J Stat Phys (2009) 134: 67–104
DOI 10.1007/s10955-008-9658-0

Spontaneous Resonances and the Coherent States of the Queuing Networks

Alexander Rybko · Senya Shlosman · Alexander Vladimirov

Received: 4 March 2008 / Accepted: 28 November 2008 / Published online: 13 December 2008
© Springer Science+Business Media, LLC 2008

Abstract We present an example of a highly connected closed network of servers, where the time correlations do not vanish in the infinite volume limit. The limiting interacting particle system behaves in a periodic manner. This phenomenon is similar to the continuous symmetry breaking at low temperatures in statistical mechanics, with the average load playing the role of the inverse temperature.

Keywords Coupled dynamical systems · Non-linear Markov processes · Stable attractor · Phase transition · Long-range order

1 Introduction

1.1 Interacting Particle Systems with Long Range Memory

The theory of phase transitions, among many results, substantiates the possibility of constructing reliable systems from non-reliable elements. As an example, consider the infinite volume stochastic Ising model at low temperature T in dimension ≥ 2, see [10]. It is well known, that if we start this system from the configuration of all pluses, then the evolution under Glauber dynamics has the property that the fraction of plus spins at any time exceeds $\frac{1+m^*(T)}{2}$, which is bigger than $\frac{1}{2}$ for $T < T_{cr}$. (Here $m^*(T)$ is the spontaneous magnetization.) On the other hand, if we consider finite volume Ising model (with empty boundary condition, say), then this property does not hold, and the system, started from the all plus state, will be found in the state with the majority of the spins to be minuses at some later (random) times. Therefore, the infinite system can remember, to some extent, its initial state, while the finite system can not.

A. Rybko · S. Shlosman (✉) · A. Vladimirov
Inst. of the Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia
e-mail: shlosman@gmail.com

S. Shlosman
Centre de Physique Theorique, UMR 6207, CNRS, Luminy, Marseille, France

There are many other examples of this kind, that belong to the theory of interacting particle systems, such as voter model, contact model, etc. In all these examples we see systems, that are capable of "remembering" their initial state for arbitrary long times.

In the present paper we are constructing a particle system that "remembers its initial phase". The rough analogy can be described as follows. Imagine a Brownian particle $\varphi(t)$, with a unit drift, which lives on a circle. Suppose the initial phase $\varphi(0) = 0$. Then the mean phase $\bar{\varphi}(t) = t \bmod (2\pi)$, but with time we know the phase $\varphi(t)$ less and less precisely, since its variance grows, and in the limit $t \to \infty$ the distribution of $\varphi(t)$ tends to the uniform one. However, one can combine infinitely many such particles by introducing suitable interaction between them in such a way that the memory of the initial phase does not vanish and persists in time. Namely, one has to put such particles at the sites of $\mathbb{Z}^3$ and to introduce the attractive interaction between them. If the initial state is chosen to be coherent, then the phase of every particle will grow linearly, while its variance will stay bounded.

This is roughly what we do in the present paper. We consider a network of simple servers which are processing messages. Since the service time of every message is random, in the course of time each single server loses the memory of its initial state. So, in particular, the network of non-interacting servers, started in the same state, becomes de-synchronized after a finite time. However, if one introduces certain natural interconnection between servers, then it can happen that they are staying synchronized after an arbitrary long time, thus breaking some generally believed properties of large networks. We have to add here that such a phenomenon is possible only if the mean number of particles per server is high enough; otherwise the infinite network becomes de-synchronized, no matter which kind of interaction between servers takes place. So the parameter of the mean number of particles per server, called hereafter *the load*, plays the same role as the temperature in the statistical mechanics.

In other words, the transition we describe happens due to the fact that at low load the behavior of our system is governed by the fixed point of the underlying dynamical system, while at high load the dominant role is played by its periodic attractor. A similar phenomenon was described by Hepp and Lieb in [6].

Below we present the simplest example of the above behavior. But we believe that the phenomenon we describe is fairly general. Its origin lies in the fact that any large network of the general type possesses some kind of the continuous symmetry in the infinite limit, and it is breaking of that symmetry at high load that causes the long-range order behavior of the network. In our case this is the rotation symmetry, corresponding to the periodic orbit of the limiting dynamical system.

1.2 Information Networks and Their Collective Behavior

Now we will describe one pattern of behavior of certain large networks, which was assumed to be universal. It is known under the name of Poisson Hypothesis.

The Poisson Hypothesis is a device to predict the behavior of large queuing networks. It was formulated first by L. Kleinrock in [7], and concerns the following situation. Suppose we have a large network of servers, through which many customers are traveling, being served at different nodes of the network. If the node is busy, the customers wait in the queue. Customers are entering into the network from the outside via some nodes, and these external flows of customers are Poissonian, with constant rates. The service time at each node is random, depending on the node, and the customer. The PH prediction about the (long-time, large-size) behavior of the network is the following:

- consider the total flow $\mathcal{F}$ of customers to a given node $\mathcal{N}$. Then $\mathcal{F}$ is approximately equal to a Poisson flow, $\mathcal{P}$, with a time dependent rate function $\lambda_{\mathcal{N}}(T)$.

- The exit flow from $\mathcal{N}$—*not* Poissonian in general—has a rate function $\gamma_{\mathcal{N}}(T)$, which is smoother than $\lambda_{\mathcal{N}}(T)$ (due to averaging, taking place at the node $\mathcal{N}$).
- As a result, the flows $\lambda_{\mathcal{N}}(T)$ at various nodes $\mathcal{N}$ should tend to a constant limit $\bar{\lambda}_{\mathcal{N}} \approx \frac{1}{T}\int_0^T \lambda(t)dt$, as $T \to \infty$, the flows to different nodes being almost independent.
- The above convergence is uniform in the size of the network.

Note that the distributions of the service times at the nodes of the network can be arbitrary, so PH deals with quite a general situation. The range of validity of PH is supposed to be the class of networks where the internal flow to every node $\mathcal{N}$ is a union of flows from many other nodes, and each one of these flows constitutes only a small fraction of the total flow to $\mathcal{N}$. If true, PH provides one with means to make easy computations of quantities of importance in network design.

The rationale behind this conjectured behavior is natural: since the inflow is a sum of many small inputs, it is approximately Poissonian. And due to the randomness of the service time the outflow from each node should be "smoother" than the total inflow to this node. (This statement was proven in [15] under quite general conditions.) In particular, the variation of the latter should be smaller than that of the former, and so all the flows should converge to corresponding constant values.

In the paper [14] the Poisson Hypothesis is proven for simple networks in the infinite volume limit, under some natural conditions. For systems with constant service times it was proven earlier in [17].

The purpose of the present paper is to construct a network that satisfies the above assumption—that the flow to every given node is an "infinite" sum of "infinitesimally small" flows from other nodes—but has coherent states. That means that the states of the servers are evolving in a synchronous manner, and the "phase" of a given server behaves (in the thermodynamic limit—i.e. in the limit of infinite network) as a periodic non-random function, the same for different servers.

We have to stress that our network exhibits these coherent states only in the regime when the average number N of the customers per server—called in what follows *the load*—is large. For low load we expect the convergence to the unique stationary state. This "high temperature" kind of behavior will be the subject of the forthcoming work.

Our network ∇_∞ is constructed from infinitely many elementary "triangular" networks ∇ (described below, in Sect. 2.1). A single triangle network $\nabla = \nabla_1$ with N customers is just a Markov continuous time ergodic jump process with finitely many states. As N becomes large, this Markov process tends (in the appropriate "Euler" limit) to a (5-dimensional) dynamical system Δ, possessing a periodic trajectory $\mathcal{C}$, which turns out to be a stable (local) attractor. The coordinate φ parameterizing that attractor $\mathcal{C}$ is the "phase" alluded to in the previous subsection. The combined network corresponds in the same sense to the coupled family Δ_∞ of dynamical systems Δ. We establish the synchronization property of that coupled family Δ_∞, and that allows us to construct coherent states of the network ∇_∞.

The networks ∇_M, composed of M triangle networks ∇, are ergodic. Their evolution is given by irreducible finite state Markov processes with continuous time. Let π_M be the invariant measure of the process ∇_M. As $M \to \infty$, the sequence of Markov processes ∇_M converges weakly on finite time intervals to a certain limiting (non-linear Markov) process ∇_∞. By the theorem of Khasminsky—see Theorem 1.2.14 in [10]—any accumulation point of the sequence π_M is a stationary measure of ∇_∞. The special measure χ_∞, describing "the Poisson Hypothesis behavior", is also a stationary measure of ∇_∞. If χ_∞ is a global attractor of ∇_∞, then, of course, the Poisson Hypothesis holds. The proof of the Poisson Hypothesis in [14] was based on this argument. The existence of an accumulation point of the sequence π_M that differs from χ_∞ would be the strongest counterexample to the Poisson Hypothesis.

