THE EQUILIBRIUM STATES OF LARGE NETWORKS OF ERLANG QUEUES

DAVIT MARTIROSYAN,* AND PHILIPPE ROBERT, ** INRIA

Abstract

The equilibrium properties of allocation algorithms for networks with a large number of nodes with finite capacity are investigated. Every node receives a flow of requests. When a request arrives at a saturated node, i.e. a node whose capacity is fully utilized, an allocation algorithm may attempt to reallocate the request to a non-saturated node. For the algorithms considered, the reallocation comes at a price: either extra capacity is required in the system, or the processing time of a reallocated request is increased.

The paper analyzes the properties of the equilibrium points of the associated asymptotic dynamical system when the number of nodes gets large. At this occasion the classical model of Gibbens, Hunt, and Kelly (1990) in this domain is revisited. The absence of known Lyapunov functions for the corresponding dynamical system significantly complicates the analysis. Several techniques are used. Analytic and scaling methods are used to identify the equilibrium points. We identify the subset of parameters for which the limiting stochastic model of these networks has multiple equilibrium points. Probabilistic approaches are used to prove the stability of some of them. A criterion of exponential stability with the spectral gap of the associated linear operator of equilibrium points is also obtained.

Keywords: Multiple equilibrium points; scaled Markov processes

2010 Mathematics Subject Classification: 60G55; 92D20

1. Introduction

In this paper we study the time evolution properties of stochastic networks with a large number of nodes. Each node receives a flow of jobs/requests. A node has a maximal number of requests which can be present at the same time; this is called its capacity. The saturation of a node refers to the fact that its current number of requests is maximal, and therefore it cannot accept new jobs. If a job arrives at a non-saturated node, it is accepted and served immediately. When a job arrives at a saturated node, it is either rejected, or else sent to a non-saturated node according to some allocation algorithm. In this case, using the terminology of communication networks, the job is said to be rerouted.

We study two classes of rerouting algorithms, defined below. For both of them, the rerouting of a request comes at a price for the network, either with a larger capacity required or with a longer processing/sojourn time. If a request is accepted at its arrival node, it is processed at
rate $\mu_1 > 0$. Otherwise, if a request cannot be accommodated at its arrival node, one of the following two models is applied.

1. **The Routing with Increased Sojourn Time (RIST) algorithm.**

   A non-saturated node is chosen at random to accommodate the request, which is processed at a lower rate $\mu_2$, $0 < \mu_2 \leq \mu_1$. On average, a rerouted job stays longer in the network in this algorithm. If all nodes are saturated, the request is rejected.

   This type of model is used to take into account the fact that, in some contexts, the transfer time of a rerouted job is not negligible. A special case has already been analyzed in Malyshev and Robert [25] when the capacity of each node is 1. See also Remark 2.1 of Tibi [33].

   A variant of this algorithm is investigated in Section 2. When an arriving job finds a node saturated, it picks another node at random, again and again, until it finds a non-saturated node, provided it makes fewer than $p_0$ attempts; otherwise it is rejected. The RIST algorithm corresponds to the case $p_0 = +\infty$.

2. **The Dynamic Alternative Routing (DAR) algorithm**

   When a request arrives at a saturated node, two other nodes are chosen at random. If both of them are non-saturated, the request takes one place in each of them. Otherwise, the request is rejected. This algorithm was initially considered by Gibbens et al. [16] in 1990 to cope with congestion in communication networks. The nodes are links of the network and connections are established on links. When a connection requires an already saturated link $(AB)$ connecting two vertices $A$ and $B$, the algorithm attempts to establish the connection between $A$ and $B$ by taking another vertex $D$ at random and using the two links $(AD)$ and $(DB)$. See also Kelly [20] and Marbukh [26].

The main goal of the mathematical study of these networks is to quantify the potential benefit of rerouting mechanisms, in order to determine whether it is worthwhile to design routing algorithms rather than do nothing (i.e. immediately reject jobs that arrive at saturated nodes). For this purpose, the main quantities of interest are the probabilities that, at equilibrium, a request is

(a) accepted without rerouting;

(b) rejected (i.e. unable to be accommodated even by rerouting).

The main problem with rerouting is the following. If there are a significant number of saturated nodes, an important fraction of the resources of the network (capacity, processing time) may be consumed by rerouted jobs due to their extra cost, making rerouting even more likely, to the detriment of the criterion associated to (a). Furthermore, if there are too many saturated nodes, even the loss rate may be nonnegligible, affecting the criterion associated to (b).

It has been shown by Gibbens et al. [16], through approximations and numerical experiments, that in some cases the DAR algorithm exhibits an unpleasant property. It may happen that the network can stay for a very large amount of time in different regimes (sets of states): some where most requests/jobs are accepted without rerouting and others in which a significant fraction of jobs are rerouted. The performance of the system does not stabilize in this case, since it depends on the current set of states of the network. This is the metastability property which is well known in statistical physics. Though this property is closely linked to the existence of multiple equilibrium points, it will not be discussed in this paper. (See den Hollander [11], Bovier and den Hollander [4], and Olivieri et al. [28], for example.)
The 1990 paper by Gibbens et al. [16] attracted a lot of attention, mainly because of the original stability properties it suggested, at least in a stochastic network context. (See also Marbukh [26].) It had a strong impact in the sense that it stressed the undesirable phenomena that can happen without some care in the design of allocation algorithms. Nevertheless, apart from the mean-field result of Graham and Méléard [17] in 1993, there have been few rigorous mathematical results on this important class of models since the appearance of [16]. (See Section 4.3 of Kelly [20].)

We now introduce the mathematical framework used to study these two classes of algorithms.

Mathematical context: mean-field convergence

It is assumed that requests arrive at each of the $N$ nodes according to a Poisson process with rate $\lambda$. Each node has a finite capacity $C$. The sojourn times of the requests at the node are exponentially distributed, with rate $\mu_1$ when they are accepted without rerouting, and with rate $\mu_2 \in (0, \mu_1]$ for the RIST algorithm if they have been rerouted. If the state of the $i$th node, $1 \leq i \leq N$, at time $t \geq 0$ is given by $Z^N_i(t)$, then, for both classes of algorithms, the process of empirical distribution $(\Lambda^N(t))$, with

$$\Lambda^N(t) = \frac{1}{N} \sum_{i=1}^{N} \delta_{Z^N_i(t)}, \quad t \geq 0,$$

where $\delta_a$ is the Dirac mass at $a$, has the Markov property. The state space $\mathcal{X}$ of the process $(Z^N_i(t))$ is finite:

$$\mathcal{X} = \{ (x, y) \in \mathbb{N}^2 : x + y \leq C \} \quad \text{(RIST)},$$

$$\mathcal{X} = \{ 0, \ldots, C \} \quad \text{(DAR)}.$$

A node in state $(x, y) \in \mathcal{X}$ for the RIST algorithm has $x$ jobs accepted without rerouting and $y$ rerouted jobs from another node. Consequently, $\Lambda^N(t)$, the state of a random node, belongs to $\mathcal{P}(\mathcal{X})$, the set of probability distributions on $\mathcal{X}$.

Under some mild conditions on the initial state, and with some restrictions for the RIST algorithm (see Subsection 2.1), it can be shown that the sequence of stochastic processes $(\Lambda^N(t))$ converges in distribution to a deterministic process $(\Lambda(t))$. As a consequence, the propagation-of-chaos property holds: in the limit, the states of a finite subset of nodes become independent. (See Sznitman [31].) It can be shown that $(\Lambda(t))$ is the solution of a nonlinear Fokker-Planck equation (see Frank [15], for example):

$$\frac{d}{dt} \Lambda(t) = \Lambda(t) \cdot Q_\Lambda(t),$$

where, for $m \in \mathcal{P}(\mathcal{X})$, $Q_m$ is the Q-matrix of the reversible irreducible Markov process with the invariant distribution $\pi_m$. For example, for the DAR algorithm, $Q_m$ is the Q-matrix of an $M/M/C/C$ queue with service rate $\mu_1$ and arrival rate $\lambda h(m(C))$, where $h$ is a quadratic function. (See (49) below.)

We now review the main problems in this context.

1.1. Existence, number, and locations of equilibrium points.

An element $m \in \mathcal{P}(\mathcal{X})$ is an equilibrium point of the dynamical system (2) if $m = \pi_m$, or, equivalently,

$$m \cdot Q_m = 0.$$
This equation does not in general have an explicit solution, and, worse, it is even quite difficult to determine the number of solutions. For example, for the DAR algorithm, Gibbens et al. [16] have made the striking observation that, for some specific parameters, numerical experiments seem to indicate that there may be three solutions. But to the best of our knowledge, this statement does not seem to have been established in a more formal way. (See p. 375 of Hunt and Kurtz [18].)

It should be noted that, in the experiments of Gibbens et al. [16], the authors were able to find convenient numerical values for the ratio of the average load per node to capacity, such that the associated dynamical system has three equilibrium points. This is remarkable, since, as we will prove in Section 3, this phenomenon occurs only if this ratio lies in an interval of width .063!

Equation (3) can be reduced to a polynomial equation of degree C involving partial sums of the exponential series. This has some (formal) similarities with the celebrated Erlang fixed point equation for loss networks. In this case there is also an asymptotic independence property, but it is due to a stochastic averaging principle rather than a mean-field convergence. (See Kelly [20].)

For Erlang systems, this part of the study seems to rely more on analytic methods than probabilistic arguments. This is probably one of the difficulties of these problems: little intuition can be extracted, a priori, from these polynomial equations. See Antunes et al. [2], Dawson [10], Muzychka [27], and Rybko and Shlosman [30] for the analysis of other ‘large’ queueing models.

1.2 Stability properties of equilibrium points.

Concerning the properties of equilibrium points of the dynamical system (2), there are two overlapping aspects.

THE CONVERGENCE OF (∆(t)) TO THE EQUILIBRIUM POINT.

Even when Equation (3) has a unique solution, i.e. the nonlinear dynamical system has a unique equilibrium distribution, proving the convergence of the dynamical system (∆(t)) is a challenging issue. The dynamical system (∆(t)) is associated to a nonlinear Markov process (Z(t)) with values in \( \mathcal{X} \); due to the time-inhomogeneity of these dynamics, the classical results on convergence of Markov processes cannot be used.

For a large class of examples of nonlinear diffusion processes, related to the Langevin evolution equation, there are nevertheless numerous results concerning the rate of convergence to equilibrium. Furthermore, an exponential decay is proved with explicit bounds for the Wasserstein distance between two solutions starting from different initial states. Several key ingredients are used in this context: some geometric properties related to curvature to prove an exponential decay of the time evolution of the relative entropy with respect to the equilibrium measure, and some functional inequalities. (See Carrillo et al. [7] and references therein.)

For discrete state spaces but in a time-homogeneous setting, Caputo et al. [6] and Dai Pra and Posta [9] have adapted some of the methods of the diffusion framework to get explicit bounds on the rate of exponential convergence to equilibrium. Erbar and Maas [12] and Maas [24] have recently developed some tools, the analogue of the geometric characteristics used in the diffusive case, to use in a general approach to these problems in a discrete state space. Some interesting but specific examples of random walks have been already investigated with these methods; see for example Erbar et al. [13]. Their use in practice, to get explicit constants on the exponential rate of convergence to equilibrium, is, as can be expected, limited for the
moment. In a nonlinear setting, examples are even more rare. See Thai [32], which investigates the case of birth-and-death processes whose birth and death rates satisfy a convexity relation.

THE LOCAL STABILITY OF AN EQUILIBRIUM.

Given a solution $m_0$ of Equation (3), the stability property of $m_0$ is the fact that if the initial point of the dynamical system $(\Lambda(t))$ defined by Equation (2) is in a sufficiently small neighborhood of $m_0$, then $(\Lambda(t))$ converges to $m_0$, possibly exponentially fast. This is also important from the point of view of qualitative properties of the algorithms, since it asserts that the equilibrium point is meaningful. It suggests that there is a set of states where the network will stay ‘for some time’. A more ambitious goal would be to determine also the basin of attraction of the stable points, i.e. the set of initial states from which they can be reached. We are, in fact, far from that here.

Unfortunately, even in a simpler setting, when there is a unique equilibrium, it is surprisingly difficult to prove such a stability property for the models of this paper. The case of the nonlinear $M/M/1$ process in Subsection 3.4 is striking from this point of view. In a finite state space, if $m_0$ is an equilibrium point and if the eigenvalues of a Jacobian matrix associated to Equation (2) at $m_0$ have negative real part, then, under some mild regularity conditions, $m_0$ is a locally stable equilibrium point by the Poincaré–Lyapunov theorem. (See Section 7.1 of Verhulst [35], for example.) For a fixed capacity $C$, where this theorem could a priori be applied, we have not been able to obtain explicit stability results. Furthermore, in an infinite-dimensional context, like the nonlinear $M/M/1$ queue analyzed in Section 3, this is a more delicate phenomenon to analyze; the choice of the norm between probability distributions, in particular, has an important impact.

In the stochastic networks literature, proofs of local stability of specific examples are rare. Antunes et al. [2] establishes, through a dimension reduction, a criterion of local stability for equilibrium points of an Erlang network where jobs move from one node to another after completing their services. See also Tibi [33] and Budhiraja et al. [5] for a discussion on this topic.

STABILITY VIA LYAPUNOV FUNCTIONS.

In some cases, dynamical systems associated to Equation (2) may admit a Lyapunov function, i.e. a function $F: \mathcal{P}(\mathbb{N}) \to \mathbb{R}_+$ such that

$$\frac{d}{dt} F(\Lambda(t)) \bigg|_{t=0} < 0,$$

with $\Lambda(0) = m \in \mathcal{P}(\mathbb{N})$, and $F(m) \neq 0$. In the case of a unique equilibrium, the existence of such a function may give the desired convergence to equilibrium. An explicit representation of such a function $F$ contains in fact a lot of information about the associated dynamical system. The state space being the set of probability distributions on a finite or countable set, examples with a Lyapunov function in such a context are quite rare. Antunes et al. [2] has such a function for Erlang networks with a Jackson-type routing. The Lyapunov function is expressed in terms of a relative entropy with respect to the invariant distribution of some single Erlang queue and a complementary term related to the nonlinear dynamics.

Tibi [33] and Budhiraja et al. [5] have investigated the conditions under which a Lyapunov function based on relative entropy can be constructed. It turns out that, in practice, the possibilities are in fact limited among ‘classical’ models. Both papers mention the fact that their respective (equivalent) conditions ‘local balance’ or ‘local Gibbs’ for having a Lyapunov
function cannot hold for the DAR algorithm. This seems also to be the case for the RIST algorithm. This may partially explain the lack of progress in the mathematical analysis of the DAR algorithm since the appearance of the original article [16].

1.3. Contributions

The original motivation of this paper can be summarized as follows: In the absence of Lyapunov functions, how can we study the equilibrium properties of these algorithms? We have used a set of quite diverse techniques.

1. Scaling Methods for the Number of Equilibrium Points.

For the RIST algorithm with an infinite number of retrials, we are able to determine exactly the set of equilibrium points of the dynamical system associated to the mean-field limit of this system. (See Propositions 3 and 7.) Because of a singularity in the coefficients of the asymptotic dynamical system, a solution may die in finite time. It turns out that there may be one, two, or three equilibria and that one of them is related to this singularity. For some of the equilibrium points, the state of the network has a significant fraction of rerouted requests; see Proposition 7 of Subsection 2.4 for a more precise description. In Subsection 2.3, we show that, for a variant of the RIST algorithm with one retrial, the corresponding system of ordinary differential equations (ODEs) does not have any singularities, but there may still be three equilibrium points.

For the DAR algorithm, by taking the global input rate proportional to the capacity $C$, $\lambda = \nu C$ for some $\nu$, it is shown that, with a convenient scaling, as $C$ gets large, the dynamical system converges to a system of ODEs in an infinite-dimensional state space. This asymptotic dynamical system is described in terms of a nonlinear $M/M/1$ queue; see Subsection 3.3.

With this scaling analysis of the equilibrium equations, we can give conditions on the parameters of the network so that for a fixed, but sufficiently large, capacity $C$, there are three equilibrium points, as was suggested and conjectured by Gibbens et al. [16] but, to the best of our knowledge, has not been proved until now. Additionally, the limiting values of these equilibrium points when $C$ goes to infinity are identified. (See Theorem 2.)

2. Probabilistic Approach to Prove Stability.

We show the stability of some of the equilibrium points associated to congested regimes of the RIST and DAR algorithms. Coupling methods are used, by constructing an ad-hoc order relation and by deriving several technical estimates. See Propositions 7 and 9.

