t)} = \overline{Y^N_*(0)} + \mu_0 \int_0^t \overline{Y^N(s)} \mathbb{1}_{\{Z^N(s) = 0\}} \, ds$$

$$- \frac{\mu_2 C^2_2}{N} \int_0^t \mathbb{1}_{\{\overline{Y^N_*(s)} > 0\}} \, ds + \overline{M^N_{Y_*}(t)}, \quad (10)$$

$$\overline{Y^N(t)} = \overline{Y^N(0)} - \mu_0 \int_0^t \overline{Y^N(s)} \left( 1 - p \mathbb{1}_{\{Z(s) > 0\}} \right) \, ds$$

$$+ p \mu_2 \frac{C^2_2}{N} \int_0^t \mathbb{1}_{\{\overline{Y^N_*(s)} > 0\}} \, ds + p \mu_1 \int_0^t (1 - \overline{Y^N_*(s)}) \, ds - \overline{Y^N(s)} ds + \overline{M^N_Y(t)}, \quad (11)$$

$$\overline{Z^N(t)} = \overline{Z^N(0)} - \mu_0 \int_0^t \overline{Y^N(s)} \mathbb{1}_{\{Z^N(s) > 0\}} \, ds$$

$$+ \mu_2 \int_0^t \left( \frac{C^2_2}{N} - \overline{Z^N(s)} \right) \mathbb{1}_{\{\overline{Y^N_*(s)} = 0\}} \, ds + \overline{M^N_Z(t)}, \quad (12)$$

where, for $V \in \{Y_*, Y, Z\}$, $(\overline{M^N_V(t)})$ is a martingale. We complete this section with a tightness result.

**Proposition 1.** The sequence of processes $(\overline{X^N(t)})$ defined by Relation (9) is tight and any of its limiting points is a continuous process.
Proof Since, for \( t \geq 0 \), one has \( Y^*_N(t) + Y^N(t) \leq N \) and \( Z^N(t) \leq C^N_2 \), the variables \( Y^*_N(t) \), \( Y^N(t) \) and \( Z^N(t) \) are uniformly bounded. By using a similar procedure as in the proof of Theorem 6.13, page 159 of Robert [21], one can show that the expected value of the previsible increasing process of the martingales \( \overline{M}_V^N(t) \), \( V \in \{ Y_* \, Y \, Z \} \), is of the order of \( 1/N \) and thus converges to 0. By Doob’s Inequality, one gets that, for any \( \eta > 0 \) and \( T > 0 \), the relation
\[
\lim_{N \to 0} \mathbb{P} \left( \sup_{0 \leq t \leq T} |\overline{M}_V^N(t)| \geq \eta \right) = 0
\] (13)
holds. Denote by \( w_{f,T} \) the modulus of continuity of a function \( (f(t)) \) on \([0, T]\), i.e., for \( \delta > 0 \)
\[
w_{f,T}(\delta) = \sup \{|f(t) - f(s)| : 0 \leq s \leq t \leq T, |t - s| \leq \delta \}.
\]
By using again that \( (Y^*_N(t)), (Y^N(t)) \) and \( (Z^N(t)) \) are bounded, and by Relation (13), Eqs. (10), (11) and (12) show that, for any \( \varepsilon > 0 \) and \( \eta > 0 \), there exist \( N_0 \geq 1 \) and \( \delta_0 > 0 \) such that if \( N \geq N_0 \) and \( \delta < \delta_0 \) then
\[
\mathbb{P} \left( w_{V,T}(\delta) \geq \eta \right) \leq \varepsilon, \quad V \in \{ Y^*_N, Y^N, Z^N \}.
\]
One concludes with Theorem 15.1 of Billingsley [7]. \(\square\)

3 Analysis of auxiliary processes

To study the asymptotic evolution of blocked customers, it is convenient to introduce two important stochastic processes. The first one describes the behavior of the system when the second level is permanently full, and the second one corresponds to the situation when there are no blocked class 0 customers at level 1.

3.1 A process with saturation of level 2

The corresponding process is denoted by \( (Y^N_a(t), Y^N_d(t)) \); it describes a system when level 2 is always saturated by class 0 jobs. The process \( (Y^N_a(t)) \) [resp. \( (Y^N_d(t)) \)] indicates the number of class 0 jobs [resp. blocked] at level 1. For this system, blocked class 0 jobs are served at rate \( \mu_{02} C^N_2 \), otherwise the statistical assumptions are the same as before.

This is a Markov process with transition rates defined by
\[
(y_*, y) \mapsto \begin{cases}
(y_* + 1, y - 1) & \text{at rate } \mu_{01} y, \\
(y_* - 1, y) & \text{" } (1 - p)\mu_{02} C^N_2 \mathbb{1}_{\{y_*>0\}}, \\
(y_* - 1, y + 1) & \text{" } p\mu_{02} C^N_2 \mathbb{1}_{\{y_*>0\}}, \\
(y_*, y + 1) & \text{" } p\mu_{11}(N - y_* - y).
\end{cases}
\] (14)
The first transition is for a 0 job being blocked after its service at level 1. The second one corresponds to a 0 job leaving level 2, allowing a blocked 0 job to go to level 2 and a new 1 job to be added at level 1. The third transition is similar except that a new 0 job enters level 1. The last transition corresponds to a 1 job leaving level 1, allowing a 0 job to enter level 1.

As long as \( Y_0^{N}(t) > 0 \), this Markov process has the same transition rates as the process \( (Y_0^{N}(t), Y_1^{N}(t)) \); see Relation (5).

**Proposition 2** If the initial condition of \( (Y_0^{N}(t), Y_1^{N}(t)) \) is such that

\[
\lim_{N \to +\infty} \frac{1}{N} (Y_0^{N}(0), Y_1^{N}(0)) = (y_0^0, y_1^0) \in [0, 1]^2,
\]

with \( 0 \leq y_0^0 + y_1^0 \leq 1 \), then, for the convergence in distribution, the relation

\[
\lim_{N \to +\infty} \frac{1}{N} (Y_0^{N}(t), Y_1^{N}(t)) = (y_0(t), y_1(t))
\]

holds, where \((y_0(t), y_1(t))\) is a pair of continuous functions such that

\[
y_0(t) + y_1(t) \geq h(t) \overset{\text{def.}}{=} (y_0^0 + y_1^0) e^{-p \mu_{11} t} + \left( 1 - \frac{(1-p)\mu_{02} r}{p \mu_{11}} \right) \left( 1 - e^{-p \mu_{11} t} \right)
\]

and \((y_0(t), u(t))\) is the unique solution of the following Skorokhod problem:

\[
y_0(t) = y_0^0 + \mu_{01} \int_0^t y_1(s) \, ds - \mu_{02} r t + u(t),
\]

where \((u(t))\) is a non-decreasing continuous function such that \(u(0) = 0\) and

\[
\int_0^{+\infty} y_0(s) \, du(s) = 0.
\]

Concerning the Skorokhod problem in dimension 1, see Skorokhod [22] and Chaleyat-Maurel and El Karoui [11]. The main trick is to express the pair \((y_0(t), u(t))\) of Eq. (17) as a regular functional of the free process

\[
\left( y_0^0 + \mu_{01} \int_0^t y_1(s) \, ds - \mu_{02} r t \right).
\]

Note that, in our case, this free process depends on \((y_0(t), y_1(t))\).

**Proof** We will proceed as follows: First show that any limiting point \((y_0(t), y_1(t))\) of \((Y_0^{N}(t), Y_1^{N}(t))\) is such that \((y_0(t))\) can be seen as the first coordinate of the solution of a Skorokhod problem associated with a free process. In a second step, we will show
that the latter process can be expressed as a regular functional of \((y_{a*}(t))\). One has then to use uniqueness results of Anderson and Orey [3] to conclude the proof.