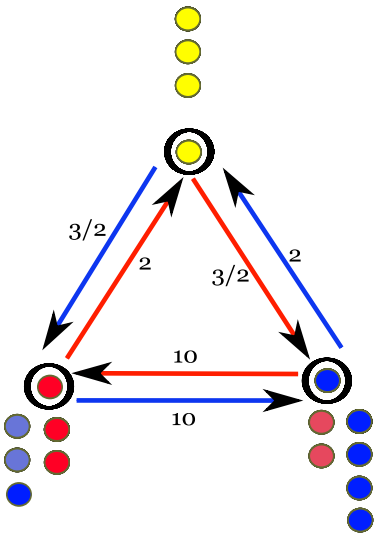

Fig. 1 (Color online) The elementary network

This problem will be addressed in forthcoming papers. Here we prove a weaker statement that χ_∞ is not a global attractor for ∇_∞.

In [16] Rybko and Stolyar observed that the condition that the workload at every node of a multiclass open queueing network is less than 1 is not sufficient for the network to be ergodic. In connection with this, they introduced a new approach to the analysis of ergodicity of networks, which reduces the problem to the question of stability of the associated fluid models. It was shown by them that the two-node priority network, considered in [16], is ergodic if and only if for every initial state of the corresponding fluid model the total amount of fluid vanishes eventually. Analogous deterministic examples were found by Kumar and his collegues, see, for instance, [8]. This approach was further developed by Dai [3], Stolyar [18], and Puhalsky and Rybko [13], who proved that stability of the fluid model is necessary and sufficient for ergodicity of a certain class of general networks. Interesting instances of non-ergodic queueing networks with mean load being smaller than the capacity, where considered by Bramson [1, 2]. Our construction will be based on the following open network introduced by Rybko and Stolyar (RS-network) in [16], see Fig. 1.

This queuing network with four types of customers is represented by the following 4-dimensional Markov process. Customers arrive to the network according to Poisson inflows of constant rate λ. The network consists of two nodes—$\bar{A}$ and $\bar{B}$. All the service times are exponential, hence the network is defined by the rates, the evolution of types of the customers and the priorities. The customer of type A (respectively, B) arrives to the node $\bar{A}$ (respectively, $\bar{B}$). The customer A is served with the rate γ_A, then is sent to $\bar{B}$, with type AB. There he is served with the rate γ_{AB} and leaves the network. Symmetrically, $\gamma_B = \gamma_A$, and $\gamma_{BA} = \gamma_{AB}$. Each customer AB is served before all the customers B, while each customer BA is served before all the customers A. The nominal workload at nodes $\bar{A}$ and $\bar{B}$ equals $\rho = \lambda(\gamma_A^{-1} + \gamma_{BA}^{-1})$. The service rates satisfy the conditions $\gamma_{AB} < 2\lambda$ and $\rho < 1$. It is proved in [16] that for certain values of the parameters the resulting Markov process is transient. The fluid limit (or the Euler limit) of this network evolves in the following non-trivial manner: each node is empty during a positive fraction of time, but at other moments it is non-empty, and, moreover, the total amount of the fluid in the network tends linearly to infinity.

The rest of the paper is organized as follows. In Sect. 2 we define our networks ∇_M^N. Here M is the size of the network and N is the load per node. We formulate the preliminary version of our Main Result. In Sect. 3 we study the limiting network, ∇_∞^N, and prove the convergence $\nabla_M^N \to \nabla_\infty^N$. In Sect. 4 we introduce the fluid networks, Δ_M, which are coupled dynamical systems, and their limit, Δ_∞, which turns out to be a non-linear dynamical system, in the sense made precise in this section. In particular, we show that Δ_∞ is not ergodic. In the next Sect. 5 we prove the convergence of the Non-Linear Markov Process ∇_∞^N to its Euler fluid limit, Δ_∞, as $N \to \infty$. The last Sect. 6 contains the formulation and the proof of our main result, Theorem 17.

To save on notation, we consider throughout this paper the simplest elementary symmetric model, depending on 3 parameters. We stress the fact that this (discrete) symmetry is not essential in our case, and our results are valid for any small 6D-perturbation of our model.

2 Mean-Field Network and Its Limit

2.1 Basic Network

We will consider the following 5-dimensional Markov process, ∇^N. It describes a closed queuing network with N customers. It consists of three nodes: $\bar{O}$, $\bar{A}$ and $\bar{B}$, through which the customers go. All the service times are exponential, so we only need to specify the rates, the evolution of types of the customers and the priorities. To simplify the presentation we will make a specific choice of these rates. The node $\bar{O}$ serves all the customers on the FIFO basis, with the rate $\gamma_O = 3$. After being served, the customer goes to the node $\bar{A}$ or to $\bar{B}$, choosing one of them with probability $\frac{1}{2}$. If he arrives to $\bar{A}$, he gets the type A, otherwise B. The customer A is served with the rate $\gamma_A = 10$, then is sent to $\bar{B}$, with type AB. There he is served with the rate $\gamma_{AB} = 2$ and goes back to $\bar{O}$. Symmetrically, $\gamma_B = 10$, and $\gamma_{BA} = 2$. Each customer AB is served before all the customers B, and each customer BA is served before all the customers A. More precisely, if an AB customer arrives to the $\bar{B}$ node, while the node is serving some B customer, his service is stopped and is resumed only at the moment when the service of all AB customers is over.

Of course, the above choice of the rates is not the only possible. Any other choice would be as good, provided the corresponding fluid network, which can be associated to our queuing network, has some specific property—namely, we need this fluid network to have a *cyclic regime*. The fluid network will be described in details in Sect. 4.2 below.

2.2 M Coupled Processes

Let ∇_M^N be the Markov process, obtained from M copies of ∇^N, interconnected in the mean-field manner, see Fig. 2. We take the total number of customers to be NM.

The mean field network is defined as follows. Each node $\bar{O}_i$, $i = 1, \dots, M$, is connected to *all* of the nodes $\bar{A}_j$, $\bar{B}_j$, $j = 1, \dots, M$, and each customer, leaving the node $\bar{O}_i$, goes to each of the $2M$ nodes $\bar{A}_j$, $\bar{B}_j$ with the same probability $\frac{1}{2M}$. The rate of leaving the node $\bar{O}_i$ is the same, as above, i.e. equals to $\gamma_O = 3$. In a similar way, the A customers of every node $\bar{A}_i$ are exiting it with the rate $\gamma_{AB} = 10$, and then choose one of the $\bar{B}_j$ nodes with probability $\frac{1}{M}$, and so on. The priorities are kept the same: if the node $\bar{A}_i$, say, is in the state with customers of both kinds—A and BA—present, then the BA customers are served first, with no delay.

The configuration of the process is given by the number of customers of each type at each of the $3M$ nodes, that is by an integer point in $(\mathbb{R}^{5M})^+$. Due to the mean-field symmetry, we

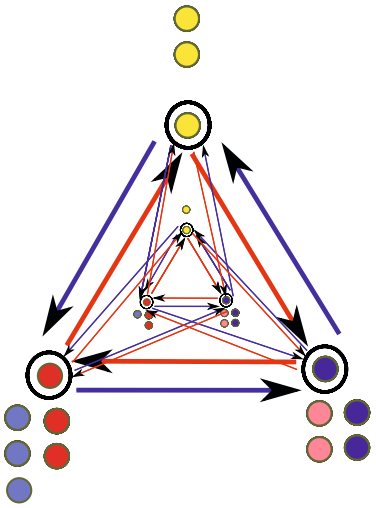

Fig. 2 (Color online) Two coupled processes

can factor the set of configurations by the product of permutation groups $S_M \times S_M \times S_M$, and still have the Markov process. The orbit of the symmetry group corresponds to a collection of M^3 integer points $\bar{x}_i \in (\mathbb{R}^5)^+$, some of which may coincide.