3. Spectral Approach to Exponential Stability.

For the other equilibrium points, the stability problem is much harder. They are defined via the solutions of some polynomial equations, which does not give much insight into their significance. This is where, when available, a Lyapunov function is useful. We have tried to derive a stability criterion in terms of the spectral gap $\kappa$ of the linear Markov process associated to the equilibrium point. This quantity is known for some classical models, such as the Erlang queue or the $M/M/1$ queue. (See Chen [8] and van Doorn [34], for example.) The (rough) idea is that if the rate of convergence to equilibrium is sufficiently large, then the nonlinear perturbations will not take the trajectory away from the neighborhood of the equilibrium point. Unfortunately this intuitive picture is difficult to establish rigorously; see Theorem 3 for an example of such a situation. By using an $L_2$-norm and with several estimates, we have nevertheless been able to establish some
exponential stability criteria in terms of the spectral gap for the RIST algorithm and for the nonlinear $M/M/1$ queueing model of the DAR algorithm. This is a surprisingly difficult problem, even when the dynamical system has a unique fixed point. A good example is the nonlinear $M/M/1$ queue of Subsection 3.4 with $\nu > 1$. Only a partial result for this case has been obtained.

2. The Routing with Increased Sojourn Time (RIST) algorithm

For $i \in \{1, \ldots, N\}$, requests arrive at node $i$ with capacity $C$ according to a Poisson process with rate $\lambda$. If node $i$ is not full at one of these instants, the corresponding request is accepted and its sojourn time is exponentially distributed with rate $\mu_1$; such a request will be referred to as a class 1 customer. If node $i$ is full, another node is picked at random. If this node is not saturated, it allocates the request. Otherwise, another node is picked at random. The maximum number of attempts is limited to $p_0 \in \mathbb{N}\cup\{+\infty\}$; if no attempt is successful, the request is rejected. We will mostly investigate the case $p_0 = +\infty$. The sojourn time of rerouted requests is exponentially distributed with rate $\mu_2 < \mu_1$; these are referred to as class 2 customers.

2.1. The associated dynamical system

We introduce some notation used throughout this section. The state space of a node is given by $\mathcal{X}$, defined as follows:

$$
\mathcal{X} = \{z = (x, y) \in \mathbb{N}^2 \mid x + y \leq C\},
$$

$$
\mathcal{X}_+ = \{z = (x, y) \in \mathbb{N}^2 \mid x + y < C\}, \text{ and } \mathcal{X}_+^N = \mathcal{X} \setminus \mathcal{X}_+ = \{z = (x, y) \in \mathbb{N}^2 \mid x + y = C\}.
$$

If $z = (x, y) \in \mathcal{X}$, $x$ (resp. $y$) is the number of class 1 (resp. class 2) customers.

The set of probability distributions on $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$, and $C(\mathbb{R}_+, \mathcal{P}(\mathcal{X}))$ is the set of continuous functions with values in $\mathcal{P}(\mathcal{X})$. If $\zeta \in \mathcal{P}(\mathcal{X})$, $f$ is a real-valued function on $\mathcal{X}$, and $A \subset \mathcal{X}$, we write

$$
(\zeta, f) = \int f(z) \zeta(\mathrm{d}z) \text{ and } \zeta(A) = \int_A \zeta(\mathrm{d}z).
$$

The state space of the process describing the whole network is $S_N \overset{\text{def.}}{=} \mathcal{X}^N$. For $t \geq 0$ and $i \in \{1, \ldots, N\}$, $(X_i^N(t))$ (resp. $(Y_i^N(t))$) denotes the number of class 1 (resp. class 2) customers at node $i$.

For $\mu \in \{\mu_1, \mu_2\}$, $(\mathcal{N}^{ij}_\mu)$ is an independent, identically distributed (i.i.d.) sequence of Poisson processes with rate $\mu$. They are associated to the service times of class 1 and 2 customers at the nodes of the networks. For $i, j \geq 1$, $\mathcal{N}^{ij}_\mu$ is the Poisson process associated to the service times of $j$th server of the $i$th node. Similarly, $(\mathcal{N}^{ij}_\lambda)$ is an i.i.d. sequence of Poisson processes with rate $\lambda$. For $i \in \mathbb{N}$, $\mathcal{N}^{ij}_\lambda$ is the arrival process at node $i$. All Poisson processes are assumed to be independent. Additionally, $(U_{nk}^i)$ is an i.i.d. sequence of uniform random variables on $[0, 1]$, also referred to as ‘marks’ of the point process $\mathcal{N}^{ij}_\lambda$ defined by

$$
\mathcal{N}^{ij}_\lambda \overset{\text{def.}}{=} \sum_{n \in \mathbb{N}} \delta_{(\mu_\lambda(U_{nk}^i \geq 1))},
$$

$\mathcal{N}^{ij}_\lambda$ is a marked Poisson point process. (See Chapter 5 of Kingman [21].) If $\mathcal{M} \overset{\text{def.}}{=} [0, 1]^\mathbb{N}$ is the space of marks, then $\mathcal{N}^{ij}_\lambda(\mathrm{d}t) = \mathcal{N}^{ij}_\lambda(\mathrm{d}t, \mathcal{M})$ is the arrival process at node $i$. 

A mark is associated to an arrival instant; it is used to determine to which queue a customer goes if the node where it arrives is saturated. For node $i$, this is (formally) done with a functional $T^N_i$ on $\mathcal{S} \times \mathcal{M}$ defined as follows. For $(z, u) = ((x_i, y_i), (u_j)) \in \mathcal{S} \times \mathcal{M}$, define $(v_j(u)) = ([N u_j])$, with 0 identified to $N$. For $i$, $n \in \mathbb{N}$, $(v_j(U^i_n))$ is an i.i.d. sequence uniformly distributed on $\{1, \ldots, N\}$. With this notation, we take

$$T^N_i (z, u) = \begin{cases} i & \text{if } x_i + y_i < C, \\ v_k(u) & \text{otherwise, if } k = \inf\{j \mid x_{v_j(u)} + y_{v_j(u)} < C\} \leq p_0, \\ +\infty & \text{otherwise.} \end{cases}$$

(5)

The variable $T^N_i (z, u)$ will be used in what follows only when node $i$ is saturated.

When the system is in a state $z \in \mathcal{S}_N$ such that $x_i + y_i = C$ for all $i \in \{1, \ldots, N\}$, if there is an arrival with a mark $u$ at node $i$, then, provided that $T^N_i (z, u)$ is less than $p_0$, the queue with index $T^N_i (z, u)$ receives the rerouted customer. When $p_0 = +\infty$, if $x_j + y_j < C$ for some $j \in \{1, \ldots, N\}$, then it is easy to check that

$$\mathbb{P}(T_i (z, U) = j) = 1 - \sum_{k=1}^{N} \mathbf{1}_{\mathcal{X}_+}(x_k, y_k),$$

where $\mathcal{X}_+$ is defined in (4).

If the addition of marks looks somewhat formal, it has the advantage of giving a neat framework to handle the stochastic calculus associated to the evolution equations of the system. The martingale property will be with respect to the filtration $(\mathcal{F}_t)$ defined as follows: for $t \geq 0$,

$$\mathcal{F}_t = \sigma \left\{ \overline{\mathcal{X}}_i^j ([0, s] \times B), \mathcal{N}_{\mu_1}^j, \mathcal{N}_{\mu_2}^j, i, j \in \mathbb{N}, s \leq t, B \in \mathcal{B}(\mathcal{M}) \right\}.$$

**Evolution equations.** We assume from now on that $p_0 = +\infty$; that is, a customer is rejected only if all nodes are saturated. We consider briefly the case of a finite $p_0$ in Proposition 2 and in Subsection 2.3.

The state of the system is represented by a process $(Z^N(t)) = ((X^N_i(t), Y^N_i(t)))$ which is càdlàg and satisfies the following stochastic differential equations (SDE): for $1 \leq i \leq N$,

$$dX^N_i(t) = \mathbf{1}_{\{(X^N_i(t-), Y^N_i(t-)) \in \mathcal{X}_+\}} \mathcal{N}^i_{\mu_1}(dt) - \sum_{\ell=1}^{\infty} \mathcal{N}^{i, \ell}_{\mu_1}(dt),$$

(6)

$$dY^N_i(t) = \sum_{j=1}^{N} \mathbf{1}_{\{(X^N_i(t-), Y^N_i(t-)) \in \mathcal{X}^+_i\}} \mathcal{N}^j_{\lambda}(dt, du) - \sum_{\ell=1}^{\infty} \mathcal{N}^{i, \ell}_{\mu_2}(dt),$$

(7)

where $U(t-)$ denotes the left limit of $U$ at $t > 0$.

The first term of the right-hand side of (7) corresponds to arrivals that find their arrival node saturated and are therefore allocated to some non-saturated node by means of repeated random
sampling of nodes until one is found that can accommodate it. Note that it is not excluded that all nodes are saturated; in this case the customer is rejected since it is not allocated anywhere.

Martingales. We recall classical results on the martingales associated to marked Poisson point processes. (See Sections 4.4 and 4.5 of Jacobsen [19], for example. See also Last and Brandt [22].) If \( h \) is a bounded function on \( \mathbb{R}_+ \times \{1, \ldots, N, +\infty\} \) such that \( h(\cdot, +\infty) \equiv 0 \), and \( \Lambda^N \) is the empirical distribution defined by Equation (1), the process

\[
\left( \int_{[0,t] \times \mathcal{M}} h(Z^N(s), T^N_i(Z^N(s), u)) \, \overline{N}_\lambda^t(\, ds, du) \right.
- \lambda \sum_{j=1}^N \int_0^t \frac{h(Z^N(s), j)}{N \Lambda^N(s)(\mathcal{X}_+)} \mathbf{1}_{\mathcal{X}_+} (X^N_j(s), Y^N_j(s)) \, ds
\]

is a martingale whose previsible increasing process (quadratic variation) is given by

\[
\left( \lambda \sum_{j=1}^N \int_0^t \frac{h^2(Z^N(s), j)}{N \Lambda^N(s)(\mathcal{X}_+)} \mathbf{1}_{\mathcal{X}_+} (X^N_j(s), Y^N_j(s)) \, ds \right).
\]

For example, the integration of SDE (7) and the compensation of the Poisson processes give the relation

\[
Y^N_i(t) = Y^N_i(0) + M^N_i(t)
+ \lambda \sum_{j=1}^N \int_0^t \frac{1}{N \Lambda^N(s)(\mathcal{X}_+)} \mathbf{1}_{\mathcal{X}_+} (X^N_j(s), Y^N_j(s)) \, ds - \mu_2 \int_0^t Y^N_i(s) \, ds,
\]

which can be written in the more compact form

\[
Y^N_i(t) = Y^N_i(0) + M^N_i(t)
+ \lambda \int_0^t \frac{1}{N \Lambda^N(s)(\mathcal{X}_+)} \Lambda^N(s)(\mathcal{X}_+) \, ds - \mu_2 \int_0^t Y^N_i(s) \, ds,
\]

(8)

where \( (M^N_i(t)) \) is a martingale whose previsible increasing process is

\[
\left( [M^N_i(\cdot)](t) \right) = \left( \lambda \int_0^t \frac{1}{N \Lambda^N(s)(\mathcal{X}_+)} \Lambda^N(s)(\mathcal{X}_+) \, ds + \mu_2 \int_0^t Y^N_i(s) \, ds \right).
\]

(9)

Empirical distributions. The empirical distribution process \( (\Lambda^N(t)) \) associated to \( (Z^N(t)) \) is defined as follows: for \( t \geq 0 \),

\[
\Lambda^N(t)(f) = \int_{\mathcal{X}} f(z) \Lambda^N(t)(\, dz) \overset{\text{def.}}{=} \frac{1}{N} \sum_{i=1}^N \int f(X^N_i(t), Y^N_i(t)) \, dz
\]

(10)

for any nonnegative function \( f \) on \( \mathcal{X} \). It is a stochastic process with values in the set \( \mathcal{P}(\mathcal{X}) \) of probability distributions on \( \mathcal{X} \).
As in the derivation of (8), the integration of (6) and (7) and the compensation of Poisson processes give the relation
\[
\{\Lambda^N(t), f\} = \{\Lambda^N(0), f\} + M^N_f(t)
\]
\[
+ \lambda \int_0^t \int_{X_+} \nabla_1^+(f)(z) \Lambda^N(s)(dz) \ ds - \mu_1 \int_0^t \int_{z = (x, y) \in X} x \nabla_1^-(f)(z) \Lambda^N(s)(dz) \ ds
\]
\[
+ \lambda \int_0^t \int_{X_+} \nabla_2^+(f)(z) \frac{\Lambda^N(s)(\lambda^c)}{\Lambda^N(s)(\lambda^c)} \Lambda^N(s)(dz) \ ds
\]
\[
- \mu_2 \int_0^t \int_{z = (x, y) \in X} y \nabla_2^-(f)(z) \Lambda^N(s)(dz) \ ds
\]
for a real-valued function \(f\) on \(X_+\), where, if \(z = (x, y)\), then \(\nabla_1^\pm(f)(z) = f(x \pm 1, y) - f(x, y)\) and \(\nabla_2^\pm(f)(z) = f(x, y \pm 1) - f(x, y)\) are the gradient operators, and \((M^N_f(t))\) is a martingale.

**Proposition 1.** (Dynamical system with an unbounded number of retrials) If \(p_0 = +\infty\) and \((\Lambda(t))\) is the unique solution of the differential equation
\[
\frac{d}{dt} \langle \Lambda(t), f \rangle = \lambda \langle \Lambda(t), \nabla_1^+(f) \mathbf{1}_{X^c_+} \rangle + \mu_1 \langle \Lambda(t), I_1 \nabla_1^-(f) \rangle
\]
\[
+ \lambda \frac{\Lambda(t)(\lambda^c)}{\Lambda(t)(\lambda^c)} \langle \Lambda(t), \nabla_2^+(f) \mathbf{1}_{X^c_+} \rangle + \mu_2 \langle \Lambda(t), I_2 \nabla_2^-(f) \rangle
\]
for \(t < H_0(\Lambda)\), where \(I_1(x, y) \equiv x\) and \(I_2(x, y)\) \(\equiv y\), and
\[
H_0(\xi) \equiv \inf \{t > 0 \mid \xi(t)(\lambda^c) = 0\} \quad \text{for } (\xi(t)) \in C(\mathbb{R}_+, \mathcal{P}(X)),
\]
then, for the convergence in distribution of processes,
\[
\lim_{N \to +\infty} \langle \Lambda^N(t), t < H_0(\Lambda^N) \rangle = (\Lambda(t), t < H_0(\Lambda)).
\]

Note that the sequence of processes \((\Lambda^N(t))\) is in fact a sequence of finite-dimensional processes with dimension \(\text{card}(X_+)\). The convergence in distribution of the proposition refers to the case when the space of càdlàg functions with values in \(\mathbb{R}_+^{\text{card}(X_+)}\) is endowed with the uniform norm.

The variable \(H_0(\Lambda)\) is the blow-up time of the dynamical system (12). If finite, it amounts to the fact that the system is completely saturated at the fluid scale. It will be seen in Subsection (2.4) that the saturated state is a stable equilibrium of the network. Note that, because of its singular aspect, it cannot be really defined through the ODEs associated to (12).

**Proof.** For the martingale \((M^N_f(t))\) of (11), using calculations similar to the ones used in the derivation of (9), one gets the existence of a constant \(K_T\) such that \(E(\langle M^N_f(T) \rangle) < K_T/N\), as well as, by Doob’s inequality, the convergence in distribution to (0) of this martingale for the topology associated to uniform convergence on \([0, T]\).

Note that, for \(s < t\),
\[
\left| \int_s^t \int_{X_+} \nabla_2^+(f)(z) \frac{\Lambda^N(u)(\lambda^c)}{\Lambda^N(u)(\lambda^c)} \Lambda^N(u)(dz) \right| \leq 2\|f\|_\infty (t - s).
\]
By using the criterion of the modulus of continuity and Equation (11), we get that the sequence of processes \((\Lambda_N(t))\) is tight in distribution for the topology of uniform convergence on compact sets. (See Theorem 7.3 of Billingsley [3], for example.)

The solution of (12) lives in a finite-dimensional state space, the set of probability distributions on \(\mathcal{X}\). The system (12) can be seen as a set of ODEs. It has in particular a unique solution up to time \(H_0(\Lambda)\). If
\[
H_\varepsilon(\xi) = \inf\{t > 0 \mid \xi(t)(\mathcal{X}_+) > \varepsilon\}
\]
then \(H_\varepsilon(\Lambda_N)\) converges in distribution to \(H_\varepsilon(\Lambda)\). By the continuous mapping theorem used in (11), one sees that on the event \(\{H_\varepsilon > t\}\), the relation
\[
\langle \Lambda(t), f \rangle = \langle \Lambda(0), f \rangle + \lambda \int_0^t \langle \Lambda(s), \nabla_1^+(f) 1_{\mathcal{X}_+} \rangle \, ds + \mu_1 \int_0^t \langle \Lambda(s), I_1 \nabla_1^-(f) \rangle \, ds
\]
\[
+ \lambda \int_0^t \frac{\Lambda(s)(\mathcal{X}_+^c)}{\Lambda(s)(\mathcal{X}_+)} \langle \Lambda(s), \nabla_2^+(f) 1_{\mathcal{X}_+} \rangle \, ds + \mu_2 \int_0^t \langle \Lambda(s), I_2 \nabla_2^-(f) \rangle \, ds
\]
holds. These equations can be seen as a system of ODEs which clearly has a unique solution up to blow-up time \(H_0\). The proposition is proved.

