A STOCHASTIC ANALYSIS OF RESOURCE SHARING WITH LOGARITHMIC WEIGHTS

PHILIPPE ROBERT AND AMANDINE VÉBER

Abstract. The paper investigates the properties of a class of resource allocation algorithms for communication networks: if a node of this network has x requests to transmit, then it receives a fraction of the capacity proportional to $\log(1+x)$, the logarithm of its current load. A detailed fluid scaling analysis of such a network with two nodes is presented. It is shown that the interaction of several time scales plays an important role in the evolution of such a system, in particular its coordinates may live on very different time and space scales. As a consequence, the associated stochastic processes turn out to have unusual scaling behaviors. A heavy traffic limit theorem for the invariant distribution is also proved. Finally, we present a generalization to the resource sharing algorithm for which the log function is replaced by an increasing function. Possible generalizations of these results with $J > 2$ nodes or with the function log replaced by another slowly increasing function are discussed.

Contents

1. Introduction 1
2. Presentation of Results 6
3. The Stochastic Model 9
4. The Initial Phase 11
5. A Local Equilibrium 15
6. The Fluid Time Scale 21
7. Heavy Traffic Regime 28
8. General Functions for Resource Sharing 31
9. The Network with $J > 2$ Nodes 35
References 37

1. Introduction

The resource allocation problem considered in this paper involves J nodes which have access to a common shared resource, for example a communication channel or a processing unit. The resource is assumed to have a fixed capacity, say 1. The resource is shared among nodes in the following way: for $1 \leq j \leq J$, if node j has n_j requests pending, it receives the instantaneous fraction of capacity

$$\frac{f(n_j)}{f(n_1) + f(n_2) + \cdots + f(n_J)} \tag{1}$$

Date: July 1, 2014.
1991 *Mathematics Subject Classification.* Primary:60K25, 60K30, 60F05; Secondary:68M20, 90B22 .

from the resource. The algorithm is thus defined by the function $x \mapsto f(x)$. There are several situations where the capacity is allocated in this way. It should be noted that our results are proved in the case where $J = 2$. The general case $J \geq 2$ is briefly sketched. See Section 9 for the conjectured behavior of this system.

1.1. **Saturated Node of the Internet.** In this context, the nodes correspond to TCP flows with different sources and destinations. The resource here is the processing time of a fixed router on the path of these flows. The packets of a flow are queued in the buffer of the router until they are routed to the next stage of their path. Congestion is simply the situation when the buffer is full and therefore incoming packets are lost. Because of the TCP protocol, a given flow will increase or decrease the rate at which it sends the packets, depending on the level of congestion of the routers on its path. There are several ways to represent this phenomenon. It should be kept in mind that the following descriptions are mathematical models of the way TCP is *thought* to allocate bandwidth, not of the TCP algorithm itself. See Massoulié and Roberts [20].

a) **Processor-Sharing disciplines.**
A popular, simplified, stochastic model of this situation consists in considering that the router allocates its processing power to each flow according to a slight generalization of the allocation policy given by Relation (1) with a function f depending on the node j and of the form $w_j n$, where $1/w_j$ can be the round trip time between the source and the destination. This allocation algorithm corresponds to the *Discriminatory Processor-Sharing Policy.* Node j has an instantaneous fraction of capacity given by

$$\frac{w_j n_j}{w_1 n_1 + w_2 n_2 + \cdots + w_J n_J}. \tag{2}$$

See Altman et al. [2] and references therein for a survey. When all the w_j's are 1, we obtain the classical processor-sharing policy: node j receives the fraction of capacity $n_j/(n_1 + \cdots + n_J)$, and the bandwidth is equally divided among the current requests. Different classes of stochastic models of processor-sharing policies have been extensively used to describe the congestion in IP networks. See Bonald et al. [5], Kelly et al. [15] and Graham and Robert [14] and references therein.

b) **Alpha-fair disciplines.**
These policies have also been introduced to describe the allocation of bandwidth in IP networks (see Mo and Walrand [22]), in terms of an optimization problem (cf. Kelly et al. [15]). In our context, a related policy would correspond to the case $f(n) = n^\alpha$, $n \geq 0$, so that a non-empty node j has an instantaneous fraction of capacity given by

$$\frac{n_j^\alpha}{n_1^\alpha + n_2^\alpha + \cdots + n_J^\alpha}. \tag{3}$$

The case $\alpha = 1$ is the processor-sharing discipline presented above.

In the wireless section below, the situation is quite different since the bandwidth allocation algorithm is defined explicitly by relations similar to (1).

1.2. **Wireless Networks.** This is again a simplified, but meaningful stochastic model of bandwidth allocation, this time in wireless networks. The resource here is a radio channel in a region where there are J stations/mobiles. At a given time, because of interferences, only one station can transmit successfully in this region. A station with n_j messages waiting for transmission can detect if there is a communication going on or not. If not, a classical backoff mechanism is used: the station starts transmitting after an exponentially distributed amount of service with parameter $f(n_j)$. If another station starts a transmission before that time, the attempt of transmission is canceled. Consequently, if initially there is no transmission, node j will be the first to access the channel with probability $f(n_j)/(f(n_1)+\cdots+f(n_J))$. See Abramson [1] and Metcalf and Boggs [21] for historical references, and Tassiulas and Ephremides [30]. If a small quantum δ is transmitted at each access, it is not difficult to see that, provided that backoff times are small (which is the case if one of the components is large), as δ goes to 0 the effective capacity allocated to station j is indeed given by

$$\frac{f(n_j)}{f(n_1)+\cdots+f(n_J)}. \tag{4}$$

This is the analogue of the approximation of the round robin policy by the processor-sharing discipline.

Fair Access to Resource: The Choice of the Function $\mathbf{x} \mapsto \mathbf{f(x)}$.
The function f should clearly be increasing, so that the fraction of the capacity allocated may grow with the number of requests. This is the case if $f(x) = x^\alpha$ which corresponds to the Alpha-fair disciplines already mentioned. However, these policies may have a serious drawback. Indeed, if a station j has a large number of requests pending while the other stations are lightly loaded, the latter will receive a negligible fraction of the bandwidth. The station j will therefore capture the channel for its own benefit, until the instant when some of the other stations reach a *comparable level of congestion*. This is a highly undesirable property for a network where fairness issues (for nodes, not requests) are of primary importance. See Bonald and Massoulié [6].

A possible way of solving this problem is to consider increasing functions f which grow slowly to infinity like, for example, the concave function $x \mapsto \log(1+x)$, or $x \mapsto \log\log(e+x)$. In this way, one can expect to reduce significantly the impact of saturated nodes even if they still receive a sizable fraction of the available capacity. Related algorithms have been considered in the context of wireless networks, see Shah and Wischik [29] and references therein. Bouman et al. [7] and Ghaderi et al. [13] investigate the impact of the growth of the function f on the stability and on the delays for several wireless network architectures with a related bandwidth allocation scheme.

In this paper, we mainly investigate the case $f(x) = \log(1+x)$. The general case is sketched in Section 8. The instantaneous fraction of capacity of the jth node is therefore given by

$$\frac{\log(1+n_j)}{\log(1+n_1)+\log(1+n_2)+\cdots+\log(1+n_J)}. \tag{5}$$

Note also that since the limit of x^α/α when α goes to 0 is $\log x$ for $x > 0$, this allocation mechanism can be seen as a limiting case of Alpha-fair disciplines described above.

The log function moderates the rate at which a saturated station tries to access the resource, which is a desirable property in an heterogeneous network where connections may have very different characteristics. In the context of wireless networks, a related algorithm was used to show that an optimal stability region is possible in a quite general network. The growth properties of the log function play an important role in the proof of the result. Basically, the log of the states of the saturated stations being quite stable on some large time intervals, the schedule (the set of stations that can transmit at some time) quickly reaches some equilibrium and stays around it. A Lyapounov function argument can then be used to prove ergodicity (see Shah and Shin [28]). Up to now, apart from these stability results, little is known about the quantitative and qualitative properties of these algorithms. As we shall see below, the mathematical analysis of this class of algorithms presents some challenging and unusual problems (see also Wischik [31]). We first achieve a fluid limit scaling analysis, which gives a very precise description of the qualitative behavior of these algorithms. Additionally, we derive a heavy traffic limit theorem result for the invariant distribution of the associated Markov process. Before presenting our main results, we briefly recall the main definitions of the fluid limit scaling. The interested reader will find an extended presentation in Bramson [9] or in Chapter 8 of Robert [25].

Fluid Limits. Throughout the paper it will be assumed that, for every $1 \le j \le J$, the requests arriving at the jth node form a Poisson process with rate $\lambda_j > 0$. Each request at node j leaves the network when it has received an exponentially distributed amount of service with parameter μ_j from the common resource. The average load of the jth node is denoted by $\rho_j = \lambda_j/\mu_j$.

The fluid limit scaling of a stochastic process $(Z(t))$ in $\mathbb{R}^J$ consists in speeding up time and space in proportion to the norm of its initial state:

$$\overline{Z}_N(t) = \frac{Z(Nt)}{N}, \text{ with } N = \|Z(0)\|.$$

A possible limit in distribution of the sequence of processes $(\overline{Z}_N(t))$ is called a fluid limit of the process $(Z(t))$. Hence, in some sense fluid limits give a first order description of $(Z(t))$. This is a convenient tool to investigate multi-dimensional processes for which general results are scarce. In the context of Markov processes, there is an additional interest since the ergodicity of the process can be connected to the fact that fluid limits (whose initial states lie on the unit sphere) return to the origin. See Rybko and Stolyar [27] and Dai [10]. Note however that the fluid limit scaling is well suited for processes that behave locally like random walks. For other processes, different scalings may have to be considered.

For certain choices of f a fluid limit is obtained by standard techniques. For instance in the case of generalized processor-sharing policy defined by Relation (2), for every $1 \le j \le J$ and $t \ge 0$, let $X_j(t)$ denote the number of jobs waiting at the jth node. The evolution of this process can be represented as

$$X_j(t) = X_j(0) + M_j(t) + \lambda_j t - \mu_j \int_0^t \frac{w_j X_j(s)}{w_1 X_1(s) + w_2 X_2(s) + \cdots + w_J X_J(s)} \, \mathrm{d}s,$$

where $(M_j(t))$ is a martingale. The scaled process with $N = \|X(0)\|$ is thus given by

$$\overline{X}_j^N(t) = \overline{X}_j^N(0) + \frac{M_j(Nt)}{N} + \lambda_j t - \mu_j \int_0^t \frac{w_j \overline{X}_j^N(s)}{w_1 \overline{X}_1^N(s) + w_2 \overline{X}_2^N(s) + \cdots + w_J \overline{X}_J^N(s)} \, \mathrm{d}s. \tag{6}$$

A standard tightness criterion and the fact that the martingale

$$((M_j(Nt)/N), 1 \le j \le J)$$

converges in distribution to 0 imply that any fluid limit $((x_j(t)), 1 \le j \le J)$ should satisfy the ordinary differential equations

$$\frac{\mathrm{d}x_j}{\mathrm{d}t}(t) = \lambda_j - \mu_j \frac{w_j x_j(t)}{w_1 x_1(t) + w_2 x_2(t) + \cdots + w_J x_J(t)}, \quad 1 \le j \le J, \tag{7}$$

on the interval $[0, t_0]$, provided that the vector $(x_j(t), 1 \le j \le J)$ does not hit 0 before t_0. See Ben Tahar and Jean-Marie [3], Ramanan and Reiman [24], and references therein.

Similarly, for alpha-fair disciplines the corresponding fluid model $(x_j(t))$ is the solution to the ODE

$$\frac{\mathrm{d}x_j}{\mathrm{d}t}(t) = \lambda_j - \mu_j \frac{x_j(t)^\alpha}{x_1(t)^\alpha + x_2(t)^\alpha + \cdots + x_J(t)^\alpha}, \quad 1 \le j \le J. \tag{8}$$

Observe that, for these two choices of function f, the scaled process satisfies an autonomous ODE, like (8), with a stochastic noise component that vanishes as N gets large. Secondly, a remarkable feature of these convergence results is that all coordinates of the scaled process are of order N. That is, as long as $x(t)$ is not the vector 0, one has $x_j(t) > 0$ *for all* $1 \le j \le J$. However as we shall now discuss, for instances of interest where f increases slowly (eg. $f(x) = \log x$) such standard techniques are not applicable.

Problem of Fluid Limits for Algorithms with Logarithmic Weights. Let $(L_j(t), 1 \le j \le J)$ denote the Markov process associated to the policy with logarithmic weights, i.e. associated to Relation (5). Due to the log function, the scaled process $(\overline{L}_j(t))$ does not have an autonomous representation analogous to (6). In fact, it is easy to see that the corresponding stochastic equations involve both $L_j(Nt)$ and $\log(1 + L_j(Nt))$, two quantities which evolve on very different scales. For this reason, there is no way of guessing a system of plausible "fluid equations" corresponding to System (7) [resp. to (8)] for discriminatory processor-sharing policy [resp. Alpha-fair policies]. See Wischik [31] and Ghaderi et al. [13]. Note that the question of stability of the system is not an issue here. Indeed, because of the work conserving property of these policies, a necessary and sufficient condition for the ergodicity of $(L_j(t), 1 \le j \le J)$ is simply given by

$$\rho_1 + \cdots + \rho_J < 1, \quad \text{with } \rho_j = \frac{\lambda_j}{\mu_j}, \quad 1 \le j \le J.$$

To the best of our knowledge, there is no explicit expression known for the invariant distribution. The fluid scaling gives a first order description of the behavior of this policy. As we shall see, an interesting convergence result for the invariant distribution just below saturation can also be derived from these results.

Additionally, an interesting phenomenon in this domain is presented. For most of the queueing networks investigated up to now, the classic general scheme for the fluid scaling of the associated Markov process $(X(t))$ is as follows: there is a subset of the coordinates whose values are of the order of $N = \|X(0)\|$ and the other coordinates form an ergodic Markov process whose invariant distribution determines the evolution of the large coordinates on the fluid scale. Here the situation is different. In the case of two nodes, under some appropriate conditions, then the coordinate $(L_2(t))$ is of the order of N but $(L_1(t))$ is an order of magnitude smaller, N^{α^*} for some $0 < \alpha^* < 1$. There is indeed an underlying ergodic Markov process but at the second order, namely an Ornstein-Uhlenbeck process $(Z(t))$ so that $L_1(t) \sim N^{\alpha^*} + \sqrt{N^{\alpha^*} \log N} \cdot Z(t)$. For this queueing system, the queue lengths of underloaded queues are not proportional to load but operate on a scale between $O(1)$ and $O(N)$.

Outline of the Paper. Section 2 presents the main results of the paper. Section 3 introduces the notation and the stochastic differential equations associated to the Markov process $(L_j(t), 1 \leq j \leq J)$. Sections 4, 5 and 6 are respectively devoted to the scaling properties of the time scales $t \mapsto N^t$, $t \mapsto N^{\alpha^*} \log Nt$ and $t \mapsto Nt$. There we provide a precise description of the evolution of the network, together with some estimates of hitting times. The key results on the fluid limits are presented in Section 6. In Section 7, we prove a heavy traffic limit theorem for the invariant distribution. The case of a two node network with a general f is discussed in Section 8. The corresponding time scales are identified in this case. Section 9 gives a brief sketch of the case of a network of J nodes.

Acknowledgments. The first author would like to thank Damon Wischik, whose presentation at the ICMS workshop in Edinburgh in 2010 is one of the motivations at the origin of this work. We thank the reviewer for the detailed work she/he has done on a first version of the paper.

2. Presentation of Results

2.1. **The General Picture.** The main result of the paper is that the states of the nodes of a network with two nodes may live on different space and time scales depending on the set of arrival and service rates. In the next paragraph we give a more precise description of this phenomenon. An illustration of the different scales is given in Figure 1, where the y-axis is on a $\log \cdot / \log N$ scale.

Giving an asymptotic picture of this queueing system with only two nodes is already a challenging problem. To concentrate on the most interesting case, from the point of view of mathematical difficulty, let us assume that the parameters satisfy the conditions: $\rho_1 < 1/2$ and $\rho_2 > 1/2$ and the initial state of the process is $(L_1^N(0), L_2^N(0)) = (0, N)$. The other cases are discussed further in the paper, in Proposition 8 for example.