It is convenient for us to index these configurations by the atomic measures,

$$\{\bar{x}_i\} \rightsquigarrow \frac{1}{M^3} \sum_{i=1}^{M^3} \delta_{\frac{\bar{x}_i}{N}}. \tag{1}$$

In fact, they belong to the set $\mathcal{M}((\frac{1}{N}\mathbb{Z}^5)^+)$. Note that every such measure μ factors into a product

$$\mu \equiv \left(\mu_O, \mu_{\bar{A}}, \mu_{\bar{B}}\right) \equiv \mu_O \times \mu_{\bar{A}} \times \mu_{\bar{B}} \equiv \Pi_{\bar{O}}[\mu] \times \Pi_{\bar{A}}[\mu] \times \Pi_{\bar{B}}[\mu]$$

of probability measures on $\mathbb{R}^1 = \{x_O\}$, resp. $\mathbb{R}^2 = \{x_A, x_{BA}\}$ and $\mathbb{R}^2 = \{x_B, x_{AB}\}$. Here we denote by Π_*-s the various projections (or marginals). We have $\mu_O = \frac{1}{M}\sum_{i=1}^{M} \delta_{\frac{\bar{x}_i}{N}}$ for some (not necessarily distinct) $\bar{x}_i \in \mathbb{Z}^1$, $i = 1, \ldots, M$, likewise $\mu_{\bar{A}} = \frac{1}{M}\sum_{i=1}^{M} \delta_{\frac{\bar{x}'_i}{N}}$, $\mu_{\bar{B}} = \frac{1}{M}\sum_{i=1}^{M} \delta_{\frac{\bar{x}''_i}{N}}$, $\bar{x}'_i, \bar{x}''_i \in \mathbb{Z}^2$. We will denote the set of all such measures by $\mathcal{M}_M$. The state ν_M^N of our Markov process is then an element from $\mathcal{M}(\mathcal{M}_M)$, i.e. a measure on the measure space. Among these there are configurations of the process ∇_M, namely, the δ-measures δ_m, with $m \in \mathcal{M}_M$, so we can define in this way the embedding $\mathcal{M}_M \subset \mathcal{M}(\mathcal{M}_M)$. However, even if the initial state $\nu_M^N(0)$ of ∇_M happens to be such a measure δ_m, i.e. $\nu_M^N(0) \in \mathcal{M}_M$, then at any positive t we have only that $\nu_M^N(t) \in \mathcal{M}(\mathcal{M}_M)$, while in general $\nu_M^N(t) \notin \mathcal{M}_M$.

For the future use we will write down the rates of the factor-process. Let v be some measure of the form (1), while v' be the measure obtained from v after a single jump of the initial process. For example, let us consider the case when the jump in question is of $AB \to O$ type, from $\bar{B}$-type server to $\bar{O}$ server (with the rate γ_{AB}). That means that for some unique well-defined (by the pair v, v') elements $x_{\bar{B}} = (x_B, x_{AB}) \in (\frac{1}{N}\mathbb{Z}^2)^+$, $x_O \in (\frac{1}{N}\mathbb{Z}^1)^+$ we have:

$$v'\left(x_{\bar{B}}\right) = v\left(x_{\bar{B}}\right) - \frac{1}{M}, \qquad v'\left(x_O\right) = v\left(x_O\right) - \frac{1}{M}. \tag{2}$$

Of course, for another pair: $\tilde{x}_{\bar{B}} = (x_B, x_{AB} - 1)$, $\tilde{x}_O = x_O + 1$, we have

$$v'\left(\tilde{x}_{\bar{B}}\right) = v\left(\tilde{x}_{\bar{B}}\right) + \frac{1}{M}, \qquad v'\left(\tilde{x}_O\right) = v\left(\tilde{x}_O\right) + \frac{1}{M}, \tag{3}$$

while at all other locations the two measures are the same. Since there are $Mv(x_{\bar{B}})$ locations where the jump could originate, and the fraction of sites with the desirable outcome is $v(x_O)$, we have for the rate $c(v, v') \equiv c_{AB}(v, v')$ the expression

$$c\left(v, v'\right) = \gamma_{AB} M v\left(x_{\bar{B}}\right) v\left(x_O\right).$$

If the measures v, v' are not related by (2)–(3), then the rate $c_{AB}(v, v') = 0$.

We will keep the notation ∇_M for the factor-process.

2.3 $M \to \infty$ Limit: Non-Linear Markov Process

Suppose that a sequence of initial states $\nu_M(0) \in \mathcal{M}(\mathcal{M}_M)$ of the Markov processes ∇_M is given, which satisfy $\nu_M(0) = \delta_{m_M}$, with $m_M \in \mathcal{M}_M$, and moreover the weak limit $\nu = \lim_{M\to\infty} m_M$ exists. Then the weak limits $\nu(t) = \lim_{M\to\infty} \nu_M(t)$ exist for every t, and, moreover, for every t we have $\nu(t) \in \mathcal{M}$. This is the Non-Linear Markov Process, NLMP, ∇_∞. The process is called Non-Linear since the transition mechanism to evolve from a given configuration depends not only on that configuration, but also on the measure from which this configuration was drawn. Such processes were introduced in [11, 12], see also [14]. The above limiting NLMP-s depend on the parameter N, which is the number of clients per basic queuing network. We want to study the dependence on N, so we explicitly (re)introduce the index N in our notation. Thus, $\nu^N(t)$ refers to the states of the process ∇^N_∞.

We will describe the limiting NLMP in the next Sect. 3. Now we can formulate the preliminary version of our main result.

Theorem 1 *Consider the Non-Linear Markov Process ∇_∞, started from the measure ν_0^N, which is close enough to the atomic measure with the single atom at vector $\bar{X}(A, N) \in (\frac{1}{N}\mathbb{Z}^5)^+$, having coordinate $x_A = 1$ and all other coordinates zero. "Close enough" here means that for some $\varepsilon > 0$ small enough we have $\rho_{KROV}(\nu_0^N, \delta_{\bar{X}(A,N)}) < \varepsilon$. Suppose additionally that the α-exponential moment of the measure ν_0^N is less than a certain quantity E; $\alpha = \alpha(\varepsilon)$, $E = E(\varepsilon)$. Then the measure ν_t^N does not converge to any limit as $t \to \infty$, provided N is large enough.*

More precisely, there exists a sequence of times $t'_k \to \infty$, such that

$$\nu^N_{t'_k}\left[U_N\left(\bar{X}\left(A, N\right)\right)\right] > 1 - \delta_N,$$

with $\delta_N \to 0$ as $N \to \infty$. Here $U_N(\bar{X}(A, N))$ is a neighborhood of $\bar{X}(A, N)$ of radius $\varkappa_N$, with $\varkappa_N \to 0$ as $N \to \infty$. At the same time, there exists another sequence $t''_k \to \infty$, for which $\nu^N_{t''_k}[U_N(\bar{X}(B, N))] > 1 - \delta_N$. In words, the measure ν_t exhibits oscillations.

Accordingly, the states of finite size networks, $(\nu^N_M)_t$, exhibit oscillations for long times, before converging to their limits. The duration of the oscillation regime diverges with M. Different components of $(\nu^N_M)_t$ are oscillating almost coherently, for large M.

3 The Convergence $\nabla_M^N \to \nabla_\infty^N$: Application of the Trotter-Kurtz Theorem

Here we prove the convergence of the Markov processes ∇_M^N to the Non-Linear Markov process ∇_∞^N. We will do that by writing down their generators A_M and A, and by subsequent application of the Trotter-Kurtz theorem (Proposition 1.3.3 in [5]), which we formulate now.

Let $A_M, A : X \to X$ are (unbounded) operators on the Banach space X, and $X_0 \subset X$ is a dense subspace, belonging to the domains of definition of all A_M-s and A. The following two conditions are sufficient for the convergence of the semigroups $\exp\{tA_M\} \to \exp\{tA\}$ on X as $M \to \infty$:

1. $\forall \psi \in X_0$ we have $A_M(\psi) \to A(\psi)$ as $M \to \infty$;
2. there exists a dense subspace $X_1 \subset X_0$, such that $\forall \psi \in X_1$ we have $\exp\{tA\}(\psi) \in X_0$.

Such subspace X_0 is called a *core* of A.

3.1 Equation for the Evolution ∇_∞^N

Here we study the limiting process ∇_∞^N. We write down its generator, and we exhibit its core.

Let ν_t be the evolution of the measure under ∇_∞^N. To find it we have to specify the initial measure ν_0 and then to solve the Cauchy problem for the differential equation, which equation we will write now.