From the transition rates (14), the process \((Y_{a*}^N(t), Y_a^N(t))\) can be seen as the solution of the stochastic differential equations (SDEs)

\[
dY_{a*}^N(t) = \sum_{i=1}^{+\infty} \mathbb{1}_{\{i \leq Y_{a*}^N(t-), i \leq N\}} a_i \mu_{01,i}(dr,\{0, 1\}) - \sum_{i=1}^{+\infty} \mathbb{1}_{\{Y_{a*}^N(t-), i \leq N\}} a_i \mu_{02,i}(dr,\{0, 1\}).
\]

\[
dY_a^N(t) = -\sum_{i=1}^{+\infty} \mathbb{1}_{\{i \leq Y_a^N(t-), i \leq N\}} a_i \mu_{01,i}(dr,\{0, 1\}) + \sum_{i=1}^{+\infty} \mathbb{1}_{\{i \leq N - Y_{a*}^N(t-), i \leq N\}} a_i \mu_{11,i}(dr,\{0\}).
\]

With the notation

\[
\left(\bar{Y}_{a*}^N(t), \bar{Y}_a^N(t)\right) = \frac{1}{N} \left(Y_{a*}^N(t), Y_a^N(t)\right),
\]

by integrating the above SDE, one gets the relations

\[
\bar{Y}_{a*}^N(t) = \bar{Y}_{a*}^N(0) + \mu_{01} \int_0^t \bar{Y}_a^N(s)ds + \mu_{02} \frac{C^N_2}{N} \int_0^t \mathbb{1}_{\{Y_{a*}^N(s) > 0\}} ds + M_{a*}^N(t), \quad (18)
\]

\[
\bar{Y}_a^N(t) = \bar{Y}_a^N(0) - \mu_{01} \int_0^t \bar{Y}_a^N(s)ds + p \mu_{11} \int_0^t \left(1 - \bar{Y}_{a*}^N(s) - \bar{Y}_a^N(s)\right) ds + p \mu_{02} \frac{C^N_2}{N} \int_0^t \mathbb{1}_{\{Y_{a*}^N(s) > 0\}} ds + M_a^N(t), \quad (19)
\]

where \((M_{a*}^N(t))\) and \((M_a^N(t))\) are local martingales. In the same way as in the proof of Proposition 1 of Sect. 2, one can easily prove that the sequence of processes \((\bar{Y}_{a*}^N(t), \bar{Y}_a^N(t))\) is tight and that any of its limiting points is a continuous process.

Let \((y_{a*}(t), y_a(t))\) be a limiting point, i.e., for some subsequence \((N_k)\) the relation

\[
\lim_{k \to +\infty} \left(\bar{Y}_{a*}^{N_k}(t), \bar{Y}_a^{N_k}(t)\right) = (y_{a*}(t), y_a(t))
\]

holds for the convergence in distribution of processes. Denote

\[
F_a^N(t) = \bar{Y}_{a*}^N(0) + \mu_{01} \int_0^t \bar{Y}_a^N(s)ds - \mu_{02} \frac{C^N_2}{N} t + M_{a*}^N(t). \quad (20)
\]

Equation (18) can be written as

\[
\bar{Y}_{a*}^N(t) = F_a^N(t) + \mu_{02} \frac{C^N_2}{N} \int_0^t \mathbb{1}_{\{Y_{a*}^N(s) = 0\}} ds,
\]
so that the pair
\[
\left( Y^N_{a*}(t), \mu_02 - N \int_0^t 1_{Y^N_{a*}(s)=0} \, ds \right)
\]
is the solution of the Skorokhod problem associated with the free process \( F^N_a(t) \).
See Skorokhod [22] and Appendix D of Robert [21] for a brief account.
For the convergence in distribution of processes, one has
\[
\lim_{k \to +\infty} (F^N_k(t)) = (fa(t)) \overset{\text{def}}{=} \left( y^0_{a*} + \mu_01 \int_0^t y_a(s) \, ds - \mu_02rt \right).
\] (21)
since, as before, the martingales are vanishing as \( N \) gets large. From Proposition D.4 of the appendix of Robert [21], one gets that \( (ya^*_{a*}(t)) \) is the first coordinate of the solution of the Skorokhod problem associated with \( (fa(t)) \) and \( (ya_{a*}(t)) \) is differentiable almost everywhere for the Lebesgue measure on \( \mathbb{R}_+ \). In particular, Relation (17) holds.
Since the free process \( (fa(t)) \) depends on \( (ya_{a*}(t), ya(t)) \), there is no guarantee of the uniqueness of such a limit point \( (ya_{a*}(t), ya(t)) \). We now give a representation of \( (fa(t)) \) in terms of \( (ya^*_{a*}(t)) \). We proceed by getting rid of the process \( (Y^N_{a*}(t)) \) in the expression (20) of \( (F^N_a(t)) \). From Eqs. (18) and (19), we get the relation
\[
pY^N_{a*}(t) + Y^N_{a}(t) = pY^N_{a*}(0) + Y^N_{a}(0) + p\mu_{11}t - p\mu_{11} \int_0^t Y^N_{a*}(s) \, ds
\]
\[
- ((1 - p)\mu_{01} + p\mu_{11}) \int_0^t Y^N_{a}(s) \, ds + pM^N_{a*}(t) + M^N(t).
\] (22)
By reordering the terms, one gets the relation
\[
\overline{Y}^N_{a}(t) + \overline{\mu} \int_0^t \overline{Y}^N_{a}(s) \, ds = \left( pY^N_{a*}(0) + \overline{Y}^N_{a}(0) \right) + p\mu_{11}t - pY^N_{a*}(t)
\]
\[
- p\mu_{11} \int_0^t \overline{Y}^N_{a*}(s) \, ds + pM^N_{a*}(t) + M^N(t),
\] (22)
with \( \overline{\mu} \overset{\text{def}}{=} (1 - p)\mu_{01} + p\mu_{11} \). Hence, by denoting
\[
K^N(t) \overset{\text{def}}{=} \frac{1}{\overline{\mu}} \left( pY^N_{a*}(0) + \overline{Y}^N_{a}(0) \right) \left( e^{\overline{\mu}t} - 1 \right) + \frac{p\mu_{11}}{\overline{\mu}^2} \left( 1 + (\overline{\mu}t - 1)e^{\overline{\mu}t} \right),
\]
and \( (k(t)) \) its limit,
\[
k(t) \overset{\text{def}}{=} \frac{1}{\overline{\mu}} \left( pya^0_{a*} + y^0_{a} \right) \left( e^{\overline{\mu}t} - 1 \right) + \frac{p\mu_{11}}{\overline{\mu}^2} \left( 1 + (\overline{\mu}t - 1)e^{\overline{\mu}t} \right),
\]
from Eq. (22), trite calculations give the representation

\[
\int_0^t \overline{Y}_a^N(s) \, ds = K^N(t) e^{-\overline{\mu}t} - p \int_0^t \left( \overline{Y}_a^N(s) + \mu_{11} \int_0^s \overline{Y}_a^N(u) \, du \right) e^{-\overline{\mu}(t-s)} \, ds \\
+ \int_0^t \left( p M^N_n(s) + M^N_n(s) \right) e^{-\overline{\mu}(t-s)} \, ds.
\]

Therefore, we can write the free process \((F^N_a(t))\) as

\[
F^N_a(t) = G(\overline{Y}_a^N(t) + \overline{Y}_a^N(0)) + \mu_{01} K^N(t) e^{-\overline{\mu}t} - \mu_{02} \frac{C^N_n}{N} t \\
+ \mu_{01} \int_0^t \left( p M^N_n(s) + M^N_n(s) \right) e^{-\overline{\mu}(t-s)} \, ds + M^N_n(t),
\]

where \(G(\cdot)\) is a functional on Borelian functions \(x(t)\) defined by

\[
G(x)(t) = -p \mu_{01} \int_0^t \left( x(s) + \mu_{11} \int_0^s x(u) \, du \right) e^{-\overline{\mu}(t-s)} \, ds.
\]