We state the analogous result when the number of retrials is finite; the notation is the same as in Proposition 1. The proof of the proposition being simpler in this case, it is skipped.

**Proposition 2.** (Dynamical system with a maximum of \(p_0\) retrials) If \((\Lambda(t))\) is the unique solution of the differential equations
\[
\frac{d}{dt} \langle \Lambda(t), f \rangle = \lambda \langle \Lambda(t), \nabla_1^+(f) 1_{\mathcal{X}_+} \rangle + \mu_1 \langle \Lambda(t), I_1 \nabla_1^-(f) \rangle
\]
\[
+ \lambda \Lambda(t)(\mathcal{X}_+^c) \left( \frac{1 - [\Lambda(t)(\mathcal{X}_+^c)]^{p_0}}{1 - \Lambda(t)(\mathcal{X}_+^c)} \right) \langle \Lambda(t), \nabla_2^+(f) 1_{\mathcal{X}_+} \rangle + \mu_2 \langle \Lambda(t), I_2 \nabla_2^-(f) \rangle
\]
for \(t > 0\), then, for the convergence in distribution of processes, the relation
\[
\lim_{N \to +\infty} \langle \Lambda_N(t) \rangle = \langle \Lambda(t) \rangle
\]
holds.

The nonlinear term in the right-hand side of (13) is bounded by \(p_0\), and is therefore without singularity.

**2.2. The number of equilibrium points**

We investigate the fixed points of the linearized version \((\Lambda_R(t))\) of the dynamical system (12) when the nonlinear term \(\Lambda(t)(\mathcal{X}_+^c)\) is replaced by a constant \(R \in (0, 1)\). This dynamical system satisfies the following equations:
\[
\langle \Lambda_R(t), f \rangle = \langle \Lambda_R(0), f \rangle + \lambda \int_0^t \langle \Lambda_R(s), \nabla_1^+(f) 1_{\mathcal{X}_+} \rangle \, ds
\]
\[
+ \mu_1 \int_0^t \langle \Lambda_R(s), I_1 \nabla_1^-(f) \rangle \, ds + \lambda R \int_0^t \langle \Lambda_R(s), \nabla_2^+(f) 1_{\mathcal{X}_+} \rangle \, ds
\]
\[
+ \mu_2 \int_0^t \langle \Lambda_R(s), I_2 \nabla_2^-(f) \rangle \, ds
\]
(14)
for any real-valued function $f$ on $\mathcal{X}$, with the notation of Proposition 1. Recall that for any function $f$ and measure $\mu$ on $\mathcal{X}$,

$$
\langle \mu, f \rangle = \sum_{(x, y) \in \mathcal{X}} f(x, y) \mu((x, y)).
$$

The process $(\Lambda(t))$ describes the evolution of the law of a classical Erlang model with capacity $C$ where two classes of customers arrive at rates $\lambda$ and $\lambda R/(1 - R)$ respectively and are served at rates $\mu_1$ and $\mu_2$. Its invariant distribution on $\mathcal{X}$ is given by

$$
\pi_R(x, y) = \frac{1}{Z_R} \rho_1^x \rho_2^y \frac{R^y}{1 - R}, \quad (x, y) \in \mathcal{X},
$$

where $\rho_1 = \lambda / \mu_1$ and $\rho_2 = \lambda / \mu_2$, and $Z_R$ is the normalization constant

$$
Z_R = \sum_{m=0}^{C} \sum_{x+y=m} \frac{\rho_1^x \rho_2^y}{x! \cdot y!} \frac{R^y}{(1 - R)} = \sum_{m=0}^{C} \frac{1}{m!} \left( \rho_1 + \rho_2 \frac{R}{1 - R} \right)^m.
$$

We have

$$
\pi_R(\mathcal{X}_+^C) = \pi_R((x, y) \in \mathcal{X}, x + y = C) = \frac{1}{Z_R} \frac{1}{C!} \left( \rho_1 + \rho_2 \frac{R}{1 - R} \right)^C.
$$

Hence $\pi_R$ as defined by (15) is a fixed point of the dynamical system (12) if and only if $R$ satisfies the relation

$$
R = \pi_R(\mathcal{X}_+^C). \tag{16}
$$

The goal of this section is to characterize completely the solutions of (16).

Note that (16) can be rewritten as $\Phi_{\rho_1, \rho_2}(z_R) = 0$, with, for $\rho_1, \rho_2 > 0$,

$$
\Phi_{\rho_1, \rho_2}(z) \overset{\text{def}}{=} \frac{1}{z} \frac{(\rho_1 + \rho_2 z)^C}{C!} - \sum_{m=0}^{C-1} \frac{(\rho_1 + \rho_2 z)^m}{m!}, \tag{17}
$$

and $z_R = R/(1 - R)$. The next proposition determines the number of roots of the function $\Phi_{\rho_1, \rho_2}$. It gives all the nonsingular equilibrium points of the ODEs (12). Subsection 2.4 gives a formal presentation of another regime which yields a stable equilibrium when $\rho_2 > C$; see the remark below.

**Proposition 3.** (Nonsingular equilibrium points.) Assume $C \geq 2$ and $\rho_1 < C$.

1. For $\rho_2 \in [0, C)$, there is a unique root $z(\rho_1, \rho_2) > 0$ of $\Phi_{\rho_1, \rho_2}$. The function $z(\rho_1, \cdot) : [0, C) \rightarrow \mathbb{R}_+$, $\rho_2 \mapsto z(\rho_1, \rho_2)$ is increasing. If

$$
z(\rho_1, C) \overset{\text{def}}{=} \lim_{\rho_2 \nearrow C} z(\rho_1, \rho_2),
$$

then $z(\rho_1, C)$ is the unique root of $\Phi_{\rho_1, C}$ if $\rho_1 < C - 1$, and $z(\rho_1, C) = +\infty$ otherwise.

2. For $\rho_2 > C$, there exists a non-increasing function $\phi_C : (C, +\infty) \rightarrow (0, C - 1)$ such that

$$
\lim_{z \rightarrow +\infty} \phi_C(z) = 0, \lim_{z \searrow C} \phi_C(z) = C - 1,
$$
The equilibrium states of large Erlang networks

exists a unique \( z \)

\[ \text{if } \rho_1 \in (0, \phi_C(\rho_2)), \Phi_{\rho_1, \rho_2} \text{ has two roots in } \mathbb{R}_+; \]

\[ \text{if } \rho_1 \in (\phi_C(\rho_2), C), \Phi_{\rho_1, \rho_2} \text{ does not have any roots.} \]

In the case \( \rho_1 = \phi_C(\rho_2) \), \( \Phi_{\rho_1, \rho_2} \) has a unique root.

Remarks

- The location of \( \rho_2 \) with respect to \( C \) and of \( \rho_1 \) with respect to \( \phi_C(\rho_2) \) determines the number of solutions to (16). The function \( \phi_C \) can in fact be (formally) defined by

\[ \phi_C(\rho_2) \overset{\text{def.}}{=} \sup\{\rho_1 : m(\rho_1, \rho_2) < 0\}, \text{ with } m(\rho_1, \rho_2) \overset{\text{def.}}{=} \min\{\Phi_{\rho_1, \rho_2}(z) : z \geq 0\}. \quad (18) \]

- As will be seen in Subsection 2.4, when \( \rho_2 > C \) there is another, singular, equilibrium which is not mentioned in this proposition; it corresponds to \( R = 1 \). Starting from some suitable initial states, the dynamical system (12) converges to the Dirac measure \( \delta_{(0,C)} \); i.e. most of the nodes are saturated. This situation corresponds to the case where, in the limit, all requests are rerouted or rejected. Mathematically, this is a consequence of the possibility that the dynamical system (12) may blow up in finite time, i.e. may degenerate.

- Note that the case \( C = 1 \) analyzed in Malyshev and Robert [25] does not exhibit multiple equilibria.

We will refer to the different cases of the proposition as, respectively, underloaded if \( \rho_2 < C \), critical if \( \rho_2 = C \), and overloaded if \( \rho_2 > C \).

Proof. Once the results to be proved have been properly formulated, the proofs of the statements will be done via real analysis. First, note that

\[ \lim_{z \downarrow 0} z \Phi_{\rho_1, \rho_2}(z) = \frac{\rho_1^C}{C!} > 0 \text{ and } \lim_{z \to +\infty} \frac{\Phi_{\rho_1, \rho_2}(z)}{z^{C-1}} = \frac{\rho_2^{C-1}}{(C-1)!} \left( \frac{\rho_2}{C} - 1 \right). \quad (19) \]

The underloaded case \( \rho_2 \in (0, C) \)

Define, for \( z \geq 0 \),

\[ f_{\rho_1, \rho_2}(z) = \Phi_{\rho_1, \rho_2}(z)e^{-\rho_2 z}. \]

After some simple calculations with telescoping sums, we get that

\[ f'_{\rho_1, \rho_2}(z) = \frac{1}{z^2} \frac{(\rho_1 + \rho_2 z)^{C-1}}{C!} e^{-\rho_2 z} \left( \rho_2 (C - \rho_2) z^2 + \rho_2 (C - \rho_1 - 1) z - \rho_1 \right). \quad (20) \]

The last term of the right-hand side of the expression for \( f'_{\rho_1, \rho_2}(z) \) is a polynomial of degree 2. Its value at \( z = 0 \) is negative and, since \( \rho_2 < C \), it converges to \( +\infty \) as \( z \) gets large; hence there exists a unique \( z_0 > 0 \) such that \( f'_{\rho_1, \rho_2}(z_0) = 0 \). It necessarily corresponds to a unique extremum of \( f_{\rho_1, \rho_2} \) on \( \mathbb{R}_+ \), which is a minimum, given the variations of \( f_{\rho_1, \rho_2} \). (See (19).) The function \( f_{\rho_1, \rho_2} \) is decreasing on \((0, z_0)\) and increasing on \([z_0, +\infty)\), and since \( f_{\rho_1, \rho_2} \) converges to 0 at infinity, the relation \( f_{\rho_1, \rho_2}(z) < 0 \) holds for \( z \geq z_0 \). Thus, there exists a unique root \( z(\rho_1, \rho_2) \)
of \( f_{\rho_1, \rho_2} \) in \( \mathbb{R}_+ \), located in \((0, z_0)\). We have therefore the equivalence of the two relations \( z < z(\rho_1, \rho_2) \) and \( \Phi_{\rho_1, \rho_2}(z) > 0 \). The latter relation can be expressed as

\[
z \sum_{m=0}^{C-1} \frac{C!}{m! (\rho_1 + \rho_2 z)^{C-m}} < 1.
\]

Hence, if \( \rho_2' > \rho_2 \), then \( \Phi_{\rho_1, \rho_2'}(z) > 0 \) holds if \( z < z(\rho_1, \rho_2) \). The function \( z(\rho_1, \cdot) : \mathbb{R}_+ \rightarrow \mathbb{R}_+ \), \( \rho_2 \mapsto z(\rho_1, \rho_2) \), is therefore increasing.

**The critical case \( \rho_2 = C \)**

We easily get the relation

\[
\lim_{z \rightarrow +\infty} \frac{z^2 \Phi_{\rho_1, C}(z)}{(\rho_1 + C z)^C} = \frac{\rho_1 - C + 1}{C}.
\]

If \( \rho_1 < C - 1 \), then \( \Phi_{\rho_1, C} \) has negative values and, by (19), a root \( z_C > 0 \). The root is unique; otherwise \( f_{\rho_1, C} \) would have two distinct extrema, which is impossible by (20). The limit \( z(\rho_1, C) \) is necessarily a root of \( \Phi_{\rho_1, C} \); hence \( z(\rho_1, C) = z_C \).

If \( \rho_1 \geq C - 1 \), then (20) gives that \( f_{\rho_1, C} \) is strictly decreasing. Since it converges to 0, we conclude that \( \Phi_{\rho_1, C} \) does not have a root in this case, and thus, necessarily, \( z(\rho_1, C) = +\infty \).

**The overloaded case \( \rho_2 > C \)**

By (19), the function \( \Phi_{\rho_1, \rho_2} \) converges to \( +\infty \) at 0 and at \( +\infty \). To determine the number of roots of \( \Phi_{\rho_1, \rho_2} \), one has therefore to obtain the sign of \( m(\rho_1, \rho_2) \) as defined in (18).

Clearly, \( m(\cdot, \rho_2) : \mathbb{R}_+ \rightarrow \mathbb{R}_+ \), \( \rho_1 \mapsto m(\rho_1, \rho_2) \), is a continuous function. For \( \rho_1^0 < \rho_1^1 \) and \( z > \rho_1^1 / \rho_2 \), the relation

\[
\Phi_{\rho_1^0, \rho_2}(z - \rho_1^0 / \rho_2) < \Phi_{\rho_1^1, \rho_2}(z - \rho_1^1 / \rho_2)
\]

implies that \( m_{\rho_1} \leq m_{\rho_1^1} \); thus \( m(\cdot, \rho_2) \) is an increasing function. Note that, for \( z > 0 \),

\[
\lim_{\rho_1 \rightarrow 0} \Phi_{\rho_1, \rho_2}(\rho_1 z) = -1;
\]

hence if \( \rho_1 \) is sufficiently small then \( m(\rho_1, \rho_2) < 0 \). If \( \rho_1 > C - 1 \), then from (20) we get that the function \( f_{\rho_1, \rho_2} \) is decreasing and converges to \( 0 \) at infinity. Consequently \( \Phi_{\rho_1, \rho_2} \) is positive on \( \mathbb{R}_+ \), and therefore \( m(\rho_1, \rho_2) > 0 \). Hence \( \phi_C \) (as defined in (18)) satisfies \( \phi_C(\rho_2) \in (0, C - 1) \). By continuity of \( m(\cdot, \rho_2) \), we have \( m_{\phi_C(\rho_2)} = 0 \); the function \( \Phi_{\phi_C(\rho_2), \rho_2} \) has therefore a unique root.

By definition, \( \rho_1 > \phi_C(\rho_2) \) if and only if \( m(\rho_1, \rho_2) < 0 \), or, equivalently, since the minimum of \( \Phi_{\rho_1, \rho_2} \) is necessarily achieved in some compact interval of \((0, +\infty)\), if the relation

\[
\sup_{z > 0} \left( z \sum_{m=0}^{C-1} \frac{C!}{m! (\rho_1 + \rho_2 z)^{C-m}} \right) < 1
\]

holds. We deduce that \( \phi_C \) is a non-increasing function on \((C, +\infty)\) and that \( \phi_C(\rho_2) \) converges to \( 0 \) as \( \rho_2 \) gets large. If \( \delta_C \) def. \( \lim_{\rho_2 \searrow C} \phi_C(\rho_2) < C - 1 \),
one can take $\rho_1 \in (\delta_C, C - 1)$. From the critical case, we know that $\Phi_{\rho_1, C}$ has exactly one root and that there exists $z_1 > 0$ such that $\Phi_{\rho_1, C}(z_1) < 0$. One can fix $\rho_2 > C$ sufficiently close to $C$ so that $\Phi_{\rho_1, \rho_2}(z_1) < 0$. Hence $\Phi_{\rho_1, \rho_2}$ has a root, and consequently $\rho_1 \leq \phi_C(\rho_2)$, which contradicts the fact that $\rho_1 > \delta_C \geq \phi_C(\rho_1 \rho_2)$. The proposition is proved.

Examples.

1. When $C = 2$ and $\rho_1 < 2 < \rho_2$, one has

$$2\Phi_{\rho_1, \rho_2}(z) = \rho_2 (\rho_2 - 2) z^2 + 2 (\rho_1 \rho_2 - \rho_1 - 1) z + \rho_1^2.$$  

The function $\Phi_{\rho_1, \rho_2}$ has two roots on $\mathbb{R}_+$ as long as $\Phi_{\rho_1, \rho_2}'(0) < 0$—that is, $\rho_1 < 1/(\rho_2 - 1)$—and the minimum of $\Phi_{\rho_1, \rho_2}$ is negative—that is,

$$\rho_1^2 + 2(1 - \rho_2)\rho_1 + 1 > 0.$$  

It is then easy to deduce that

$$\phi_2(\rho_2) = \rho_2 - 1 - \sqrt{\rho_2(\rho_2 - 2)}.$$  

2. When $C = 3$ and $\rho_1 < 3 < \rho_2$, one has

$$6\Phi_{\rho_1, \rho_2}(z) = \rho_2^2 (\rho_2 - 3) z^3 + 3 \rho_2 (\rho_1 \rho_2 - 2 \rho_1 - 2) z^2 + 3 (\rho_1^2 \rho_2 - \rho_1^2 - 2 \rho_1 - 2) z + \rho_1^3.$$  

The discriminant of this polynomial (in $z$) is

$$H_{\rho_2}(u) \overset{\text{def}}{=} 3 \rho_1^4 + 2 (6 - 5 \rho_2) u^3 + 3 \left(3 \rho_2^2 - 8(\rho_2 - 1)\right) u^2 - 12 (\rho_2 - 2) u - 8 \rho_2 + 12.$$  

Since $H_{\rho_2}(0) = -12(\rho_2 - 2) < 0$ and $H_{\rho_2}(C - 1) = 4(9 \rho_2 - 25)(\rho_2 - 3) > 0$ for $\rho_2 > 3$, it is then easily seen that

$$\phi_3(\rho_2) = \inf\{u : H_{\rho_2}(u) = 0\}.$$  

2.3. The RIST algorithm with one retrial

Proposition 3 shows that the dynamical system associated to the RIST algorithm with an infinite number of retrials has at most two equilibrium points. As noted earlier, there is another equilibrium which is not described here; it is investigated in Subsection 2.4.