To do it we first introduce the (Poisson) rates $\bar{\lambda}(t) = (\lambda_O(t), \lambda_A(t), \lambda_B(t), \lambda_{AB}(t), \lambda_{BA}(t))$, corresponding to the state ν_t:

$$\lambda_a(t) = \gamma_a \sum_{x: x_a > 0} \nu_t(x), \qquad \text{for } a = O, AB, BA, \tag{4}$$

$$\lambda_A(t) = \gamma_A \sum_{x: x_A > 0, x_{BA} = 0} \nu_t(x), \qquad \lambda_B(t) = \gamma_B \sum_{x: x_B > 0, x_{AB} = 0} \nu_t(x). \tag{5}$$

We also introduce the 5D vectors Δ_a to be the basis vectors of the lattice $\mathbb{Z}^5$. Then

$$\begin{aligned}
\frac{d\nu_t(x)}{dt} &= -\nu_t(x) \left(\sum_{a=O,A,B,AB,BA} \lambda_a(t) \right) \\
&\quad - \nu_t(x) \left(\sum_{a=O,AB,BA} \gamma_a \left(1 - \delta_{x_a}\right) + \gamma_A \left(1 - \delta_{x_A}\right) \delta_{x_{BA}} + \gamma_B \left(1 - \delta_{x_B}\right) \delta_{x_{AB}} \right) \\
&\quad + \nu_t(x - \Delta_O) \left(1 - \delta_{x_O}\right) (\lambda_{AB}(t) + \lambda_{BA}(t)) \\
&\quad + \sum_{a=A,B} \nu_t(x - \Delta_a) \left(1 - \delta_{x_a}\right) \frac{\lambda_O(t)}{2} \\
&\quad + \nu_t(x - \Delta_{AB}) \left(1 - \delta_{x_{AB}}\right) \lambda_A(t) + \nu_t(x - \Delta_{BA}) \left(1 - \delta_{x_{BA}}\right) \lambda_B(t) \\
&\quad + \sum_{a=O,AB,BA} \nu_t(x + \Delta_a) \gamma_a \\
&\quad + \nu_t(x + \Delta_A) \gamma_A \delta_{x_{BA}} + \nu_t(x + \Delta_B) \gamma_B \delta_{x_{AB}}.
\end{aligned} \tag{6}$$

This is the value of the function $A\varphi_x$, where A is the generator of the Markov semigroup $S_t = \exp\{tA\}$ of the process ∇_∞^N, and the function φ_x, which on every measure $\nu \in \mathcal{M}(\mathbb{Z}^{5+})$

takes value $\nu(x)$, computed at the point ν_t. As we will show below, the system (6) has a unique solution.

For reasons which will be explained later, it will be more convenient for us to use another basis in the space of functions on measures. Namely, for every $\nu \in \mathcal{M}(\mathbb{Z}^{5+})$ and every $x \in \mathbb{Z}^{5+}$ we define the function $u(x) = \sum_{y \geq x} \nu(y)$, where the summation goes over all sites y such that all the coordinates of the difference $y - x$ are non-negative. Then the functions $\lambda_a(t)$ (see (4)) are given in the new variables as

$$\lambda_a(t) = \gamma_a u_t(\Delta_a),$$

while the action of the generator A on the function u is given by the following (simpler) equation:

$$\begin{aligned}
\frac{du_t(x)}{dt} = &- \sum_{a=O,AB,BA} (u_t(x) - u_t(x+\Delta_a))\,\gamma_a\left(1-\delta_{x_a}\right) \\
&- (u_t(x) - u_t(x+\Delta_A) - u_t(x+\Delta_{AB}) \\
&+ u_t(x+\Delta_A+\Delta_{AB}))\,\gamma_A\left(1-\delta_{x_A}\right)\delta_{x_{BA}} \\
&- (u_t(x) - u_t(x+\Delta_B) - u_t(x+\Delta_{BA}) \\
&+ u_t(x+\Delta_B+\Delta_{BA}))\,\gamma_B\left(1-\delta_{x_B}\right)\delta_{x_{AB}} \\
&+ (u_t(x-\Delta_O) - u_t(x))\left(1-\delta_{x_O}\right)(\gamma_{AB}u_t(\Delta_{AB}) + \gamma_{BA}u_t(\Delta_{BA})) \\
&+ \sum_{a=A,B} (u_t(x-\Delta_a) - u_t(x))\left(1-\delta_{x_a}\right)\frac{\gamma_O u_t(\Delta_O)}{2} \\
&+ (u_t(x-\Delta_{AB}) - u_t(x))\left(1-\delta_{x_{AB}}\right)\gamma_A u_t(\Delta_A+\Delta_{BA}) \\
&+ (u_t(x-\Delta_{BA}) - u_t(x))\left(1-\delta_{x_{BA}}\right)\gamma_B u_t(\Delta_B+\Delta_{AB}).
\end{aligned} \tag{7}$$

The first five lines correspond to the second line of (6), while the last four—to the lines 3–5; the remaining lines of it disappear from the equations for u. The advantage of (7) over (6) is that the equation for $\frac{du_t(x)}{dt}$ contains only $u_t(y)$-s with y-s in some finite set $Y(x)$, and moreover $\max_x |Y(x)| = 20$.

Of course, the coordinates $u(\cdot)$-s on $\mathcal{M}(\mathbb{Z}^{5+})$ are not independent. There are two kinds of relations between them:

1. every value $\nu(x)$ equals to $L_x(u)$, where L is a certain linear form, depending on $u(y)$ with y-s having form $y = x + \sum e_a \Delta_a$, $e_a = 0, 1$; we need that for all x $L_x(u) \geq 0$;
2. $$\lim_{x \to \infty} u(x) = 0. \tag{8}$$

For technical reasons we will extend the action of our Markov semigroup to the space $\mathcal{M}(\mathbb{K})$ of measures on the compactification $\mathbb{K}$ of the lattice $\mathbb{Z}^{5+}$, where

$$\mathbb{K} = \left\{\mathbb{Z}^+ + \infty\right\}^5.$$

The functions $\{u(x), x \in \mathbb{Z}^{5+}\}$ on $\mathcal{M}(\mathbb{K})$ also play the role of coordinates there, provided that the relation (8) is dropped. The evolution of the measures is given by the same set of (7).

We supply $\mathcal{M}(\mathbb{K})$ with the topology of weak convergence. (We repeat for clarity that the subset $\mathcal{M}(\mathbb{Z}^{5+}) \subset \mathcal{M}(\mathbb{K})$ is invariant under our semigroup.) Let $\mathcal{C}^0 = \mathcal{C}(\mathcal{M}(\mathbb{K}))$ be the space of functions on $\mathcal{M}(\mathbb{K})$, continuous with respect to this topology.

Theorem 2 *The semigroup S_t acts on the space $\mathcal{C} = \mathcal{C}(\mathcal{M}(\mathbb{K}))$ of continuous functions on $\mathcal{M}(\mathbb{K})$, and is strongly continuous and contracting.*

Proof 1. Let us show the existence of the solutions to (7). The equations (7) are describing the evolution of the "closed" system. Consider now the corresponding "open" system, defined by the (arbitrary) rates $\bar{\lambda} = \lambda_a(t)$ of the Poisson inflows and the initial state u_0. It evolves according to the equations

$$\begin{aligned}
\frac{du_t(x)}{dt} = &- \sum_{a=O,AB,BA} \left(u_t(x) - u_t(x+\Delta_a)\right)\gamma_a\left(1-\delta_{x_a}\right) \\
&- \left(u_t(x) - u_t(x+\Delta_A) - u_t(x+\Delta_{AB})\right. \\
&\left. + u_t(x+\Delta_A+\Delta_{AB})\right)\gamma_A\left(1-\delta_{x_A}\right)\delta_{x_{BA}} \\
&- \left(u_t(x) - u_t(x+\Delta_B) - u_t(x+\Delta_{BA})\right. \\
&\left. + u_t(x+\Delta_B+\Delta_{BA})\right)\gamma_B\left(1-\delta_{x_B}\right)\delta_{x_{AB}} \\
&+ \left(u_t(x-\Delta_O) - u_t(x)\right)\left(1-\delta_{x_O}\right)\left(\lambda_{AB}(t)+\lambda_{BA}(t)\right) \\
&+ \sum_{a=A,B} \left(u_t(x-\Delta_a) - u_t(x)\right)\left(1-\delta_{x_a}\right)\frac{\lambda_O(t)}{2} \\
&+ \left(u_t(x-\Delta_{AB}) - u_t(x)\right)\left(1-\delta_{x_{AB}}\right)\lambda_A(t) \\
&+ \left(u_t(x-\Delta_{BA}) - u_t(x)\right)\left(1-\delta_{x_{BA}}\right)\lambda_B(t).
\end{aligned}$$

The corresponding exit rates b_a are given by the natural relations

$$b_a^{\bar{\lambda}}(t) = \gamma_a u_t(\Delta_a).$$

Consider the function $\bar{d}^{\bar{\lambda}}(t)$:

$$\begin{aligned}
d_O^{\bar{\lambda}}(t) &= b_{AB}^{\bar{\lambda}}(t) + b_{BA}^{\bar{\lambda}}(t), \\
d_A^{\bar{\lambda}}(t) &= d_B^{\bar{\lambda}}(t) = \frac{1}{2} b_O^{\bar{\lambda}}(t), \\
d_{AB}^{\bar{\lambda}}(t) &= b_A^{\bar{\lambda}}(t), \qquad d_{BA}^{\bar{\lambda}}(t) = b_B^{\bar{\lambda}}(t).
\end{aligned}$$

The closed system is a fixed point of the map $\bar{\lambda} \overset{\psi_{u_0}}{\rightsquigarrow} \bar{d}^{\bar{\lambda}}$, i.e. a solution of the equation

$$\bar{\lambda} = \bar{d}^{\bar{\lambda}}.$$

To see the existence of a fixed point, let us introduce the functions $\bar{\Lambda}$, $\bar{B}^{\bar{\Lambda}}$ and $\bar{D}^{\bar{\Lambda}}$:

$$\Lambda_a(t) = \int_0^t \lambda_a(t)dt, \qquad B_a^{\bar{\Lambda}}(t) = \int_0^t b_a^{\bar{\lambda}}(t)(t)dt, \qquad D_a^{\bar{\Lambda}}(t) = \int_0^t d_a^{\bar{\lambda}}(t)dt,$$

and the corresponding mapping $\bar{\Lambda} \overset{\Psi_{u_0}}{\rightsquigarrow} \bar{D}^{\bar{\Lambda}}$. The functions $\bar{\Lambda}$, $\bar{B}^{\bar{\Lambda}}$ and $\bar{D}^{\bar{\Lambda}}$ are monotone continuous, moreover, the functions $\bar{B}^{\bar{\Lambda}}$ and $\bar{D}^{\bar{\Lambda}}$ have uniformly bounded derivatives. Let $\mathfrak{c}$ be the upper bound for these derivatives, and $\mathfrak{C}$ be the space of all continuous monotone 5D vector-functions on $[0, T]$, vanishing at zero, with the derivatives bounded by $\mathfrak{c}$ once they exist. Then $\mathfrak{C}$ is compact and convex, therefore the map $\Psi_{u_0} : \mathfrak{C} \to \mathfrak{C}$ has at least one fixed point.