This gives us an alternative representation of \((f_a(t))\) as

\[
f_a(t) = \overline{G}(y_{a*})(t) \overset{\text{def.}}{=} G(y_{a*})(t) + y_{a*}^0 + \mu_{01} k(t) e^{-\overline{\mu}t} - \mu_{02} r t.
\]

We have shown that \((y_{a*}(t))\) is the first coordinate of \((y_{a*}(t), u(t))\), the solution of a generalized Skorokhod problem associated to the functional \(\overline{G}\),

\[
y_{a*}(t) = \overline{G}(y_{a*})(t) + u(t) \text{ and } \int_0^{+\infty} y_{a*}(s) \, du(s) = 0,
\]

with the usual assumptions on \((y_{a*}(t))\) and \((u(t))\). See Anderson and Orey [3]. For any Borelian functions \((a(t))\) and \((b(t))\) on \(\mathbb{R}_+\), it is not difficult to check that

\[
\|G(a) - G(b)\|_{\infty,t} \overset{\text{def.}}{=} \sup_{0 \leq s \leq t} \|G(a)(s) - G(b)(s)\| \leq C_t \int_0^t \|a - b\|_{\infty,s} \, ds,
\]

with \(C_t = p \mu_{01} (1 + \mu_{11} t)\). Anderson and Orey [3] show that such \((y_{a*}(t))\) is unique. The convergence in distribution follows:

\[
\lim_{N \to +\infty} \left( \overline{Y}_a^N(t), \mu_{02} \frac{C^N_n}{N} \int_0^t \mathbb{I}_{\{\overline{Y}_a^N(s) = 0\}} \, ds \right) = (y_{a*}(t), u(t)).
\]
Consequently, Relations (18) and (19) give the relations

\[
\begin{align*}
    y_{a*}(t) &= y_{a*}^0 + \mu_0 \int_0^t y_a(s) \, ds - \mu_{02}rt + u(t), \\
    y_a(t) &= y_a^0 - \mu_0 \int_0^t y_a(s) \, ds + p\mu_{11} \int_0^t (1 - y_{a*}(s) - y_a(s)) \, ds \\
    &\quad + p\mu_{02}rt - pu(t). 
\end{align*}
\]

By using Relations (21) and (24), one deduce the uniqueness of \((y_{a*}(t))\) and, therefore, the convergence in distribution of the sequence of processes \((Y_{a*}^N(t), Y_a^N(t))\).

We now prove that the limit \((y_{a*}(t), y_a(t))\) satisfies necessarily \(y_{a*}(t) + y_a(t) \geq \overline{h}(t)\) for all \(t\), where \(\overline{h}\) is the solution of

\[
\overline{h}(t) = (y_{a*}^0 + y_a^0) - (1 - p)\mu_{02}rt + p\mu_{11} \int_0^t (1 - \overline{h}(s)) \, ds,
\]

that is,

\[
\overline{h}(t) = (y_{a*}^0 + y_a^0)e^{-p\mu_{11}t} + \left(1 - \frac{(1 - p)\mu_{02}r}{p\mu_{11}}\right)(1 - e^{-p\mu_{11}t}).
\]

First note that, for any \(N\), the process \(\overline{Y}_{a}^N\) is bounded above by 1, so that \(F_{a}^N\) is Lipschitz. Hence, again by Proposition D.4 of the appendix of Robert [21], \(u\) is also Lipschitz, and thus continuous.

From Relations (25), one gets that the identity

\[
(y_{a*}(t) + y_a(t)) = (y_{a*}^0 + y_a^0) - (1 - p)\mu_{02}rt + p\mu_{11} \int_0^t (1 - (y_{a*}(s) + y_a(s))) \, ds + (1 - p)u(t)
\]

holds, so that the difference \(y_{a*} + y_a - \overline{h}\) satisfies the system

\[
x(t) + p\mu_{11} \int_0^t x(s) \, ds = (1 - p)u(t), \text{ with } x(0) = 0.
\]

Any continuous solution \((x(t))\) of this system is nonnegative. Suppose that there exists \(t_1 > 0\) such that \(x(t_1) < 0\). Then, by continuity of \((x(t))\), there exists \(t_0 < t_1\) such that \(x(t_0) = 0\) and \(x(t) < 0\) for \(t_0 < t < t_1\). But

\[
x(t_1) - x(t_0) = x(t_1) = (1 - p)(u(t_1) - u(t_0)) - p\mu_{11} \int_{t_0}^{t_1} x(s) \, ds
\]

and the right-hand side of the equality is positive because \((u(t))\) is non-decreasing and \((x(t))\) is negative on this interval, which is a contradiction. Relation (16) is established. The proposition is proved. \(\Box\)
Proposition 3 Under Condition (2), there exists \( t_0 \geq 0 \), independent of the initial state (15), such that, for \( t \geq t_0 \), the functions \( (y_a^*(t)) \) and \( (y_a(t)) \) introduced in Proposition 2 are differentiable at \( t \) and

\[
\begin{align*}
\frac{d}{dt} y_a^*(t) &= \mu_01 y_a(t) - \mu_02 r, \\
\frac{d}{dt} y_a(t) &= -(\mu_01 + p \mu_{11}) y_a(t) - p \mu_{11} y_a^*(t) + p(\mu_02 r + \mu_{11}).
\end{align*}
\]

Any solution \( (y_a^*(t), y_a(t)) \) of the differential system (26) converges to

\[
\left( 1 - \frac{\mu_02 r}{\mu_01} \left( \frac{(1 - p) \mu_01}{p \mu_{11}} + 1 \right), \frac{\mu_02 r}{\mu_01} \right).
\]

Proof The above proposition shows that

\[
\liminf_{t \to +\infty} y_a^*(t) + y_a(t) \geq 1 - \frac{(1 - p) \mu_02 r}{p \mu_{11}}.
\]

Let

\[
\varepsilon_0 = \left( 1 - \frac{(1 - p) \mu_02 r}{p \mu_{11}} \right) - \frac{\mu_02 r}{\mu_01},
\]

then \( \varepsilon_0 > 0 \) by Condition (2). Let \( t_0 \) be such that if \( t \geq t_0 \) then

\[
y_a^*(t) + y_a(t) \geq 1 - \frac{(1 - p) \mu_02 r}{p \mu_{11}} - \frac{\varepsilon_0}{2}.
\]

The classical representation of the solution of a one-dimensional Skorokhod problem, see Relation (D.1), p.376 of Robert [21], gives the identity

\[
y_a^*(t) = \left( y_{a^*}^0 + \mu_01 \int_0^t y_a(u) du - \mu_02 r t \right) \vee \sup_{0 \leq s \leq t} \left( \mu_01 \int_s^t y_a(u) du - \mu_02 r (t - s) \right).
\]

If \( t_1 > t_0 \) is such that \( y_a^*(t_1) = 0 \), then by continuity of \( y_a(t) \), one gets the relation

\[
\mu_01 y_a(t_1) \leq \mu_02 r,
\]

and, by using Relation (28),

\[
1 - \frac{(1 - p) \mu_02 r}{p \mu_{11}} - \frac{\varepsilon_0}{2} \leq y_a^*(t_1) + y_a(t_1) = y_a(t_1) \leq \frac{\mu_02 r}{\mu_01}.
\]
which leads to a contradiction. One concludes that \( t \mapsto y_{a^*}(t) \) is positive for \( t > t_0 \) and consequently that the measure \( d\mu(t) \) vanishes on the interval \((t_0, +\infty)\). The proposition is proved.

\[ \square \]

3.2 A system without blocked jobs

A second auxiliary process is introduced, and it is denoted by \((Y^N_b(t), Z^N_b(t))\). It describes the situation when there are no blocked jobs at level one: if a class 0 job finishes while level two is saturated, i.e., \( Z^N_b(t) = 0 \), then it leaves the system instead of being blocked. If there are free servers at level two, the process behaves in the same way as the main process under study. The process \((Y^N_b(t), Z^N_b(t))\) is a Markov process, with the following transition rates:

\[
(y, z) \mapsto \begin{cases} 
(y - 1, 0) & \text{at rate } (1 - p)\mu_{01}y \mathbb{1}_{[z=0]}, \\
(y - 1, z - 1) & \text{at rate } (1 - p)\mu_{01}y \mathbb{1}_{[z>0]}, \\
(y, z - 1) & \text{at rate } p\mu_{01}y \mathbb{1}_{[z>0]}, \\
(y + 1, z) & \text{at rate } p\mu_{11}(N - y), \\
(y, z + 1) & \text{at rate } \mu_{02}(C^N_b - z).
\end{cases}
\]

(29)

Note that, when \( z > 0 \), this Markov process has the same transition rates as the process \((X^N(t))\) (see (5)).