Three Time Scales.

a) *The time scale $t \to N^t$.*
A convergence result, Proposition 2, shows the convergence in distribution, for any $0 < s_0 < t_0 < \alpha^*$,

$$\lim_{N \to +\infty} \left(L_1^N(N^t), u < t < v\right) = \left(\left(\lambda_1 - \mu_1 \frac{t}{t+1}\right) N^t, s_0 < t < t_0\right)$$

with

$$\alpha^* \stackrel{\text{def.}}{=} \frac{\rho_1}{1-\rho_1}.$$

The process L_2^N stays at N on this time scale. The condition $\rho_1 < 1/2$ implies that $\alpha^* < 1$. Note that the pre-factor of N^t vanishes at $t = \alpha^*$. For this reason, this convergence result does not prove that values of the order of N^{α^*} can be reached. In fact, an extra $(\log N)$ factor is required and Proposition 3 shows that if $\delta < 1$, the average hitting time of the value $\lfloor \delta N^{\alpha^*} \rfloor$ by $(L_1^N(t))$ is bounded above by $K_1 N^{\alpha^*} \log N$ for some $K_1 > 0$. On the other hand, reaching the value $\lfloor N^{\alpha^*} \rfloor$ is slightly longer: the average hitting time of $\lfloor N^{\alpha^*} \rfloor$ is bounded by $C N^{\alpha^*} (\log N)^2 \log\log(N)$ for some $C > 0$, see Relation (29) in Section 5.

b) *The time scale $t \to N^{\alpha^*}(\log N)\, t$.*
We now assume that $L_1^N(0) = \lfloor N^{\alpha^*} \rfloor$ and $L_2(0) = N$. Theorem 1 proves the following convergence in distribution

$$\lim_{N\to+\infty} \left(\frac{L_1^N\big(N^{\alpha^*}(\log N)t\big) - N^{\alpha^*}}{\sqrt{N^{\alpha^*}\log N}}, t \geq 0 \right) = (Z(t)), \tag{9}$$

where $(Z(t))$ is an Ornstein-Uhlenbeck process. In other words, on this time scale L_1^N is stabilized around the value N^{α^*}. Again, the process L_2^N stays at N on this time scale.

c) *The fluid time scale $t \to Nt$.*
Theorem 3 shows the convergence in distribution,

$$\lim_{N\to+\infty} \left(\frac{L_1^N(Nt)}{N^{\alpha^*}}, \frac{L_2^N(Nt)}{N} \right) = \big(\gamma(t)^{\alpha^*}, \gamma(t)\big) \tag{10}$$

for the convergence in distribution of processes, with

$$\gamma(t) = (1 + (\lambda_2 - \mu_2(1-\rho_1))t)^+.$$

Consequently, as long as the fluid limit of $(L_2(t))$ is not 0, the process L_1^N lives on the space scale N^{α^*}.

2.2. **Properties of Resource Sharing with Logarithmic Weights.** The bandwidth allocation with logarithmic weights exhibits some interesting properties. For the two node network described above, when the initial state is $(0, N)$ we prove that the fluid limit is given by

$$\Big((0, 1 + (\lambda_2 - \mu_2(1-\rho_1))t)^+ \Big).$$

This shows that node 2 receives the capacity $1-\rho_1$, which is another way of saying that node 1 is stable at the fluid level. The simplicity of this expression somewhat hides the complexity of the situation, since the quantity α^* does not show up. Yet, the quantity α^* has a crucial impact on the equilibrium distribution and on the transient behavior.

Equilibrium: Heavy-Traffic Regime. Under the stability condition $\rho_1 + \rho_2 < 1$, let $(L_{1,\rho}, L_{2,\rho})$ denote random variables with the equilibrium distribution of the

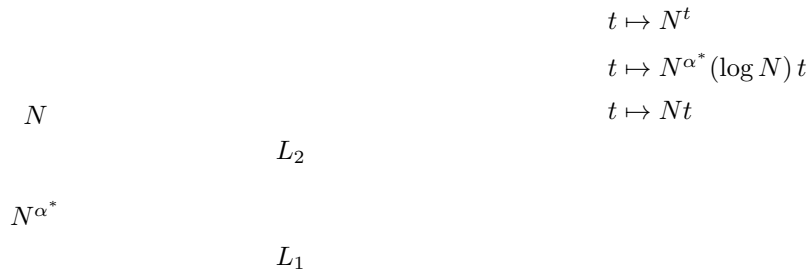

FIGURE 1. A first order picture of the network with $\rho_1 + \rho_2 < 1$, $\rho_1 < 1/2$ and $(L_1(0), L_2(0)) = (0, N)$. The first queue grows according to step a), then the queue behaves as an OU process for a while according to step b), then finally the system converges to zero according to step c).

Markov process $(L_1(t), L_2(t))$. If $\rho_1 < 1/2$ is fixed, the heavy traffic-limit Theorem 5 shows the convergence in distribution

$$\lim_{\rho_2 \nearrow 1-\rho_1} \Big((1-\rho_1-\rho_2)^{\alpha^*} L_{1,\rho}, (1-\rho_1-\rho_2) L_{2,\rho} \Big) = (X^{\alpha^*}, X),$$

where X is an exponential random variable. Hence at equilibrium and in the heavy traffic regime, the relation $L_1 \sim L_2^{\alpha^*}$ also holds as in Relation (10) for fluid limits. Note that X^{α^*} has a Weibull distribution.

Transient Case. If the system is overloaded ($\rho_1 + \rho_2 > 1$) and if $\rho_1 < 1/2$, the size of the queue of class 1 requests grows at rate proportional to t^{α^*}, with $\alpha^* < 1$. This implies that queue 1 is stable at the fluid level, i.e. that $L_1(t)/t$ goes to 0 in distribution as t becomes large. Hence, without any priority mechanism among nodes, if a node has a light load, $\rho_1 < 1/2$, then most of its messages will be transmitted with success even in the case where the system is globally saturated. Recall that in the transient case of the processor-sharing policy or even with the Alpha-fair disciplines, this is not true at all: the states of the nodes diverge to infinity at the same speed, linearly in time.

This is an interesting feature from the point of view of fairness issues among nodes. Indeed, if the node is not too aggressive, $\rho_1 < 1/2$, this result implies that it will be able to transmit most of its traffic *independently* of the load of the other node. It can be shown that an analogous property is valid for the network with J nodes, see Section 9.

It is unlikely that a standard fluid analysis, i.e. deriving directly some equations similar to Relation (7) for example, can be done to investigate the qualitative behavior of more complex networks. This is where the consideration of the various time scales is useful. It gives a tool to explain, via a dynamic picture, the multiple orders of magnitude of the state variables at equilibrium.

2.3. **An Interaction of Time Scales.** There is an unconventional property for the fluid scaling of a queueing system. For most of the queueing networks investigated up to now, the classic general scheme for the fluid scaling of the associated Markov process $(X(t))$ is as follows: there is a subset of the coordinates whose values are of the order of $N = \|X(0)\|$ and the other coordinates form an ergodic Markov process whose invariant distribution determines the evolution of the large coordinates on the fluid scale. See Malyshev [19] and Bramson [9] for some examples.

Here the situation is different. If the initial state is $(0, N)$ and if $\rho_1 < 1/2$, then the large coordinate L_2 is of the order of N but L_1 is an order of magnitude smaller, N^{α^*} with $0 < \alpha^* < 1$. There is indeed an underlying ergodic Markov process but at the second order, namely an Ornstein-Uhlenbeck process scaled by a factor $\sqrt{N^{\alpha^*} \log N}$. See Relation (9).

The associated stochastic model exhibits a *stochastic averaging principle* at the origin of the second expansion in Relation (10). The key technical result of the paper, Theorem 2, states that when $\rho_1 < 1/2$, on *the fluid time scale*, $t \mapsto Nt$, L_1^N is *uniformly* of the order of $(L_2^N)^{\alpha^*}$ on any finite time interval with high probability. Recall that

a) If $L_2^N(0) = N$ and $L_1^N(0) = N^{\alpha^*}$, then on the time scale $t \mapsto N^{\alpha^*}(\log N)t$ we have $L_2^N \sim L_2^N(0)$ and L_1^N is of the order of $(L_2^N(0))^{\alpha^*}$. Additionally, L_1^N can be represented by an Ornstein-Uhlenbeck process around N^{α^*}, see previous point, point b) in Subsection 2.1.
b) On the fluid time scale, $L_2^N(Nt)$ is of the order of $\gamma(t)L_2^N(0)$ with $\gamma(t)$ defined above by Relation (10).

The problem lies in proving that on the fluid time scale L_1^N adapts sufficiently quickly to preserve the relation $L_1^N \sim (L_2^N)^{\alpha^*}$. A central limit result, Proposition 5, suggests that this is not the case on the timescale of the Ornstein-Uhlenbeck process, at least for a second order description. On the other hand, on the fluid time scale Theorem 2 shows that this separation of time scales holds. Its proof uses several estimates related to average hitting times of reflected random walks and some coupling arguments. One of the problems encountered is that the potential natural stochastic fluctuations of the fluid time scale, of the order of $\sqrt{N}$, can be large compared to N^{α^*} (if $\alpha^* < 1/2$ for instance). In particular, standard stochastic calculus cannot be used as such to prove the result. It turns out that the potentially large fluctuations are reduced by the strong ergodicity properties of the underlying Ornstein-Uhlenbeck process. Thus, it does not seem that the classical techniques for proving stochastic averaging results can be used here. See Khasminskii [16], Freidlin and Wentzell [12] and Papanicolau et al. [23] for a general presentation of methods to prove stochastic averaging principles.

3. The Stochastic Model

In this section, we introduce the main stochastic processes and some notation. If h is a non-negative Borelian function on $\mathbb{R}_+$, we let $\mathcal{N}_h$ denote a Poisson process with rate $x \mapsto h(x)$ on $\mathbb{R}_+$. This process can be defined as follows. If $\mathcal{P}$ is a homogeneous Poisson point process on $\mathbb{R}_+^2$ with rate 1 and f is some nonnegative

Borelian function on $\mathbb{R}_+$, then $\mathcal{N}_h(f)$ is defined by

$$\int f(u)\mathcal{N}_{h(u)}(\mathrm{d}u) = \int_{\mathbb{R}_+^2} f(u)\mathcal{P}([0,h(u)]\times \mathrm{d}u).$$

For $\xi \geq 0$, $\mathcal{N}_\xi$ denotes the Poisson process with rate ξ on $\mathbb{R}_+$, i.e. corresponding to the constant function equal to ξ. In addition, for any $0 \leq a \leq b$, $\mathcal{N}_\xi([a,b])$ stands for the number of points of $\mathcal{N}_\xi$ in the interval $[a,b]$. Throughout the paper, the various Poisson processes used will be assumed independent.

We consider two classes of customers. The arrival process of class j customers is a Poisson process with rate λ_j, the distribution of the duration of the required service is exponential with rate μ_j, and ρ_j denotes the ratio λ_j/μ_j. Each class of customers has a dedicated queue and there is a single server working at unit speed. If the state of the system is $(x_1,x_2) \in \mathbb{N}^2$, where x_j is the number of jobs in queue j, then customers of class j receive the fraction of service

$$W_i(x_1,x_2) \stackrel{\text{def.}}{=} \frac{\log(1+x_i)}{\log(1+x_1)+\log(1+x_2)}, \quad i=1,2, \tag{11}$$

from the server, with the convention that $0/0$ is 0. The process of the number of jobs in queue $j \in \{1,2\}$ is denoted by $(L_j(t))$. Since we are only interested in the total number of customers of each class, there is no need to specify the service discipline for each queue. It can be Processor-Sharing or FIFO (First In First Out), for example.

Stochastic Differential Equation. The stochastic process $(L_1(t), L_2(t))$ can be expressed as the solution to the following stochastic differential equation (SDE):

$$\mathrm{d}L_i(t) = \mathcal{N}_{\lambda_i}(\mathrm{d}t) - \mathcal{N}_{\mu_i W_i(L_1(t-),L_2(t-))}(\mathrm{d}t), \quad i=1,2, \tag{12}$$

where $L_i(t-)$ denotes the left limit of L_i at t and W_i is the function defined by Relation (11).

A Saturated System. For $N \in \mathbb{N}$, λ, $\mu > 0$, it will be convenient to introduce a one-dimensional process $(X_N(t))$ describing the evolution of the number of customers in a given queue when the number of jobs in the other queue is "large", i.e., of the order of N. The process $(X_N(t))$ is thus defined as the solution to the SDE

$$\mathrm{d}X_N(t) = \mathcal{N}_\lambda(\mathrm{d}t) - \mathcal{N}_{\mu W_1(x,N-1)}(\mathrm{d}t). \tag{13}$$

From a Markov process point of view, $(X_N(t))$ is simply a *birth and death process* on $\mathbb{N}$ whose Q-matrix $(q(x,y))$ is defined by

$$\begin{cases} q(x,x+1) &= \lambda \\ q(x,x-1) &= \mu\,\dfrac{\log(1+x)}{\log(1+x)+\log N}, \quad x>0. \end{cases}$$

As we shall see, when $(L_1(0),L_2(0)) = (0,N-1)$, $X_N(0)=0$, $\lambda=\lambda_1$, and $\mu=\mu_1$, the two processes $(L_1(t))$ and $(X_N(t))$ are close enough (for our purposes). Note that this is not completely clear since the process $(L_2(t))$ may drift away from N and therefore change the service rate received by each class. It turns out that, because of the slow increase of the log function, this property will hold at least at the beginning of the sample paths.

By integrating the SDE (13) one obtains that, for any $t \geq 0$,

$$X_N(t) = X_N(0) + \mathcal{N}_\lambda([0,t]) - \int_0^t \mathcal{N}_{\mu W_1(X_N(u-),N-1)}(\mathrm{d}u) \tag{14}$$
$$= X_N(0) + \lambda t - \mu \int_0^t \frac{\log(1+X_N(u))}{\log(1+X_N(u)) + \log N}\,\mathrm{d}u + M_N(t),$$

where $(M_N(t))$ is the martingale

$$M_N(t) = \mathcal{N}_\lambda([0,t]) - \lambda t + \int_0^t \left[\mathcal{N}_{\mu W_1(X_N(u-),N-1)}(\mathrm{d}u) - \mu W_1(X_N(u), N-1)\,\mathrm{d}u\right],$$

whose increasing process is given by

$$\langle M_N \rangle (t) = \lambda t + \mu \int_0^t \frac{\log(1+X_N(u))}{\log(1+X_N(u)) + \log N}\,\mathrm{d}u. \tag{15}$$

4. The Initial Phase

This section is devoted to the very beginning of the evolution of the first component $(L_1^N(t))$, when it starts from 0 while $L_2^N(0) = N$. To start with, we have the following asymptotic result on the initial growth rate of the process $(X_N(t))$ defined by Equation (13). Here and later, we write $a \wedge b$ for the quantity $\min(a,b)$.

Proposition 1. *If $X_N(0) = 0$ and*

$$\alpha^* \stackrel{def.}{=} \frac{\rho}{1-\rho}, \text{ where } \rho = \frac{\lambda}{\mu},$$

then, for any $0 < s_0 < t_0 < \alpha^* \wedge 1$, the convergence in distribution of stochastic processes

$$\lim_{N\to+\infty} \left(\frac{X_N(N^t)}{N^t}, s_0 \leq t \leq t_0\right) = \left(\lambda - \mu\frac{t}{t+1},\ s_0 \leq t \leq t_0\right)$$

holds.

See Chapter 2 and 3 of Billingsley [4] on the convergence in distribution of a sequence of processes to a continuous stochastic processes.