The purpose of this section is to show that in the case when only one attempt to accommodate a request is allowed, there are also cases with three equilibrium points, but all of them are equilibria of the “smooth” dynamical system.

In view of Proposition 2, and in the same way as in the derivation of (17), an equilibrium point is of the form $\pi_S$, with

$$\pi_S(x, y) = \frac{1}{Z_S} \cdot \frac{\rho_1^x}{x!} \cdot \frac{(\rho_2 S)^y}{y!}, \quad (x, y) \in \mathcal{X},$$  

where $S$ is a solution $z \in (0, 1)$ of

$$\frac{(\rho_1 + \rho_2 z)^C}{C!} - z \sum_{m=0}^{C} \frac{(\rho_1 + \rho_2 z)^m}{m!} = 0,$$  

and $Z_S$ is the normalization constant.
With the change of coordinates \( z \mapsto (z - \rho_1)/\rho_2 \), this amounts to finding the roots \( z \in (\rho_1, \rho_1 + \rho_2) \) of \( g \), where

\[
g(z) \overset{\text{def.}}{=} \frac{\rho_2 z^C}{z - \rho_1} e^{-z} - \sum_{m=0}^{C} \frac{z^m}{m!} e^{-z}.
\]

Note that \( g(s) \to +\infty \) when \( s \searrow \rho_1 \), and \( g(\rho_1 + \rho_2) < 0 \). Simple calculations give the relation

\[
g'(z) = \frac{z C}{C!} e^{-z} f(z),
\]

with \( f(z) = z^3 - (2\rho_1 + \rho_2)z^2 + (\rho_1^2 + \rho_1 \rho_2 + \rho_2(C - 1))z - C \rho_1 \rho_2 \). This shows in particular that Equation (23) cannot have more than three solutions.

We give a scaling picture of the fixed point equation (23). As with the DAR algorithm, which is investigated in the next section, we study the case when the capacity \( C \) is a scaling parameter going to infinity and \( \rho_2 \) is of the order of \( C \), i.e. \( \rho_2 = \nu_2 C \) for some \( \nu_2 > 0 \).

Under some conditions, there is always a solution of Equation (23) close to 0. For that we do the change of variable \( z \mapsto \rho_1 z \); the relation becomes

\[
\psi_{1,C}(z) \overset{\text{def.}}{=} -z + \rho_1^{C-1} \frac{1+\nu_2 C z}{C!} - z \sum_{m=1}^{C} \frac{\rho_1^m (1 + \nu_2 C z)^m}{m!} = 0.
\]

Using Stirling’s formula, it is easily seen that, for \( \varepsilon > 0 \),

\[
\limsup_{C \to +\infty} \psi_{1,C}(\varepsilon) \leq -\varepsilon + \lim_{C \to +\infty} \rho_1^{C-1} \frac{(1+\nu_2 C \varepsilon)^C}{C!} = -\varepsilon,
\]

provided that \( \nu_2 \rho_1 \geq e \). Since \( \psi_{1,C}(0) > 0 \), we get that, for \( C \) sufficiently large, Equation (23) has a solution in the interval \((0, (\rho_1 \varepsilon) \land 1)\).

Returning to Equation (23), it can be written as

\[
\psi_{2,C}(z) \overset{\text{def.}}{=} z \sum_{k=0}^{C} \frac{1}{(\rho_1 + \nu_2 C z)^k (C - k)!} - 1 = 0.
\]

Assuming that \( \nu_2 z > 1 \), we can check that

\[
\lim_{C \to +\infty} \psi_{2,C}(z) = \Psi_{2,\infty}(z) \overset{\text{def.}}{=} \frac{\nu_2 z^2 - \nu_2 z + 1}{\nu_2 z - 1},
\]

and the convergence is uniform on any compact interval of \((1/\nu_2, 1] \). If \( \nu_2 > 4 \), the function \( \Psi_{2,\infty} \) has two zeroes in the interval \((1/\nu_2, 1)\), given by

\[
z_* \overset{\text{def.}}{=} \frac{1 \pm \sqrt{1 - 4/\nu_2}}{2}.
\]

We now summarize this result in the following proposition.
Proposition 4. Under the assumption that $\rho_2 = v_2 C$ and $v_2 > \max (4, e/\rho_1)$, there exists $C_0 > 0$ such that if $C \geq C_0$, the dynamical system (13) associated to the RIST algorithm with one retrial has exactly three equilibrium points, which converge respectively to

$$0, \quad \frac{1 - \sqrt{1 - 4/v_2}}{2}, \quad \frac{1 + \sqrt{1 - 4/v_2}}{2}$$

as $C$ goes to infinity.

The proof is skipped since most of the arguments have been given and the proof of Theorem 2 in the next section is similar and slightly more technical.

2.4. Stability of saturation

In this section it is assumed that $\rho_2 > C$. The main result is that if the initial state is sufficiently congested, so is the state of the network on any finite time interval.

We describe the general strategy of our approach. Note first that the two-dimensional process of the total number of empty places and total number of class 1 customers in the network does not have the Markov property. Indeed, in state $(m, n)$ with $m > 0$, when an external job arrives it is not possible to determine if the transition to $(m - 1, n)$ or to $(m - 1, n + 1)$ occurs, i.e. if the job is blocked at its arrival queue or not. This process can in fact be compared with an ergodic Markov process in $\mathbb{N}^2$ by using a convenient coupling and a specific order relation in $\mathbb{N}^2$. The ergodicity property is then used to show that, asymptotically, the total number of empty places and of class 1 customers in the network is negligible with respect to $N$, so that the entire system is composed of rerouted jobs in the limit.

If $(Z^N(t)) = (X^N_1(t), Y^N_1(t))$ is the solution of the SDE (6) and (7), define

$$Z^N_1(t) \overset{\text{def.}}{=} \sum_{i=1}^N X^N_i(t), \quad Z^N_2(t) \overset{\text{def.}}{=} \sum_{i=1}^N Y^N_i(t),$$

and let $Z^N_0(t) = CN - Z^N_1(t) - Z^N_2(t)$ denote the total number of empty places at time $t$

A COUPLING WITH A TWO-DIMENSIONAL MARKOV PROCESS. We introduce a Markov process in $\mathbb{N}^2$ which will be used in the analysis of the asymptotic behavior of the process of the empirical distribution associated to $(Z^N(t))$.

Definition 1. Let $(U^N(t)) \overset{\text{def.}}{=} (U^N_0(t), U^N_1(t))$ be the Markov process on $\mathbb{N}^2$ with the initial state $(Z^N_0(0), Z^N_1(0))$, and Q-matrix $Q = (q(\cdot, \cdot))$ defined on the state space $S_U = \{u = (u_0, u_1) \in \mathbb{N}^2 : u_0 + u_1 \leq N\}$ as follows:

$$q(u, u - e_0 + e_1) = \lambda u_0,$$

$$q(u, u - e_0) = \lambda (N - u_0),$$

$$q(u, u + e_0 - e_1) = \mu_1 u_1,$$

$$q(u, u + e_0) = \mu_2 (CN - u_0 - u_1),$$

with $e_0 = (1, 0)$ and $e_1 = (0, 1)$, provided that the transitions keep the process in $S_U$. Define

$$U^N_2(t) \overset{\text{def.}}{=} CN - U^N_0(t) - U^N_1(t).$$
With some abuse of notation, we will also speak of the $Z$- and $U$-systems to refer to the associated stochastic processes $(Z^N(t))$ and $(U^N(t))$, and similarly, we will speak of class 1 and 2 jobs in the U-system with the obvious meaning. Finally,
\begin{align}
S_U^N \text{ def.} & = \inf \left\{ t \geq 0 : U_0^N(t) + U_1^N(t) = N \right\} , \\
\text{and for } a \in \mathbb{N}, & \quad T^a_U \text{ def.} = \inf \left\{ t \geq 0 : U_0^N(t) \geq a \right\} .
\end{align}

The transition rates of the process $(U^N(t))$ suggest that this process would behave like $(Z^N(t))$ if all nodes had, at most, one empty place. The coupling shows that the process $(Z^N(t))$ can be bounded above in some way by such a process. A key element in the coupling is the explicit use of the fact that the service times of class 1 jobs are ‘smaller’ than the service times of class 2 jobs.

**Proposition 5.** There exists a coupling of the processes $(U_0^N(t), U_1^N(t))$ and $(Z^N(t))$ such that $(U_0^N(0), U_1^N(0)) = (Z_0^N(0), Z_1^N(0))$ and the relations
\begin{align}
\begin{cases}
U_2^N(t) \leq Z_2^N(t), \\
U_1^N(t) + U_2^N(t) \leq Z_1^N(t) + Z_2^N(t)
\end{cases}
\end{align}
hold for all $t < S_U^N$, where $S_U^N$ is defined by (24).

**Proof.** We proceed by induction on the number of jumps. One has to show that if the relation holds initially then it will also hold at the first jump of $(Z^N(t))$ or $(U_0^N(t), U_1^N(t))$.

For $j \in \{0, 1\}$, define $z_j \text{ def.} = Z_j^N(0)$ and $u_j \text{ def.} = U_j^N(0)$. By assumption,
\begin{align}
\begin{cases}
u_2 \leq z_2, \\
u_1 + u_2 \leq z_1 + z_2,
\end{cases}
\end{align}
and we can assume that $u_0 < N$ since the process is stopped at time $S_U^N$. We define
\begin{align}
\tilde{a}_0 = \sum_{i=1}^{N} \mathbb{1}_{\{x_i^N(0) + y_i^N(0) < C\}},
\end{align}
the number of non-saturated queues. Clearly,
\begin{align}
\tilde{a}_0 \leq z_0 = CN - z_1 - z_2 \leq CN - u_1 - u_2 = u_0.
\end{align}

We will take the convention that $E_{\xi}$ denotes an exponentially distributed random variable with parameter $\xi \geq 0$ and that all exponential random variables constructed are independent. The coupling is done by introducing the following random variables; the minimum of them will define the first jump of the process. For each random variable, the transition is indicated for $(Z^N(t)) = (Z_0^N(t), Z_1^N(t))$ and for $(U^N(t))$ in the case that it has the minimal value.
1. Arrivals

(a) \( E_{\lambda \tilde{a}_0} \) is the minimum of the arrivals of jobs finding a non-congested queue in the 
\( Z \)-system.
Transition: \( z \mapsto z + e_1 - e_0 \) and \( u \mapsto u + e_1 - e_0 \).

(b) \( E_{\lambda(u_0 - \tilde{a}_0)} \) is the minimum of the remaining arrivals of jobs finding a non-congested
queue in the \( U \)-system.
Transition: \( u \mapsto u + e_1 - e_0 \) and, if \( \tilde{a}_0 > 0 \), \( z \mapsto z - e_0 \).

(c) If \( u_0 > 0 \), \( E_{\lambda(N - u_0)} \) is the minimum of the arrivals of jobs finding a congested
queue in the \( Z \)-system and the \( U \)-system.
Transition: \( u \mapsto u - e_0 \) and, if \( \tilde{a}_0 > 0 \), \( z \mapsto z - e_0 \).

2. Services

(a) \( E_{\mu_2 u_2} \) is the minimum of the services of \( u_2 \) class 2 jobs of the \( U \)-system and 
\( Z \)-system.
Transition: \( z \mapsto z + e_0 \) and \( u \mapsto u + e_0 \). Recall that \( u_2 \leq z_2 \).

(b) If \( u_1 \leq z_1 \):
   i. \( E_{\mu_1 u_1} \) is the minimum of the services of \( u_1 \) class 1 jobs of the \( U \)-system and 
   \( Z \)-system.
      Transition: \( z \mapsto z + e_0 - e_1 \) and \( u \mapsto u + e_0 - e_1 \).
   ii. \( E_{\mu_1(z_1 - u_1)} \) is the minimum of the remaining services of the \( Z \)-system. In this case,
      note that one has necessarily \( z_1 > u_1 \).
      Transition: \( z \mapsto z + e_0 - e_1 \) and \( u \mapsto u \).
   iii. \( E_{\mu_2(z_2 - u_2)} \) is the minimum of the services of the remaining \( z_2 - u_2 \) class 2 jobs
      of \( Z \)-system. In this case \( z_2 > u_2 \).
      Transition: \( z \mapsto z + e_0 \) and \( u \mapsto u \).

(c) If \( u_1 > z_1 \), and we fix \( F = E_{u_1 - z_1} \):
   i. \( E_{\mu_1 z_1} \) is the minimum of the services of \( z_1 \) class 1 jobs of the \( U \)-system and 
   \( Z \)-system.
      Transition: \( z \mapsto z + e_0 - e_1 \) and \( u \mapsto u + e_0 - e_1 \).
   ii. \( F/\mu_1 \) is the minimum of the services of \( u_1 - z_1 \) remaining class 1 jobs of the 
   \( U \)-system.
      Transition: \( z \mapsto z \) and \( u \mapsto u + e_0 - e_1 \).
   iii. \( F/\mu_2 \) is the minimum of the services of some \( u_1 - z_1 \) class 2 jobs of the \( Z \)-system.
      Transition: This cannot be the next step, since \( F/\mu_1 < F/\mu_2 \); the transition of (2) (c) (ii)
      therefore occurs earlier.
   iv. \( E_{\mu_2(z_2 - u_2 - (u_1 - z_1))} \) is the minimum of the services of the remaining \( \mu_2(z_2 - u_2 - 
   (u_1 - z_1)) \) class 2 jobs of the \( Z \)-system. In this case, \( u_1 + u_2 < z_1 + z_2 \).
      Transition: \( z \mapsto z + e_0 \) and \( u \mapsto u \).
An easy, but somewhat tedious, check shows that the two processes have the correct time evolution, and furthermore that the order relation is preserved after any of the transitions mentioned above.

**AN ASYMPTOTIC ANALYSIS OF \( (U^N(t)) \).** It is assumed that the initial state of the process \((U^N(t))\) of Definition (1) satisfies the relation

\[
\lim_{N \to +\infty} \frac{1}{N} (U^N_0(0), U^N_1(0)) = (a_0, a_1).
\]  

(27)

**Proposition 6.** Under the condition \( \rho_1 < C < \rho_2 \), there exists \( \eta_0 > 0 \) such that, if the initial condition (27) satisfies the relations \( 0 \leq a_0 + a_1 \leq \eta_0 \), then for any \( \varepsilon > 0 \), there is \( t_0 > 0 \) and a constant \( K_0 \) such that, for any \( T > 0 \),

\[
\lim_{N \to +\infty} \mathbb{P} \left( \sup_{t_0 \leq t \leq t_0 + T} U^N_0(t) \leq K_0 \log N, \sup_{t_0 \leq t \leq t_0 + T} \frac{U^N_1(t)}{N} \leq \varepsilon, \quad S^N_U(t_0) \geq t_0 + T \right) = 1.
\]

Proof. For \( 0 < a_0 + a_1 < \eta < 1 \), by Definition 1 of the process \((U^N(t))\), a simple coupling shows that the process \((U^N_1(t \wedge S^N_U \wedge T^N_U))\) can be stochastically bounded above by \((L^N_\eta(t \wedge S^N_U \wedge T^N_U))\), where \((L^N_\eta(t))\) is the process of the number of jobs of an \( M/M/\infty \) queue with arrival rate \( \lambda \eta N \), service rate \( \mu_1 \), and initial point \( U^N_1(0) \), and \( T^N_U \) is defined by (25). A classical result (see for example Theorem 6.13 of Robert [29]) gives the following convergence in distribution:

\[
\lim_{N \to +\infty} \left( \frac{L^N_\eta(t)}{N} \right) = \left( \frac{\lambda}{\mu_1} \eta + \left( a_1 - \frac{\lambda}{\mu_1} \eta \right) e^{-\mu_1 t} \right).
\]  

(28)

By a similar argument, the process \((U^N_0(t \wedge S^N_U \wedge T^N_U))\) can be bounded above by \((Q(Nt \wedge S^N_U \wedge T^N_U))\), where \((Q(t))\) is the process of the number of jobs of an \( M/M/1 \) queue with respective arrival and service rates \( \mu_2 C + \eta (\mu_1 - \mu_2) \) and \( \lambda \), and initial point \( U^N_0(0) \). Again a classical result (see for example Proposition 5.16 of Robert [29]) gives the following convergence in distribution:

\[
\lim_{N \to +\infty} \left( \frac{Q(Nt)}{N} \right) = (a_0 + ((\mu_1 - \mu_2)\eta + \mu_2 C - \lambda) t)^+, \quad (29)
\]

where \( a^+ = \max (a, 0) \) for \( a \in \mathbb{R} \).