**Proposition 4** If the initial condition of \((Y^N_b(t), Z^N_b(t))\) is such that

\[
\lim_{N \to +\infty} \frac{1}{N} (Y^N_b(0), Z^N_b(0)) = (y^0_b, z^0_b) \in [0, 1] \times [0, r],
\]

then, for the convergence in distribution, the relation

\[
\lim_{N \to +\infty} \frac{1}{N} (Y^N_b(t), Z^N_b(t)) = (y_b(t), z_b(t))
\]

holds, where \((y_b(t))\) is given by

\[
y_b(t) = y^0_b e^{-(p\mu_{11} + (1-p)\mu_{01})t} + \frac{p\mu_{11}}{p\mu_{11} + (1-p)\mu_{01}} \left( 1 - e^{-(p\mu_{11} + (1-p)\mu_{01})t} \right)
\]

(31)

and \((z_b(t))\) is the unique solution of the Skorokhod problem

\[
z_b(t) = z^0_b + \mu_{02}rt - \mu_{02} \int_0^t z_b(s)ds - \mu_{01} \int_0^t y_b(s)ds + u(t),
\]

(32)

where \((u(t))\) is a non-decreasing continuous function such that \( u(0) = 0 \) and

\[
\int_0^{+\infty} z_b(s)du(s) = 0.
\]
As before, the free process associated to the Skorokhod problem is
\[
\left( z^0_b + \mu_02rt - \mu_02 \int_0^t z_b(s)ds - \mu_01 \int_0^t y_b(s)ds \right),
\]
it is also a functional of \((z_b(t))\); see the proof of Proposition 2.

**Proof** From the transition rates (29) and as in the proof of Proposition 2, if \((Y^N_b(t), Z^N_b(t)) \equiv \frac{1}{N} (Y^N_b(t), Z^N_b(t))\), then one gets the evolution equations
\[
Y^N_b(t) = Y^N_b(0) - \mu_01(1-p) \int_0^t Y^N_b(s)ds + p \mu_{11} \int_0^t (1-Y^N_b(s))ds + M^N_Y(t),
\]
\[
Z^N_b(t) = Z^N_b(0) + \mu_02 C^2_N t - \mu_02 \int_0^t Z^N_b(s)ds
\]
\[
- \mu_01 \int_0^t Y^N_b(s)ds + M^N_Z(t) + R^N_Z(t),
\]
with
\[
R^N_Z(t) = \mu_01 \int_0^t Y^N_b(s)ds \mathbb{1}_{\left\{ Z_b^N(s)=0 \right\}} ds,
\]
and \((M^N_Z(t))\) and \((M^N_Y(t))\) are local martingales. It is easily seen that these two martingales vanish when \(N\) gets large and hence that, with the criterion of the modulus of continuity, the sequence of processes \((\overline{Y}^N_b(t))\) is tight. Furthermore, any limiting point \((y_b(t))\) satisfies the integral equation
\[
y_b(t) = y^0_b - (p \mu_{11} + (1-p) \mu_01) \int_0^t y_b(s)ds + p \mu_{11} t,
\]
so that Relation (31) holds.

Clearly, \((\overline{Z}^N_b(t), R^N_Z(t))\) is the solution of a generalized Skorokhod process associated with a free process which depends itself on \(\overline{Z}^N_b(t)\). One concludes in the same way as in the proof of Proposition 2. \(\square\)

We now prove that, under Condition (1), the reflecting part of the Skorokhod problem of the last proposition vanishes for \(t\) large enough.

**Proposition 5** Under Condition (1), there exists \(t_0 \geq 0\), independent of the initial state (30), such that, for \(t \geq t_0\),
\[
\frac{d}{dt} z_b(t) = \mu_02rt - \mu_02z_b(t) - \mu_01y_b(t).
\]
Furthermore,
\[
\lim_{t \to +\infty} (y_b(t), z_b(t)) = \left( \frac{p \mu_{11}}{p \mu_{11} + (1 - p) \mu_{01}}, r - \frac{p \mu_{01} \mu_{11}}{\mu_{02} (p \mu_{11} + (1 - p) \mu_{01})} \right).
\]

**Proof** By Proposition 4, \((y_a(t))\) converges to \(\bar{y}_b\) as \(t\) gets large. By Condition (1), the relation \(\bar{y}_b < \mu_{02} r / \mu_{01}\) holds. Consequently, there exists \(t_0\) such that the inequality
\[
\mu_{01} y_b(t) < \mu_{02} r
\]
holds for \(t \geq t_0\).

Theorem D.1 of Robert [21] gives the relation
\[
z_b(t) = \left( z_b^0 + \mu_{02} r t - \mu_{02} \int_0^t z_b(s) ds - \mu_{01} \int_0^t y_b(s) ds \right)
\]
\[
\vee \sup_{0 \leq s \leq t} \left( \mu_{02} r (t - s) - \mu_{02} \int_s^t z_b(u) du - \mu_{01} \int_s^t y_b(u) du \right).
\]

Suppose that \(z_b(t_1) = 0\) for some \(t_1 > t_0\), then, in particular, for \(0 \leq s \leq t_1\),
\[
\mu_{02} r (t_1 - s) - \mu_{02} \int_s^{t_1} z_b(u) du - \mu_{01} \int_s^{t_1} y_b(u) du \leq 0,
\]
by continuity of \((z_b(t))\) and \((y_b(t))\), this gives that
\[
\mu_{02} r - \mu_{01} y_b(t_1) = \mu_{02} r - \mu_{02} z_b(t_1) - \mu_{01} y_b(t_1) \leq 0,
\]
a contradiction. This implies that \(z_b(t) > 0\) holds for \(t \geq t_0\) and, consequently, the measure \(\mu du(t)\) vanishes on \([t_0, +\infty)\). The proposition is proved.

\(\Box\)

### 4 Asymptotic study of the blocking phenomenon

The goal of this section is to show that if the ratio \(r \sim C_2^N / C_1^N\) of the capacities of the two levels of our system is less than the quantity
\[
r_c \overset{\text{def.}}{=} \frac{p}{\mu_{02}} \left( \frac{p}{\mu_{01}} + \frac{1 - p}{\mu_{11}} \right),
\]
then there exists some fixed instant \(t_0\) such that, with a probability converging to 1 as \(N\) gets large, the number of blocked servers at level 1 is of the order of \(N\) on any finite time interval after \(t_0\). Otherwise, if \(r > r_c\), then there exists \(t_0 > 0\) such that the number of blocked servers is 0 with high probability on any finite time interval after \(t_0\). These results are, respectively, proved in Sects. 4.1 and 4.2. The proofs use the technical tools introduced in the last section and additional probabilistic arguments.
4.1 The overloaded regime

In this section, we will assume that Condition (2) holds, i.e., that \( r < r_c \). Recall that \((Y^N_*, Y^N, Z^N(t))\) describes the state of our system. The following proposition is a technical result that shows that type 0 jobs occupy at least a fixed fraction of the first level after some time.