Proof. The evolution equation (14) and a change of variables give us that for every $t \geq 0$,

$$\frac{X_N(N^t)}{N^t} = \frac{X_N(1) - \lambda - M_N(1)}{N^t} + \frac{M_N(N^t)}{N^t}$$
$$+ \lambda - \mu \int_0^t \frac{\log(1+X_N(N^u))}{\log(1+X_N(N^u)) + \log N}(\log N) N^{u-t}\,\mathrm{d}u.$$

Letting $Z_N(t) \stackrel{\text{def.}}{=} (1 + X_N(N^t))/N^t$, we thus have

$$Z_N(t) = \frac{X_N(1) + 1 - \lambda - M_N(1)}{N^t} + \frac{M_N(N^t)}{N^t} \tag{16}$$
$$+ \lambda - \mu \int_0^t \frac{\log(Z_N(t-v)) + (t-v)\log N}{\log(Z_N(t-v)) + (t-v+1)\log N}(\log N) N^{-v}\,\mathrm{d}v.$$

Let us first show that the martingale term does not play a role for this scaling. Indeed, for $0 < a < b < 1$, Doob's Inequality yields for every $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{a\le s\le b}\frac{|M_N(N^s)|}{N^s}\ge\varepsilon\right)\le\mathbb{P}\left(\sup_{N^a\le x\le N^b}\frac{|M_N(x)|}{N^a}\ge\varepsilon\right)$$
$$\le\frac{1}{\varepsilon^2N^{2a}}\mathbb{E}(|M_N(N^b)|^2)=\frac{1}{\varepsilon^2N^{2a}}\mathbb{E}(\langle M_N\rangle(N^b))\le\frac{(\lambda+\mu)}{\varepsilon^2}N^{b-2a}.$$

The last term can be made arbitrarily small when N is large by choosing $b < 2a$. Since any interval $[s_0, t_0] \subset (0, \alpha^*)$ can be covered by a finite number of such intervals, the martingale term is indeed negligible with probability tending to 1 as N goes to infinity. The relation $X_N(1) \le \mathcal{N}_\lambda([0, 1])$ implies that the first term in the r.h.s. of (16) vanishes too when divided by N^t.

Next, the inequality $X_N(s) \le \mathcal{N}_\lambda([0, s])$ gives us that

$$\int_0^t\frac{\log(Z_N(t-v))+(t-v)\log N}{\log(Z_N(t-v))+(t-v+1)\log N}(\log N)N^{-v}\,\mathrm{d}v$$
$$\le Y_N(t)\overset{\text{def.}}{=}\int_0^t\frac{\log((1+\mathcal{N}_\lambda([0,N^{t-v}]))/N^{t-v})+(t-v)\log N}{\log((1+\mathcal{N}_\lambda([0,N^{t-v}]))/N^{t-v})+(t-v+1)\log N}(\log N)N^{-v}\,\mathrm{d}v.$$

For $0 \le v \le t \le b$,

$$\left|\frac{\log((1+\mathcal{N}_\lambda([0,N^v]))/N^v)+v\log N}{\log((1+\mathcal{N}_\lambda([0,N^v]))/N^v)+(v+1)\log N}-\frac{v}{1+v}\right|$$
$$\le\frac{\log(1+S_N(b))}{\log N}\frac{1}{\log(1+S_N(b))/\log(N)+1},$$

with

$$S_N(b)=\sup_{0\le v\le b}\frac{\mathcal{N}_\lambda([0,N^v])}{N^v}.$$

By the law of large numbers for Poisson processes, for any $0 < \varepsilon < b$, the process $(\mathcal{N}_\lambda([0, N^s])/N^s,\ \varepsilon \le s \le b)$ converges in distribution to $(\lambda,\ \varepsilon \le s \le b)$. The sequence of random variables $(S_N(b))$ is therefore tight. One gets that

$$\sup_{a\le t\le b}\left|Y_N(t)-\int_0^t\frac{t-v}{1+t-v}(\log N)N^{-v}\,\mathrm{d}v\right|$$

converges to 0 in distribution. It is not difficult to check that the convergence

$$\lim_{N\to+\infty}\int_0^t\frac{t-v}{1+t-v}(\log N)N^{-v}\,\mathrm{d}v=\frac{t}{t+1},\tag{17}$$

occurs uniformly for $a \le t \le b$, one gets therefore the convergence in distribution

$$\lim_{N\to+\infty}(Y_N(t),\ a\le t\le b)=\left(\frac{t}{t+1}, a\le s\le b\right).$$

Gathering these estimates, we obtain that for any $0 < a < b < \alpha^* \wedge 1$ and any $\varepsilon > 0$,

$$\lim_{N\to+\infty}\mathbb{P}\left(\lambda-\mu\frac{b}{b+1}-\varepsilon\le\inf_{a\le s\le b}Z_N(s)\le\sup_{a\le s\le b}Z_N(s)\le\lambda+\varepsilon\right)=1,$$

note that $\lambda - \mu b/(b + 1) > 0$ for $b < \alpha^*$. Therefore, for any $0 < a_0 < a \le b \le b_0 < \alpha^*$, on some event $\mathcal{E}$ with an arbitrarily high probability, the process

$(|\log(Z_N(s))|, a_0 \le s \le b_0)$ can be bounded by some constant. Consequently, on this event, with a similar uniform convergence argument for (17), one gets that the sequence of processes

$$\left(\int_0^{t-a_0} \frac{\log(Z_N(t-v)) + (t-v)\log N}{\log(Z_N(t-v)) + (t-v+1)\log N}(\log N)N^{-v}\,\mathrm{d}v, a_0 \le t \le b_0\right)$$

converges in distribution to $(t/(t+1), a_0 \le s \le b_0)$. The remaining term contributing in the integral of the r.h.s. of Relation (16) is

$$\left\|\int_{t-a_0}^{t} \frac{\log(Z_N(t-v)) + (t-v)\log N}{\log(Z_N(t-v)) + (t-v+1)\log N}(\log N)N^{-v}\,\mathrm{d}v\right| \le \int_{t-a_0}^{t} (\log N)N^{-v}\,\mathrm{d}v$$

and thus converges to (0) as a process on $[a, b]$. The desired convergence in distribution has thus been proved.

□

The following proposition shows that, if $(L_1^N(0), L_2^N(0))$=$(0, N)$ and $X^N(0)$=0, then the two processes $(L_1^N(t))$ and $(X^N(t))$ are close on the time scale $t \mapsto N^t$, $0 < t < \alpha^*$. As a consequence, it implies that the convergence result of Proposition 1 is also valid for the process $(L_1^N(t))$.

Proposition 2. *If $(L_1^N(t), L_2^N(t))$ is the solution to the SDE (12) with initial condition $(0, N)$ and α^*=$\rho_1/(1-\rho_1)$ with $\rho_1 = \lambda_1/\mu_1$, then the convergence*

$$\lim_{N\to+\infty}\left(\frac{L_1^N(N^t)}{N^t},\ 0 < t < \alpha^* \wedge 1\right) = \left(\lambda_1 - \mu_1\frac{t}{t+1},\ 0 < t < \alpha^* \wedge 1\right)$$

holds for the uniform topology on compact sets of $(0, \alpha^* \wedge 1)$.

Proof. The idea of the proof is quite simple. First one shows that, on the time scale $t \to N^t$ with $t < 1$, the second component do not change much so that its log is equivalent to $\log N$. On this time scale the first coordinate grows linearly as long as $t < \alpha^*$ and for this reason its log is equivalent to $t \log N$, so the capacity it receives is $t/(t+1)$ which explains the result.

Since $L_2^N(0) = N$ and the number of jobs of class 2 decreases at rate at most μ_1 and increases at rate λ_1, for any $0 < b < 1$, there exist two constants $0 < \eta < \gamma$ such that if

$$\mathcal{A}_N \stackrel{\text{def.}}{=} \left\{\eta N \le \inf_{0\le s\le N^b} L_2^N(s) \le \sup_{0\le s\le N^b} L_2^N(s) \le \gamma N\right\},$$

then $\mathbb{P}(\mathcal{A}_N)$ tends to 1 as N tends to infinity. On the set $\mathcal{A}_N$, the jump rate for departures of L_1^N lies between $\mu_1 W_1^\eta(\cdot, N)$ and $\mu_1 W_1^\gamma(\cdot, N)$, where

$$W_1^\delta(x, N) = \frac{\log(1+x)}{\log(1+x) + \log(\delta N)}.$$

Now if $(X_N^\delta(t))$ denotes the solution to Equation (13) with W_1 replaced by W_1^δ, a straightforward coupling shows that on the set $\mathcal{A}_N$ the relation

$$X_N^\eta(s) \le L_1^N(s) \le X_N^\gamma(s), \quad 0 \le s < N^b$$

holds almost surely. A glance at the proof of the convergence result of Proposition 1 shows that this result also holds for both processes $(X_N^\eta(s))$ and $(X_N^\gamma(s))$, and so the proposition is proved. □

The above proposition shows that if $\alpha^* < 1$ (i.e. $\rho_1 < 1/2$), then on the time scale $t \mapsto N^t$, $0 < t < \alpha^*$, we have

$$L_1^N(N^t) \sim \left(\lambda_1 - \mu_1 \frac{t}{t+1}\right) N^t.$$

In particular, the process L_1^N reaches the quantity $N^{\alpha^* - \varepsilon}$ for any $0 < \varepsilon < \alpha^*$. Note that, when $t \nearrow \alpha^*$, the quantity multiplying N^t vanishes, so that this convergence result does not show that the value N^{α^*} is indeed reached. In Sections 5 and 6, we shall prove that the process L_1^N lives in fact in a "small" neighborhood of N^{α^*}. This local equilibrium around N^{α^*} is the key phenomenon to grasp in order to understand this bandwidth sharing policy. For now, we conclude this section by proving that for any $0 < \delta < 1$, the value $\delta N^{\alpha^* \wedge 1}$ is reached. This is done by providing an estimation of the corresponding hitting time.

Proposition 3. *With the same notation and assumptions as in Proposition 2, and if H_a denotes the hitting time of $a > 0$ by L_1^N,*

$$H_a = \inf\{t > 0 : \ L_1^N(t) \geq a\},$$

then

a) *if $\alpha^* < 1$, for any $0 < \delta < 1$ there exists a constant $C > 0$ such that for any $N \geq 1$,*

$$\mathbb{E}\left(H_{\delta N^{\alpha^*}}\right) \leq \frac{C\delta}{\log(1/\delta)} N^{\alpha^*} \log N,$$

b) *if $\alpha^* > 1$, for $\delta > 0$ sufficiently small we have*

$$\limsup_{N \to +\infty} \frac{1}{N} \mathbb{E}\left(H_{\delta N}\right) \leq \frac{\delta}{\lambda_1 - \mu_1/2}.$$

Proof. As we did for Proposition 2, let us prove these two inequalities for the process $(X_N(t))$. We use the simplified notation of Proposition 1.

Let us first assume that $\alpha^* < 1$. For $x > 0$, the elementary relation

$$\frac{\log(1+x)}{\log(1+x) + \log N} - \frac{\alpha^*}{\alpha^* + 1} = \frac{\log((1+x)/N^{\alpha^*})}{(\alpha^*+1)(\log N)(1 + \alpha^* + \log((1+x)/N^{\alpha^*})/(\log N))}, \tag{18}$$

together with the identity

$$\lambda = \mu \frac{\alpha^*}{1 + \alpha^*}$$

and Equation (14) written at time $N^{\alpha^*}(\log N)\, t$ give the representation

$$Z_N(t) \stackrel{\text{def.}}{=} \frac{X_N(N^{\alpha^*}(\log N)t)}{N^{\alpha^*}} = \frac{M_N(N^{\alpha^*}(\log N)t)}{N^{\alpha^*}} - \frac{\mu}{(1+\alpha^*)} \int_0^t \frac{\log(N^{-\alpha^*} + Z_N(u))}{\alpha^* + 1 + \log(N^{-\alpha^*} + Z_N(u))/(\log N)} \, \mathrm{d}u. \tag{19}$$

Let

$$\tau_N = \inf\left\{s > 0 : X_N\big(N^{\alpha^*}(\log N)s\big) \geq \delta N^{\alpha^*}\right\}.$$

From Doob's optional stopping time theorem and the fact that $Z_N(\tau_N \wedge t) \leq N^{-\alpha^*}\lceil \delta N^{\alpha^*}\rceil$, we obtain the inequality

$$\frac{\lceil \delta N^{\alpha^*}\rceil}{N^{\alpha^*}} + \frac{\mu}{(1+\alpha^*)}\frac{\log(\delta + 2N^{-\alpha^*})}{\alpha^* + 1 + \log(\delta + 2N^{-\alpha^*})/(\log N)}\,\mathbb{E}(\tau_N) \geq 0.$$

Since $H_{\delta N^{\alpha^*}} \leq N^{\alpha^*}(\log N)\tau_N$, item a) is proved.

Assume now that $\alpha^* > 1$, i.e. that $\rho > 1/2$. Equation (14) written at time $N\,t$ gives the relation

$$X_N(Nt) = M_N(Nt) + \lambda N t - \mu \int_0^{Nt} \frac{\log(1 + X_N(s))}{\log(1 + X_N(s)) + \log N}\,\mathrm{d}s,$$

from which we can write that

$$\lceil \delta N\rceil \geq \mathbb{E}(X_N(H_{\delta N} \wedge (Nt))) \geq \mathbb{E}(H_{\delta N} \wedge (Nt))\left(\lambda - \mu\frac{\log(1 + \lceil \delta N\rceil)}{\log(1 + \lceil \delta N\rceil)) + \log N}\right).$$

Letting t go to infinity, the monotone convergence theorem gives us that

$$\limsup_{N\to+\infty} \frac{1}{N}\,\mathbb{E}(H_{\delta N}) \leq \frac{\delta}{\lambda - \mu/2}. \tag{20}$$

If we now choose δ sufficiently small so that, with high probability, the component $L_2^N(t)$ is still of the order of N at time $\delta N/(\lambda - \mu/2)$, a straightforward coupling between L_1^N and some X_N^η (recall the notation X_N^η from the proof of Proposition 2) ensures that (20) holds as well for the process $(L_1^N(t))$. This completes the proof of Proposition 3. □

5. A Local Equilibrium

This section is essentially devoted to the behavior of our two-dimensional process on the time scale $t \mapsto N^{\alpha^*}(\log N)t$, when the initial state is $(L_1^N(0), L_2^N(0)) = (\delta N^{\alpha^*}, N)$ and $\rho_1 < 1/2$. The following result, Proposition 4, shows that on this time scale the sample paths of $(L_1^N(t))$ have values of the order of xN^{α^*}, where $0 < x < 1$. When the initial value of $(L_1^N(t))$ is N^{α^*}, we shall prove in Theorem 1 that the process is stabilized around N^{α^*}. At first sight, the interest of Proposition 4 and of the associated central limit theorem (Proposition 5 below) may seem marginal. This is not true at all since, as we shall see in Section 6, the process is also of the order of $\gamma(t)N^{\alpha^*}$ on the fluid time scale $t \mapsto Nt$. Furthermore, the main difficulty in the key technical result of Theorem 2 is precisely connected to the interaction of these two time scales $t \mapsto N^{\alpha^*}(\log N)t$ and $t \mapsto Nt$.

Recall the notation $\alpha^* = \rho_1/(1 - \rho_1)$.

Proposition 4. *If $\rho_1 < 1/2$ and $(L_1^N(t), L_2^N(t))$ is the solution to the SDE (12) with initial conditions $L_2^N(0) = N$ and $L_1^N(0) \sim \delta N^{\alpha^*}$ for some $\delta \in (0, 1]$, then the sequence of stochastic processes*

$$\left(\frac{L_1^N(N^{\alpha^*}(\log N)t)}{N^{\alpha^*}}\right)$$

converges in distribution to $(h(t))$ defined by

$$\begin{cases} h \equiv 1 & \text{if } \delta = 1, \\ \displaystyle\int_\delta^{h(t)} \frac{1}{\log(u)}\,\mathrm{d}u = -\frac{\mu t}{(1+\alpha^*)^2} & \text{if } \delta \neq 1. \end{cases} \tag{21}$$

Proof. As we did for Proposition 2, the convergence is proved for the process $(X_N(t))$. The result for $(L_1^N(t))$ follows from a similar coupling argument.