Now, we fix \( 0 < \eta_0 < 1 \) such that

\[
(\mu_1 - \mu_2)\eta_0 + \mu_2 C < \lambda \quad \text{and} \quad \eta_0 \frac{\lambda}{\mu_1} < 1.
\]  

(30)

If \((a_0, a_1)\) is such that

\[
a_0 < \eta_0(1 - \lambda \eta_0/\mu) \quad \text{and} \quad a_1 < \lambda \eta_0/\mu_1,
\]

then, by using (28) and (29), we get that

\[
\lim_{N \to +\infty} \mathbb{P} \left( \sup_{0 \leq s \leq t} \frac{L^N_{\eta_0}(s)}{N} < \eta_0, \sup_{0 \leq s \leq t} \frac{1}{N}(L^N_{\eta_0}(s) + Q(Ns)) < 1 \right) = 1.
\]
This implies, in particular, the relation
\[
\lim_{N \to +\infty} \mathbb{P} \left( \min \left( S_{\min}^N, T_U^{\eta_0 N} \right) > t \right) = 1
\]
for all \( t \geq 0 \). Therefore, (28) and (29) hold with \((L_{\eta_0}^N(t))\) (resp. \((Q(Nt))\)) replaced by \((U_1^N(t))\) (resp. \((U_0^N(t))\)). Additionally, (29) shows that
\[
\lim_{N \to +\infty} \mathbb{P} \left( T_0^U \leq \frac{\eta_0}{\lambda - (\mu_1 - \mu_2)\eta_0 + \mu_2 C} \right) = 1.
\]

With the same coupling as before and the strong Markov property, the process \((U_0^N(T_0^U + t), 0 \leq t \leq T)\) is bounded above by \((Q(Nt), 0 \leq t \leq T)\), where \((Q(t))\) is the same \(M/M/1\) process as before but starting at 0, \(Q(0) = 0\). Let
\[
H_b \overset{\text{def.}}{=} \inf \{ t \geq 0 : Q(t) = b \}.
\]

Proposition 5.11 of Robert [29] shows that, if \(\rho \overset{\text{def.}}{=} (\mu_1 - \mu_2)\eta_0 + \mu_2 C) / \lambda < 1\), then as \(b\) goes to infinity, the sequence of random variables \((\rho^b H_b)\) converges in distribution to an exponential distribution. Define
\[
A_N \overset{\text{def.}}{=} \sup_{T_0^U \leq s \leq T_0^U + T} \frac{\eta_0}{\lambda - (\mu_1 - \mu_2)\eta_0 + \mu_2 C}
\]
by choosing \(C > -\log \rho\); one then has the relation
\[
\lim_{N \to +\infty} \mathbb{P} (A_N) \geq \lim_{N \to +\infty} \mathbb{P} \left( \sup_{0 \leq s \leq NT} Q(s) \leq C \log N \right) = 1.
\]
The proposition is proved.

Propositions 5 and 6 yield the following proposition.

**Proposition 7.** (Stability of saturation.) Under the condition \(\rho_1 < C < \rho_2\), there exist some \(\eta_0 > 0\) and \(t_0 > 0\) such that, if the initial condition is such that
\[
\lim \inf_{N \to +\infty} \frac{1}{N} \sum_{i=1}^{N} Y_i^N(0) \geq C - \eta_0,
\]
then, for any \(\varepsilon > 0\) and \(T \geq 0\),
\[
\lim_{N \to +\infty} \mathbb{P} \left( \inf_{t_0 \leq s \leq t_0 + T} \frac{1}{N} \sum_{i=1}^{N} Y_i^N(s) \geq C - \varepsilon \right) = 1.
\]

**Proof.** With the above notation,
\[
Z_1^N(t) \overset{\text{def.}}{=} \sum_{i=1}^{N} X_i^N(t), \quad Z_2^N(t) \overset{\text{def.}}{=} \sum_{i=1}^{N} Y_i^N(t),
\]
and \(Z_0^N(t) = CN - Z_1^N(t) - Z_2^N(t)\). Without loss of generality, by taking a subsequence for example, we can assume that
\[
\lim_{N \to +\infty} \frac{Z_1^N(0)}{N} = a_1 \text{ and } \lim_{N \to +\infty} \frac{Z_0^N(0)}{N} = a_0,
\]
for some \( a_0, a_1 \in [0, C] \). Proposition 5 gives a coupling of the process \((Z^N_0(t), Z^N_1(t))\) with the process \((U^N_0(t), U^N_1(t))\), with the same initial conditions, such that the relation \(U^N_2(t) \leq Z^N_2(t)\) holds for \( t < S^N_U \).

Proposition 6 shows that there exist \( \eta_0 > 0 \) and \( t_0 \geq 0 \) such that if (32) holds, then \( a_0 + a_1 \leq \eta_0 \), so that, for \( T > 0 \),

\[
\lim_{N \to +\infty} \mathbb{P} \left( \sup_{t_0 \leq t \leq t_0 + T} \frac{U^N_0(t)}{N} \leq \varepsilon, \sup_{t_0 \leq t \leq t_0 + T} \frac{U^N_1(t)}{N} \leq \varepsilon, S^N_U \geq t_0 + T \right) = 1.
\]

With the coupling, we obtain the relation

\[
\lim_{N \to +\infty} \mathbb{P} \left( \inf_{t_0 \leq t \leq t_0 + T} \frac{Z^N_2(t)}{N} \geq C - 2\varepsilon \right) = 1.
\]

The proposition is proved.

The next result gives a stability property for the limiting points of the sequence of empirical distributions \((\Lambda^N(t))\).

**Corollary 1.** Under the assumptions of Proposition 7, any limiting point \((\Lambda(t))\) of the sequence of empirical processes \((\Lambda^N(t))\) of (10) satisfies the following relation for convergence in distribution: for any \( \varepsilon > 0 \) there exists \( t_0 > 0 \) such that

\[
\Lambda(t)((0, C)) > 1 - \varepsilon \quad \forall t \geq t_0.
\]

**Proof.** This is simply due to the above proposition and the fact that

\[
\frac{1}{N} \sum_{i=1}^{N} 1\{Y^N_i(s) \neq C\} \leq \frac{1}{N} \sum_{i=1}^{N} (C - Y^N_i(s)) = C - \frac{Z^N_2(t)}{N}.
\]

### 2.5. A spectral criterion of stability

In this section we investigate the stability properties of the nonlinear dynamical system \((\Lambda(t))\) defined by (12). The corresponding linear system \((\Lambda_R(t))\) is the solution of (14). The goal of this section is to show that if the linear process \((\Lambda_R(t))\) is converging sufficiently fast to equilibrium, the nonlinear process will converge to this invariant distribution provided its initial state is sufficiently close to it.

As before, \( \pi_R \) is the probability distribution on \( \mathcal{X} \) defined by (15); it is the invariant measure of \((\Lambda_R(t))\). If \( R \) is a solution of Equation (16) (which was studied in Subsection 2.2), \( \pi = \pi_R \) is an invariant measure of \((\Lambda(t))\). Following Aldous and Fill [1], the ‘distance’ between \( \mu \in \mathcal{P}(\mathcal{X}) \) and a fixed probability \( \pi \) is defined as

\[
\| \mu - \pi \|_2^2 \overset{\text{def}}{=} \sum_{z \in \mathcal{X}} \left( \frac{\mu(z)}{\pi(z)} - 1 \right)^2 \pi(z) = \sum_{z \in \mathcal{X}} \frac{(\mu(z) - \pi(z))^2}{\pi(z)}.
\]

Lemma 3.26 of Aldous and Fill [1] shows that there exists a maximal \( \kappa_R > 0 \) such that, for all \( t \geq 0 \),

\[
\frac{d}{dt} \| \Lambda_R(t) - \pi \|_2^2 \leq -2\kappa_R \| \Lambda_R(t) - \pi \|_2^2.
\]

(33)
The equilibrium states of large Erlang networks

This is the classical exponential convergence to equilibrium for finite Markov processes; the distance \( \| \cdot \|_2 \) gives the nice inequality (33) describing this phenomenon. The quantity \( \kappa_R \) is the spectral gap of the process \( (\Lambda_R(t)) \); see Theorem 3.25 of Aldous and Fill [1] for a variational characterization.

**Theorem 1.** Let \( R \) be a solution of Equation (16) and assume that the spectral gap \( \kappa_R \) of \( (\Lambda_R(t)) \) satisfies the condition

\[
\kappa_R > \frac{\lambda}{1 - R} \sqrt{\frac{C}{\rho_2}}.
\]

Then there exist positive constants \( q \) and \( \varepsilon_0 \) such that the relation

\[
\frac{d}{dt} \| \Lambda(t) - \pi \|_2^2 \leq -q \| \Lambda(t) - \pi \|_2^2 \quad \text{for all } t \geq 0
\]

holds when \( \| \Lambda(0) - \pi \|_2 \leq \varepsilon_0 \), with \( \pi = \pi_R \) defined by (24). In particular, \( \pi \) is an exponentially stable equilibrium point of \( (\Lambda(t)) \).

**Strategy of the Proof.** We first describe the main ideas. Our technique is based on a variation of an argument coming from the theory of attractors. The original approach consists of splitting the half-line \( \mathbb{R}_+ \) into intervals of large length and decomposing the dynamical system into two parts on each of them. The first one takes the linear part of the equation, for which there is exponential convergence. The second one includes the nonlinearity and is issued from zero on each interval. The convergence of the first part is then used to absorb the second part of the flow originated after the splitting, which allows one to obtain the desired result. We refer the reader to the paper of Zelik [36], where this powerful approach is used in another context. Unfortunately, the direct application of this technique does not lead to good results in our case because of the strong nonlinearity in (12). To overcome this difficulty, we shall first prove an ‘instantaneous’ absorbing and then use a bootstrap argument to propagate; see Step 2 of the proof below.

**Proof.** **Step 1: Splitting of the flow.** Let \( \Lambda_R(t) \) be the solution of Equation (14) with a fixed initial point \( \Lambda(0) \) and

\[
F(t) \overset{\text{def.}}{=} \| \Lambda(t) - \pi \|_2^2.
\]

For \( t \geq 0 \), we have

\[
F'(t) = \frac{d}{dt} \sum_{z \in \mathcal{X}} \frac{[\Lambda(t) - \pi](z)^2}{\pi(z)} = 2 \sum_{z \in \mathcal{X}} \frac{[\Lambda(t) - \pi](z) \Lambda'(t)(z)}{\pi(z)}.
\]

We will now state two claims and show how they are used to establish our theorem. The proofs of the claims conclude the proof of the theorem.

**Claim 1.** The relation

\[
\Lambda'(0)(z) = \Lambda'_{R}(0)(z) + \lambda L_{\Lambda,R} \left( \Lambda(0)(z) - e_2 \right) - 1_{\{z \in \mathcal{X}_+\}} \Lambda(0)(z)
\]

holds, where

\[
L_{\Lambda,R} = \frac{\Lambda(0)(\mathcal{X}_+^c)}{1 - \Lambda(0)(\mathcal{X}_+^c)} - \frac{R}{1 - R}.
\]
Taking \( t = 0 \) and using that \( \Lambda_R(0) = \Lambda(0) \) together with (36), we obtain
\[
F' (0) = 2 \sum_{z \in \mathcal{X}} \left( [\Lambda_R(0) - \pi](z) \Lambda'_R(0)(z) \right) \frac{\pi(z)}{\pi(z)} + 2\lambda L \sum_{z \in \mathcal{X}} \left( [\Lambda(0) - \pi](z) (\Lambda(0)(z) - e_2) - 1_{\{z \in \mathcal{X}_+\}} \Lambda(0)(z) \right) \frac{\pi(z)}{\pi(z)} = I_1 + I_2. \tag{38}
\]

Note that the term \( I_1 \) is the derivative of the function \( t \to \|\Lambda_R(t) - \pi\|_2^2 \) at zero. From Inequality (33), we get therefore that
\[
\frac{I_1}{2} \leq -\kappa_R \|\Lambda_R(0) - \pi\|_2^2 = -\kappa_R \|\Lambda(0) - \pi\|_2^2 = -\kappa_R F(0). \tag{39}
\]

**Claim 2.** There exist \( \varepsilon > 0 \) and \( \theta < 2\kappa_R \) such that, for \( \|\Lambda(0) - \pi\|_2 \leq \varepsilon \),
\[
I_2 \leq \theta F(0). \tag{40}
\]

Combining this with (38) and (39), we get
\[
F' (0) \leq -q F(0) \tag{41}
\]
for \( \|\Lambda(0) - \pi\|_2 \leq \varepsilon \) and \( q = 2\kappa_R - \theta > 0 \).

**STEP 2: BOOTSTRAP ARGUMENT.** First note that we may suppose that \( F(0) > 0 \). Otherwise \( \Lambda(0) = \pi \) and therefore \( \Lambda(t) = \pi \) for all \( t \geq 0 \); in this case there is nothing to prove. Let
\[
\tau \overset{\text{def.}}{=} \inf \left\{ s > 0 : F(s) > \varepsilon^2 \right\},
\]
with the convention that \( \inf \emptyset = +\infty \). Since \( 0 < F(0) \leq \varepsilon^2 \), we have by (41) that \( F'(0) \leq -q F(0) < 0 \), so that \( \tau > 0 \).

The constants \( \varepsilon \) and \( q \) in Inequality (41) do not depend on the initial point \( \Lambda(0) \) as long as \( \|\Lambda(0) - \pi\|_2 \leq \varepsilon \). Hence we can apply the same argument considering the equation as starting from \( \Lambda(s) \), for \( s \geq 0 \), as long as \( \|\Lambda(s) - \pi\|_2 \leq \varepsilon \); we can thus infer the differential inequality with 0 replaced by \( s \). We have therefore the relation
\[
F'(s) \leq -q F(s);
\]
hence \( F(s) \leq F(0) \exp (-qs) \) for all \( 0 \leq s < \tau \). This implies that \( \tau \) is infinite. Inequality (35) is established. It remains to prove our two claims, i.e. Equation (36) and Inequality (40).

**STEP 3: PROOF OF THE IDENTITY (36).**

From now on, \( z \) denotes a generic element \((x, y) \in \mathcal{X}\) of the state space, and \( e_1 = (1, 0) \), \( e_2 = (0, 1) \) are the unit vectors of \( \mathcal{X} \).

From Equation (12), we get
\[
\Lambda' (t)(z) = \lambda \left( \Lambda(t)(z) - e_2 \right) - 1_{\{z \in \mathcal{X}_+\}} \Lambda(t)(z) \right) + \Lambda(t)(\mathcal{X}_-^c) \frac{\Lambda(t)(\mathcal{X}_+^c)}{1 - \Lambda(t)(\mathcal{X}_+^c)}
+ \lambda \left( \Lambda(t)(z) - e_1 \right) - 1_{\{z \in \mathcal{X}_+\}} \Lambda(t)(z) \right) + \mu_1 \left( (x + 1) \Lambda(t)(z + e_1) - x \Lambda(t)(z) \right)
+ \mu_2 \left( (y + 1) \Lambda(t)(z + e_2) - y \Lambda(t)(z) \right), \tag{42}
\]
with the convention \( \Lambda(t)(z') = 0 \) if \( z' \notin \mathcal{X} \).
Similarly, the definition (14) gives

\[
\Lambda'_R(t)(z) = \lambda \left( \Lambda_R(t)(z - e_2) - I_{[\epsilon_2]}(z) \Lambda_R(t)(z) \right) \frac{R}{1 - R} + \lambda \left( \Lambda_R(t)(z - e_1) - I_{[\epsilon_1]}(z) \Lambda_R(t)(z) \right) + \mu_1 \left( (x + 1) \Lambda_R(t)(z + e_1) - x \Lambda_R(t)(z) \right) + \mu_2 \left( (y + 1) \Lambda_R(t)(z + e_2) - y \Lambda_R(t)(z) \right).
\]

(43)

Taking \( t = 0 \) in these two relations and using that \( \Lambda_R(t) \) and \( \Lambda(t) \) have the same initial conditions, we get the identity (36). We now establish the most intricate inequality of our theorem, namely Inequality (40).