2. We now will show that for every u_0 the map Ψ_{u_0} is a contraction; that will imply the uniqueness of the solution. Without loss of generality we can assume that T is small. Informally, the contraction takes place because the exit rates $b_a^{\bar{\lambda}}(t)$ for $t \in [0, T]$ with T small depend mainly on the initial state u_0 : the new clients, arriving during the time $[0, T]$ have no chance to be served before T, if there were clients already waiting. Therefore the "worst" case for us is when the initial state ν_0 is the measure δ_0, having a unit atom at $0 \in \mathbb{Z}^{5+}$.

So let $\lambda_1(t), \lambda_2(t), t \in [0, T]$ be the rates of two Poisson inflows to the empty server, and γ be the service rate. We want to estimate the difference $b_1(t) - b_2(t)$ of the rates of the exit flows. We can couple the two service processes in the following way: let $\lambda(t) = \min\{\lambda_1(t), \lambda_2(t)\}$. Then we write $\lambda_i(t) = \lambda(t) + \eta_i(t)$, where $\eta_i(t) = \lambda_i(t) - \lambda(t)$. We will call the clients arriving with the rate $\lambda(t)$ as colorless, and we call the $\eta_1(t)$ clients as red, while the $\eta_2(t)$ clients as blue. The colorless clients have priority in their service: if a colorless client arrives, then all the colored ones have to wait—even the one currently under the service. Then the difference $|b_1(t) - b_2(t)|$ is bounded from above by the sum of the exit rates of colored clients, which does not exceed

$$|b_1(t) - b_2(t)| \le \gamma \Pr\begin{pmatrix}\text{server is occupied by a}\\ \text{colored client at the moment } t\end{pmatrix} \le \gamma \int_0^t |\lambda_1(t) - \lambda_2(t)|\, dt.$$

Hence we have contraction with the contraction rate at most γT, which is small for small T. We denote by $\bar{\lambda}_u(t)$ the unique fixed point of Ψ_u.

3. Finally we prove that the semigroup preserves the space of continuous functions. First of all we observe that the map Ψ_u depends on the initial measure u in a continuous way. Therefore the same is true for $\bar{\lambda}_u(t)$, the fixed point of Ψ_u. Hence $u(t)$ depends continuously on $u(0)$. □

Let us consider the subspace $\mathcal{C}^2 \subset \mathcal{C}^0$ of functions f, which have the following properties:

1. for every $x \in \mathbb{Z}^{5+}$ the function f has the first derivative $\frac{\partial f}{\partial u(x)}$;
2. for every $x, y \in \mathbb{Z}^{5+}$ the function f has the second derivative $\frac{\partial^2 f}{\partial u(x)\partial u(y)}$;
3. all these derivatives are bounded, uniformly in x, y.

It is easy to see that the set of coordinate functions $\{u(x), x \in \mathbb{Z}^{5+}\}$ on $\mathcal{M}(\mathbb{K})$ can distinguish any two measures from $\mathcal{M}(\mathbb{K})$. Due to the compactness of $\mathcal{M}(\mathbb{K})$ we can apply the Stone-Weierstrass theorem, which implies that the subspace $\mathcal{C}^2 \subset \mathcal{C}^0$ is dense in $\mathcal{C}^0$. We now will show the following

Proposition 3 *For every t we have $S_t(\mathcal{C}_0^2) \subset \mathcal{C}^2$, where the subspace $\mathcal{C}_0^2 \subset \mathcal{C}^2$ consists of all functions depending only on finitely many variables $\{u(x)\}$. In particular, the subspace $\mathcal{C}^2$ is a core of the generator A.*

Proof To do this we will use Proposition 1 of the paper [4]:

Lemma 4 *Consider the infinite system of equations*

$$\frac{d}{dt}z_k(t) = \sum_i a_{ki}(t)z_i(t) + b_k(t), \quad t \geq 0.$$

Suppose that for all k

$$\sum_i |a_{ki}(t)| \leq a, \qquad |b_k(t)| \leq b_0 \exp\{bt\}, \qquad |z_k(0)| \leq c,$$

with $a < b$. Then

$$|z_k(t)| \leq c \exp\{at\} + \frac{b_0}{b-a}\left(\exp\{bt\} - \exp\{at\}\right).$$

From (7) it follows immediately, that

$$\frac{d}{dt}\left(\frac{\partial u_t(v)}{\partial u_0(x)}\right) = \sum_{w \in Y(v)} \bar{a}_{vw}(t)\left(\frac{\partial u_t(w)}{\partial u_0(x)}\right),$$

with $\sum_{w\in Y(v)} |\bar{a}_{vw}(t)| < \hat{a}$, $|\frac{\partial u_0(v)}{\partial u_0(x)}| \leq 1$, for some $\hat{a} < \infty$, uniformly in v, so Lemma 4 applies, and all the derivatives $|\frac{\partial u_t(v)}{\partial u_0(x)}|$ are uniformly bounded, provided $t < T$. Further on,

$$\frac{d}{dt}\left(\frac{\partial^2 u_t(v)}{\partial u_0(x)\,\partial u_0(y)}\right) = \sum_{w \in Y(v)} \bar{a}_{vw}(t)\left(\frac{\partial^2 u_t(w)}{\partial u_0(x)\,\partial u_0(y)}\right) + \tilde{b}_v(t),$$

with the same $\bar{a}_{vw}(t)$-s, while the term $\tilde{b}_v(t)$, consisting of the products of the first derivatives $\frac{\partial u_t(w')}{\partial u_0(x)}\frac{\partial u_t(w'')}{\partial u_0(y)}$, is also uniformly bounded, as was just shown, provided $t < T$. Therefore the derivatives $|\frac{\partial^2 u_t(v)}{\partial u_0(x)\partial u_0(y)}|$ are uniformly bounded as well.

For other functions we just use the chain rule. □

3.2 Equation for the Evolution ∇_M^N and the Convergence

Now we will write the generator A_M of the process ∇_M^N. Let $\psi(\cdot)$ be a function on $\mathcal{M}_M$. (In fact, we need it to be defined on a smaller set $\mathcal{M}_M \cap \mathcal{M}((\frac{1}{N}\mathbb{Z}^5)^+)$. Throughout this section the value of N will be fixed, and we will keep it just 1, in order to simplify the notation.) We will introduce the following notations for the increments of the measure $v = (v_{\bar{O}}, v_{\bar{A}}, v_{\bar{B}})$ (which are themselves (signed) measures on $\mathbb{Z}^1$ or $\mathbb{Z}^2$):

$$\Delta_{O,x}(y) = \begin{cases} 1 & \text{if } y = x, \\ -1 & \text{if } y = x-1, \quad x, y \in \mathbb{Z}^1, \\ 0 & \text{otherwise,} \end{cases}$$

$$\Delta_{A,x_{\bar{A}}}(y) = \begin{cases} 1 & \text{if } y = x_{\bar{A}} \equiv (x_A, x_{BA}), \\ -1 & \text{if } y = (x_A - 1, x_{BA}), \quad x_{\bar{A}}, y \in \mathbb{Z}^2, \\ 0 & \text{otherwise,} \end{cases}$$

$$\Delta_{BA,x_{\bar{A}}}(y)=\begin{cases}1 & \text{if } y=x_{\bar{A}}\equiv(x_A,x_{BA}),\\ -1 & \text{if } y=(x_A,x_{BA}-1),\\ 0 & \text{otherwise,}\end{cases}\quad x_{\bar{A}},y\in\mathbb{Z}^2,$$