Let us first show that if $X_N(0) = \lfloor \delta N^{\alpha^*} \rfloor$, then $X_N(t)/N^{\alpha^*}$ remains within $[\delta/3, 3]$ on a given time interval $[0, N^{\alpha^*}(\log N)T]$ with probability tending to 1. Indeed, writing H_a for the hitting time of a by $X_N(N^{\alpha^*}(\log N)\cdot)$, the strong Markov property of $(X_N(t))$ gives us that

(22)

$$\begin{aligned}\mathbb{P}\big(H_{3N^{\alpha^*}} \le T\big) &= \mathbb{E}\big(\mathbf{1}_{\{H_{2N^{\alpha^*}} < T\}} \mathbb{P}_{\lceil 2N^{\alpha^*} \rceil}\big(H_{3N^{\alpha^*}} \le T - H_{2N^{\alpha^*}} \mid H_{2N^{\alpha^*}} \le T\big)\big) \\ &\le \mathbb{P}\big(H_{2N^{\alpha^*}} < T\big)\, \mathbb{P}_{\lceil 2N^{\alpha^*} \rceil}\big(H_{3N^{\alpha^*}} \le T\big)\big).\end{aligned}$$

Now, due to the monotonicity properties of the service rate of $(X_N(t))$, if $X_N(0) = 2N^{\alpha^*}$ then we can couple $(X_N(t))$ with the process $(2N^{\alpha^*} + R_N(t))$ defined by

— $R_N \to R_N + 1$ at rate λ,
— $R_N \to R_N - 1$ at rate

$$\mu \frac{\log(2N^{\alpha^*})}{\log(2N^{\alpha^*}) + \log N} = \mu \frac{\alpha^*}{\alpha^* + 1} + \frac{C}{\log N},$$

for some $C > 0$,
— $R_N(0) = 0$ and $(R_N(t))$ reflects at 0,

in such a way that $X_N(t) \le 2N^{\alpha^*} + R_N(t)$ for every $t \ge 0$. Hence, there remains to prove that $(R_N(t))$ does not reach N^{α^*} in less than $N^{\alpha^*}(\log N)T$ units of time. But by Kingman's inequality (see Kingman [17] or Relation (3.3) of Theorem 3.5 in Robert [25]) and the fact that $\lambda = \mu\alpha^*/(1+\alpha^*)$, if θ_N stands for the time span of the first excursion of $(R_N(t))$ away from 0 we have

$$\mathbb{P}_1\Big(\sup_{s \in [0, \theta_N]} R_N(s) > N^{\alpha^*}\Big) \le \exp\Big(-\frac{C' N^{\alpha^*}}{\log N}\Big)$$

for some explicit $C' > 0$. Since the ith pair of consecutive excursions is separated by the amount E_i, an exponentially distributed random variable with parameter λ, and, since these exponential times are independent of each other, a Chernoff bound on Poisson random variables gives the relation

$$\begin{aligned}&\mathbb{P}(\text{ more than } 2\lambda N^{\alpha^*}(\log N)T \text{ excursions before time } N^{\alpha^*}(\log N)T\ ) \\ &\le \mathbb{P}\Big(\sum_{i=1}^{2\lambda N^{\alpha^*}(\log N)T} E_i \le N^{\alpha^*}(\log N)T\Big) \le e^{-C'' N^{\alpha^*}(\log N)T}\end{aligned}$$

for some $C'' > 0$. Coming back to (22), we obtain that

$$\mathbb{P}\big(H_{3N^{\alpha^*}} \le T\big) \le e^{-C'' N^{\alpha^*}(\log N)T} + 2\lambda N^{\alpha^*}(\log N)T \exp\Big\{-\frac{C' N^{\alpha^*}}{\log N}\Big\},$$

which tends to 0 as N tends to infinity. Finally, since the infinitesimal drift of $X_N(t)$ is positive when $X_N(t) < N^{\alpha^*}$, the same method can be used to show that $(X_N(t))$ remains above $(\delta/3)N^{\alpha^*}$ on the time interval $[0, N^{\alpha^*}(\log N)T]$, with overwhelming probability.

Now let $(w_f(\xi))$ denote the modulus of continuity of the function $(f(t))$ on $[0,T]$. That is,

$$w_f(\xi) = \sup_{\substack{s,t\le T\\ |s-t|\le\xi}} |f(s)-f(t)|\,.$$

Let us also write again

$$Z_N(t) \stackrel{\text{def.}}{=} \frac{X_N(N^{\alpha^*}(\log N)t)}{N^{\alpha^*}}.$$

Using the bounds on $(X_N(t))$ we have just obtained, we can deduce the existence of a constant A independent of N such that, with probability tending to 1, we have, for any $s<t\in[0,T]$,

$$\int_s^t \left|\frac{\log(N^{-\alpha^*}+Z_N(u))}{\alpha^*+1+\log(N^{-\alpha^*}+Z_N(u))/(\log N)}\right|\mathrm{d}u \le A(t-s).$$

Together with Relation (19) and the fact that its martingale term vanishes as N gets large, we can conclude that for any $\varepsilon>0$ and $\eta>0$, there exists $\xi>0$ such that

$$\mathbb{P}(w_{Z_N}(\xi)>\eta)\le\varepsilon$$

holds for all N. By the tightness criterion of the modulus of continuity (see Theorem 8.3 in Billingsley [4]), the sequence of processes $(Z_N(t))$ is thus tight and any limiting point h satisfies the relation

$$h(t)=\delta-\frac{\mu}{(1+\alpha^*)^2}\int_0^t \log(h(u))\,\mathrm{d}u.$$

It is easily seen that there is a unique solution to this integral equation and that its solution can be expressed as the solution to the fixed point equation (21). Proposition 4 is proved. □

The above proposition can be seen as a kind of law of large numbers, with

$$L_1^N(N^{\alpha^*}\log(N)t)\sim h(t)N^{\alpha^*}, \qquad \text{as } N\to\infty.$$

It is thus natural to expect for the corresponding central limit theorem that

$$\left(\frac{L_1^N(N^{\alpha^*}\log(N)t)-h(t)N^{\alpha^*}}{\sqrt{N^{\alpha^*}\log(N)}}\right)$$

should converge in distribution to some Gaussian process $(R(t))$. However, the following proposition shows that such a convergence cannot hold: the centering term $h(t)$ has to be replaced by a deterministic term $h_N(t)$ which depends on N. As we shall see, h_N is such that the following expansion holds

$$h_N(t)=h(t)-\frac{\mu_1}{(\alpha^*+1)^3\log N}\int_0^t \log(h(u))^2\,\mathrm{d}u+o\left(\frac{1}{\log N}\right).$$

In particular, $(h_N(t))$ converges "slowly" to $(h(t))$ at the rate $1/\log N$ instead of a rate $1/N^{\alpha^*/2+\varepsilon}$ for which a "classic" centering procedure would be enough. In the end, this second limit theorem gives the representation

$$L_1(N^{\alpha^*}\log(N)t)=h_N(t)N^{\alpha^*}+O\left(\sqrt{N^{\alpha^*}\log N}\right).$$

Proposition 5 (Central Limit Theorem)**.** *If $\rho_1 < 1/2$ and $(L_1^N(t), L_2^N(t))$ is the solution to the SDE* (12) *with the initial conditions $L_2^N(0) = N$ and $L_1^N(0)$ being such that*

$$\lim_{N\to+\infty} \frac{L_1^N(0) - \delta N^{\alpha^*}}{\sqrt{N^{\alpha^*}\log N}} = y,$$

for some $0 < \delta \le 1$ and $y \in \mathbb{R}$. Then, we have

$$\lim_{N\to+\infty} \left(\frac{L_1^N(N^{\alpha^*}(\log N)t) - h_N(t)N^{\alpha^*}}{\sqrt{N^{\alpha^*}\log N}} \right) = (R(t)),$$

for the convergence in distribution, where $(h_N(t))$ is the solution to the ordinary differential equation

$$\dot{h}_N(t) = -\frac{\mu_1}{(1+\alpha^*)} \frac{\log h_N(t)}{(\alpha^* + 1 + \log(h_N(t))/\log N)},$$

with $h_N(0) = \delta$ and $(R(t))$ is the solution to the following SDE

$$\mathrm{d}R(t) = \sqrt{2\lambda_1}\ \mathrm{d}B(t) - \frac{\mu_1}{(1+\alpha^*)^2} \frac{R(t)}{h(t)}\,\mathrm{d}t, \tag{23}$$

with $R(0) = y$, $(B(t))$ denoting standard Brownian motion on $\mathbb{R}$ and $(h(t))$ being defined in (21)*.*

Proof. From Relations (15) and (18), the increasing process of the martingale

$$\left(\overline{M}_N(t)\right) \stackrel{\text{def.}}{=} \left(\frac{M_N(N^{\alpha^*}(\log N)t)}{\sqrt{N^{\alpha^*}\log N}} \right)$$

is given by

$$\left(2\lambda t + \frac{\mu}{(1+\alpha^*)\log N} \int_0^t \frac{\log(N^{-\alpha^*} + Z_N(u))}{\alpha^* + 1 + \log(N^{-\alpha^*} + Z_N(u))/(\log N)}\,\mathrm{d}u \right).$$

By Proposition 4, this quantity converges in distribution to $(2\lambda t)$. Hence, using the martingale criterion for convergence to a Brownian motion (see Theorem 1.4 page 339 of Ethier and Kurtz [11]) we can conclude that $\left(\overline{M}_N(t)\right)$ converges in distribution to Brownian motion run at speed 2λ.

With the same notation as in the proof of the previous proposition, the relation satisfied by $(Z_N(t))$ yields

$$\begin{aligned} \overline{Z}_N(t) &\stackrel{\text{def.}}{=} \sqrt{\frac{N^{\alpha^*}}{\log N}} \left(Z_N(t) - h_N(t) \right) = \overline{Z}_N(0) + \overline{M}_N(t) \\ &\quad - \frac{\mu}{(1+\alpha^*)^2} \int_0^t \sqrt{\frac{N^{\alpha^*}}{\log N}} \left(d_N\left(N^{-\alpha^*} + Z_N(u)\right) - d_N(h_N(u)) \right) \mathrm{d}u, \end{aligned} \tag{24}$$

where

$$d_N(x) = \frac{\log x}{1 + (\log x)/((\alpha^*+1)\log N)}.$$

Let $T > 0$ and $\varepsilon > 0$, and let A_N be the event

$$A_N = \left\{ \inf_{0\le s\le T} Z_N(s) \ge \frac{\delta}{2} \right\}.$$

By Proposition 4 and the fact that $(h(t))$ is non-decreasing with $h(0) = \delta$, there exists N_0 such that for any $N \ge N_0$, $\mathbb{P}(A_N^c) \le \varepsilon$.

The properties of the derivate d'_N of d_N are used to analyze the limit of $(\overline{Z}_N(t))$. Observe that

$$d'_N(x) = \left\{ x\left(1 + \frac{\log x}{(\alpha^* + 1)\log N}\right)\right\}^{-1}. \tag{25}$$

On the event A_N, we obtain from Relation (24) and the fact that for N large enough,

$$\sup_{[\delta/2,\infty)} d'_N \leq \frac{4}{\delta}$$

that for any $t \leq T$,

$$|\overline{Z}_N(t)| \leq |\overline{Z}_N(0)| + |\overline{M}_N(t)| + \frac{4\mu}{\delta(1+\alpha^*)^2}\int_0^t |\overline{Z}_N(u)|\,\mathrm{d}u.$$

As a consequence, Gronwall's lemma gives us that on the set A_N,

$$\sup_{0\leq t\leq T} \left|\overline{Z}_N(t)\right| \leq \left(\left|\overline{Z}_N(0)\right| + \sup_{0\leq t\leq T}\left|\overline{M}_N(t)\right|\right) e^{CT},$$

where $C = 4\mu/(2(1+\alpha^*)^2)$. Using the fact that the sequence of processes $(M_N(t))$ is tight for the topology of uniform convergence, together with our assumption on $X_N(0)$, we obtain that there exists N_1 and $K > 0$ such that for any $N \geq N_1$

$$\mathbb{P}(B^c_{N,K}) \leq \varepsilon, \quad \text{where} \quad B_{N,K} = \left\{\inf_{0\leq s\leq T} Z_N(s) \geq \frac{\delta}{2}, \sup_{0\leq t\leq T}\left|\overline{Z}_N(t)\right| \leq K\right\}.$$

Next, recall the notation w_f for the modulus of continuity of the function f. On the event $B_{N,K}$, Relation (24) gives us that for any $\xi > 0$,

$$w_{\overline{Z}_N}(\xi) \leq w_{\overline{M}_N}(\xi) + C \sup_{\substack{s,t\leq T\\ |s-t|\leq\xi}} \int_s^t |\overline{Z}_N(u)|\,\mathrm{d}u \leq w_{\overline{M}_N}(\xi) + CK\xi.$$

The sequence of processes $(\overline{M}_N(t))$ being tight, this relation and the fact that $\mathbb{P}(B_{N,K}) > 1-\varepsilon$ show that for any $\eta > 0$, there exists ξ_0 such that for every $\xi \leq \xi_0$ and $N \geq N_0$,

$$\mathbb{P}(w_{Z_N}(\xi) \geq \eta) \leq 3\varepsilon.$$

Since we can apply this reasoning to any $\varepsilon > 0$, here again the tightness criterion of the modulus of continuity enables us to conclude that the sequence of processes $(\overline{Z}_N(t))$ is tight.

Let $(R(t))$ be one of the limiting points. By using Skorohod's Representation Theorem, see Ethier and Kurtz [11], up to a change of probability space, we can assume that the convergence to $(R(t))$ holds almost surely on $[0,T]$ for the uniform norm. But on the set $B_{N,K}$, (25) and Lebesgue's differentiation theorem, see Rudin [26] for example, guarantees that the integral term on the right hand side of Equation (24) converges almost surely to

$$\int_0^t \frac{R(u)}{h(u)}\,\mathrm{d}u.$$

Consequently $(R(t))$ satisfies the SDE (23) with $R(0) = y$, and the uniqueness of such a solution gives us the convergence in distribution we were seeking. □

A direct consequence of this result is that, starting from N^{α^*}, the process $(L^N_1(t))$ behaves like an Ornstein-Uhlenbeck process around N^{α^*}.

Theorem 1 (A Stable Regime for $(L_1^N(t))$)**.** *If $\rho_1 < 1/2$, $L_2^N(0) = N$ and $L_1^N(0)$ is such that, for some $y \in \mathbb{R}$,*

$$\lim_{N\to+\infty} \frac{L_1^N(0) - N^{\alpha^*}}{\sqrt{N^{\alpha^*}\log N}} = y,$$

then the sequence of processes

$$\left(\frac{L_1^N\big(N^{\alpha^*}(\log N)t\big) - N^{\alpha^*}}{\sqrt{N^{\alpha^*}\log N}}\right)$$

converges in distribution to an Ornstein-Uhlenbeck process $(Z(t))$, i.e. the solution to the SDE

$$\mathrm{d}Z(t) = \sqrt{2\lambda_1}\,\mathrm{d}B(t) - \frac{\mu_1}{(\alpha^*+1)^2}\,Z(t)\mathrm{d}t, \qquad Z(0) = y, \tag{26}$$

where $(B(t))$ denotes standard Brownian motion.

To complete this section on the time scale $t \to N^{\alpha^*}\log N$, the following proposition shows that, if $\rho_1 < 1/2$ and $L_2^N(0) = N$, then the process $(L_1^N(t))$ always reaches the stable regime around N^{α^*} described in the previous theorem. In particular, if $L_1^N(0) = 0$, the hitting time of $\lfloor N^{\alpha^*}\rfloor$ is smaller than N^β for any $\beta > \alpha^*$.