**Proposition 6** Under Condition (2), for \( \varepsilon > 0 \), there exists \( t_0 > 0 \) such that, for any initial fluid state (4) and \( T \geq t_0 \),

\[
\lim_{N \to \infty} \mathbb{P} \left( \inf_{t \in [t_0, T]} (Y^N_* (t) + Y^N (t)) \geq N (\bar{y} - \varepsilon) \right) = 1,
\]

with

\[
\bar{y} \overset{\text{def.}}{=} \frac{p \mu_{11}}{p \mu_{11} + (1 - p) \mu_{01}}.
\]

**Proof** Define, for \( t \geq 0 \),

\[
H^N_0 (t) = N \bar{y} - Y^N_* (t) - Y^N (t).
\]

We want to show that \( H^N_0 (t) \) is, with high probability, below \( N \varepsilon \) on any finite time interval after some finite fixed instant.

The stochastic differential Eqs. (6) and (7) give the relation

\[
dH^N (t) = - \sum_{i=1}^{+\infty} \mathbb{1}_{i \leq N(1-\bar{y})+H^N (t-)} \bar{N} \mu_{11,i} (dt, \{0\}) + dD^N_0 (t),
\]

where \((D^N_0 (t))\) is the process associated with the positive jumps of this SDE:

\[
D^N_0 (t) \overset{\text{def.}}{=} \sum_{i=1}^{+\infty} \int_0^t \mathbb{1}_{i \leq N(1-\bar{y})+H^N (t-)} \bar{N} \mu_{02,i} (ds, \{1\})
\]

\[
+ \mathbb{1}_{i \leq N(1-\bar{y})+H^N (t-)} \bar{N} \mu_{01,i} (ds, \{1\}).
\]

In the following, we will use repeatedly, without mentioning it explicitly, the coupling of Poisson process with ordered rates described at the beginning of Sect. 2.

If, for some \( T > 0 \), one has that, for all \( 0 \leq t \leq T \), \( H^N_0 (t) \geq \varepsilon N \), then, on this time interval, \( Y^N (t) \leq N (\bar{y} - \varepsilon) \) and \( D^N_0 (t) - D^N_0 (s) \leq D^N_1 (t) \nu D^N_1 (s) \) for all \( 0 \leq s \leq t \), with

\[
D^N_1 (t) \overset{\text{def.}}{=} \int_0^t \sum_{i=1}^{+\infty} \mathbb{1}_{i \leq N(1-\bar{y})+H^N (t-)} \bar{N} \mu_{02,i} (ds, \{1\})
\]

\[
+ \mathbb{1}_{i \leq N(1-\bar{y})+H^N (t-)} \bar{N} \mu_{01,i} (ds, \{1\}).
\]
By using classical results on superposition and thinning of independent Poisson processes, see Kingman \cite{15} for example, we have \((D^N_1(t)) \overset{\text{dist.}}{=} (D^N_2(t))\), with
\[
D^N_2(t) \overset{\text{def.}}{=} \int_0^t \mathbb{1}_{(Y^N_*(s) > 0)} \mathcal{N}_{\mu_{02}(1-p)C^N_2}(ds) + \mathbb{1}_{(Z^N(s) > 0)} \mathcal{N}_{\mu_{01}(1-p)N(\bar{\gamma} - \epsilon)}(ds).
\]

It is easily seen that Condition (2) is equivalent to the relation \(r \mu_{02} < \bar{\gamma} \mu_{01}\), hence for \(\eta > 0\) there exists some \(N_0\) such that
\[
\mu_{02}C^N_2 < \mu_{01}(\bar{\gamma} + \eta)N, \quad \text{for } N \geq N_0.
\]

By using this inequality and the relation \([Y^N_*(t- > 0)] \subset [Z^N(t- = 0)]\), one has
\[
\int_0^t \mathbb{1}_{(Y^N_*(s) > 0)} \mathcal{N}_{\mu_{02}(1-p)C^N_2}(ds) \leq \int_0^t \mathbb{1}_{(Z^N(s) = 0)} \mathcal{N}_{\mu_{01}(1-p)N(\bar{\gamma} + \eta)}(ds).
\]

Moreover, since
\[
\int_0^t \mathbb{1}_{(Z^N(s) > 0)} \mathcal{N}_{\mu_{01}(1-p)N(\bar{\gamma} - \epsilon)}(ds) \leq \int_0^t \mathbb{1}_{(Z^N(s) > 0)} \mathcal{N}_{\mu_{01}(1-p)N(\bar{\gamma} + \eta)}(ds),
\]
\[
D^N_2(t) - D^N_2(s) \leq D^N_3(t) - D^N_3(s) \text{ for } 0 \leq s \leq t,
\]
with
\[
D^N_3(t) \overset{\text{def.}}{=} \int_0^t \mathbb{1}_{(Z^N(s) = 0)} \mathcal{N}_{\mu_{01}(1-p)N(\bar{\gamma} + \eta)}(ds) + \mathbb{1}_{(Z^N(s) > 0)} \mathcal{N}_{\mu_{01}(1-p)N(\bar{\gamma} + \eta),2}(ds),
\]
where
\[
\mathcal{N}_{\mu_{01}(1-p)N(\bar{\gamma} + \eta)} \text{ and } \mathcal{N}_{\mu_{01}(1-p)N(\bar{\gamma} + \eta),2}
\]
are two independent Poisson point processes. The integer valued process \((D^N_3(t))\) has jumps of size 1 and it is easily checked that
\[
\left( D^N_3(t) - \mu_{01}N(1 - p)(\bar{\gamma} + \eta)t \right)
\]
is a martingale. From Watanabe’s Theorem, see Watanabe \cite{23}, one gets that \((D^N_3(t))\) is a Poisson process on \(\mathbb{R}_+\) with rate \(\lambda_HN\), with \(\lambda_H \overset{\text{def.}}{=} \mu_{01}(1 - p)(\bar{\gamma} + \eta)\).

If, for all \(0 \leq t \leq T\), \(H^N(t) \geq \lfloor \epsilon N \rfloor\), then
\[
D^N_0(t) - D^N_0(s) \leq D^N_3(t) - D^N_3(s) \text{ for } 0 \leq s \leq t, \quad (35)
\]
where \((D^N_3(t))\) is a Poisson process with rate \(\lambda_HN\).
Assume that $H^N(0) \geq [\varepsilon N]$ and define the process $(\overline{H}^N(t))$, with the initial condition $\overline{H}^N(0) = H^N(0) - [\varepsilon N]$ and such that

$$d\overline{H}^N(t) = N\lambda [\overline{H}^N(t)]^{-1} \sum_{i=1}^{+\infty} 1_{[i \leq N(1-\bar{y})+\varepsilon N]} \frac{\varepsilon}{\mu_{11,i}} (dr,\{0\})$$

holds for $t \geq 0$. Clearly, $(\overline{H}^N(t))$ has the same distribution as $(L_H(Nt))$, where $(L_H(t))$ is the Markov process associated with an $M/M/1$ queue whose arrival and service rates are, respectively, $\lambda_H$ and $\mu_H \overset{\text{def}}{=} p\mu_{11}(1-\bar{y}+\varepsilon)$. Note that, by the definition of $\bar{y}$,

$$\lambda_H - \mu_H = \mu_{01}(1-p)\eta - p\mu_{11}\varepsilon.$$ 

One can choose $\eta > 0$ such that $\lambda_H - \mu_H < -p\mu_{11}\varepsilon/2$ holds, hence $(L_H(t))$ is an ergodic Markov process in this case.