and similar definitions for the remaining measures $\Delta_{B,x_{\bar{B}}}$ and $\Delta_{AB,x_{\bar{B}}}$. Then

$$\begin{aligned}(A_M\psi)(v)&=\sum_{v'}c\left(v,v'\right)\left[\psi\left(v'\right)-\psi\left(v\right)\right]\\
&=\sum_{x_O\geq 1}M\frac{\gamma_O}{2}v_{\bar{O}}(x_O)\sum_{x_{\bar{A}}}v_{\bar{A}}\left(x_{\bar{A}}\right)\\
&\quad\times\left[\psi\left(v_{\bar{O}}-\frac{\Delta_{O,x_O}}{M},v_{\bar{A}}+\frac{\Delta_{A,x_{\bar{A}}+(1,0)}}{M},v_{\bar{B}}\right)-\psi\left(v_{\bar{O}},v_{\bar{A}},v_{\bar{B}}\right)\right]\\
&\quad+\sum_{x_O\geq 1}M\frac{\gamma_O}{2}v_{\bar{O}}(x_O)\sum_{x_{\bar{B}}}v_{\bar{B}}\left(x_{\bar{B}}\right)\\
&\quad\times\left[\psi\left(v_{\bar{O}}-\frac{\Delta_{O,x_O}}{M},v_{\bar{A}},v_{\bar{B}}+\frac{\Delta_{B,x_{\bar{B}}+(1,0)}}{M}\right)-\psi\left(v_{\bar{O}},v_{\bar{A}},v_{\bar{B}}\right)\right]\\
&\quad+\sum_{x_{\bar{A}}:x_{BA}\geq 1}M\gamma_{BA}v_{\bar{A}}\left(x_{\bar{A}}\right)\sum_{x_O}v_{\bar{O}}(x_O)\\
&\quad\times\left[\psi\left(v_{\bar{O}}+\frac{\Delta_{O,x_O+1}}{M},v_{\bar{A}}-\frac{\Delta_{BA,x_{\bar{A}}}}{M},v_{\bar{B}}\right)-\psi\left(v_{\bar{O}},v_{\bar{A}},v_{\bar{B}}\right)\right]\\
&\quad+\sum_{x_{\bar{B}}:x_{AB}\geq 1}M\gamma_{AB}v_{\bar{B}}\left(x_{\bar{B}}\right)\sum_{x_O}v_{\bar{O}}(x_O)\\
&\quad\times\left[\psi\left(v_{\bar{O}}+\frac{\Delta_{O,x_O+1}}{M},v_{\bar{A}},v_{\bar{B}}-\frac{\Delta_{AB,x_{\bar{B}}}}{M}\right)-\psi\left(v_{\bar{O}},v_{\bar{A}},v_{\bar{B}}\right)\right]\\
&\quad+\sum_{x_A\geq 1}M\gamma_Av_{\bar{A}}(x_A,0)\sum_{x_{\bar{B}}}v_{\bar{B}}\left(x_{\bar{B}}\right)\\
&\quad\times\left[\psi\left(v_{\bar{O}},v_{\bar{A}}-\frac{\Delta_{A,(x_A,0)}}{M},v_{\bar{B}}+\frac{\Delta_{AB,x_{\bar{B}}+(0,1)}}{M}\right)-\psi\left(v_{\bar{O}},v_{\bar{A}},v_{\bar{B}}\right)\right]\\
&\quad+\sum_{x_B\geq 1}M\gamma_Bv_{\bar{B}}(x_B,0)\sum_{x_{\bar{A}}}v_{\bar{A}}\left(x_{\bar{A}}\right)\\
&\quad\times\left[\psi\left(v_{\bar{O}},v_{\bar{A}}+\frac{\Delta_{BA,x_{\bar{A}}+(0,1)}}{M},v_{\bar{B}}-\frac{\Delta_{B,(x_B,0)}}{M}\right)-\psi\left(v_{\bar{O}},v_{\bar{A}},v_{\bar{B}}\right)\right].\end{aligned}\tag{9}$$

Suppose now that the function ψ is differentiable in each of the variables $v_{\bar{O}}(x_O)$, $x_O\in\mathbb{Z}^1$, $v_{\bar{A}}(x_{\bar{A}})$, $x_{\bar{A}}\in\mathbb{Z}^2$, $v_{\bar{B}}(x_{\bar{B}})$, $x_{\bar{B}}\in\mathbb{Z}^2$. Then each of the six increments in (9) equals to the corresponding derivative ψ' of ψ, computed at some intermediate point. If moreover the function ψ is continuously differentiable (which is implied by the twice differentiability), we can take a limit as $M\to\infty$, obtaining the convergence to the limiting operator

$$
\begin{aligned}
(A\psi)(v) = & \sum_{x_O \geq 1, x_{\bar{A}}} \frac{\gamma_O}{2} v_{\bar{O}}(x_O)\, v_{\bar{A}}\left(x_{\bar{A}}\right)\left[\frac{\partial \psi(v)}{\partial(v_{\bar{O}}(x_O-1))} - \frac{\partial \psi(v)}{\partial(v_{\bar{O}}(x_O))}\right. \\
& \left. + \frac{\partial \psi(v)}{\partial(v_{\bar{A}}(x_A+1, x_{BA}))} - \frac{\partial \psi(v)}{\partial(v_{\bar{A}}(x_A, x_{BA}))}\right] \\
& + \sum_{x_O \geq 1, x_{\bar{B}}} \frac{\gamma_O}{2} v_{\bar{O}}(x_O)\, v_{\bar{B}}\left(x_{\bar{B}}\right)\left[\frac{\partial \psi(v)}{\partial(v_{\bar{O}}(x_O-1))} - \frac{\partial \psi(v)}{\partial(v_{\bar{O}}(x_O))}\right. \\
& \left. + \frac{\partial \psi(v)}{\partial(v_{\bar{B}}(x_B+1, x_{AB}))} - \frac{\partial \psi(v)}{\partial(v_{\bar{B}}(x_B, x_{AB}))}\right] \\
& + \sum_{x_{\bar{A}}: x_{BA} \geq 1, x_O} \gamma_{BA} v_{\bar{A}}\left(x_{\bar{A}}\right) v_{\bar{O}}(x_O)\left[\frac{\partial \psi(v)}{\partial(v_{\bar{O}}(x_O+1))} - \frac{\partial \psi(v)}{\partial(v_{\bar{O}}(x_O))}\right. \\
& \left. + \frac{\partial \psi(v)}{\partial(v_{\bar{A}}(x_A, x_{BA}-1))} - \frac{\partial \psi(v)}{\partial(v_{\bar{A}}(x_A, x_{BA}))}\right] \\
& + \sum_{x_{\bar{B}}: x_{AB} \geq 1, x_O} \gamma_{AB} v_{\bar{B}}\left(x_{\bar{B}}\right) v_{\bar{O}}(x_O)\left[\frac{\partial \psi(v)}{\partial(v_{\bar{O}}(x_O+1))} - \frac{\partial \psi(v)}{\partial(v_{\bar{O}}(x_O))}\right. \\
& \left. + \frac{\partial \psi(v)}{\partial(v_{\bar{B}}(x_B, x_{AB}-1))} - \frac{\partial \psi(v)}{\partial(v_{\bar{B}}(x_B, x_{AB}))}\right] \\
& + \sum_{x_{\bar{B}}, x_{\bar{A}}: x_A \geq 1} \gamma_A v_{\bar{A}}(x_A, 0)\, v_{\bar{B}}\left(x_{\bar{B}}\right)\left[\frac{\partial \psi(v)}{\partial(v_{\bar{A}}(x_A-1, 0))} - \frac{\partial \psi(v)}{\partial(v_{\bar{A}}(x_A, 0))}\right. \\
& \left. + \frac{\partial \psi(v)}{\partial(v_{\bar{B}}(x_B, x_{AB}+1))} - \frac{\partial \psi(v)}{\partial(v_{\bar{B}}(x_B, x_{AB}))}\right] \\
& + \sum_{x_{\bar{A}}, x_{\bar{B}}: x_B \geq 1,} \gamma_B v_{\bar{B}}(x_B, 0)\, v_{\bar{A}}\left(x_{\bar{A}}\right)\left[\frac{\partial \psi(v)}{\partial(v_{\bar{B}}(x_B-1, 0))} - \frac{\partial \psi(v)}{\partial(v_{\bar{B}}(x_B, 0))}\right. \\
& \left. + \frac{\partial \psi(v)}{\partial(v_{\bar{A}}(x_A, x_{BA}+1))} - \frac{\partial \psi(v)}{\partial(v_{\bar{A}}(x_A, x_{BA}))}\right].
\end{aligned}
$$