Proposition 6. *Let*

$$T_N \overset{def.}{=} \inf\left\{s{>}0\colon L_1^N(s) \in N^{\alpha^*} + \left[-\sqrt{N^{\alpha^*}\log N}, \sqrt{N^{\alpha^*}\log N}\right]\right\}.$$

If $\rho_1 < 1/2$, $L_2^N(0) = N$ and $L_1^N(0) \le N^\beta$ for some $\beta \in (\alpha^*, 1)$, then

$$\lim_{N\to+\infty} \mathbb{P}\left(\frac{T_N}{N^\beta(\log N)^2} \le 1\right) = 1.$$

Proof. As before, the result is proved for the process $(X_N(t))$ instead of $(L_1^N(t))$. Suppose that $X_N(0) = \lfloor N^\beta\rfloor$. The SDE (14) and Relation (18) show that for any stopping time τ, one has

$$\begin{aligned}\mathbb{E}(X_N(t\wedge\tau)) &= X_N(0)\\ &\quad - \frac{\mu}{(\alpha^*+1)\log N}\,\mathbb{E}\left(\int_0^{t\wedge\tau} \frac{\log((1+X_N(u))/N^{\alpha^*})}{1+\alpha^*+\log((1+X_N(u))/N^{\alpha^*})/\log N}\,\mathrm{d}u\right).\end{aligned} \tag{27}$$

Defining again

$$H_x \overset{\text{def.}}{=} \inf\{s > 0 : L_1(s) = \lfloor x\rfloor\},$$

and setting $x_0^N = 2\lceil N^{\alpha^*}\rceil$, the above relation gives

$$0 \le X_N(0) - \frac{\mu}{(\alpha^*+1)\log N}\,\frac{\log 2}{1+\alpha^*+(\log 2)/\log N}\,\mathbb{E}\big(t\wedge H_{x_0^N}\big).$$

Consequently, letting t and then N go to infinity yields

$$\limsup_{N\to+\infty} \frac{\mathbb{E}(H_{x_0^N})}{N^\beta\log N} < +\infty.$$

We can therefore assume that $X_N(0) = 2\lceil N^{\alpha^*}\rceil$. Setting this time $x_1^N = \lfloor N^{\alpha^*}\rfloor + \lfloor N^{\alpha^*-\varepsilon}\rfloor$, where $\varepsilon > 0$ is such that $\alpha^* + \varepsilon < \beta$, the same argument gives us that

$$x_1^N \le X_N(0) - \frac{\mu}{(\alpha^*+1)\log N}\,\frac{\log(1+1/N^\varepsilon)}{1+\alpha^*+\log(1+1/N^\varepsilon)/\log N}\,\mathbb{E}\big(H_{x_1^N}\big),$$

and thus

$$\limsup_{N\to+\infty} \frac{\mathbb{E}(H_{x_1^N})}{N^{\alpha^*+\varepsilon}\log N} < +\infty. \tag{28}$$

Similarly, if $x_2^N = \lfloor N^{\alpha^*}\rfloor + \lfloor N^{\alpha^*-2\varepsilon}\rfloor$ and if we choose $X_N(0) = x_1^N$, the above equation gives for $\tau = H_{x_2^N}$

$$\lfloor N^{\alpha^*-2\varepsilon}\rfloor \le \lfloor N^{\alpha^*-\varepsilon}\rfloor - \frac{\mu}{(\alpha^*+1)\log N}\frac{\log(1+1/N^{2\varepsilon})}{1+\alpha^*+\log(1+1/N^{2\varepsilon})/\log N}\,\mathbb{E}\big(H_{x_2^N}\big),$$

so that Relation (28) also holds for $H_{x_2^N}$. Setting $x_i^N = N^{\alpha^*} + N^{\alpha^*-i\varepsilon}$, we can proceed by induction until the smallest integer i^* such that $i^*\varepsilon > \alpha^*/2$. Finally, we obtain that as $N\to\infty$,

$$\frac{\mathbb{E}(T_N)}{N^\beta(\log N)^2} = \frac{\mathbb{E}(H_{x_0^N})}{N^\beta(\log N)^2} + \sum_{i=1}^{i^*}\frac{\mathbb{E}(H_{x_i^N} - H_{x_{i-1}^N})}{N^\beta(\log N)^2} \to 0.$$

We conclude by using the Markov inequality.

Up to now we have been dealing with the case $X_N(0) > N^{\alpha^*}$. There remains to consider the case $X_N(0) < N^{\alpha^*}$. First, Proposition 3 shows that we can assume directly that $X_N(0) = \lfloor xN^{\alpha^*}\rfloor$ for some $x \in (0,1)$. We can then proceed as before by estimating $H_{N^{\alpha^*}-N^{\alpha^*-\varepsilon}}$ for ε sufficiently small and by decreasing the exponent by ε at each step until it falls below $\alpha^*/2$. This completes the proof of Proposition 6. □

Remark.
Proposition 6 completes the results of Section 4. Indeed it shows in particular that if $L_1^N(0) = 0$ and $L_2^N(0) = 0$, then the average hitting time $\mathbb{E}(T_N)$ of the neighborhood of N^{α^*} is, up to a constant, upper bounded by N^β for any $\beta > \alpha^*$. With the same arguments as in the previous proof, it can be shown in fact that there exists a constant $C > 0$, such that

$$\mathbb{E}_{(0,N)}(T_N) \le C_1\frac{N^{\alpha^*}\log(N)^2\log\log(N)}{\log\log\log(N)}. \tag{29}$$

6. The Fluid Time Scale

Recall that the fluid scaling of $(L(t)) = (L_1(t), L_2(t))$ consists in speeding up the time scale of the Markov process proportionally to the norm of its initial state and by scaling the state variable by the same quantity. Hence, if $\|L^N(0)\| = \max(L_1^N(0), L_2^N(0)) = N$, we are interested in the process

$$\big(\overline{L}_N(t)\big) \overset{\text{def.}}{=} \frac{1}{N}\left(L(Nt)\right).$$

Without loss of generality, it can be assumed that

$$\lim_{N\to+\infty}\overline{L}_N(0) = \lim_{N\to+\infty}\left(\frac{L_1(0)}{N}, \frac{L_2(0)}{N}\right) = (x, 1-x),$$

for some $0 \le x \le 1$. See for example Bramson [9] and Robert [25].

The initial fluid state considered up to now in Propositions 4 and 6 corresponds to the case $x = 0$,

$$\lim_{N\to+\infty}\overline{L}_N(0) = (0,1).$$

It has been shown in Proposition 6 of Section 5 that in this setting the hitting time of N^{α^*} by $(L_1(t))$ is negligible compared to N. Consequently, on the fluid time scale, $L_1(t)$ is immediately of the order of $L_2^{\alpha^*}$.

The following proposition completes this result. It shows that for any initial fluid state, then the first time L_1 is close to $L_2^{\alpha^*}$ is of the order of N and therefore that this event occurs on the fluid time scale.

Proposition 7. *Suppose that $\rho_1 < 1/2$, $\rho_2 > 1/2$, and that $(L_1^N(t), L_2^N(t))$ is the solution to the SDE (12) with initial conditions $L_2^N(0) = N$ and $L_1^N(0)$ such that*

$$\lim_{N\to+\infty} L_1^N(0)/N = x \in \mathbb{R}_+.$$

Let T_N be defined by

$$T_N = \inf\left\{s>0\colon\ L_1^N(s) \in L_2^N(s)^{\alpha^*} + \left[-\sqrt{L_2^N(s)^{\alpha^*}\log L_2^N(s)}, \sqrt{L_2^N(s)^{\alpha^*}\log L_2^N(s)}\right]\right\}.$$

Then, for the convergence in distribution we have

$$\lim_{N\to+\infty} \frac{T_N}{N} = t_0(x) \stackrel{def.}{=} \frac{2x}{\mu_1 - 2\lambda_1}.$$

Proof. To start with, note that for both $i \in \{1,2\}$ and all $t \geq 0$, $L_i(t) \leq L_i(0) + \mathcal{N}_{\lambda_i}([0,t])$. Hence, by the law of large numbers for Poisson processes, for every $\eta > 0$ and $K > 0$ there exists $N_0 \in \mathbb{N}$, such that with probability greater than $1-\eta$, the relations

$$L_1(Nt) \leq (1 + 2\lambda_1 K)N \qquad \text{and} \qquad L_2(Nt) \leq (1 + 2\lambda_2 K)N$$

hold for all $t \leq K$ and $N \geq N_0$.

Let us now define $\tau_0^N = \inf\{s \geq 0 :\ L_1^N(s) \leq N/(\log N)^2\}$. Of course, this time is 0 when $L_1^N(0) \leq N/(\log N)^2$. By the remark made in the previous paragraph, with probability at least $1-\eta$ we have for every $s \leq \tau_0^N \wedge (KN)$

$$\frac{\log N - 2\log\log N}{\log(1+2\lambda_2 K) - 2\log\log N + 2\log N} \leq \frac{\log L_1^N(s)}{\log L_1^N(s) + \log L_2^N(s)} \tag{30}$$

and

$$\frac{\log L_2^N(s)}{\log L_1^N(s) + \log L_2^N(s)} \leq \frac{\log(1+2\lambda_2 K) + \log N}{\log(1+2\lambda_2 K) - 2\log\log N + 2\log N}.$$

Hence, on this time interval the service capacity offered to class 1 jobs (resp., class 2 jobs) is at least (resp., at most) 1/2 in the limit. Since $\rho_2 > 1/2$, this implies that the process $(L_2^N(s))$ is increasing on the time interval $[0, \tau_0^N \wedge (KN)]$. That is, for N sufficiently large we have with probability greater than $1-\eta$

$$\inf_{s \leq \tau_0^N \wedge (KN)} L_2^N(s) \geq N - \log N.$$

Consequently, Relations (30) can be completed by the inequality

$$\frac{\log L_1^N(s)}{\log L_1^N(s) + \log L_2^N(s)} \leq \frac{\log(x + 2\lambda_1 K) + \log N}{\log(x + 2\lambda_1 K) + \log(1 - (\log N)/N) + 2\log N}.$$

This shows that, with high probability, as N gets large the two queues receive the capacity 1/2. As a straightforward consequence, we have the convergence in distribution

$$\lim_{N\to\infty} \frac{\tau_0^N}{N} = \frac{2x}{\mu_1 - 2\lambda_1} = t_0(x).$$

Next, let us suppose that $L_1^N(0) \leq \lfloor N/(\log N)^2 \rfloor$ and let us define

$$\tau_1^N = \inf\left\{s > 0 : L_1^N(s) \leq 2N^{\alpha^*}\right\},$$

Since $L_2^N(Nt)$ grows at linear rate (recall that $\rho_2 > 1/2$), we can again compare L_1^N to $(X_N(t))$ and conclude from Relation (27) applied to the stopping time τ_1^N that

$$\frac{\mu_1}{(\alpha^*+1)\log N}\frac{\log(2)}{1+\alpha^*+\log(2)/\log N}\,\mathbb{E}\left(\tau_1^N\right) \leq L_1^N(0).$$

Consequently,

$$\lim_{N\to+\infty} \mathbb{E}\left(\tau_1^N\right)/N = 0.$$

We can thus assume that $L_1^N(0) = 2\lfloor N^{\alpha^*}\rfloor$, and Proposition 6 shows that in this case,

$$\lim_{N\to\infty} \mathbb{E}(T_N)/N = 0.$$

Coming back to the initial question (with an arbitrary x) and using the strong Markov property of L^N combined with the last two limits, we obtain that $T_N = \tau_0^N + (\tau_1^N - \tau_0^N) + (T_N - \tau_1^N)$, where

$$\frac{\tau_0^N}{N} \overset{(d)}{\to} t_0(x), \quad \frac{\tau_1^N - \tau_0^N}{N} \overset{(d)}{\to} 0 \quad \text{and} \quad \frac{T_N - \tau_1^N}{N} \overset{(d)}{\to} 0$$

as $N \to \infty$. By Slutzky's Lemma, we can conclude that $T_N \to t_0(x)$ in law. □

The following theorem is a key result in the analysis of the fluid limits of this system. It states that if $\rho_1 < 1/2$ and $L_1(0)$ is of the order of $L_2(0)^{\alpha^*}$, there is non-empty time interval *on the fluid time scale* on which the relation $L_1 \sim L_2^{\alpha^*}$ holds.

Theorem 2. *Suppose that $\rho_1 < 1/2$, $\rho_1 + \rho_2 < 1$ and let $\kappa > 0$. Then, there exists $\eta_0 > 0$ such that for any sequence (l_1^N) satisfying*

$$\limsup_{N\to+\infty}\left|\frac{l_1^N}{N^{\alpha^*}} - 1\right| \leq \kappa,$$

if $(L_1^N(t), L_2^N(t))$ is the solution to the stochastic differential equation (12) *with initial condition $(L_1(0), L_2(0)) = (l_1^N, N)$, then*

$$\lim_{N\to+\infty} \mathbb{P}\left(\sup_{0\leq s\leq \eta_0 N}\left|\frac{L_1^N(s)}{(L_2^N(s))^{\alpha^*}} - 1\right| > \kappa\right) = 0. \tag{31}$$

Proof. The main argument consists in controlling the upward jumps of L_1 on sufficiently small fluid time scale intervals by an ergodic reflected random walk. It gives the excursions of L_1 on the interval are upper bounded by the excursions of the reflected random walk and therefore cannot be too large by Kingman's Inequality. The same argument applies to the downward jumps. Since L_1 remains close to $L_2^{\alpha^*}$ at the end of the time interval it gives finally the result.

Let us write $a_0 = 1 + \kappa$, and let us prove that there exists $\eta_0 > 0$ such that

$$\lim_{N\to+\infty} \mathbb{P}\left(\sup_{0\leq s\leq \eta_0 N}\frac{L_1^N(s)}{(L_2^N(s))^{\alpha^*}} > a_0\right) = 0. \tag{32}$$

The other inequality can then be shown by using the same technique, and so we omit the details.

The process $(L_2^N(s))$ is stochastically bounded from above by $(\overline{Q}_2^N(s))$ describing the number of class 2 jobs in the priority system where class 1 jobs have the priority of service, i.e. queue 2 is served only when queue 1 is empty. For this system, it is not difficult to show that the convergence in distribution

$$\lim_{N\to+\infty}\left(\frac{\overline{Q}_2^N(Ns)}{N},\ s<1\right)=(1+\mu_2(\rho_1+\rho_2-1)s,\ s<1)$$

holds. Similarly, the process $(L_2^N(s))$ is stochastically bounded from below by $(\underline{Q}_2^N(s))$ describing the number of class 2 jobs when they have priority, and one has the corresponding convergence in distribution

$$\lim_{N\to+\infty}\left(\frac{\underline{Q}_2^N(Ns)}{N},\ s<1\right)=(1+\mu_2(\rho_2-1)s,\ s<1).$$

Fix $0<\eta_0<1$ such that

$$a_0\left(1+2\mu_2(\rho_2-1)\eta_0\right)^{\alpha^*}>1.$$

Let also $\mathcal{E}$ be the event defined by

$$\mathcal{E}=\left\{\sup_{0\le s\le\eta_0 N}\frac{L_2^N(s)}{N}\le\frac{3}{2},\quad \inf_{0\le s\le\eta_0 N}\frac{L_2^N(s)}{N}\ge 1+\frac{3}{2}\mu_2(\rho_2-1)\eta_0\right\}.$$

The above convergence in distribution results show in particular that, for any $\varepsilon>0$, there exists N_0 such that for every $N\ge N_0$,

$$\mathbb{P}\left(\mathcal{E}^c\right)\le\mathbb{P}\left(\sup_{0\le s\le\eta_0 N}\frac{\overline{Q}_2^N(s)}{N}>\frac{3}{2}\right)$$
$$+\mathbb{P}\left(\inf_{0\le s\le\eta_0 N}\frac{\underline{Q}_2^N(s)}{N}<1+\frac{3}{2}\mu_2(\rho_2-1)\eta_0\right)\le\varepsilon.$$

By Relation (18), the service rate of queue 1 at time s is given by

$$\Delta(s)\stackrel{\text{def.}}{=}\lambda_1+\mu_1\frac{\log[(1+L_1^N(s))/(L_2^N(s))^{\alpha^*}]}{(\alpha^*+1)((\alpha^*+1)\log(L_2^N(s))+\log[(1+L_1^N(s))/(L_2^N(s))^{\alpha^*}])}.$$

If for some $y>1$ and some $s<1$ $L_1^N(s)\ge y(L_2^N(s))^{\alpha^*}$, then

$$\Delta(s)\ge\lambda_1+\mu_1\frac{\log(y)}{(\alpha^*+1)((\alpha^*+1)\log(L_2^N(s))+\log(y))}.$$

Furthermore,

$$\Delta(s)\ge\mu_N(y)\stackrel{\text{def.}}{=}\lambda_1+\mu_1\frac{\log(y)}{(\alpha^*+1)((\alpha^*+1)(\log(3/2)+\log(N))+\log(y))}\tag{33}$$

holds on the event $\mathcal{E}$.

Denote by $(X_y(u))$ the birth and death process on $\mathbb{Z}$ whose $+1$ [resp. -1] jumps have rate λ_1 [resp. $\mu_N(y)$] and starting at 0.

Define

$$a_1=a_0\Big(1+\frac{3}{2}\mu_2(\rho_2-1)\eta_0\Big)^{\alpha^*}\text{ and }a_2=\frac{1+a_1}{2}.$$

Since $\rho_2<1$, the definition of η_0 gives that $a_1>1$ and therefore $a_2>1$. Note that $a_0>a_2$. Suppose first that $l_1^N/N^{\alpha^*}\to a_0$.

A simple coupling argument gives that the processes $(L_1^N(s))$ and $(X_{a_2}(s))$ can be constructed so that, on the event $\mathcal{E}$, the relation

$$L_1^N(s) \leq l_1^N + X_{a_2}(s) \tag{34}$$

holds for every $s \leq \inf\{u : L_1^N(u)/(L_2^N(u))^{\alpha^*} \leq a_2\}$.