**STEP 4: A BOUND FOR \( I_2 \).** From the expression (37) of \( L_{A,R} \), we get

\[
|L_{A,R}| \leq \frac{1}{(1 - R)^2} \frac{|\Lambda(0)(X^c) - R|}{1 - |\Lambda(0)(X^c) - R|/(1 - R)},
\]

and, by the Cauchy-Schwartz inequality,

\[
|\Lambda(0)(X^c) - R| = \left( \sum_{z \in X^c} \left| \Lambda(0) - \pi \right| (z) \right) \left( \sum_{z \in X^c} \frac{\left| \Lambda(0) - \pi \right| (z)}{\pi (z)} \right) \leq \sqrt{\sum_{z \in X^c} \pi (z)} \frac{\left| \Lambda(0) - \pi \right| (z)}{\sqrt{\pi (z)}}\]

\[
= \sqrt{\sum_{z \in X^c} \pi (z)} \frac{\left| \Lambda(0) - \pi \right| (z)}{\pi (z)} = \sqrt{RF(0)}.
\]

Combining these two relations, we get the inequality

\[
|L_{A,R}| \leq \frac{1}{(1 - R)^2} \frac{\sqrt{RF(0)}}{1 - \sqrt{RF(0)}/(1 - R)}.
\]

(44)

Since

\[
\sum_{z \in X} \frac{\left| \Lambda(0) - \pi \right| (z) [\Lambda(0)(z - e_2) - I_{[\epsilon_2]} \Lambda(0)(z)]}{\pi (z)} = \sum_{z \in X, y \geq 1} \frac{\left| \Lambda(0) - \pi \right| (z) [\Lambda(0) - \pi ](z - e_2)}{\pi (z)} + \sum_{z \in X} \frac{\left| \Lambda(0) - \pi \right| (z) [\pi (z - e_2) - I_{[\epsilon_2]} \pi (z)]}{\pi (z)}
\]

\[
+ \sum_{z \in X^c} \frac{\left| \Lambda(0) - \pi \right| (z) [\pi - \Lambda(0)](z)}{\pi (z)} \equiv J_1 + J_2 + J_3,
\]

(45)

we obtain from (44) and (45) that

\[
I_2 \leq 2\lambda |L_{A,R}| (|J_1| + |J_2| + |J_3|).
\]

(46)
STEP 5: Estimates for $J_1$ and a final bound for $I_2$.

Clearly $|J_3| \leq \|\Lambda(0) - \pi\|^2_2$. By using (15) and the Cauchy-Schwartz inequality, we get

$$|J_1| \leq \max_{z \in \mathcal{X}} \sqrt{\frac{\pi(z-e_2)}{\pi(z)}} \sum_{z \in \mathcal{X}, y \geq 1} \frac{|[\Lambda(t) - \pi](z)| |[\Lambda(0) - \pi](z-e_2)|}{\sqrt{\pi(z)}}$$

$$\leq \sqrt{\frac{C}{R\rho^2}} \left( \sum_{z \in \mathcal{X}, y \geq 1} \frac{([\Lambda(0) - \pi](z))^2}{\pi(z)} \right)^{1/2} \left( \sum_{z \in \mathcal{X}, y \geq 1} \frac{([\Lambda(0) - \pi](z-e_2))^2}{\pi(z-e_2)} \right)^{1/2}$$

Another application of the Cauchy-Schwartz inequality gives

$$|J_2| \leq \left( \sum_{z \in \mathcal{X}} \frac{(\pi(z-e_2) - 1_{\mathcal{X}_e}(z)\pi(z))^2}{\pi(z)} \right)^{1/2} \|\Lambda(0) - \pi\|_2.$$ 

On the other hand, by using (15) and (16), we obtain

$$\sum_{z \in \mathcal{X}} \frac{(\pi(z-e_2) - 1_{\mathcal{X}_e}(z)\pi(z))^2}{\pi(z)} \leq \sum_{z \in \mathcal{X}} \frac{(\pi(z-e_2))^2}{\pi(z)}$$

$$= \sum_{z \in \mathcal{X}} \pi(z-e_2)^2 (1-R) \frac{1}{R\rho^2} \leq \frac{C(1-R)^2}{R\rho^2},$$

so that

$$|J_2| \leq (1-R) \left( \frac{C(0)}{R\rho^2} \right)^{1/2}.$$ 

Inequalities (44) and (46) imply

$$I_2 \leq 2\lambda \frac{1}{(1-R)^2} \frac{\sqrt{R\rho^2}}{1 - \sqrt{R\rho^2}/(1-R)}$$

$$\times \left( (1-R) \left( \frac{C(0)}{R\rho^2} \right)^{1/2} + \left( \frac{\sqrt{C}}{\sqrt{R\rho^2} + 1} \right) F(0) \right);$$

consequently,

$$\limsup_{\varepsilon \to 0} \sup_{0 < F(0) < \varepsilon^2} \frac{I_2}{F(0)} \leq \frac{2\lambda}{1-R} \sqrt{\frac{C}{\rho^2}} \leq 2\kappa_R$$

by the assumption (34). Inequality (40) is thus established. The theorem is proved.

3. The Dynamic Alternative Routing (DAR) algorithm

Recall that, in this algorithm, when a request cannot be accommodated at its arriving node, two other nodes are chosen at random. If both of them are non-saturated, the request takes one place in each of them. Otherwise, the request is rejected. This algorithm was initially considered (by Gibbens et al. [16] in 1990, and in subsequent papers) to cope with congestion in telephone networks.
3.1. The basic ODEs

We recall briefly the technical background for this algorithm. (See Gibbens et al. [16].) There is a set $V$ of vertices, and for each couple $(A, B) = (B, A)$ of vertices, referred to as a link or node, there is a Poisson flow of requests with rate $\lambda$, referred to as calls or jobs, to establish a connection between $A$ and $B$. The capacity constraint is that there can be at most $C$ jobs at a given time on any node. The state of the process is given by the number of jobs in the nodes of the network. The algorithm works as follows. If a node $(A, B)$ has strictly fewer than $C$ jobs at some instant, then a request arriving at $(A, B)$ is accepted. Otherwise, a vertex $D \notin \{A, B\}$ is chosen at random, and if both links $(A, D)$ and $(D, B)$ have strictly fewer than $C$ jobs, then the job occupies a place in $(A, D)$ and in $(D, B)$ during an exponentially distributed amount of time with parameter 1. If one of the links $(A, D), (D, B)$ is saturated, the request is rejected.

The mean-field result of Graham and Méléard [17], described below, shows that from the point of view of the convergence of the empirical distribution process $(X^N(t))$ defined by Equation (1), the DAR algorithm has the same limiting behavior as the following allocation algorithm. There is a set of $N = |V|(|V| - 1)/2$ nodes with finite capacity $C$; each node receives a Poisson flow of jobs with parameter $\lambda$ to be processed at rate 1. When a request arrives at a saturated node $\ell$, two other nodes are chosen at random. If neither of them is saturated, a new request is added to each of them. Otherwise the initial request to node $\ell$ is rejected.

For $1 \leq \ell \leq N$, $L_N^\ell(t)$ denotes the number of jobs in node $\ell$ at time $t \geq 0$. Note that $(L_N^\ell(t), 1 \leq \ell \leq N)$ is not a Markov process. The mean-field result of Graham and Méléard [17] (which was conjectured by Gibbens et al. [16]) is as follows. The initial state is given by i.i.d. random variables with some distribution $\mu$ on $\{0, \ldots, C\}$ for the $L_N^\ell(0)$, $1 \leq \ell \leq N$, and without any request using two nodes. It has been shown in [17] that the convergence in distribution

$$\lim_{N \to +\infty} (L_N^\ell(t)) = (\bar{L}_C(t))$$

holds for any $\ell \geq 1$, where, for $t \geq 0$, the distribution of $\bar{L}_C(t)$, given by the vector $(\mathbb{P}(\bar{L}_C(t) = k), 0 \leq k \leq C) = (x^C_k(t), 0 \leq k \leq C)$, is the solution of the following ODEs: for $1 \leq j < C$,

$$\begin{align*}
\frac{dx^0_C}{dt}(t) &= x^1_C(t) - \lambda h \left(x^C_C(t)\right) x^0_C(t), \\
\frac{dx^j_C}{dt}(t) &= \lambda h \left(x^C_C(t)\right) x^C_{j-1}(t) + (j + 1) x^C_{j+1}(t) - (\lambda h \left(x^C_C(t)\right) + j) x^C_j(t), \\
\frac{dx^C_C}{dt}(t) &= \lambda h \left(x^C_C(t)\right) x^C_{C-1}(t) - C x^C_C(t),
\end{align*}$$

where $h(x) \equiv (1 + 2x(1 - x))$, with initial condition $(x^C_k(0)) = (\mu(k))$.

We give an intuitive explanation of this system, to explain the role of the function $h$ in particular. For $t \geq 0$, $x^C_C(t)$ is the fraction of saturated nodes, i.e. the number of nodes with $C$ jobs. The mean-field limit expresses an asymptotic independence property: the numbers of jobs at a fixed finite subset of nodes are, in the limit, independent, and $x^C_C(t)$ is the probability that an arbitrary node is saturated. Each non-saturated node accepts external requests arriving at rate $\lambda$. It may also be occupied with a request which has arrived at a saturated node, if this request has picked (at random) this empty node and another non-saturated node. With the independence approximation, this occurs with probability

$$2x^C_C(t)(1 - x^C_C(t)).$$
The equations of the system (49) can then easily be explained. The equations (49) can be seen as the set of Fokker-Planck equations for a non-homogeneous $M/M/C/C$ queue for which the arrival rate at time $t$ is $\lambda h(x^c_C(t))$ and the service rate is 1. Equivalently, from a probabilistic point of view, the process $(\bar{L}_C(t))$ has the same distribution as the solution of the following McKean-Vlasov SDE:

$$
d\bar{L}_C(t) = 1_{[\bar{L}_C(t) < C]} \mathcal{P}_1 \left( \left[ 0, \lambda h \left( \mathbb{P} \left( \bar{L}_C(t) = C \right) \right) \right] \times dt \right) - \mathcal{P}_2 \left( \left[ 0, \bar{L}_C(t) \right] \times dt \right),
$$

with $\bar{L}_C(0) \overset{\text{dist.}}{=} \mu$, where $\mathcal{P}_i, i = 1, 2$, are independent Poisson processes on $\mathbb{R}_+^2$ with rate 1.

An equilibrium point $(x^*_k)$ of the dynamical system (49) is given by

$$
x^*_k = \frac{1}{Z_C} \left( \frac{\lambda h(x^*_C)}{k!} \right)^k, \quad 0 \leq k < C,
$$

where $Z_C$ is the normalization constant and $z = x^*_C$ is a positive solution of the fixed point equation

$$
\frac{(\lambda h(z))^C}{C!} - z \sum_{k=0}^{C} \frac{(\lambda h(z))^k}{k!} = 0.
$$

There always exists a solution to this equation, since the left-hand side of (51) is positive for $z = 0$ and negative for $z = 1$. The rest of this section is devoted to determining the conditions under which there may exist several solutions for this equation and to investigating their stability properties for the dynamical system (49).

The insightful Gibbens et al. [16] suggests, through approximations and numerical experiments, that this equation may have in fact several solutions:

*Observe the possibility of multiple solutions for $x^*_C$, for $C$ large enough and for a narrow range of the ratio $\lambda/C$. The upper and lower solutions correspond to stable fixed points for the system of equations (2)-(5), while the middle solution corresponds to an unstable fixed point.*

(The notation has been adapted.) To the best of our knowledge, these statements do not seem to have been established in a more formal way. The rest of this section is devoted to a scaling analysis of the set of ODEs (49). As suggested by these numerical experiments, we will study the case of a large capacity $C$. Concerning the stability results of this assertion, we have not been able to prove them as such. Subsections 3.3 and 3.4 give only partial results in this domain.

### 3.2. An asymptotic dynamical system

We denote by $(x^C_C(t))$ the solution of the ODE (49) when $\lambda$ is replaced by $\lambda C$ and $h(x) = (1 + ax(1 - x))$ for some $a > 1$ and $x \in (0, 1)$. For this algorithm too, there is a kind of analogue of the regime analyzed in Subsection 2.4, in the sense that it has some intuitive explanation. In this regime, in the limit, all jobs are accommodated without rerouting provided that the initial state is not already saturated. As will be seen, for the same parameters, there are nevertheless
The equilibrium states of large Erlang networks

Theorem 2. (The solutions of a fixed point equation.) For the fixed point equation (51) with \( \lambda = \nu C \) for some \( \nu > 0 \) and \( h(x) = 1 + ax(1-x) \) for \( a > 1 \), there exists \( C_0 > 0 \) such that, for all \( C \geq C_0 \), the following statements hold.

1. \( \nu \in (0, 1) \), then there exists a solution \( x^*_{C,1} \in (0, 1) \) of Equation (51) such that \( vh(x^*_{C,1}) \leq 1 \), and
\[
\lim_{C \to +\infty} x^*_{C,1} = 0. \tag{52}
\]

2. If \( \nu \in (\nu_a, 1) \), with
\[
\nu_a \overset{\text{def.}}{=} 3 \left( 1 + \frac{9}{36}a + \frac{2}{3} (a + 3)^{3/2} \right) \tag{53}
\]
then there are three solutions \( x^*_{i,C,1} \in (0, 1) \), \( i \in \{1, 2, 3\} \), of Equation (51), with the following properties: \( vh(x^*_{C,1}) < 1 \), and \( vh(x^*_{C,i}) > 1 \) for \( i = 2, 3 \). The limiting values of \( (x^*_{i,C,2}) \) and \( (x^*_{i,C,3}) \) are the two solutions of the polynomial equation
\[
ax^3 - 2az^2 + (a - 1)z + 1 = \frac{1}{\nu} \tag{54}
\]
in \((0,1)\), and \( (x^*_{C,1}) \) satisfies (52).

3. If \( \nu > 1 \), there exists a unique solution \( x^*_{C,1} \in (0, 1) \) of Equation (51), and the sequence \( (x^*_{C,1}) \) converges to the unique solution of Equation (54) in the interval \((0, 1)\).

For the Gibbens et al. model, which corresponds to the case \( a = 2 \), this gives \( \nu_2 \sim 0.937 \); hence \((0.937, 1)\) is the ‘narrow range of the ratio \( \lambda/C \)’ suggested by these authors (see page 41) for which there are three solutions to the fixed point equation.

According to Part (1), when \( \nu < 1 \), there is an equilibrium in the light-load regime \( (x^*_{C,1} \sim 0) \). We will see a more precise result, Proposition 9, concerning the asymptotic local stability of this equilibrium. When \( \nu_a < \nu < 1 \), there are two other equilibrium points, but in a saturated regime: \( x^*_{C,i} \geq \eta > 0 \) for \( i = 2, 3 \) and \( C \) sufficiently large.

**Proof.** The fixed point equation (51) with \( \lambda \) replaced by \( \nu C \) can be rewritten as \( \Psi_C(z) = 0 \), with, for \( z > 0 \),
\[
\Psi_C(z) \overset{\text{def.}}{=} 1 - z \sum_{k=0}^{C} \frac{C!}{(C - k)!} \frac{1}{(vCh(z))^k} = 1 - z \sum_{k=0}^{C} \prod_{i=0}^{k-1} \left( 1 - \frac{i}{C} \right) \frac{1}{(vCh(z))^k}. \]

Note that the function \( C \mapsto \Psi_C(z) \) is decreasing.

If \( \nu < 1 \), we can choose \( \varepsilon < 1/2 \) sufficiently small so that \( \delta = vh(\varepsilon) < 1/2 \) holds. For \( z \leq \varepsilon \), we have
\[
\sum_{k=0}^{C} \prod_{i=0}^{k-1} \left( 1 - \frac{i}{C} \right) \frac{1}{(vCh(z))^k} \geq \sum_{k=0}^{\lfloor \delta C \rfloor} \prod_{i=0}^{k-1} \left( 1 - \frac{i}{C} \right) \frac{1}{(vCh(z))^k} \geq \sum_{k=0}^{\lfloor \delta C \rfloor} \left( 1 - \frac{\delta}{\delta} \right)^k.
\]

In particular, for \( 0 < z < \varepsilon \), \( (\Psi_C(z)) \) converges to \( -\infty \) as \( C \) goes to infinity. Since \( \Psi_C(0) = 1 \), one can find \( C_0 > 0 \) such that if \( C \geq C_0 \), then there is a zero of \( \Psi_C \) in the interval \((0, \varepsilon)\).
For \( \delta > 1 \), it is easily checked that the convergence
\[
\lim_{C \to +\infty} \Psi_C(z) = \Psi(z) \overset{\text{def.}}{=} 1 - z \frac{1}{1 - 1/(v h(z))} = \frac{v(1 - z) h(z) - 1}{v h(z) - 1} \tag{55}
\]
holds uniformly for all \( z \in (0, 1) \) such that \( v h(z) > \delta \). Equation (51) becomes, in the limit,
\[
(1 - z) h(z) = \frac{1}{v} \tag{56}
\]
Note that any solution \( z < 1 \) of such an equation satisfies \( v h(z) > 1 \).

The quantity \((1 - z) h(z)\) is the polynomial \(a z^3 - 2 a z^2 + a z - z + 1\). This polynomial is increasing from 1 on the interval \([0, x_0]\) and decreasing on \([x_0, 1]\), where
\[
x_0 \overset{\text{def.}}{=} \frac{1}{3} \left( 2 - \sqrt[3]{\frac{a + 3}{a}} \right), \quad (1 - x_0) h(x_0) = 1/v_a = \frac{1}{3} \left( 1 + \frac{2}{9} a + \frac{2}{3} (a + 3)^{3/2} \right).
\]
Equation (56) therefore has two solutions if and only if \( v \in (v_a, 1) \), one solution when \( v = v_a \) or \( v > 1 \), and none if \( v < v_a \).