Let us apply this formula to the "coordinate" function $\psi(\cdot) = \phi_y(\cdot)$, where $\phi_y(v) = v(y) \equiv v_O(y_O) v_{\bar{A}}(y_{\bar{A}}) v_B(y_{\bar{B}})$. The result is the one given by (6):

$$
\begin{aligned}
(A\phi_x)(v) = & -v(x)\left(\sum_{a=O,A,B,AB,BA} \lambda_a(t)\right) \\
& - v(x)\left(\sum_{a=O,AB,BA} \gamma_a\left(1-\delta_{x_a}\right) + \gamma_A\left(1-\delta_{x_A}\right)\delta_{x_{BA}} + \gamma_B\left(1-\delta_{x_B}\right)\delta_{x_{AB}}\right) \\
& + v(x-\Delta_O)\left(1-\delta_{x_O}\right)\left(\lambda_{AB}(t) + \lambda_{BA}(t)\right) \\
& + \sum_{a=A,B} v(x-\Delta_a)\left(1-\delta_{x_a}\right)\frac{\lambda_O(t)}{2} \\
& + v(x-\Delta_{AB})\left(1-\delta_{x_{AB}}\right)\lambda_A(t) + v(x-\Delta_{BA})\left(1-\delta_{x_{BA}}\right)\lambda_B(t)
\end{aligned}
$$

$$+ \sum_{a=O,AB,BA} v\,(x+\Delta_a)\,\gamma_a$$
$$+ v\,(x+\Delta_A)\,\gamma_A\delta_{x_{BA}} + v\,(x+\Delta_B)\,\gamma_B\delta_{x_{AB}}.$$

Since the space of functions of finitely many v-s coincides with that depending of finitely many u-s, that finishes the proof.

4 Fluid Networks

One of the key ingredients of the proof of our Main result is the investigation of the fluid (Euler) limits of various networks. They are introduced in the present section.

4.1 Fluid Systems with One Fluid

The fluid systems with one fluid are the following simple dynamical systems. Consider the containers $V_1, V_2, \ldots, V_n$, filled (partially) with water. Suppose that some pairs of these containers are connected by (directed) pipes, through which the water can flow. On every pipe i, j there is a pump ρ_{ij} working, with the capacity of sending $\gamma_{ij} \geq 0$ units of water per unit time from V_i to V_j. The pumps are working constantly, and if the container V_i has less water than the pump ρ_{ij} can handle, the result is that the pump sucks in whatever there is. For example, if the network is given by the graph

$$V_1 \overset{\gamma_{12}}{\to} V_2 \overset{\gamma_{23}}{\to} V_3,$$

with $\gamma_{12} = \frac{1}{2}$, $\gamma_{23} = 1$, then in a while the container V_2 will be empty, and the flow along the pipe 23 will be $\frac{1}{2}$ (provided the water supply in V_1 lasts). If a container V_i has several pipes ij_k attached, then the water is shared by the pumps ρ_{ij_k} proportionally to the capacities γ_{ij_k} (this is relevant only in the situation when the level of water in V_i is zero, and all the incoming water immediately leaves it).

Suppose now that at the moment $T = 0$ all the containers are filled with water in the amounts of $v_i(0)$, and then we turn on all the pumps. We are looking on the levels $v_i(t)$, as they are changing in time. It turns out that there exists a time T' (depending on the system), after which the levels $v_i(t)$ become stable and do not change anymore. (Some of them can in fact be zero.) In particular, such a system can never exhibit a cyclic behavior. Of course, the stability of the levels does not imply that the water in the network does not flow. It just means that for every container the amount of fluid entering is equal to the amount of fluid leaving.

We will not provide the proof of this known statement (see [3] and the references there), since we will not use this fact. Below we will consider fluid networks which do exhibit cyclic behavior, and we find such examples among the fluid networks with several kinds of fluids.

4.2 Basic Fluid Network

We will consider the following 5-dimensional dynamical system, Δ, which is a closed version of the open RS-network. It consists of three nodes: $\bar{O}$, $\bar{A}$ and $\bar{B}$, through which various fluids are passing, see Fig. 3. The node $\bar{O}$ has one type of the fluid, in the amount $x_O \geq 0$, and that fluid flows into the nodes $\bar{A}$ and $\bar{B}$ in equal amounts. The rate of this flow $\gamma_O = 3$,

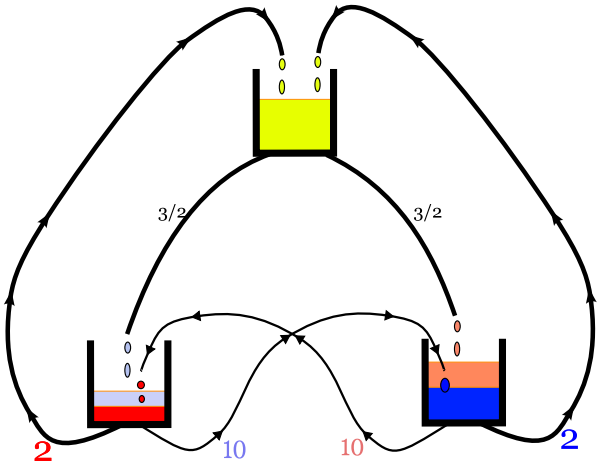

Fig. 3 (Color online) Basic fluid network

which means that three units of the fluid x_O leave $\bar{O}$ per unit time (provided the supply lasts, of course), so each of the two nodes $\bar{A}$ and $\bar{B}$ gets $\frac{3}{2}$ units of incoming fluids, A and B, per unit time. The amounts of these fluids are denoted by x_A and x_B. The fluid A then goes to the node $\bar{B}$, where it turns into the fluid AB, while the fluid B goes to $\bar{A}$ and turns there into BA. The corresponding rates are $\gamma_A = \gamma_B = 10$. The fluids AB and BA then go back to $\bar{O}$, with the rates $\gamma_{AB} = \gamma_{BA} = 2$. The last thing which has to be specified is the following priority: if the node $\bar{A}$ is in the state with both amounts x_A and x_{BA} positive, then *the fluid BA goes first*. One can think that the fluid BA is heavier and of higher viscosity than A, so it goes to the bottom of the node $\bar{A}$ and flows out first (and relatively slow). The same applies to the node $\bar{B}$.

As stated, the system is not well-defined. For example, the dynamics is not specified if we try to start it from the configuration

$$x_A = a > 0, \qquad x_B = b > 0, \qquad x_{BA} = x_{AB} = 0, \qquad x_O = 1 - a - b. \tag{10}$$

The reason is that if the fluid A "starts first", then it will create some amount of the heavy fluid AB in $\bar{B}$, so the fluid B in $\bar{B}$ will be blocked. The same holds for the fluid B "starting first".

The precise definition of the system, given below, follows [3, 16, 18]. Consider first the simplest case of a single node with capacity γ per unit time. Let the initial amount of fluid $x(0)$ be given, an let $Y(t)$ be the net amount of fluid that arrives to the node during the time interval $[0, t]$. In what follows we will call it the *net inflow*. The function $Y(t)$ is monotone non-decreasing function, $Y(0) = 0$, and we assume that $Y(t)$ is Lipschitz continuous, $t \geq 0$. It is easy to check that the evolution of the amount $x(t)$ is given by

$$x(t) = V(t) + U(t), \tag{11}$$

where $V(t) = x(0) + Y(t) - \gamma t$ is called the *virtual level* of the fluid, while $U(t) = \max\{0, -\inf_{s\in[0,t]} V(s)\}$ is the *unused service capacity* of our node.

We introduce for the correspondence (11) the notation

$$x(\cdot) = W(\gamma, x(0), Y(\cdot)), \tag{12}$$

that is, W maps the initial fluid level and the inflow function to the evolution of fluid level. As we will show later (in Lemma 13), the map W is Lipschitz continuous map from $R \times C[0,\infty)$ to $C[0,\infty)$.

Further on, let $Z(t)$ be the total amount of fluid that leaves the node during the time interval $[0,t]$: $Z(t) = x(0) + Y(t) - x(t)$. Again, the function $Z(t)$—the *net outflow*—is monotone non-decreasing Lipschitz continuous, with $Z(0) = 0$. By assumption, the derivatives $z(t) = \dot{Z}(t)$, $y(t) = \dot{Y}(t)$, existing a.e., satisfy

$$z(t) = \begin{cases} \gamma & \text{if } x(t) > 0, \\ y(t) & \text{otherwise.} \end{cases}$$

This property is the reason to call our discipline *work-conserving*; the server is always working at its full capacity.