For every $x \geq 0$, Kingman's inequality (see again Kingman [17] or Relation (3.3) of Theorem 3.5 in Robert [25]) gives us the estimate

$$\mathbb{P}\left(\sup_{s\geq 0} X_{a_2}(s) \geq x\right) \leq \exp\left(-x(\mu_N(a_2) - \lambda_1)\right). \tag{35}$$

In particular, the random variable

$$\frac{1}{N^{\alpha^*}} \sup_{s\geq 0} X_{a_2}(s)$$

converges in distribution to 0 since

$$\mu_N(a_2) \sim \lambda_1 + \frac{\mu_1 \log(a_2)}{(\alpha^*+1)^2 \log N},$$

as N goes to infinity. Let

$$T_N = \inf\left\{s > 0 : l_1^N + X_{a_2}(s) \leq a_2 N^{\alpha^*}\right\}.$$

In a similar way as in the proof of Proposition 6 for example, a simple drift analysis shows that

$$\mathbb{E}(T_N) \leq \frac{(\alpha^*+1)((\alpha^*+1)(\log(3/2)+\log(N)) + \log(a_2))}{\mu_1 \log(a_2)}\, l_1^N,$$

and consequently

$$\lim_{N\to+\infty} \mathbb{P}(T_N > N/(\log N)) = 0.$$

From Relation (34), one gets that with probability tending to 1, $(L_1^N(t)/N^{\alpha^*})$ does not grow above $(l_1^N(t)/N^{\alpha^*})$ until the time T_N, which itself occurs much before N. As a consequence, it is enough to prove Identity (32) with the assumption that $L_2^N(0) = N$ and

$$\lim_{N\to+\infty} \frac{L_1^N(0)}{N^{\alpha^*}} = a_2.$$

The reflected process of the birth and death process $(X_y(u))$ is denoted by $(X_y^+(u))$. This is in fact an $M/M/1$ queue with input rate λ_1 and service rate $\mu_N(y)$. As before, with Inequality (33), one can construct a coupling such that the relation

$$L_1^N(s) \leq L_1^N(0) + X_{a_2}^+(s), \tag{36}$$

holds for every $s \leq \eta_0 N$ on the event $\mathcal{E}$.

For every $y > 0$, let us define

$$\tau_y = \inf\left\{s \geq 0 :\ X_{a_2}^+(s) \geq y\right\} \text{ and } H = \inf\left\{s \geq 0 :\ \frac{L_1^N(s)}{(L_2^N(s))^{\alpha^*}} > a_0\right\}.$$

The proposition will be proved if one shows that $\mathbb{P}(H \leq \eta_0 N)$ converges to 0 as N goes to infinity.

On the event $\mathcal{E}$, if $0 \leq s \leq \eta_0 N$ is such that $L_1^N(s) \geq a_0 (L_2^N(s))^{\alpha^*}$ then

$$L_1^N(s) \geq a_0\left(1 + \frac{3}{2}\mu_2(\rho_2 - 1)\eta_0\right)^{\alpha^*} N^{\alpha^*} = a_1 N^{\alpha^*}.$$

Consequently, Equation (36) and the definition of a_2 give that for every $N \geq N_0$,

$$\mathbb{P}(H \leq \eta_0 N) \leq \varepsilon + \mathbb{P}\big(\tau_{N^{\alpha^*}(a_1-1)/2} \leq \eta_0 N\big). \tag{37}$$

But using the same technique as in the first part of the proof of Proposition 4, we can show that the probability that $X^+_{a_2}$ reaches $N^{\alpha^*}(a_1-1)/2$ in one excursion away from 0 is so low, that the probability that it reaches this height during one of the $\mathcal{O}(N)$ excursions it does in the interval $[0, \eta_0 N]$ tends to 0 as N tends to infinity. Therefore, we can conclude that

$$\mathbb{P}(H \leq \eta_0 N) \leq 2\varepsilon,$$

for N large enough. Since this property holds for every $\varepsilon > 0$, Theorem 2 is proved. □

Corollary 1. *Under the assumptions of Theorem 2, the convergence in distribution*

$$\lim_{N\to+\infty} \left(\frac{L_2^N(Nt)}{N},\, 0 \leq t < t_0 \right) = (\gamma(t),\, 0 \leq t < t_0)$$

holds, with $t_0 = 1/(\mu_2(1-\rho_1-\rho_2))$ *and* $\gamma(t) = 1 + \mu_2(\rho_1+\rho_2-1)t$.

In addition, for every $t < t_0$ *we have*

$$\lim_{N\to+\infty} \mathbb{P}\left(\sup_{0\leq s\leq t} \left| \frac{L_1^N(Ns)}{(L_2^N(Ns))^{\alpha^*}} - 1 \right| > \kappa \right) = 0. \tag{38}$$

Proof. Let us first prove the convergence

$$\lim_{N\to+\infty} \left(\frac{L_2^N(Nt)}{N}, 0 \leq t \leq \eta_0 \right) = (\gamma(t), 0 \leq t \leq \eta_0)\,. \tag{39}$$

The SDE (12) is used in the same way as before, and so we only sketch the proof. By Theorem 2, we have

$$\lim_{N\to+\infty} \mathbb{P}\left(\sup_{0\leq s\leq \eta_0} \left| \frac{L_1^N(Ns)}{(L_2^N(Ns))^{\alpha^*}} - 1 \right| \leq \kappa \right) = 1.$$

Thus, $L_1^N(Ns)$ is at most of the order of N^{α^*} with arbitrarily large probability. This implies that for any $s \leq \eta_0$, all the arrivals at queue 1 up to time Ns are processed. Hence, queue 1 uses the fraction ρ_1 of the capacity of the server, and the remaining capacity is devoted to queue 2. The convergence (39) is proved. Furthermore, from Relation (31) we obtain that

$$\lim_{N\to+\infty} \mathbb{P}\left(\left| \frac{L_1^N(N\eta_0)}{(L_2^N(N\eta_0))^{\alpha^*}} - 1 \right| \leq \kappa \right) = 1,$$

and $L_2^N(N\eta_0) \sim \gamma(\eta_0)N$. Consequently, Theorem 2 applied with the initial condition $(L_1^N(N\eta_0), L_2^N(N\eta_0))$ shows that the convergence (39) and Relation (31) can be extended to the interval $[0, \eta_0(1+\gamma(\eta_0))]$. That is,

$$\lim_{N\to+\infty} \mathbb{P}\left(\sup_{0\leq s\leq \eta_0(1+\gamma(\eta_0))} \left| \frac{L_1^N(Ns)}{(L_2^N(Ns))^{\alpha^*}} - 1 \right| > \kappa \right) = 0. \tag{40}$$

Proceeding by induction, as long as $\gamma(x) \neq 0$ if the convergence (39) and Inequality (40) hold on $[0, x]$, these relations can be extended to $[0, x + \gamma(x)\eta_0]$. The corollary is proved. □

The following theorem is the main result of this section. Proposition 4 and Proposition 6 show that if $L_2^N(0) = N$ then quickly, on a time scale faster than $t \to N^\alpha \log(N)^2 \cdot t$, the first coordinate is very close to N^{α^*}. Since the component L_2 does not change much on this time scale, this can be rephrased as follows: very quickly $L_1 \sim L_2^{\alpha^*}$. This theorem establishes this property on the fluid time scale: L_2 is decreasing linearly on this time scale and L_1 adapts very quickly to the new values of L_2 so that $L_1 \sim L_2^{\alpha^*}$. This is of course a much stronger result than Proposition 4 and Proposition 6.

Theorem 3. *Suppose that $\rho_1 < 1/2$ and $\rho_1 + \rho_2 < 1$. If $(L_1^N(t), L_2^N(t))$ is the solution to the SDE* (12) *with $L_2^N(0) = N$ and*

$$\lim_{N\to+\infty} \frac{L_1^N(0)}{N^{\alpha^*}} = 1,$$

then we have for the convergence in distribution

$$\lim_{N\to+\infty} \left(\left[\frac{L_1^N(Nt)}{N^{\alpha^*}}, \frac{L_2^N(Nt)}{N} \right], 0 \le t < t_0 \right) = \left([\gamma(t)^{\alpha^*}, \gamma(t)], 0 \le t < t_0 \right),$$

where $t_0 = 1/(\mu_2(1-\rho_1-\rho_2))$ and $\gamma(t) = 1 + \mu_2(\rho_1 + \rho_2 - 1)t$.

Proof. From Corollary 1, we obtain that Relation (38) holds for any $\kappa > 0$. Hence, the process

$$\left(\frac{L_1^N(Ns)}{(L_2^N(Ns))^{\alpha^*}},\ 0 \le s < t_0 \right)$$

converges in distribution to the process constant equal to 1 on $[0, t_0)$ as N goes to infinity. We conclude with the convergence of $(L_2^N(Nt)/N,\ 0 \le t < t_0)$. □

We can now state a fluid limit result concerning this network under the assumption $\rho_1 < 1/2$ and $\rho_2 > 1/2$ and with the more general initial conditions considered in Section 4. Analogous statements for the other cases are available, but for the sake of simplicity we do not give them here.

Theorem 4 (Fluid Limits)**.** *Suppose that $\rho_1 < 1/2$ and $\rho_2 > 1/2$. If $(L_1^N(t), L_2^N(t))$ is the solution to the SDE* (12) *with initial conditions such that*

$$\lim_{N\to+\infty} \left(\frac{L_1^N(0)}{N}, \frac{L_2^N(0)}{N} \right) = (x, 1-x),$$

for some $x \in [0,1]$ then, for the convergence in distribution, we have

$$\lim_{N\to+\infty} \left(\frac{L_1^N(Nt)}{N}, \frac{L_2^N(Nt)}{N} \right) = (\ell_1(t), \ell_2(t)),$$

where the pair (ℓ_1, ℓ_2) is defined as follows: if $t_1(x) = 2x/(\mu_1 - 2\lambda_1)$,

$$(\ell_1(t), \ell_2(t)) = \begin{cases} \left(x + \left[\lambda_1 - \frac{\mu_1}{2} \right] t,\ (1-x) + \left[\lambda_2 - \frac{\mu_2}{2} \right] t \right), & t \le t_1, \\ \left(0, (\ell_2(t_1) + [\lambda_2 - \mu_2(1-\rho_1)](t - t_1))^+ \right), & t \ge t_1. \end{cases}$$

Proof. We give only a sketch of the proof. Until the time Nt_1, both queues are of the same asymptotic order for the log function. Consequently, as it has been already seen in the proof of Proposition 7, they both receive half of the capacity. Notice that on the time interval $[0, Nt_1]$, because of our assumptions the variable

$(L_1^N(t))$ decreases whereas $(L_2^N(t))$ increases. After time Nt_1, from Proposition 7 and Theorem 3, the variable $(L_1^N(t))$ is of the order of N^{α^*} and is therefore negligible for the fluid scaling. □

The following proposition completes the description of fluid limits of the system. Its proof follows the same line as the above proposition, it is omitted.

Proposition 8. *If $(L_1^N(t), L_2^N(t))$ is the solution to the SDE (12) with initial conditions such that*

$$\lim_{N\to+\infty}\left(\frac{L_1^N(0)}{N}, \frac{L_2^N(0)}{N}\right) = (x, 1-x),$$

then, for the convergence in distribution, we have

$$\lim_{N\to+\infty}\left(\frac{L_1^N(Nt)}{N}, \frac{L_2^N(Nt)}{N}\right) = (\ell_1(t), \ell_2(t)),$$

where the pair (ℓ_1, ℓ_2) is defined as follows

— *If $\rho_1 > 1/2$ and $\rho_2 > 1/2$, then*

$$(\ell_1(t), \ell_2(t)) = \left(x + \left[\lambda_1 - \frac{\mu_1}{2}\right] t,\ (1-x) + \left[\lambda_2 - \frac{\mu_2}{2}\right] t\right), t \geq 0;$$

— *If $\rho_1 < 1/2$ and $\rho_2 < 1/2$, when x is such that*

$$t_1 \stackrel{def.}{=} \frac{x}{\mu_1/2 - \lambda_1} \leq t_2 \stackrel{def.}{=} \frac{1-x}{\mu_2/2 - \lambda_2}$$

then

$$(\ell_1(t), \ell_2(t)) = \begin{cases} \left(x + \left[\lambda_1 - \frac{\mu_1}{2}\right] t,\ (1-x) + \left[\lambda_2 - \frac{\mu_2}{2}\right] t\right), & t \leq t_1, \\ \left(0, (\ell_2(t_1) + [\lambda_2 - \mu_2(1-\rho_1)](t - t_1))^+\right), & t \geq t_1, \end{cases}$$

and similarly for $t_2 \leq t_1$.

Remark.
Note that the constant α^* does not play a role in the fluid limit, but clearly enough this result is much weaker that Theorem 3. As we shall see in Section 7, the constant α^* plays an important qualitative role in the expression of the invariant distribution when the regime is close to saturation.

7. Heavy Traffic Regime

When $\rho_1 + \rho_2 < 1$, since the queueing system is work conserving the Markov process $(L_1(t), L_2(t))$ has an invariant distribution π_ρ, where $\rho = (\rho_1, \rho_2)$. In the following, $(L_{\rho,1}, L_{\rho,2})$ stands for a random variable with distribution π_ρ. Obtaining explicit expressions to describe π_ρ seems to be quite difficult: because of the logarithmic weights, the double generating function of $(\pi_\rho(m,n), (m,n) \in \mathbb{N})$ does not satisfy an autonomous equation as it is often the case, e.g., in classic two-dimensional reflected random walks. A precise result on the asymptotic behavior of π_ρ when $\rho_1 + \rho_2$ is close to 1 can nevertheless be obtained. Its proof relies heavily on the scaling results of the previous sections.

Theorem 5 (Heavy Traffic Limit of Invariant Distribution)**.** *If $\rho = (\rho_1, \rho_2)$ and $\rho_1 < 1/2$ is fixed, then, as $\bar{\rho} = \rho_1 + \rho_2$ tends to 1, we have the convergence in distribution*

$$\lim_{\lambda_2 \nearrow \mu_2(1-\rho_1)} \Big((1-\bar{\rho})^{\alpha^*} L_{\rho,1}, (1-\bar{\rho}) L_{\rho,2} \Big) = (E_\eta^{\alpha^*}, E_\eta),$$

where here again $\alpha^* = \rho_1/(1-\rho_1)$ and E_η denotes an exponentially distributed random variable whose parameter η is given by

$$\eta^{-1} = \frac{\mu_2}{\sqrt{2}} \sqrt{1 + \frac{(\mu_1 - \mu_2)^2}{\mu_1 \mu_2} \rho_1 (1-\rho_1)}.$$

Proof. The proof uses the fact that the load of this system, i.e. the sum of the services to be processed, is the same as for a classical $M/G/1$ FIFO queue for which a classical heavy traffic limit results holds. It gives that, for $\bar{\rho} \sim 1$, the relation $L_{\rho,1}/\mu_1 + L_{\rho,2}/\mu_2 \sim X/(1-\bar{\rho})$ holds in distribution for a convenient exponential random variable. This shows that at least one of the variables must be large. Then one considers the process $(L_1(t), L_2(t))$ with initial state given by $(L_{\rho,1}, L_{\rho,2})$ so that it is stationary. The results of the previous sections show that if the initial state is large, very quickly L_1 will be of the order of $L_2^{\alpha^*}$, but since the process is stationary it implies that this is also true for $L_{\rho,1}$ and $L_{\rho,2}^{\alpha^*}$.