If \( v \in (v_a, 1) \), for \( \epsilon > 0 \) sufficiently small, there exist \( z_0 < z_1 < z_2 \) such that \( v h(z_i) > 1 \) for \( i \in \{0, 1, 2\} \), and \( \Psi(z_0) < -\epsilon, \, \Psi(z_1) > \epsilon \) and \( \Psi(z_2) < -\epsilon \). Consequently, there exists \( K_0 \) such that, if \( C \geq K_0 \), then the last three inequalities hold with \( \Psi \) replaced by \( \Psi_C \) and \( \epsilon \) by \( \epsilon/2 \).
Hence, we get that there are two solutions of the relation \( \Psi_C(z) = 0 \) such that \( v h(z) > 1 \) for \( C \geq K_0 \). Assertion (2) is proved.

If \( v > 1 \), the convergence (55) is uniform for \( z \in [0, 1] \). Since \( \Psi(0) > 0 \) and \( \Psi(1) < 0 \), by the same argument as before, there exists some \( K_1 \) such that if \( C \geq K_1 \) then there is a solution \( x_{C,1}^v \) of Equation (51). A simple calculation gives
\[
\frac{d}{dz} \Psi(z) = -\frac{v}{(v h(z) - 1)^2} \left( v z^2 (z - 1)^2 a^2 - z (2 v (z - 1) - 3 z + 2) a + v - 1 \right),
\]
and one has
\[
v z^2 (z - 1)^2 a^2 - z (2 v (z - 1) - 3 z + 2) a \geq a z (3 z - 2 + \left( z^3 - 2 z^2 - z + 2 \right) v)
\]
\[
\geq a z (3 z - 2 + \left( z^3 - 2 z^2 - z + 2 \right)) = a z^2 (z^2 - 2 z + 2) \geq 0
\]
by using successively that \( a \geq 1 \), that \( z^3 - 2 z^2 - z + 2 = (1 - z)(2 + 2 z - z^2) \geq 0 \) when \( z \in [0, 1] \), and finally that \( v \geq 1 \). Consequently, we get that
\[
\frac{d}{dz} \Psi(z) \leq -\frac{v(v - 1)}{(v h(z) - 1)^2} \leq -\frac{v(v - 1)}{(v(1 + a/4) - 1)^2} < 0 \tag{57}
\]
holds for all \( z \in [0, 1] \).

Note that, as in (55), since \( v > 1 \), the convergence
\[
\lim_{C \to +\infty} \frac{d}{dz} \Psi_C(z) = \frac{d}{dz} \Psi(z)
\]
holds uniformly for \( z \in [0, 1] \).
To prove uniqueness, we assume that there is a subsequence \((C_n)\) converging to infinity for which the equation \(\Psi_{C_n}(z) = 0\) has two solutions. This implies in particular that we have a sequence \((z_n)\) in \((0,1)\) such that \(\Psi_{C_n}(z_n) = 0\). Due to the uniform convergence, this is in contradiction to (57). The proposition is proved.

When \(\nu < 1\), with our method based on the asymptotic behavior of \(\Psi_C(z)\) as \(C\) gets large, we have not been able to prove that all solutions of Equation (51) are identified, though this is very likely the case.

A scaled version of the dynamical system

For the moment we have given a scaled version of the fixed point equations. It turns out that one can also get some insight from a scaled version of the dynamical system (49) converging to a nontrivial dynamical system whose fixed points are described in Theorem 2.

Let us introduce some notation. The set of bounded sequences is denoted by \(B(\mathbb{N})\), and it is endowed with the following norm: for \(z \in B(\mathbb{N})\),

\[
\|z\| = \sum_{k=0}^{+\infty} \frac{1}{2^k} |z_k|.
\]

For \(T > 0\) and \((z(t)) = (z_k(t)) \in C(\mathbb{R}_+, B(\mathbb{N}))\) a continuous function on \(\mathbb{R}_+\), we define

\[
\|z\|_T \overset{\text{def.}}{=} \sup_{0 \leq t \leq T} \|z(t)\|.
\]

Additionally, \(\mathcal{P}(\mathbb{N})\) is the set of probability distributions on \(\mathbb{N}\).

The scaling consists of slowing down the time scale by a factor \(C\) and looking at the number of empty places for the McKean-Vlasov process.

**Proposition 8.** (Asymptotic dynamical system.) If \((x^C_k(t))\) is the solution of the set of ODEs defined by (49) with \(\lambda = \nu C\), with an initial point such that

\[
\lim_{C \to +\infty} \|x^C_{C-k}(0), k \in \mathbb{N} \) \) - \( (q_0(k), k \in \mathbb{N}) \| = 0
\]

for some probability distribution \(q_0 \in \mathcal{P}(\mathbb{N})\), then, as \(C\) goes to infinity, the process

\[
(y^C_k(t), k \in \mathbb{N}) \overset{\text{def.}}{=} \left( x^C_{C-k} (t/C), k \in \mathbb{N} \right)
\]

converges in distribution for the uniform norm \(\|\cdot\|_T\) to \((\Gamma(t)) \in C(\mathbb{R}_+, \mathcal{P}(\mathbb{N}))\), which is the unique solution of the set of differential equations

\[
\begin{align*}
\frac{d}{dt} \Gamma_0(t) &= \nu h (\Gamma_0(t)) \Gamma_1(t) - \Gamma_0(t), \\
\frac{d}{dt} \Gamma_k(t) &= \nu h (\Gamma_0(t)) \Gamma_{k+1}(t) + \Gamma_{k-1}(t) - (\nu h (\Gamma_0(t)) + 1) \Gamma_k(t), \quad k \geq 1,
\end{align*}
\]

with \(\Gamma(0) = q_0\).

**Proof.** Note that the process \((y^C_k(t))\) can be seen as a version of the empirical distribution process of empty places in the nodes of the network with the slowed-down time scale \(t \to t/C\).
It is not difficult to check that \((y^C_k(t))\) satisfies the following system of ODEs:

\[
\begin{align*}
\frac{d}{dt} y^C_0(t) &= \nu h \left( y^C_0(t) \right) y^C_1(t) - y^C_0(t), \\
\frac{d}{dt} y^C_k(t) &= \nu h \left( y^C_0(t) \right) y^C_{k+1}(t) + \left( 1 - \frac{k + 1}{C} \right) y^C_{k-1}(t) \\
&\qquad - \left( \nu h \left( y^C_0(t) \right) + \left( 1 - \frac{k}{C} \right) \right) y^C_k(t), \\
\frac{d}{dt} y^C_C(t) &= \frac{1}{C} y^C_1(t) - \nu h \left( y^C_0(t) \right) y^C_C(t).
\end{align*}
\] (59)

The proposition is proved using a classical consequence of the compactness-uniqueness argument: the sequence of functions \((y^C_k(t))\), \(C \geq 1\), is tight, and any limiting point satisfies the equations (59). If this system of equations has a unique solution, then the convergence is established, since all subsequences have a subsequence converging to this limit. (See Chapter 10 and 11 of Ethier and Kurtz [14], for example.)

The tightness is due to the Arzelà-Ascoli theorem (see Theorem 7.2 of Billingsley [3], for example). The equations (59) show that for \(k \in \mathbb{N}\), the sequence of functions \((y^C_k(t))\) is equicontinuous and therefore is relatively compact for the uniform norm on bounded intervals. The integral form of the ODEs (59) shows that any limiting point satisfies (58). Let \((x_k(t))\) and \((y_k(t))\) be two solutions of (58) with the same initial condition. Using the fact that the function \(h\) is Lipschitz on \([0, 1]\) with parameter 6, the integral form of (58) gives the following inequality: for \((u_k(t) = (x_k(t) - y_k(t)))\) and \(t > 0\),

\[
|u_k(t)| \leq 6\nu \int_0^t |u_0(s)| (x_k(s) + x_{k+1}(s)) \, ds \\
+ 6\nu \int_0^t (|u_{k-1}(s)| + |u_k(s)| + |u_{k+1}(s)|) \, ds.
\]

Since \((x_k(t))\) and \((y_k(t))\) are probability distributions on \(\mathbb{N}\), \((u_k(t))\) is a convergent series and

\[
\sum_{k=0}^n |u_k(t)| \leq 30\nu \int_0^t \sum_{k=0}^{n+1} |u_k(t)| \, ds,
\]

so that

\[
U(t) \overset{\text{def}}{=} \sum_{k=0}^{+\infty} |u_k(t)| \leq 30\nu \int_0^t U(s) \, ds.
\]

Grönwall’s inequality gives the relation \((U(t)) = (0)\), i.e. the uniqueness of the solution of (58), and, consequently, the desired convergence.

A probabilistic translation of this result can be stated as the fact that if the process \((\bar{L}_C(t))\) defined by (50) satisfies the relation

\[
\lim_{C \to +\infty} C - \bar{L}_C(0) = q_0
\]

for convergence in distribution, then

\[
\lim_{N \to +\infty} (Q_C(t)) \overset{\text{def}}{=} (C - \bar{L}_C(t/C)) = (\bar{Q}(t)),
\]
where \((\overline{Q}(t))\) is the solution of the McKean-Vlasov SDE
\[
d\overline{Q}(t) = N_1(dt) - 1_{\{\overline{Q}(t) > 0\}}P_1([0, \nu h(\{\overline{Q}(t) = 0\})]) \times dt.
\] (60)

with \(\overline{Q}(0) \overset{\text{dist.}}{=} q_0\), where \(N_1\) and \(P_1\) are independent homogeneous Poisson processes with rate 1 on \(\mathbb{R}_+\) and \(\mathbb{R}_+^2\) respectively.

The process \((\overline{Q}(t))\) is a nonlinear \(M/M/1\) queue with the following jump rates at time \(t\):
\[
\begin{cases}
+1 & 1 \\
-1 & \nu h(\{\overline{Q}(t) = 0\})
\end{cases}
\] (61)

It should be noted that this scaling is convenient for studying regimes where the number of empty places is small, i.e. when the system has some level of saturation. Subsection 3.3 studies the case when there is an equilibrium regime with a large number of empty places.

**Remark 1.** An invariant distribution \(\pi\) of the nonlinear Markov process \((\overline{Q}(t))\) defined by (60) is the invariant distribution of an \(M/M/1\) queue with arrival rate 1 and service rate \(\nu h(\pi(0))\). Thus \(\pi\) is a geometric distribution with parameter \(1/(\nu h(\pi(0)))\); in particular,
\[
\pi(0) = 1 - \frac{1}{\nu h(\pi(0))},
\]
which is Equation (56), as can be expected. Indeed, Equation (56) is the limit, as \(C\) gets large, of the fixed point equation (51) for \(x_C\), which is the equilibrium probability that a node has zero empty places.

### 3.4 Stability of the underloaded regime

In Theorem 2 we have seen that, under the condition \(\nu < 1\), there is a root of the fixed point equation (51) that is arbitrarily close to 0 as \(C\) gets large. This result suggests that the underloaded regime is stable for the dynamical system (49). In this regime most requests are accepted at their arrival node. The following proposition gives a formal characterization of this property.

**Proposition 9.** (Stability of underloaded regime.) If \(\lambda = \nu C\) for some \(\nu < 1\), there exists \(\eta \in (0, 1)\) such that if the initial state of the dynamical system (49) satisfies the relation
\[
\lim_{C \to +\infty} \sum_{k \geq \eta C} x^C_k(0) = 0,
\]
then there exists \(\eta^* \in (0, 1)\) such that
\[
\lim_{C \to +\infty} \sup_{t \geq 0} \left( \sum_{k \geq \eta^* C} x^C_k(t) \right) = 0. \tag{62}
\]

**Proof.** We fix \(\delta_0 > 0\) such that \(\eta_0 = \nu(1 + a\delta_0) \in (\eta, 1)\), and we choose \(\eta_1, \eta_2 \in (\eta_0, 1)\) with \(\eta_1 < \eta_2\). For \(\varepsilon < \delta_0/2\), we take \(C_0\) such that
\[
\sum_{k \geq \eta_0 C_0} x^C_k(0) \leq \varepsilon \quad \text{and} \quad \left( \frac{\eta_0}{\eta_1} \right)^{[\eta_2 - \eta_1]C_0} \leq \varepsilon.
\]
Let \((Q(t))\) be an \(M/M/1\) queue with arrival rate \(\eta_0\) and service rate \(\eta_1\) with \(Q(0) = 0\), and let \((L_C(t))\) be the processes defined by Equation (50). A simple coupling of \((Q(t))\) with a stationary version of the process \((Q(t))\) gives the stochastic monotonicity property of the process \((Q(t))\). (See the proof of Proposition 5.8 of Robert [29], for example.) We thus obtain that, for \(C \geq C_0\),

\[
\mathbb{P}(Q(t) \geq (\eta_2 - \eta_1)C) \leq \mathbb{P}(Q(\infty) \geq (\eta_2 - \eta_1)C) \leq \left(\frac{\eta_0}{\eta_1}\right)^{\lfloor (\eta_2 - \eta_1)C \rfloor} \leq \varepsilon,
\]

where \(Q(\infty)\) is a geometrically distributed random variable with parameter \(\eta_0/\eta_1\). If \((Q_1(t))\) def. = \(\eta_1 C + Q(t)\), then, for all \(t \geq 0\),

\[
\mathbb{P}(Q_1(t) \geq \eta_2 C) \leq \varepsilon \leq \frac{\delta_0}{2}.
\]

Note that \((Q_1(Ct))\) is a birth-and-death Markov process with birth rate \(\eta_0 C\) and death rate \(\eta_1 C\).

We can construct a coupling of the processes \((L_C(t))\) and \((Q(t))\) such that, for all \(t \geq 0\),

\[
\sum_{k \geq \eta_2 C} x^C_k(t) = \mathbb{P}(L_C(t) \geq \eta_2 C) \leq \mathbb{P}(L_C(t) \geq \eta_2 C | L_C(0) \leq \eta_0 C) + \varepsilon \leq \mathbb{P}(Q_1(Ct) \geq \eta_2 C) + \varepsilon \leq 2\varepsilon.
\]

The proposition is proved.

### 3.4. Nonlinear \(M/M/1\) queues

The equations (58) defining the asymptotic process \((\Gamma(t)) = (\Gamma_k(t), k \in \mathbb{N})\) can be written in a vectorial form as follows: for any function \(f : \mathbb{N} \to \mathbb{R}_+\) with finite support,

\[
\frac{d}{dt} (\Gamma(t), f) = \left[\Gamma(t), \nabla^+(f) \right] + v h(\Gamma(t)(0)) \left[\Gamma(t), \nabla^-(f) \right],
\]

with

\[
\nabla^+(f)(x) \text{ def.} = f(x + 1) - f(x), \quad \nabla^-(f)(x) \text{ def.} = (f(x - 1) - f(x)) 1_{\{x > 0\}}.
\]

In (58), we had \(h(x) = 1 + ax(1 - x)\) with \(a > 1\).

We assume that \(h\) is a continuously differentiable function from \([0, 1]\) to \([1, +\infty)\), and that the relation \(\nu > 1\) holds. If there is an equilibrium \(\pi \in \mathcal{P}(\mathbb{N})\) for the dynamical system (63), it is the equilibrium of the linear Markov process \((\Gamma_S(t))\), where \(S = \pi(0)\):

\[
\frac{d}{dt} (\Gamma_S(t), f) = \left[\Gamma_S(t), \nabla^+(f) \right] + v h(S) \left[\Gamma_S(t), \nabla^-(f) \right].
\]

As noted before, this is a classical \(M/M/1\) queue which is ergodic since the service rate \(v h(S)\) is greater than 1, the arrival rate, by assumption. We have therefore a representation for the invariant distribution

\[
\pi_S(n) \text{ def.} = \left(\frac{1}{v h(S)}\right)^n \left(1 - \frac{1}{v h(S)}\right), \quad n \in \mathbb{N},
\]

where \(Q(\infty)\) is a geometrically distributed random variable with parameter \(\eta_0/\eta_1\). If \((Q_1(t))\) def. = \(\eta_1 C + Q(t)\), then, for all \(t \geq 0\),

\[
\mathbb{P}(Q_1(t) \geq \eta_2 C) \leq \varepsilon \leq \frac{\delta_0}{2}.
\]

Note that \((Q_1(Ct))\) is a birth-and-death Markov process with birth rate \(\eta_0 C\) and death rate \(\eta_1 C\).

We can construct a coupling of the processes \((L_C(t))\) and \((Q(t))\) such that, for all \(t \geq 0\),

\[
\sum_{k \geq \eta_2 C} x^C_k(t) = \mathbb{P}(L_C(t) \geq \eta_2 C) \leq \mathbb{P}(L_C(t) \geq \eta_2 C | L_C(0) \leq \eta_0 C) + \varepsilon \leq \mathbb{P}(Q_1(Ct) \geq \eta_2 C) + \varepsilon \leq 2\varepsilon.
\]

The proposition is proved.