Let $T^N_\varepsilon = \inf\{t \geq 0 : H^N(t) < \varepsilon N\}$, in particular $H^N(t) \geq [\varepsilon N]$ for $0 \leq t \leq T^N_\varepsilon$. From Relations (34), (35) and (36), we get therefore the inequality

$$H^N(t) - [\varepsilon N] \leq \overline{H}^N(t),$$

for $t < T^N_\varepsilon$, hence $T^N_\varepsilon \leq \tau^N_L$, with

$$\tau^N_L = \inf\{t > 0 : \overline{H}^N(t) = 0\} = \inf\{t > 0 : L_H(Nt) = 0\}.$$ 

Since $\overline{H}^N(0) \leq H^N(0) \leq N$, Proposition 5.16 of Robert [21] gives that, for any $t_0$ such that $t_0 > 1/(\mu_H - \lambda_H)$,

$$\lim_{N \to +\infty} P\left(T^N_\varepsilon \leq t_0\right) \geq \lim_{N \to +\infty} P\left(\tau^N_L \leq t_0\right) = 1.$$ 

By using the strong Markov property, up to a change of time origin, one can assume that $H^N(0) \leq [\varepsilon N]$. If $(\overline{H}^N(t))$ is defined as before with the initial condition $\overline{H}^N(0) = 0$, then it is not difficult to show that Relation (37) holds for all $t \geq 0$. For the excursions of $(H^N(t))$ below $[\varepsilon N]$ this is clear, and for the excursions above this level it has just been proved. In particular, for any $T > 0$,

$$P\left(\inf_{0 \leq t \leq T} Y^N_*(t) + Y^N(t) \leq N(\bar{y} - 2\varepsilon)\right) = P\left(\sup_{0 \leq t \leq T} \overline{H}^N(t) \geq 2\varepsilon N\right) \leq P_0\left(\sup_{0 \leq t \leq NT} L_H(t) \geq [\varepsilon N]\right),$$

and the last quantity is the probability that the hitting time of $[\varepsilon N]$ by an $M/M/1$ queue starting from 0 is less than $NT$. Proposition 5.11 of Robert [21] shows that this
hitting time is of the order of \((\mu_H/\lambda_H)^{[\epsilon N]}\) and therefore exponentially large in \(N\) (recall that \(\lambda_H < \mu_H\)). In particular, the last term of the right-hand side of the above relation is converging to 0 as \(N\) gets large. The proposition is proved.

The above proof relies on the comparisons of point processes associated to the counting processes \((D_i(t)), i \in \{0, 1, 2, 3\}\). We have, for example, that the point process associated to \((D_0(t))\) is “smaller” that the one associated to \((D_1(t))\) in the following sense: \(D_0^N(t) - D_0^N(s) \leq D_1^N(t) - D_1^N(s)\) holds for all \(0 \leq s \leq t\). In the following, for convenience, we will use the notation \(D_0(dt) \leq D_1(dt)\).

**Proposition 7** Under Condition (2), for any \(\epsilon > 0\) small enough, there exists \(t_1 > 0\) such that, for any initial fluid state (4), and for any \(T \geq t_1\), the relation

\[
\lim_{N \to \infty} \mathbb{P} \left( \inf_{t \in [t_1, T]} Y_N^N(t) \geq \epsilon N \right) = 1
\]

holds.

**Proof** From Eqs. (6) and (8), we get that

\[
d \left( Y^N_\ast - Z^N \right)(t) = \sum_{i=1}^{\infty} \mathbb{1}_{\{i \leq Y^N(t-)\}} \mathcal{N}^{\mu_{01},i}(dr) - \sum_{i=1}^{\infty} \mathbb{1}_{\{i \leq C^N_2 - Z^N(t-)\}} \mathcal{N}^{\mu_{02},i}(dr).
\]

By Proposition 6, there exists \(t_0\) such that, for \(T \geq t_0\), for the events

\[
\mathcal{A}_N \overset{\text{def}}{=} \left\{ \inf_{t \in [t_0, T]} (Y^N_\ast(t) + Y^N(t)) \geq N(\tilde{\gamma} - \epsilon) \right\},
\]

the sequence \((\mathbb{P}(\mathcal{A}_N))\) converges to 1.

Suppose that, for some instant \(t \in (t_0, T)\), one has \(Y^N_\ast(t) < 2\epsilon N\), then, on the event \(\mathcal{A}_N\), the relation \(Y^N(t) \geq N(\tilde{\gamma} - 3\epsilon)\) holds, and consequently

\[
d \left( Y^N_\ast - Z^N \right)(t) \geq \sum_{i=1}^{\infty} \mathbb{1}_{\{i \leq Y^N(\tilde{\gamma} - 3\epsilon)\}} \mathcal{N}^{\mu_{01},i}(dr) - \sum_{i=1}^{\infty} \mathbb{1}_{\{i \leq C^N_2 \}} \mathcal{N}^{\mu_{02},i}(dr)
\]

\[
\overset{\text{dist.}}{=} \mathcal{N}^{\mu_N}(dr) - \mathcal{N}^{\lambda_N}(dr) \tag{38}
\]

with \(\mu_N = [\mu_{01}N(\tilde{\gamma} - 3\epsilon)]\) and \(\lambda_N = \mu_{02}C^N_2\). As noted before, Condition (2) is equivalent to the relation \(\tilde{\gamma}\mu_{01} - r\mu_{02} > 0\). Since

\[
\lim_{N \to +\infty} \frac{\mu_N - \lambda_N}{N} = (\mu_{01}\tilde{\gamma} - \mu_{02}r) - \mu_{01}3\epsilon,
\]

one can find \(0 < \lambda < \mu\) and \(\epsilon\) sufficiently small such that, for \(N\) sufficiently large, \(\lambda_N \geq \lambda N\) and \(\mu_N \leq \mu N\) hold. If

\[
T_N = \inf\{t > 0 : Y^N_\ast(t) \geq 2\epsilon N\},
\]

\(\Box\) Springer
then since \((Y_N^* - Z_N^*) (0) \geq - C_2^N\), Relation (38) gives the existence of \(t_1 > 0\) such that

\[
\lim_{N \to +\infty} \mathbb{P}(T_N \leq t_1) = 1.
\]

We now assume that \((Y_N^* - Z_N^*) (0) = [2\varepsilon N]\). By taking \(T > t_1\), using Relation (38) and the estimates for \(\lambda_N\) and \(\mu_N\), we get that, on the event \(A_N\),

\[
(Y_N^* - Z_N^*)(t) \geq [2\varepsilon N] - X(Nt)
\]

holds for all \(t > 0\), where \((X(t))\) is an \(M/M/1\) queue with arrival [resp. service] rate \(\lambda\) [resp. \(\mu\)] starting at 0. With the same argument as at the end of the previous proof, since \(\lambda < \mu\), we have that, for any \(T > 0\),

\[
\lim_{N \to +\infty} \mathbb{P}\left(\sup_{0 \leq t \leq NT} X(s) \geq \varepsilon N\right) = 0,
\]

and consequently

\[
\lim_{N \to +\infty} \mathbb{P}\left(\left\{\sup_{0 \leq t \leq T} (Y_N^* - Z_N^*)(t) \geq [\varepsilon N]\right\} \cap A_N\right) = 1.
\]

We conclude the proof by using the fact that \(Y_N^*(t) > 0\) implies that \(Z_N^*(t) = 0\). \(\square\)

The following corollary gives a more precise statement concerning the asymptotic behavior of the \(Z\)-component of the state vector. It is a simple consequence of the fact that \(Y_N^*(t) > 0\) implies \(Z_N^*(t) = 0\).

**Corollary 1** Under Condition (2), there exists \(t_1 > 0\) such that

\[
\lim_{N \to +\infty} \mathbb{P}\left(Z_N^*(t) = 0, \forall t \in [t_1, T]\right) = 1
\]

holds for any initial fluid state (4).