For every $\xi > 0$, let E_ξ denote, as before, an exponential random variable with parameter ξ. All the variables used here will be assumed to be independent. The total workload of the system is independent of the service allocation and so has the same (invariant) distribution as the workload of an $M/G/1$ queue, with arrival rate $\lambda_1 + \lambda_2$ and with the same service distribution as

$$\sigma \stackrel{\text{def.}}{=} B E_{\mu_1} + (1-B) E_{\mu_2},$$

where B is a Bernoulli random variable with parameter $\lambda_1/(\lambda_1 + \lambda_2)$. Hence, Kingman's heavy traffic result for the workload at equilibrium (see Kingman [18] or Proposition 3.10 of Robert [25]) gives the convergence in distribution

$$\lim_{\lambda_2 \nearrow \mu_2(1-\rho_1)} (1-\bar{\rho}) \left(\frac{L_{\rho,1}}{\mu_1} + \frac{L_{\rho,2}}{\mu_2} \right) = E_{\eta_0}, \tag{41}$$

where η_0 is the constant given by

$$\eta_0 = \lim_{\lambda_2 \nearrow \mu_2(1-\rho_1)} \frac{2}{(\lambda_1 + \lambda_2)\sqrt{\mathrm{Var}(\sigma - E_{\lambda_1+\lambda_2})}} = \sqrt{2} \Bigg/ \sqrt{1 + \frac{(\mu_1 - \mu_2)^2}{\mu_1 \mu_2} \rho_1 (1-\rho_1)}.$$

In particular, the family of random variables

$$[(1-\bar{\rho})(L_{\rho,1}), (1-\bar{\rho})(L_{\rho,2})] \tag{42}$$

is tight as $\bar{\rho} \nearrow 1$. Let us denote a possible limit by (X_1, X_2), corresponding to a sequence $(\bar{\rho}_n)$ converging to 1. For every $n \geq 1$, let us define $(L_{\rho_n,1}(t), L_{\rho_n,2}(t))$ as the process with initial condition $(L_{\rho_n,1}, L_{\rho_n,2})$, which is thus a stationary process.

The strategy of the proof can be described as follows: use the results of Sections 4 and 6 to prove that, for a fixed time t_n, $L_{\rho_n,2}(t_n)$ is large and $L_{\rho_n,1}(t_n)$ is close to $(L_{\rho_n,2}(t_n))^{\alpha^*}$. Consequently, by stationarity, the equivalence $L_{\rho_n,1} \sim L_{\rho_n,2}^{\alpha^*}$ holds and Relation (41) then gives us the desired result.

To start with, let us assume that $\mathbb{P}(X_2 = 0) > 0$. Then, there exists $\delta > 0$ with the property that for any $\eta > 0$, there exists $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$,

$$\mathbb{P}\left(L_{\rho_n,2} \leq \frac{\eta}{1-\bar{\rho}_n}\right) \geq \delta. \tag{43}$$

Now, Relation (41) tells us that $\mathbb{P}(X_1 + X_2 > 0) = 1$. Consequently, for any $\varepsilon > 0$ there exists $\eta_0 > 0$ and n_0 such that if $n \geq n_0$ and $\eta_1 + \eta_2 < 2\eta_0$,

$$\mathbb{P}\left(L_{\rho_n,1} \leq \frac{\eta_1}{1-\bar{\rho}_n}, L_{\rho_n,2} \leq \frac{\eta_2}{1-\bar{\rho}_n}\right) \leq \varepsilon. \tag{44}$$

Let $s_0 > 0$ and $\eta_1 > 0$, and set $t_n = s_0/(1-\bar{\rho}_n)$. For n large enough, we have

$$\begin{aligned} \delta \leq \mathbb{P}\left(L_{\rho_n,2} \leq \frac{\eta_1}{1-\bar{\rho}_n}\right) &= \mathbb{P}\left(L_{\rho_n,2}(t_n) \leq \frac{\eta_1}{1-\bar{\rho}_n}\right) \\ &\leq \mathbb{P}\left(L_{\rho_n,2}(t_n) \leq \frac{\eta_1}{1-\bar{\rho}_n}, \sup_{[0,t_n]} L_{\rho_n,2}(s) > \frac{\eta_0}{3(1-\bar{\rho}_n)}\right) + \varepsilon \\ &+ \mathbb{P}\left(L_{\rho_n,2}(t_n) \leq \frac{\eta_1}{1-\bar{\rho}_n}, \sup_{[0,t_n]} L_{\rho_n,2}(s) \leq \frac{\eta_0}{3(1-\bar{\rho}_n)}, L_{\rho_n,1}(0) > \frac{\eta_0}{1-\bar{\rho}_n}\right). \end{aligned} \tag{45}$$

Clearly enough since $(L_2(t))$ decreases at rate at most μ_2, as n increases the first term on the right hand side of (45) can be made arbitrarily small by choosing η_1 and s_0 in such a way that

$$\eta_1 < \frac{\eta_0}{3} + (\lambda_2 - \mu_2)s_0 = \frac{\eta_0}{3} - \mu_2(1-\rho_2)s_0. \tag{C_1}$$

The third term in the right-hand side of Relation (45) can be upper bounded by

$$\mathbb{P}\left(L_{\rho_n,1}(0) > \frac{\eta_0}{1-\bar{\rho}_n}, \inf_{[0,t_n]} L_{\rho_n,1}(s) < \frac{2\eta_0}{3(1-\bar{\rho}_n)}\right) \tag{46}$$

$$+\mathbb{P}\left(L_{\rho_n,2}(t_n) \leq \frac{\eta_1}{1-\bar{\rho}_n}, \sup_{[0,t_n]} L_{\rho_n,2}(s) \leq \frac{\eta_0}{3(1-\bar{\rho}_n)},\right.$$

$$\left. L_{\rho_n,1}(0) > \frac{\eta_0}{1-\bar{\rho}_n}, \inf_{[0,t_n]} L_{\rho_n,1}(s) \geq \frac{2\eta_0}{3(1-\bar{\rho}_n)}\right). \tag{47}$$

As before, the quantity in (46) will tend to 0 as $n \to \infty$ if we choose s_0 such that

$$\frac{\eta_0}{3} - \mu_1(1-\rho_1)s_0 > 0. \tag{C_2}$$

Finally, if

$$\sup_{[0,t_n]} L_{\rho_n,2}(s) \leq \frac{\eta_0}{3(1-\bar{\rho}_n)} \text{ and } \inf_{[0,t_n]} L_{\rho_n,1}(s) \geq \frac{2\eta_0}{3(1-\bar{\rho}_n)},$$

the infinitesimal drift $\Delta_{\rho_n,2}(s)$ of $L_{\rho_n,2}$ satisfies on $[0,t_n]$

$$\Delta_{\rho_n,2}(s) \geq \mu_2\left(\rho_2 - \frac{\log\left(\eta_0/(3(1-\bar{\rho}_n))\right)}{\log\left(\eta_0/(3(1-\bar{\rho}_n))\right) + \log\left(2\eta_0/(3(1-\bar{\rho}_n))\right)}\right) \sim \mu_2\left(\rho_2 - \frac{1}{2}\right),$$

as n goes to infinity. But $\rho_{n,2}$ converges to $1-\rho_1 > 1/2$ by assumption, which means that the quantity in (47) will tend to 0 as n gets large whenever

$$\eta_1 < \mu_2(1/2 - \rho_1)s_0. \tag{C_3}$$

Choosing first s_0 small enough to match (C_1) and (C_2), and then η_1 small enough to satisfy C_1 and C_3, we can conclude that the right hand side of Relation (45) can be made smaller than $\delta/2$, which yields a contradiction. We have thus proved that $\mathbb{P}(X_2 > 0) = 1$.

As a consequence, for any $\varepsilon > 0$, one can find η_0 and $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$,

$$\mathbb{P}\left(L_{\rho_n,2} \geq \frac{\eta_0}{1-\bar{\rho}_n}\right) \geq 1 - \varepsilon.$$

The last step follows an analogous path, in that we choose a convenient time t_n to first show that, for any $\beta > \alpha^*$, the quantity

$$\mathbb{P}\left(L_{\rho_n,2} \geq \frac{\eta_1}{1-\bar{\rho}_n}, L_{\rho_n,1} \geq \frac{1}{(1-\bar{\rho}_n)^\beta}\right)$$

is arbitrarily small with the help of Proposition 6, and next that for K sufficiently large,

$$L_{\rho_n,1} \in (L_{\rho_n,2})^{\alpha^*} + \left[-K\sqrt{(L_{\rho_n,2})^{\alpha^*}\log(L_{\rho_n,2})}, K\sqrt{(L_{\rho_n,2})^{\alpha^*}\log(L_{\rho_n,2})}\right]$$

with high probability by Proposition 7.

Since $L_{\rho_n,1}$ is of the order of $(L_{\rho_n,2})^{\alpha^*}$ and that $\alpha^* < 1$, Relation (41) gives the desired convergence result. □

8. General Functions for Resource Sharing

In this section, we consider the case where the function $(\log x)$ describing the access to the resource is replaced by some other function $(f(x))$. For a two node network, the corresponding Q-matrix is given by: for every $x \in \mathbb{N}_+^2$,

$$\begin{cases} q(x, x+e_i) & = \lambda_i \\ q(x, x-e_i) & = \mu_i \dfrac{f(x_i)}{f(x_1)+f(x_2)}. \end{cases} \tag{48}$$

To concentrate on the most interesting case, throughout this section we assume that $\rho_1 < 1/2$. We analyze only the first two time scales in order to stress the difference with the log function, at least concerning the scaling parameters. The proofs of the results presented here are similar to those in the log case, and so most of the time we only sketch them.

8.1. The $\log\log$ Function.

Here we consider the function $f(x) = \log(\log(e+x))$. To simplify the notations, we use instead the function $x \to \log_2(x) \stackrel{\text{def.}}{=} \log(\log x)$. This, of course, does not change the convergence results obtained in the following paragraphs.

Initial Phase. The first time scale is $t \mapsto \phi_N(t)$ with

$$\phi_N(t) = \exp\left[(\log N)^t\right].$$

The stochastic evolution equation is in this case

$$\begin{aligned} &L_1(\phi_N(t)) \\ &\quad = L_1(0) + \lambda_1\phi_N(t) - \mu_1 \int_0^{\phi_N(t)} \frac{\log_2(L_1(u))}{\log_2(L_1(u)) + \log_2(N)}\,\mathrm{d}u + M_N(\phi_N(t)), \end{aligned} \tag{49}$$

where $(M_N(t))$ is a local martingale. Since $\phi_N(t) \ll N$ as long as $t < 1$, the second coordinate L_2 stays at N at least up to $t \approx 1$. This justifies the fact that in the above expression, the term $\log_2(L_2(u))$ has been replaced by $\log_2(N)$.
Let us now define $Z_N(t) = L_1(\phi_N(t))/\phi_N(t)$. We have

$$Z_N(t)=Z_N(1)+\lambda_1-\frac{1}{\phi_N(t)} - \frac{\mu_1}{\phi_N(t)}\int_0^t \frac{\log_2(Z_N(u)\phi_N(u))}{\log_2(Z_N(u)\phi_N(u))+\log_2(N)}\phi_N'(u)\,\mathrm{d}u + \frac{M_N(\phi_N(t))}{\phi_N(t)}.$$

Note that, for every $u \geq 0$,

$$\log_2(Z_N(u)\phi_N(u)) = \log_2(\phi_N(u)) + \log\left(1+\frac{\log(Z_N(u))}{\log(\phi_N(u))}\right),$$

and

$$\log_2(\phi_N(u)) = u\log_2(N).$$

Hence, using the same methods as in Section 4, we obtain the following equivalence for the convergence in distribution of processes, uniformly over any compact subset of $(0, \alpha^*)$:

$$(Z_N(t)) \sim \left(\lambda_1 - \frac{\mu_1}{\phi_N(t)}\int_0^t \frac{u}{u+1}\,\phi_N'(u)\,\mathrm{d}u\right) \sim \left(\lambda_1 - \mu_1\frac{t}{t+1}\right).$$

The following proposition is the analogue of Proposition 2 for the log log function. Recall that $\alpha^* = \rho_1/(1-\rho_1) < 1$ since $\rho_1 < 1/2$.

Proposition 9. *If $(L_1^N(t), L_2^N(t))$ is the Markov process with Q-matrix* (48) *and with initial condition* $(0, N)$*, then the convergence in distribution*

$$\lim_{N\to+\infty}\left(\frac{L_1^N\left(\exp\left[(\log N)^t\right]\right)}{\exp\left[(\log N)^t\right]}, 0 < t < \alpha^*\right) = \left(\lambda_1 - \mu_1\frac{t}{t+1}, 0 < t < \alpha^*\right)$$

holds for the uniform topology on compact sets of $(0, \alpha^*)$.

The Local Equilibrium. The last proposition together with the results obtained in Section 5 suggest that, if $\rho_1 < 1/2$, the process should remain stable around the value $\exp\left[(\log N)^{\alpha^*}\right]$. As in Section 5, let us assume that $L_1^N(0) = \delta\phi_N(\alpha^*)$ for some $\delta \leq 1$, while $L_2^N(0) = N$. Using the relation $\rho_1 = \alpha^*/(1+\alpha^*)$ and the evolution equation (49), we obtain that for every $t \geq 0$

$$L_1^N(t) = L_1^N(0) - \frac{\mu_1}{\alpha^*+1}\int_0^t \frac{\log[\log(L_1^N(u))/(\log N)^{\alpha^*}]}{\log[\log(L_1^N(u))/(\log N)^{\alpha^*}] + (\alpha^*+1)\log_2(N)}\,\mathrm{d}u + M_N(t).$$

As we shall see, the appropriate scaling of time around $\phi_N(\alpha^*)$ turns out to be $t \mapsto \psi_N t$, where

$$\psi_N \stackrel{\text{def}}{=} \phi_N(\alpha^*)\,(\log N)^{\alpha^*}\log_2(N) = \exp\left[(\log N)^{\alpha^*}\right](\log N)^{\alpha^*}\log\log(N).$$

Indeed, let $\widehat{Z}_N(t) = L_1^N(\psi_N t)/\phi_N(\alpha^*)$. For every $u > 0$, we have

$$\log\left[\frac{\log(L_1^N(\psi_N u))}{(\log N)^{\alpha^*}}\right] = \log\left[1+\frac{\log(\widehat{Z}_N(u))}{(\log N)^{\alpha^*}}\right],$$

so that

$$\begin{aligned}
&\widehat{Z}_N(t) - \widehat{Z}_N(0) - \frac{M_N(\psi_N t)}{\phi_N(\alpha^*)} + \frac{M_N(1)}{\phi_N(\alpha^*)} \\
&= -\frac{\mu_1}{(\alpha^*+1)\phi_N(\alpha^*)} \int_0^{\psi_N t} \frac{\log[\log(L_1^N(u))/(\log N)^{\alpha^*}]}{\log[\log(L_1^N(u))/(\log N)^{\alpha^*}] + (\alpha^*+1)\log_2(N)}\,\mathrm{d}u \\
&\sim -\frac{\mu_1 (\log N)^{\alpha^*} \log_2(N)}{(\alpha^*+1)} \int_0^t \frac{\log(\widehat{Z}_N(u))}{\log(\widehat{Z}_N(u)) + (\alpha^*+1)(\log N)^{\alpha^*}\log_2(N)}\,\mathrm{d}u.
\end{aligned}$$

With the same tightness argument as in the proof of Proposition 4, we obtain the corresponding convergence result detailed below. It is remarkable that the limit is the same as in Proposition 4.

Proposition 10. *If $(L_1^N(t), L_2^N(t))$ is the Markov process with Q-matrix* (48)*, and initial conditions $L_2^N(0) = N$ and $L_1^N(0) \sim \delta \exp[(\log N)^{\alpha^*}]$ for some $\delta \in (0,1]$, then for the convergence in distribution of processes we have*

$$\lim_{N\to+\infty} \left(L_1^N(\psi_N t) e^{-(\log N)^{\alpha^*}} \right) = (h(t)),$$

where

$$\psi_N = \exp\left[(\log N)^{\alpha^*}\right] (\log N)^{\alpha^*} \log\log N$$

and $(h(t))$ is the function defined by Relation (21)*.*

Furthermore, if $L_1^N(0) \sim \exp[(\log N)^{\alpha^*}] + y\sqrt{\psi_N}$ for some $y \in \mathbb{R}$, then the sequence of processes

$$\left(\frac{L_1^N(\psi_N t) - e^{(\log N)^{\alpha^*}}}{\sqrt{\psi_N}} \right)$$

converges in distribution to the Ornstein-Uhlenbeck process defined by Relation (26)*.*

Remark.
If $\rho_1 < 1/2$ and if the initial condition is $(0, N)$, we obtain that L_1 is of the order of

$$\exp[(\log N)^{\alpha^*}]$$

on the time scale $t \mapsto \psi_N t$. This is much smaller that the quantity N^{α^*} corresponding to the log policy. One can then take a function f growing more slowly to infinity, such as $\log\log\log$. Under the same assumptions, the variable L_1^N live in a region with an even smaller order of magnitude. By pushing this scheme a little further, we would obtain a policy similar to the Head of the Line Processor-Sharing (see Bramson [8]), where node j receives the bandwidth

$$\frac{\mathbf{1}_{\{n_i \neq 0\}}}{\mathbf{1}_{\{n_1 \neq 0\}} + \cdots + \mathbf{1}_{\{n_J \neq 0\}}}.$$

The main drawback of this policy is that a node with many jobs has the same fraction of the capacity as a node with only a few of them. For this reason, it is more likely that local congestion will occur more frequently.