### 3.4. Nonlinear \(M/M/1\) queues

The equations (58) defining the asymptotic process \((\Gamma(t)) = (\Gamma_k(t), k \in \mathbb{N})\) can be written in a vectorial form as follows: for any function \(f : \mathbb{N} \to \mathbb{R}_+\) with finite support,

\[
\frac{d}{dt} (\Gamma(t), f) = \left[\Gamma(t), \nabla^+(f) \right] + v h(\Gamma(t)(0)) \left[\Gamma(t), \nabla^-(f) \right],
\]

with

\[
\nabla^+(f)(x) \text{ def.} = f(x + 1) - f(x), \quad \nabla^-(f)(x) \text{ def.} = (f(x - 1) - f(x)) 1_{\{x > 0\}}.
\]

In (58), we had \(h(x) = 1 + ax(1 - x)\) with \(a > 1\).

We assume that \(h\) is a continuously differentiable function from \([0, 1]\) to \([1, +\infty)\), and that the relation \(\nu > 1\) holds. If there is an equilibrium \(\pi \in \mathcal{P}(\mathbb{N})\) for the dynamical system (63), it is the equilibrium of the linear Markov process \((\Gamma_S(t))\), where \(S = \pi(0)\):

\[
\frac{d}{dt} (\Gamma_S(t), f) = \left[\Gamma_S(t), \nabla^+(f) \right] + v h(S) \left[\Gamma_S(t), \nabla^-(f) \right].
\]

As noted before, this is a classical \(M/M/1\) queue which is ergodic since the service rate \(v h(S)\) is greater than 1, the arrival rate, by assumption. We have therefore a representation for the invariant distribution

\[
\pi_S(n) \text{ def.} = \left(\frac{1}{v h(S)}\right)^n \left(1 - \frac{1}{v h(S)}\right), \quad n \in \mathbb{N},
\]
and, consequently, the fixed point equation for $S$,

$$S = 1 - \frac{1}{\nu h(S)},$$

which we have already seen (see Equation (54)). It is well known that for the standard $M/M/1$ process ($\Gamma_S(t)$), for any initial condition $\Gamma_S(0) \in \mathcal{P}(\mathbb{N})$, the following inequality holds for all $t \geq 0$:

$$\frac{d}{dt} \|\Gamma_S(t) - \pi_S\|^2 \leq -2\kappa_S \|\Gamma_S(t) - \pi_S\|^2,$$

where $\kappa_S$, the spectral gap of the process, has the explicit representation

$$\kappa_S = \left(\sqrt{\nu h(S)} - 1\right)^2.$$

See Chen [8] and Liu and Ma [23], for example.

**Theorem 3.** If $h : [0, 1] \to [1, +\infty)$ is a $C^1$ function and $S \in (0, 1)$ is a solution of (66) such that

$$|h'(S)| < \frac{1}{\nu S} \left(\frac{1}{\sqrt{1-S}} - 1\right)^2,$$

then the probability distribution $\pi_S$ defined by (65) is an exponentially stable equilibrium point of the dynamical system defined by (63): there exist positive constants $q$ and $\varepsilon$ such that if $\|\Gamma(0) - \pi_S\|_2 \leq \varepsilon$, the relation

$$\|\Gamma(t) - \pi_S\|_2^2 \leq \|\Gamma(0) - \pi_S\|_2^2 \cdot e^{-qt},$$

holds for all $t \geq 0$.

**Proof.** Identity (63) gives

$$\Gamma'(t)(x) = \left(\Gamma(t)(x - 1) - \Gamma(t)(x)\right) 1_{\{x > 0\}} + \nu h(\Gamma(t)(0)) \left(\Gamma(t)(x + 1) - \Gamma(t)(x)\right);$$

in the same way as with Equation (64), $(\Gamma_S(t))$ is defined as the solution of

$$\Gamma'_S(t)(x) = \left(\Gamma_S(t)(x - 1) - \Gamma_S(t)(x)\right) 1_{\{x > 0\}} + \nu h(S) \left(\Gamma_S(t)(x + 1) - \Gamma_S(t)(x)\right),$$

with the same initial condition $\Gamma_S(0) = \Gamma(0)$. We have

$$\Gamma'(0)(x) = \Gamma'_S(0)(x) + \nu (h(\Gamma(0)(0)) - h(S)) \left(\Gamma(0)(x + 1) - \Gamma(0)(x)\right).$$

As in Subsection 2.5, setting

$$F(t) \overset{\text{def.}}{=} \|\Gamma(t) - \pi_S\|^2_2,$$

we have

$$F'(0) = 2 \sum_{x \in \mathbb{N}} \frac{[\Gamma_S(0) - \pi_S](x)\Gamma'_S(0)(x)}{\pi_S(x)} + 2\nu (h(\Gamma(0)(0)) - h(S)) \sum_{x \in \mathbb{N}} \frac{[\Gamma(0) - \pi_S](x)(\Gamma(0)(x + 1) - \Gamma(0)(x))}{\pi_S(x)}$$

$$\overset{\text{def.}}{=} \mathcal{I}_1 + 2\nu (h(\Gamma(0)(0)) - h(S)) \cdot \mathcal{I}_2.$$

(69)
By (67) and the fact that \((\Gamma_S(t))\) and \((\Gamma(t))\) have the same initial condition, we get
\[
\frac{I_1}{2} \leq -2\kappa S F(0) = -2 \left( \sqrt{v} h(S) - 1 \right)^2 F(0).
\] (70)

Furthermore, we have
\[
\mathcal{I}_2 = \sum_{x \in \mathbb{N}} \frac{[\Gamma(0) - \pi_S(x)](\Gamma(0)(x + 1) - \Gamma(0)(x))}{\pi_S(x)}
\]
\[
= \sum_{x \in \mathbb{N}} \frac{[\Gamma(0) - \pi_S(x)](\pi_S(x + 1) - \pi_S(x))}{\pi_S(x)}
+ \sum_{x \in \mathbb{N}} \frac{[\Gamma(0) - \pi_S(x)](\Gamma(0) - \pi_S)(x + 1)}{\pi_S(x)} - \|\Gamma(0) - \pi_S\|^2_2.
\]

The Cauchy-Schwartz inequality and (65) give
\[
|\mathcal{I}_2| \leq \left( \sum_{x \in \mathbb{N}} \frac{(\pi_S(x + 1) - \pi_S(x))^2}{\pi_S(x)} \right)^{1/2} \|\Gamma(0) - \pi_S\|_2
\]
\[
+ \sup_{x \in \mathbb{N}} \left( \frac{\pi_S(x + 1)}{\pi_S(x)} + 1 \right) \|\Gamma(0) - \pi_S\|^2_2
\]
\[
= \left( 1 - \frac{1}{\sqrt{v} h(S)} \right) \|\Gamma(0) - \pi_S\|_2 + \left( \frac{1}{\sqrt{v} h(S)} + 1 \right) \|\Gamma(0) - \pi_S\|^2_2.
\]

Combining this with Equation (66) and Inequalities (69) and (70), as well as the fact that 
\(|\Gamma(0)(0) - S| \leq \|\Gamma(0) - \pi_S\|_2\), we obtain that if \(F(0) > 0\),
\[
\frac{F'(0)}{2F(0)} \leq -\kappa_S + \frac{|h(\Gamma(0)(0)) - h(S)|}{|\Gamma(0)(0) - S|} \left( S + \frac{1}{\sqrt{v} h(S)} + 1 \right) \sqrt{F(0)};
\]

hence
\[
\limsup_{\varepsilon \to 0} \sup_{F(0) \in (0, \varepsilon^2)} \frac{F'(0)}{F(0)} \leq -q = \frac{2}{\left( \sqrt{v} h(S) - 1 \right)^2 + v|\Gamma'(S)|S} < 0
\]
by the assumption (68). To complete the proof, it remains to apply a similar bootstrap argument
as in the proof of Theorem 1 to get the above inequality for the ratio \(F'(s)/F(s)\). The theorem
is proved.

We now apply this result to the asymptotic dynamical system of the DAR algorithm.

**Corollary 2.** When \(h(x) = 1 + ax(1 - x)\) with \(a > 1\), there exists a neighborhood \(I\) of \(u_0 = 8/(4 + a)\) such that, if \(\nu \in I\), then the unique fixed point of the dynamical system (58) of the
corresponding nonlinear \(M/M/1\) queue is exponentially stable.

**Proof.** The fixed point equation (66) is
\[
(1 - S)(1 + aS(1 - S)) = \frac{1}{\nu},
\] (71)
and since \(h'(x) = av(1 - 2x)\), the condition (68) is equivalent to
\[
aS|1 - 2S| < \left( 1 - \sqrt{1 - S} \right)^2 (1 + aS(1 - S)).
\] (72)

To conclude, note that \(S = 1/2\) satisfies the condition.
Remarks.

1. It is easy to check that, for $a > 4$, $u_a \in (v_a, 1)$, where $v_a$ is defined by (53). In this case there are two positive fixed points for the asymptotic dynamical system; the above corollary gives that one of them is locally stable. We have not been able to prove that the other one is unstable, as suggested by Gibbens et al. [16] (see the quotation on page 41).

2. A little more work can give more precise conditions on $\nu$ for the stability of the fixed point. Let $x = \sqrt{1 - S}$. If $S \in [0, 1/2]$, Condition (72) amounts to

\[ P_1(x) \overset{\text{def.}}{=} ax^5 - ax^4 - 3ax^3 - ax^2 + (a - 1)x + a + 1 > 0. \]

Notice that $P_1(\sqrt{2}/2) = (2 - \sqrt{2})(1 + a/4)/2$ and $P_1(1) = -2a$. If $S \in [1/2, 1]$, the condition is

\[ P_2(x) \overset{\text{def.}}{=} ax^5 - ax^4 + ax^3 + 3ax^2 - (1 + a)x + 1 - a > 0, \]

with $P_2(0) = 1 - a < 0$ and $P_2(\sqrt{2}/2) = P_1(\sqrt{2}/2) > 0$. It is not difficult to check that $P_1$ (resp. $P_2$) is concave (resp. convex) on $[0, 1]$; hence there exists a unique root $z_{a,1}$ of $P_1$ in $(\sqrt{2}/2, 1)$ (resp. $z_{a,2}$ of $P_2$ in $(0, \sqrt{2}/2)$). Hence the condition (68) is satisfied when $S \in (1 - z_{a,1}^2, 1 - z_{a,2}^2)$, and, by Equation (71), this holds if $\nu \in (Q(z_{a,2}), Q(z_{a,1}))$, where $Q(z) = 1/[z^3(1 + az^2 - az^4)]$.

3. For the precise case of Gibbens et al. [16], $h(x) = (1 + 2x(1 - x))$, this shows that when $\nu \in (1.2068, 1.5978)$, the unique fixed point is a locally stable equilibrium.

A toy example with an arbitrary number $n$ of stable equilibrium points.

We fix ($u_k$), a set of $n$ distinct points in $(0,1)$. Let $f$ be a $C^1$ function such that, for each $k \in \{1, \ldots, n\}$, the relation $f(u) = 1 + \ln(1 - u) - \ln(1 - u_k)$ holds in a small neighborhood of $u_k$. Note that since $f(u_k) = 1 > 1 - u_k$ for $1 \leq k \leq n$, we can choose $f$ in such a way that $f(u) > 1 - u$ holds for all $u \in (0, 1)$. If we define $h(u) = f(u)/(1 - u)$ for $u \in (0, 1)$, then $h$ maps $(0,1)$ to $(1, +\infty)$. Taking $\nu = 1$, we have that each $u_k$, $1 \leq k \leq n$, is clearly a fixed point, and the condition (68) is satisfied since $h'(u_k) = 0$. It is therefore locally stable.

Acknowledgement

During the writing of this paper, in August 2018, we learned of the sad news that Richard Gibbens passed away. His remarkable paper with his colleagues in 1990 was the main motivation of this work. We would like to pay tribute to his memory.

References

[1] ALDOUS, D. AND FILL, J. A. (2002). Reversible Markov chains and random walks on graphs. Unfinished monograph, recompiled 2014. Available at http://www.stat.berkeley.edu/~aldous/RWG/book.html.
[2] ANTUNES, N., FRICKER, C., ROBERT, P. AND TIBI, D. (2008). Stochastic networks with multiple stable points. Ann. Prob. 36, 255–278.
[3] BILLINGSLEY, P. (1999). Convergence of Probability Measures, 2nd edn. John Wiley, New York.
[4] BOVIER, A. AND DEN HOLLANDER, F. (2015). Metastability: A Potential-Theoretic Approach. Springer, Cham.
[5] BUDHIRAJA, A., DUPUIS, P., FISCHER, M. AND RAMANAN, K. (2015). Local stability of Kolmogorov forward equations for finite state nonlinear Markov processes. Electron. J. Prob. 20, 30 pp.
[6] CAPUTO, P., DAI PRA, P. AND POSTA, G. (2009). Convex entropy decay via the Bochner–Bakry–Émery approach. Ann. Inst. H. Poincaré Prob. Statist. 45, 734–753.
[7] Carrillo, J. A., McCann, R. J. and Villani, C. (2003). Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. Rev. Mat. Iberoam. 19, 971–1018.

[8] Chen, M.-F. (2010). Speed of stability for birth-death processes. Front. Math. China 5, 379–515.

[9] Dai Pra, P. and Posta, G. (2013). Entropy decay for interacting systems via the Bochner-Bakry-Émery approach. Electron. J. Prob. 18, 21 pp.

[10] Dawson, D. A., Tang, J. and Zhao, Y. Q. (2005). Balancing queues by mean field interaction. Queueing Systems 49, 335–361.

[11] Den Hollander, F. (2004). Metastability under stochastic dynamics. Stoch. Process. Appl. 114, 1–26.

[12] Erbar, M. and Maas, J. (2012). Ricci curvature of finite Markov chains via convexity of the entropy. Arch. Ration. Mech. An. 206, 997–1038.

[13] Erbar, M., Maas, J. and Tetali, P. (2015). Discrete Ricci curvature bounds for Bernoulli-Laplace and random transposition models. Ann. Fac. Sci. Toulouse Math. Ser. 6, 24, 781–800.

[14] Ethier, S. N. and Kurtz, T. G. (1986). Markov Processes: Characterization and Convergence. John Wiley, New York.

[15] Frank, T. D. (2005). Nonlinear Fokker-Planck Equations: Fundamentals and Applications. Springer, Berlin, Heidelberg.

[16] Gibbens, R. J., Hunt, P. J. and Kelly, F. P. (1990). Bistability in communication networks. In Disorder in Physical Systems, Oxford University Press, New York, pp. 113–127.

[17] Graham, C. and Méleard, S. (1993). Propagation of chaos for a fully connected loss network with alternate routing. Stoch. Process. Appl. 44, 159–180.

[18] Hunt, P. and Kurtz, T. (1994). Large loss networks. Stoch. Process. Appl. 53, 363–378.

[19] Jacobsen, M. (2006). Point Process Theory and Applications. Birkhäuser, Boston, MA.

[20] Kelly, F. P. (1991). Loss networks. Ann. Appl. Prob. 1, 319–378.

[21] Kingman, J. F. C. (1993). Poisson processes. Oxford University Press, New York.

[22] Last, G. and Brandt, A. (1995). Marked Point Processes on the Real Line. Springer, New York.

[23] Liu, W. and Ma, Y. (2009). Spectral gap and convex concentration inequalities for birth-death processes. Ann. Inst. H. Poincaré Prob. Statist. 45, 58–69.

[24] Maas, J. (2017). Entropic Ricci curvature for discrete spaces. In Modern Approaches to Discrete Curvature, eds. L. Najman and P. Romon, Springer, Cham, pp. 159–174.

[25] Malyshev, V. and Robert, P. (1994). Phase transition in a loss load sharing model. Ann. Appl. Prob. 4, 1161–1176.

[26] Marbukh, V. (1993). Loss circuit switched communication network: performance analysis and dynamic routing. Queueing Systems 13, 111–141.

[27] Muzyczka, S. (2015). A class of nonlinear processes admitting complete study. Mosc. U. Math. Bull. 70, 141–143.

[28] Olivieri, E. and Vares, M. E. (2005). Large Deviations and Metastability (Encyclopedia Math. Appl. 100). Cambridge University Press.

[29] Robert, P. (2003). Stochastic Networks and Queues. Springer, New York.

[30] Rybko, A. and Shlosman, S. (2005). Poisson hypothesis for information networks. I. Mosc. Math. J. 5, 679–704, 744.

[31] Sznitman, A. (1991). Topics in propagation of chaos. In École d’Été de Probabilités de Saint-Flour XIX – 1989 (Lecture Notes Math. 1464), Springer, Berlin, pp. 167–243.

[32] Thai, M.-N. (2015). Birth and death process in mean field type interaction. Preprint. Available at http://arxiv.org/abs/1510.03238.

[33] Tibi, D. (2010). Metastability in communication networks. Preprint. Available at http://arxiv.org/abs/1002.0796.

[34] Van Doorn, E., Zeifman, A. and Panfilova, T. (2010). Bounds and asymptotics for the rate of convergence of birth-death processes. Theory Prob. Appl. 54, 97–113.

[35] Verhulst, F. (1990). Nonlinear Differential Equations and Dynamical Systems. Springer, Berlin.

[36] Zelik, S. (2004). Asymptotic regularity of solutions of a nonautonomous damped wave equation with a critical growth exponent. Commun. Pure Appl. Anal. 3, 921.