**Theorem 1** (Saturated regime) Under Condition (2), there exists \(t_1 > 0\) such that, for any initial fluid state (4), any limiting point \((y_\infty^*(t), y_\infty(t), z_\infty(t))\) of the sequence \((X_N^*(t))\) defined by Relation (9) satisfies the following relations: for all \(t \geq t_1\), \(z_\infty(t) = 0\) and the differential equations

\[
\begin{align*}
\frac{dy_\infty^*}{dt}(t) &= \mu_0 y_\infty(t) - \mu_0 z^r, \\
\frac{dy_\infty}{dt}(t) &= -\mu_0 y_\infty(t) + p(\mu_0 z^r + \mu_1 (1 - y_\infty^*(t) - y_\infty(t)))
\end{align*}
\]
hold. Furthermore
\[
\lim_{t \to +\infty} \left( y_*^N(t), y_\infty^N(t) \right) = \left( 1 - \frac{\mu_02p}{\mu_01} \left( \frac{(1 - p)\mu_01}{p\mu_11} + 1 \right), \frac{\mu_02p}{\mu_01} \right).
\]

**Proof** By Proposition 7, for some \( \varepsilon_0 \) sufficiently small, there exists \( t_1 \) such that, for any \( T > t_1 \), the event
\[
\mathcal{E}_N \overset{\text{def.}}{=} \left\{ Y_*^N(t) \geq \varepsilon_0 N : \forall t \in [t_1, T] \right\}
\]
has a probability arbitrarily close to 1 as \( N \) gets large. Let the process \((Y_*^N(t), Y_\infty^N(t))\) defined in Sect. 3.1 start from the initial state \((Y_*^N(t_1), Y_\infty^N(t_1))\), then, by checking the Q-matrix of each process, it is easily seen that, on the event \( \mathcal{E}_N \), the relation
\[
\left( (Y_*^N(t), Y_\infty^N(t)), t_1 \leq t \leq T \right) \overset{\text{dist.}}{=} \left( (Y_*^N(t), Y_\infty^N(t)), t_1 \leq t \leq T \right)
\]
holds. By using the fact that the sequence of random variables
\[
\left( \frac{1}{N} (Y_*^N(t_1), Y_\infty^N(t_1)) \right) \in [0, 1]
\]
is tight, one has only to use Proposition 2 to conclude the proof of the theorem. \( \square \)

### 4.2 The underloaded regime

In this section, it will be assumed that Condition (1) holds.

**Proposition 8** Under Condition (1), there exists \( \eta_0 > 0 \) and \( t_1 > 0 \) such that, for any initial fluid state (4) and for \( T > t_1 \),
\[
\lim_{N \to \infty} \mathbb{P} \left( \sup_{t \in [t_1, T]} \left( Y_*^N(t) + Y_\infty^N(t) \right) \leq N \left( y - \eta_0 \right) \right) = 1,
\]
with \( y = r\mu_{02}/\mu_{01} \).

Note that \( y < 1 \) by Condition (1).

**Proof** By using the SDEs (6) and (7), we get that
\[
d \left( Y_*^N + Y_\infty^N \right)(t) = \sum_{i=1}^{+\infty} 1_{[i \leq N - Y_*^N(t-), Y_\infty^N(t-)]} \overline{\nu}_{\mu_{11},i}(dr, \{0\})
- 1_{[Y_*^N(t-) > 0]} \overline{\nu}_{\mu_{02},E_2}(dr, \{1\})
- \sum_{i=1}^{+\infty} 1_{[i \leq Y_\infty(t-), Z_\infty(t-) > 0]} \overline{\nu}_{\mu_{01},i}(dr, \{1\})
\]
(39)
holds. We are going to give an upper bound for the process \((Y^*_N(t) + Y^N(t))\). As before we will work on the differential terms of the above relation.

We choose \(\eta > 0\) sufficiently small such that, for \(N\) large enough, the relation
\[
\frac{\mu_{02}}{\mu_{01}} C_2^N > \left\lfloor N \left( r \frac{\mu_{02}}{\mu_{01}} - \eta \right) \right\rfloor
\]
holds. Under this condition one has, with a convenient coupling of Poisson processes,
\[
\mathcal{N}_{\mu_{02} C_2^N}(dt, du) \geq \sum_{i=1}^{+\infty} \mathbb{1}_{\{i \leq N(\mu_{02} r/\mu_{01} - \eta)\}} \mathcal{N}_{\mu_{01},i}(dt, \{1\}).
\] (40)

The relation \(Y^*_N(t-) > 0\) implies \(Z^N(t-) = 0\), consequently we get the inequality
\[
\mathbb{1}_{\{Y^*_N(t-) > 0\}} \mathcal{N}_{\mu_{02} C_2^N}(dt, \{1\}) \geq \mathbb{1}_{\{Z^N(t-) = 0\}} \sum_{i=1}^{+\infty} \mathbb{1}_{\{i \leq N(\mu_{02} r/\mu_{01} - \eta)\}} \mathcal{N}_{\mu_{01},i}(dt, \{1\}).
\]

If the relation \(Y^N(t-) + Y^*_N(t-) \geq N(\mu_{02} r/\mu_{01} - \eta)\) holds, then
\[
\mathbb{1}_{\{Z^N(t-) > 0\}} \sum_{i=1}^{+\infty} \mathbb{1}_{\{i \leq Y^N(t-)\}} \mathcal{N}_{\mu_{01},i}(dt, \{1\}) \geq \mathbb{1}_{\{Z^N(t-) = 0\}} \sum_{i=1}^{+\infty} \mathbb{1}_{\{i \leq N(\mu_{02} r/\mu_{01} - \eta)\}} \mathcal{N}_{\mu_{01},i}(dt, \{1\}),
\] (41)
since \(Y^*_N(t-) = 0\) if \(Z^N(t-) > 0\).

By plugging Relations (40) and (41) into the SDE (39), we get that
\[
d \left( Y^*_N + Y^N \right)(t) \leq \mathcal{N}_{\lambda_N}(dt) - \mathcal{N}_{\mu_N}(dt)
\] (42)
holds on the event \(Y^N(t-) + Y^*_N(t-) \geq N(\mu_{02} r/\mu_{01} - \eta)\), with
\[
(\lambda_N, \mu_N) \overset{\text{def.}}{=} \left( p \mu_{11} N \left( 1 - r \frac{\mu_{02}}{\mu_{01}} + \eta \right), \left( 1 - p \right) \mu_{01} N \left( r \frac{\mu_{02}}{\mu_{01}} - \eta \right) \right).
\]

By Condition (1) we can take \(\eta = \eta_0 > 0\) to be such that
\[
2\eta_0 < r \frac{\mu_{02}}{\mu_{01}} - \frac{p \mu_{11}}{p \mu_{11} + (1 - p) \mu_{01}}.
\]
In this case, if \((\lambda, \mu)\) is the limit of the sequence \(((\lambda_N, \mu_N))/N\), then \(\lambda < \mu\). There exist \(0 < \lambda_0 < \mu_0\), such that, for \(N\) sufficiently large, the relations \(\lambda_N \leq \lambda_0 N\) and
Hold. Let \((X(t))\) be the (ergodic) \(M/M/1\) queue with input [resp. service] rate given by \(\lambda_0\) [resp. \(\mu_0\)] and \(X(0) = N\). Then Eq. (42) gives the coupling relation
\[
(Y_*^N + Y^N(t)) \leq X(Nt),
\]
for \(t\) less than the hitting time of \(N(\mu_0r/\mu_0 - 2\eta_0)\). Consequently, by ergodicity, there exists some \(t_1 \geq 0\) such that this hitting time is, with high probability, less than \(t_1 N\). Now, by taking the initial conditions \((Y_*^N + Y^N)(0) = X(0) = [N(\mu_0r/\mu_0 - 2\eta_0)]\), with the same argument as in the proof of Proposition 7, one gets that, for \(T > 0\), the process \(((Y_*^N + Y^N)(t))\) remains below \(N(\mu_0r/\mu_0 - \eta_0)\) on the time interval \([0, T]\) with high probability. The proposition is proved. \(\blacksquare\)

The following result is the analogue of Proposition 7 for the underloaded regime.