8.2. **The Time Scales for a General Function.** Finally, let us return to the general case. Let us assume that the function $x \mapsto f(x)$ is an increasing continuous function on $\mathbb{R}_+$, tending to infinity as $x \to \infty$, and suppose that there exist two functions $x \mapsto A_f(x)$ and $x \mapsto B_f(x)$ on $\mathbb{R}_+$ such that for any $t > 0$ and $z \geq 0$,

$$(F_1) \qquad \lim_{x\to+\infty} \frac{f^{-1}(tf(x))}{x} = 0, \qquad \lim_{x\to+\infty} \frac{f(zx) - f(x)}{A_f(x)} = B_f(z).$$

The first time scale for the initial phase is given by

$$\phi_N : t \mapsto f^{-1}(tf(N)), \tag{50}$$

where f^{-1} denotes the inverse function of f. If $(L_1^N(t), L_2^N(t))$ is the Markov process with Q-matrix (48) and initial condition $(0, N)$, then the convergence in distribution

$$\lim_{N\to+\infty} \left(\frac{L_1^N(\phi_N(t))}{\phi_N(t)}, 0 < t < \alpha^* \right) = \left(\lambda_1 - \mu_1 \frac{t}{t+1}, 0 < t < \alpha^* \right)$$

holds for the uniform topology on compact sets of $(0, \alpha^*)$. Indeed, the first relation in Condition (F_1) ensures that the second coordinate L_2^N stays at N while L_1^N is of the order of $\phi_N(t)$, so that exactly the same techniques as in Section 4 can be applied.

The second time scale of interest is $t \mapsto \psi_N t$, where

$$\psi_N = \frac{\phi_N(\alpha^*) f(N)}{A_f(\phi_N(\alpha^*))}. \tag{51}$$

If $(L_1^N(t), L_2^N(t))$ is the Markov process with Q-matrix (48) and initial conditions $L_2(0) = N$ and $L_1(0) \sim \delta\phi_N(\alpha^*)$ for some $\delta \in (0, 1]$, then we conjecture that the convergence in distribution of processes

$$\lim_{N\to+\infty} \left(\frac{L_1^N(\psi_N t)}{\phi_N(\alpha^*)} \right) = (h(t)),$$

should hold, where $(h(t))$ is the function defined by

$$\begin{cases} h \equiv 1 & \text{if } \delta = 1, \\ \displaystyle\int_\delta^{h(t)} \frac{1}{B_f(u)}\, \mathrm{d}u = -\frac{\mu t}{(1+\alpha^*)^2} & \text{if } \delta \neq 1. \end{cases} \tag{52}$$

Some regularity properties (required to use Lebesgue's Differentiation Theorem for instance) are clearly necessary to justify this convergence, but then the rest of the proof should follow the lines of the proof of the corresponding result for the log case.

Some Examples.

a) $f(x) = \log\log(x)$.
In this case $A_f(x) = 1/\log x$ and $B_f(x) = \log x$, and we recover the expression of the time scales obtain in Section 8.1 for the $\log\log$ function, that is

$$\phi_N(t) = \exp[(\log N)^t] \quad \text{and} \quad \psi_N = (\log N)^{\alpha^*} \log\log(N) \exp\left((\log N)^{\alpha^*} \right).$$

b) $f(x) = (\log x)^\beta$.
Here $A_f(x) = 1/[\beta(\log x)^{\beta-1}]$ and $B_f(x) = \log x$, we gives the two time scales

$$\phi_N(t) = N^{t^{1/\beta}} \quad \text{and} \quad \psi_N = \beta\alpha^{*1-1/\beta} N^{\alpha^{*1/\beta}} (\log N)^{2\beta-1}.$$

In particular, we recover the time scales of the log case which have already been identified.

Note that the functions $x \mapsto x^\alpha$, $\alpha > 0$, do not satisfy the first relation in Condition (F_1).

9. The Network with $J > 2$ Nodes

In this section, we briefly describe the case of $J > 2$ nodes competing for the single resource. A special case is considered to illustrate the similarities and also the differences in qualitative behaviors. The motivation of this section is to show that the analysis of the two node network gives the main ideas to start the investigation of more complicated situations. New difficulties are indicated in the text, and the proofs of these results will be the subject of a further work in a more general setting.

Let us assume that

$$\rho_1 < \rho_2 < \cdots < \rho_J \text{ and } \sum_1^J \rho_j < 1.$$

If the state of the system is $(n_j)_{1\le j\le J}$, the kth station receives the fraction of service

$$\frac{\log(1+n_k)}{\log(1+n_1)+\log(1+n_2)+\cdots+\log(1+n_J)}.$$

From now on, we consider the case where $L_j(0) = 0$ for $j = 1, \ldots, J-1$, and $L_J(0) = N$. As before $(L_j^N(t))$ denotes the Markov process describing the number of requests in each queue and with this initial condition.

9.1. Initial Phase.

The following proposition is analogous to Proposition 2, and can be proved in the same way.

Proposition 11. *If $(L_j^N(t))$ is the solution of the SDE (12) with the initial condition $L_j(0) = 0$ for $1 \le j \le J-1$, and $L_J = N$, and if*

$$t_1 \stackrel{def.}{=} \frac{\rho_1}{1-(J-1)\rho_1} < 1, \qquad i.e. \qquad \rho_1 < \frac{1}{J},$$

then, for every $1 \le j \le J-1$, the convergence

$$\lim_{N\to+\infty}\left(\frac{L_j^N(N^t)}{N^t}, 0 < t < t_1\right) = \left(\lambda_j - \mu_j \frac{t}{(J-1)t+1}, 0 < t < t_1\right)$$

holds for the uniform topology on compact sets of $(0, t_1)$.

Observe that the quantity t_1 is precisely the first time t for which

$$\lambda_1 = \mu_1 \frac{t}{1+(J-1)t}$$

9.2. **Second Phase.** Let us now give a more heuristic description of the evolution of the network after "time" N^{t_1}, but still on the time scale $t \mapsto N^t$. As we shall see, for the second phase the exponent in N of the random variables $(L_j(N^t), 2 \leq j \leq J-1)$ are still t, but the exponent $\alpha_{2,1}(t)$ of L_1^N becomes a linear function of t with slope less than 1.
Let us use the different phenomena observed in the two node case to infer the behavior of the J-node system.

First, since $\rho_j > \rho_1$ for every $j \geq 2$, the infinitesimal drift of $L_j(N^t)$ remains positive at least for a small amount of time after t_1. Hence, one should have the following convergence in distribution: for $2 \leq j \leq J-1$,

$$\lim_{N \to +\infty} \left(\frac{L_j^N(N^t)}{N^t}, t_1 < t < t_2 \right) = \left(\lambda_j - \mu_j \frac{t}{\alpha_{1,2}(t) + (J-2)t + 1}, t_1 < t < t_2 \right),$$

where t_2 is the first time t at which

$$\lambda_2 = \mu_2 \frac{t}{\alpha_{1,2}(t) + (J-2)t + 1}.$$

Second, the station 1 should remain at a local equilibrium in the sense that the coefficient $\alpha_{1,2}(t)$ should be determined as follows:

$$\lambda_1 = \mu_1 \frac{\alpha_{1,2}(t)}{\alpha_{1,2}(t) + (J-2)t + 1}.$$

Combining these relations, we obtain that

$$\alpha_{1,2}(t) = \frac{\rho_1}{1-\rho_1}(1 + (J-2)t) \text{ and } t_2 = \frac{\rho_2}{1 - \rho_1 - (J-2)\rho_2}.$$

Of course, t_2 has to be strictly less than 1.

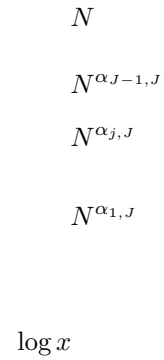

FIGURE 2. The network with $J > 2$ nodes on the $t \mapsto N^t$ time scale when $t_{J-1} < 1$.

9.3. **Subsequent Phases.** Let us now give a, still heuristic, description of the kth phase for $1 < k < J-1$. The first $k-1$ stations are at a "local" equilibrium. Denoting the exponent of the jth station in the kth phase by $\alpha_{j,k}(t)$, the equilibrium is characterized by the relation

$$\frac{\alpha_{j,k}(t)}{\alpha_{1,k}(t) + \alpha_{2,k}(t) + \cdots + \alpha_{k-1,k}(t) + (J-k)t + 1} = \rho_j.$$

for $1 \leq j \leq k-1$. The end of the kth phase, t_k, corresponds to the situation when the kth station reaches an equilibrium, i.e. t_k satisfies

$$\frac{t_k}{\alpha_{1,k}(t_k) + \alpha_{2,k}(t_k) + \cdots + \alpha_{k-1,k}(t_k) + (J-k)t_k + 1} = \rho_k.$$

Consequently, we obtain that

$$\alpha_{j,k}(t) = \frac{\rho_j}{1 - \sum_{i=1}^{k-1} \rho_i}(1 + (J-k)t), \quad 1 \leq j \leq k-1,$$

and

$$t_k = \frac{\rho_k}{1 - \sum_{i=1}^{k-1} \rho_i - (J-k)\rho_k},$$

The time t_k is such that $t_k < 1$ if

$$\sum_{i=1}^{k-1} \rho_i + (J-k+1)\rho_k < 1.$$

9.4. **Final Phase.** Provided that t_{J-1} is strictly less that 1, at the time $N^{t_{J-1}}$ all the stations are at a local equilibrium around $N^{\alpha_{j,J}}$ where $\alpha_{j,J}$ is given by

$$\alpha_{j,J} = \frac{\rho_j}{1 - \sum_{i=1}^{J-1} \rho_i}.$$

Keep in mind that strictly speaking, the local equilibrium of the jth station is on the time scale $t \mapsto N^{\alpha_{j,J}} \log Nt$. Hence, the whole process can be thought as a collection of stationary processes evolving on different time scales. See Figure 2.

The difficult technical problem to solve here is for the J phases in between, during which some of the exponents depend on time, adapting to the linear growth of the other exponents. One of our main problems throughout this work has been that we did not succeed in proving convergence of the $\log(L_j)/\log N$ variables without showing a more demanding result, namely a convergence result for the variables L_j. This difficulty is even more serious here since the exponents depend on time.

References

1. N. Abramson, *The Aloha system*, FJCC AFIPS Conf. Proc. (Montvale, N.J.), vol. 37, AFIPS Press,, 1970, pp. 281–285.
2. Eitan Altman, Konstantin Avrachenkov, and Urtzi Ayesta, *A survey on discriminatory processor sharing*, Queueing Systems. Theory and Applications **53** (2006), no. 1-2, 53–63.
3. Abdelghani Ben Tahar and Alain Jean-Marie, *The fluid limit of the multiclass processor sharing queue*, Queueing Systems. Theory and Applications **71** (2012), no. 4, 347–404.
4. Patrick Billingsley, *Convergence of probability measures*, second ed., Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons Inc., New York, 1999, A Wiley-Interscience Publication.
5. T. Bonald, L. Massoulié, A. Proutière, and J. Virtamo, *A queueing analysis of max-min fairness, proportional fairness and balanced fairness*, Queueing Systems **53** (2006), no. 1-2, 65–84.
6. Thomas Bonald and Laurent Massoulié, *Impact of fairness on Internet performance*, SIGMETRICS '01: Proceedings of the 2001 ACM SIGMETRICS international conference on Measurement and modeling of computer systems (New York, NY, USA), ACM, 2001, pp. 82–91.
7. Nick Bouman, Sem Borst, Johan van van Leeuwaarden, and Alexandre Proutière, *Backlog-based random access in wireless networks: Fluid limits and delay issues*, 23rd International Teletraffic Congress (ITC), September 2011, pp. 39–46.
8. Maury Bramson, *Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks*, Queueing Systems **23** (1996), 1–26.

9. Maury Bramson, *Stability of queueing networks*, Probability Surveys **5** (2008), 169–345.
10. J. G. Dai, *A fluid limit model criterion for instability of multiclass queueing networks*, Annals of Applied Probability **6** (1996), no. 3, 751–757.
11. Stewart N. Ethier and Thomas G. Kurtz, *Markov processes: Characterization and convergence*, John Wiley & Sons Inc., New York, 1986.
12. M. I. Freidlin and A. D. Wentzell, *Random perturbations of dynamical systems*, second ed., Springer-Verlag, New York, 1998, Translated from the 1979 Russian original by Joseph Szücs.
13. Javad Ghaderi, Sem Borst, and Phil Whiting, *Backlog-based random access in wireless networks: Fluid limits and instability issues*, 46th Annual Conference on Information Sciences and Systems (CISS), March 2012, pp. 1–6.
14. Carl Graham and Philippe Robert, *Interacting multi-class transmissions in large stochastic networks*, Annals of Applied Probability **19** (2009), no. 6, 2334–2361.
15. F. Kelly, A. Maulloo, and D. Tan, *Rate control in communication networks: shadow prices, proportional fairness and stability*, Journal of the Operational Research Society, vol. 49, 1998.
16. R. Z. Khasminskii, *Stochastic stability of differential equations*, Sijthoff & Noordhoff, Alphen aan den Rijn, 1980, Translated from the Russian by D. Louvish.
17. J. F. C. Kingman, *Inequalities in the theory of queues*, Journal of the Royal Statistical Society B **32** (1970), 102–110.
18. 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.
19. V. A. Malyshev, *Networks and dynamical systems*, Advances in Applied Probability **25** (1993), no. 1, 140–175.
20. Laurent Massoulié and James Roberts, *Bandwidth sharing: Objectives and algorithms*, INFOCOM '99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies, 1999, pp. 1395–1403.
21. R. Metcalf and D. Boggs, *Ethernet: Distributed packet switching for local computer networks*, Communications of the ACM **19** (1976), no. 7, 395–403.
22. Jeonghoon Mo and Jean Walrand, *Fair end-to-end window-based congestion control*, IEEE Trans. Netw. **8** (2000), no. 5, 556–567.
23. G. C. Papanicolaou, D. Stroock, and S. R. S. Varadhan, *Martingale approach to some limit theorems*, Papers from the Duke Turbulence Conference (Duke Univ., Durham, N.C., 1976), Paper No. 6, Duke Univ., Durham, N.C., 1977, pp. ii+120 pp. Duke Univ. Math. Ser., Vol. III. MR 0461684 (57 #1669)
24. Kavita Ramanan and Martin I. Reiman, *Fluid and heavy traffic diffusion limits for a generalized processor sharing model*, The Annals of Applied Probability **13** (2003), no. 1, 100–139.
25. Philippe Robert, *Stochastic networks and queues*, Stochastic Modelling and Applied Probability Series, vol. 52, Springer, New-York, June 2003.
26. Walter Rudin, *Real and complex analysis*, third ed., McGraw-Hill Book Co., New York, 1987.
27. A. N. Rybko and A. L. Stolyar, *On the ergodicity of random processes that describe the functioning of open queueing networks*, Problems on Information Transmission **28** (1992), no. 3, 3–26.
28. Devavrat Shah and Jinwoo Shin, *Randomized scheduling algorithm for queueing networks*, The Annals of Applied Probability **22** (2012), no. 1, 128–171.
29. Devavrat Shah and Damon Wischik, *Log-weight scheduling in switched networks*, Queueing Systems, Theory and Applications **71** (2012), no. 1-2, 97–136.
30. L. Tassiulas and A. Ephremides, *Stability Properties of Constrained Queueing Systems and Scheduling Policies for Maximum Throughput in Multihop Radio Networks*, IEEE Trans. Automat. Contr. **37** (1992), no. 12, 1936–1948.
31. Damon Wischik, *Queueing theory for switched networks*, ICMS workshop on stochastic processes in communication networks for young researchers (Edinburgh), June 2010, URL : http://www.cs.ucl.ac.uk/staff/ucacdjw/Talks/netsched.html.

(Ph. Robert) INRIA Paris—Rocquencourt, Domaine de Voluceau, 78153 Le Chesnay, France.
E-mail address: Philippe.Robert@inria.fr

(A. Véber) Centre de Mathématiques Appliquées, École Polytechnique, Route de Saclay, 91128 Palaiseau Cedex, France
E-mail address: Amandine.Veber@cmap.polytechnique.fr