**Proposition 9** Under Condition (1), for any \(\varepsilon > 0\) small enough, there exists a time \(t_1 \geq 0\) such that, for any initial fluid state (4) and for \(T > t_1\),
\[
\lim_{N \to \infty} \mathbb{P} \left( \inf_{t \in [t_1, T]} (Z^N(t) - Y_*^N(t)) \geq \varepsilon N \right) = 1.
\]

**Proof** Since the proof follows the same lines as the proof of Proposition 7, we sketch the main technical arguments. From the last proposition, one can chose \(\eta_0\) and \(t_1 \geq 0\) such that the event
\[
\mathcal{B}_N \overset{\text{def.}}{=} \left\{ \sup_{t \in [t_1, T]} (Y_*^N(t) + Y^N(t)) \leq N \left( \frac{r \mu_0}{\mu_0 - \eta_0} \right) \right\}
\]
has a probability converging to 1 when \(N\) gets large.

The SDEs (7) and (8) give the relation
\[
d (Z^N - Y_*^N) (t) = \sum_{i=1}^{+\infty} \mathbb{1}_{i \leq C^N_2 - (\varepsilon N)} \mathcal{N}_{\mu_0, i} (dr) - \sum_{i=1}^{+\infty} \mathbb{1}_{i \leq Y^N(t-)} \mathcal{N}_{\mu_0, i} (dr),
\]
by using again that \(Z^N(t-)\) is null if \(Y_*^N(t-)\) is positive.

One takes \(\eta_0 < \mu_0 r / (4\mu_0)\). Then on the event \(\mathcal{B}_N\), if \(Z^N(t) \leq \varepsilon N\),
\[
d (Z^N - Y_*^N) (t) \geq \sum_{i=1}^{+\infty} \mathbb{1}_{i \leq C^N_2 - \varepsilon N} \mathcal{N}_{\mu_0, i} (dr) - \sum_{i=1}^{+\infty} \mathbb{1}_{i \leq N \mu_0 (r - \varepsilon / 2) / \mu_0} \mathcal{N}_{\mu_0, i} (dr).
\]

Hence, the process \((Z^N - Y_*^N(t))\) can be compared with a (scaled) ergodic \(M/M/1\) queue with arrival rate \([C^N_2 - \varepsilon N] \mu_0\) and service rate \([N(r - \varepsilon / 2) \mu_0 / \mu_0] \mu_0\). We conclude the proof in the same way as in the proof of Proposition 6. \(\blacksquare\)
**Corollary 2** Under Condition (1), there exists $t_1 > 0$ such that
\[
\lim_{N \to +\infty} \mathbb{P} \left( Y^N_\ast (t) = 0, \forall t \in [t_1, T] \right) = 1
\]
holds for any initial fluid state (4).

**Proof** The proof follows from the mutual exclusivity of the events $\{ Y^N_\ast (t) > 0 \}$ and $\{ Z^N (t) > 0 \}$ and from Proposition 9.

We can now state the main result for the underloaded regime.

**Theorem 2** (Underloaded Regime) Under Condition (2), there exists $t_1 > 0$ such that, for any initial fluid state (4), any limiting point $(y^\infty_\ast (t), y^\infty (t), z^\infty (t))$ of the sequence $(X^N (t))$ defined by Relation (9) satisfies the following relations, for all $t \geq t_1$, $y^\infty_\ast (t) = 0$ and the differential equations
\[
\frac{dy^\infty_\ast}{dt}(t) = -(p \mu_{11} + (1-p) \mu_{01}) y^\infty_\ast (t) + p \mu_{11},
\]
\[
\frac{dz^\infty}{dt}(t) = -\mu_{02} z^\infty (t) - \mu_{01} y^\infty_\ast (t) + \mu_{02} r
\]
hold. Furthermore,
\[
\lim_{t \to +\infty} (y^\infty (t), z^\infty (t)) = \left( \frac{p \mu_{11}}{p \mu_{11} + (1-p) \mu_{01}}, r - \frac{p \mu_{01} \mu_{11}}{\mu_{02} (p \mu_{11} + (1-p) \mu_{01})} \right).
\]

**Proof** In the same way as in the proof of Theorem 1, a coupling between the processes $(Y^N_\ast (t), Z^N_\ast (t))$ defined in Sect. 3.2 and the process $(Y^N (t), Z^N (t))$ can be constructed so that the convergence results of Proposition 5 can be used.

**References**

1. Allamigeon, X., Beuf, V., Gaubert, S.: Performance evaluation of an emergency call center: tropical polynomial systems applied to timed Petri nets, FORMATS’15. Lecture Notes in Computer Science, vol. 9268. Springer (2015)
2. Allamigeon, X., Beuf, V., Gaubert, S.: Stationary solutions of discrete and continuous Petri nets with priorities. Perform. Eval. 113, 1–12 (2017)
3. Anderson, R.F., Orey, S.: Small random perturbation of dynamical systems with reflecting boundary. Nagoya Math. J. 60, 189–216 (1976)
4. Balsamo, S.: Queueing Networks with Blocking: Analysis, Solution Algorithms and Properties, pp. 233–257. Springer, Berlin (2011)
5. Best, E., Wimmel, H.: Structure Theory of Petri Nets, Transactions on Petri Nets and Other Models of Concurrency VII, pp. 162–224. Springer, Berlin (2013)
6. Bhusari, S., Koole, G.: A queueing model for call blending in call centers. IEEE Trans. Autom. Control 48(8), 1434–1438 (2003)
7. Billingsley, P.: Convergence of Probability Measures. Wiley Series in Probability and Statistics: Probability and Statistics. Wiley, New York (1999)
8. Chen, H., Mandelbaum, A.: Discrete flow networks: bottleneck analysis and fluid approximations. Math. Oper. Res. 16(2), 408–446 (1991)
9. Cohen, G., Gaubert, S., Quadrat, J.-P.: Asymptotic throughput of continuous timed Petri nets. In: 34th Conference on Decision and Control (New Orleans) (1995)
10. Dong, J., Perry, O.: Queueing models for patient-flow dynamics in inpatient wards. Available at SSRN 3246641 (2018)
11. El Karoui, N., Chaleyat-Maurel, M.: Temps locaux, vol. 52–53, ch. Un problème de réflexion et ses applications au temps local et aux équations différentielles stochastiques sur $\mathbb{R}$, pp. 117–144, Société Mathématique de France, 1978, Exposés du Séminaire J. Azéma-M. Yor, Paris (1976–1977)
12. Harrison, J.M., Reiman, M.I.: Reflected Brownian motion on an orthant. Ann. Probab. 9(2), 302–308 (1981)
13. Kelly, F.P.: Reversibility and Stochastic Networks. Wiley Series in Probability and Mathematical Statistics. Wiley, Chichester (1979)
14. Kelly, F.P.: Blocking probabilities in large circuit-switched networks. Adv. Appl. Probab. 18, 473–505 (1986)
15. Kingman, J.F.C.: Poisson Processes, Oxford studies in probability. Oxford Science Publications, The Clarendon Press, Oxford University Press, New York (1993)
16. Koole, G., Mandelbaum, A.: Queueing models of call centers: an introduction. Ann. Oper. Res. 113(1), 41–59 (2002)
17. Liggett, T.M.: Ergodic theorems for the asymmetric simple exclusion process. Trans. Am. Math. Soc. 213, 237–261 (1975)
18. Liggett, T.M.: Interacting Particle Systems. Grundlehrender mathematischen Wissenschaften. Springer, New York (1985)
19. Murata, T.: Petri nets: properties, analysis and applications. Proc. IEEE 77(4), 541–580 (1989)
20. Pang, G., Perry, O.: A logarithmic safety staffing rule for contact centers with call blending. Manag. Sci. 61(1), 73–91 (2015)
21. Robert, P.: Stochastic Networks and Queues, Stochastic Modelling and Applied Probability Series, vol. 52. Springer, New York (2003)
22. Skorokhod, A.V.: Stochastic equations for diffusion processes in a bounded region. Theory Probab. Appl. 7, 3–23 (1962)
23. Watanabe, S.: On discontinuous additive functionals and Lévy measures of a Markov process. Jpn. J. Math. 34, 53–70 (1964)
24. Yang, R., Bhulai, S., Van der Mei, R.: Optimal resource allocation for multiqueue systems with a shared server pool. Queueing Syst. 68(2), 133–163 (2011)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.