In the same way we can treat the node through which several fluids are passing. Thus we introduce the vector $\bar{Y}(t) = \{Y_1(t), \ldots, Y_n(t)\}$ of the net inflows during the time interval $[0,t]$, each $Y_i(t)$ being non-decreasing and Lipschitz continuous. The vector $\bar{Z}(t)$ will denote the corresponding collection of the net outflows. (Of course, it does depend on the priorities of the fluids.) Consider again the derivatives $y_i(t) = \dot{Y}_i(t)$ and $z_i(t) = \dot{Z}_i(t)$ (they exist for almost all t and define $Y_i(t)$ and $Z_i(t)$ in a unique way once we fix $Y_i(0)$ and $Z_i(0)$ to be zero). Introduce also the workload rate by $v(t) = \sum y_i(t)\gamma_i^{-1}$, where γ_i are the service rates. The service discipline of our node is work-conserving, if the following property holds: once $\|x(t)\| > 0$, we have

$$\sum z_i(t)\gamma_i^{-1} = 1; \tag{13}$$

otherwise

$$z_i(t) = y_i(t) \tag{14}$$

for all i. The following statement is immediate:

Proposition 5 *Let $\bar{x}(0) = 0$ and $v(t) \le 1$ for almost all $t \ge 0$. Then $\bar{x}(t) = 0$, $t \ge 0$.*

The system described in the beginning of the present subsection corresponds to the following specification of the above general formulation. We have $\bar{Y}(t) \in (\mathbb{R}^5)^+$, and in terms of the map W (see (12)) the evolution is given by the equations:

$$x_O(\cdot) = W(3, x_O(0), Y_O(\cdot)), \tag{15}$$

$$x_{BA}(\cdot) = W(2, x_{BA}(0), Y_{BA}(\cdot)), \tag{16}$$

$$x_A(\cdot) = W(10, 5x_{BA}(0) + x_A(0), 5Y_{BA}(\cdot) + Y_A(\cdot)) - 5x_{BA}(\cdot), \tag{17}$$

the equations for x_B and x_{AB} being symmetric. Then for the derivatives, whenever they exist, we derive from (15)–(17) that

$$\frac{d}{dt}x_O(t) = \begin{cases} y_O(t) - 3 & \text{if } x_O(t) > 0, \text{ or if } x_O(t) = 0 \text{ and } y_O(t) - 3 > 0, \\ 0 & \text{if } x_O(t) = 0 \text{ and } y_O(t) - 3 \le 0; \end{cases}$$

$$\frac{d}{dt}x_{AB}(t) = \begin{cases} y_{AB}(t) - 2 & \text{if } x_{AB}(t) > 0, \text{ or if } x_{AB}(t) = 0 \text{ and } y_{AB}(t) - 2 > 0, \\ 0 & \text{if } x_{AB}(t) = 0 \text{ and } y_{AB}(t) - 2 \le 0; \end{cases}$$

$$\frac{d}{dt}x_A(t) = \begin{cases} y_A(t) & \text{if } x_{BA}(t) > 0, \\ y_A(t) - 10^{\frac{2-y_{AB}(t)}{2}} & \text{if } x_A(t) > 0,\ x_{BA}(t) = 0,\ y_{AB}(t) - 2 \le 0; \text{ or} \\ & \text{if } x_A(t) = x_{BA}(t) = 0,\ y_{AB}(t) - 2 \le 0 \\ & \text{and } y_A(t) - 10^{\frac{2-y_{AB}(t)}{2}} > 0 \\ 0 & \text{if } x_A(t) = x_{BA}(t) = 0,\ y_{AB}(t) - 2 \le 0 \\ & \text{and } y_A(t) - 10^{\frac{2-y_{AB}(t)}{2}} \le 0; \end{cases}$$

the equations for x_B and x_{BA} being symmetric.

In fact, for most points in the phase space the above differential equations are sufficient to describe our dynamical system, since the set of values t where one of the derivatives does not exist is nowhere dense. However, this is not true for all points, and so we need to use more complicated equations (15–17).

The above relations define what we will call Non-Homogeneous Dynamical System, NHDS. We will denote this dynamical system by $\Delta(\bar{Y})$, since it is driven by the inflow $\bar{Y}$. This is just a usual non-autonomous dynamical system. All the non-linear dynamical systems, which will appear below, correspond to different choices of the flows $\bar{Y}$.

Let now $\bar{x}(t)$ be the trajectory of the NHDS, corresponding to the initial state $\bar{x}(0)$ and the given inflows $\bar{Y}(t)$. We define the closed fluid network evolution Δ of the point $\bar{x}(0)$ as the evolution $\bar{x}(t)$ under any dynamics $\Delta(\bar{Y}(t))$, for which the following relations between the inflows $\bar{Y}(t)$ and the outflows

$$\bar{Z}(t) \equiv \bar{Z}^{\bar{x},\bar{Y}}(t) = \bar{x}(0) + \bar{Y}(t) - \bar{x}(t) \tag{18}$$

hold:

$$Y_O(t) = Z_{AB}^{\bar{x},\bar{Y}}(t) + Z_{BA}^{\bar{x},\bar{Y}}(t), \tag{19}$$

$$Y_A(t) = \frac{1}{2} Z_O^{\bar{x},\bar{Y}}(t), \tag{20}$$

$$Y_{AB}(t) = Z_A^{\bar{x},\bar{Y}}(t), \tag{21}$$

and symmetric relations for A and BA variables. The set of all solutions $\bar{Y}(t)$ of (19)–(21) will be denoted by $\mathcal{Y}(\bar{x}(0))$.

It is well known that the set $\mathcal{Y}(\bar{x}(0))$ is not empty for any initial state $\bar{x}(0)$. We reproduce here the proof, since analogous argument will be used throughout the paper.

Proposition 6 *For any point $\bar{x}(0)$ there exists at least one trajectory of closed fluid network evolution Δ, passing through it.*

Proof It suffices to prove that a solution exists in any given bounded time interval $[0, T]$. For a given $\bar{x} = \bar{x}(0)$, consider the map $G : \bar{Y}(\cdot) \to \bar{Z}^{\bar{x},\bar{Y}}(\cdot)$. It is clearly a continuous map of $C[0, T]$ into itself. The outflow $\bar{Z}(\cdot)$ is Lipschitz continuous (by Lemma 13 below), with Lipschitz constant L independent of $\bar{x}(0)$ or $\bar{Y}(\cdot)$. Hence G takes the convex compact set of L-Lipschitz continuous functions $\bar{Y}(\cdot)$ on $[0, T]$, satisfying $\bar{Y}(0) = 0$, into itself. Therefore, by the Brouwer theorem, the map G has at least one fixed point. □

The system Δ has the following properties:

- The total amount of fluid $|\bar{x}(t)| = x_O(t) + x_A(t) + \cdots + x_B(t)$ is evidently conserved. We will assume $|\bar{x}(t)| = 1$.

- For some initial states $\bar{x}(0)$ the equations (19)–(21) have multiple solutions. This is true for all initial conditions $\bar{x}(0)$, given by (10). The equations (19)–(21) have in this case three solutions, so there are at least three dynamical systems, defined by $\bar{x}(0)$. The first solution corresponds to the flow rates $y_{AB} = 10$, $y_{BA} = 0$, the other one is symmetric to the first one: $y_{BA} = 10$, $y_{AB} = 0$, while the third one is given by the flow rates $AB \to O, BA \to O, A \to AB, B \to BA$ all equal to $\frac{10}{6}$, $O \to A$, $O \to B$ equal to $\frac{3}{2}$. One can say, having that property in mind, that at some points the uniqueness of the trajectory breaks down, so it can happen that for two trajectories $\bar{x}'(t)$, $\bar{x}''(t)$ we have $\bar{x}'(t) = \bar{x}''(t)$ for $t \le t_0$, but $\bar{x}'(t) \ne \bar{x}''(t)$ for $t > t_0$. That just means that there are two different (non-autonomous) dynamical systems, which have trajectories, coinciding for $t \le t_0$, but not for $t > t_0$.
- The curve $\bar{x}(t) \equiv \{1, 0, 0, 0, 0\}$ is a trajectory, i.e. the point $* = \{1, 0, 0, 0, 0\}$ is a fixed point of Δ. Its flow rates are constant; their values are: $y_A = y_B = y_{AB} = y_{BA} = \frac{3}{2}$, $y_O = 3$. Let us check that with these flow rates the amount of the fluids in $\bar{A}$ and $\bar{B}$ will stay zero. Indeed, $\frac{y_A}{\gamma_A} + \frac{y_{BA}}{\gamma_{BA}} = \frac{3/2}{10} + \frac{3/2}{2} = \frac{9}{10} < 1$. So our claim follows from the Proposition 5. In fact, the nodes $\bar{A}$ and $\bar{B}$ are even underloaded, which means that in the long run each node gets on the average less fluid than its serving capacity is. Note also that there are trajectories $\bar{x}(t)$ such that $\bar{x}(t) = *$ for $t \le t_0$, but $\bar{x}(t) \ne *$ for $t > t_0$, with any t_0.
- There are (not necessarily uniqueness) points in $(\mathbb{R}^5)^+$, from where (some) trajectory goes to $*$. One such family is the set of points $U = \{\bar{x} : 0 < x_A = x_B < \frac{1}{2}, x_{AB} = x_{BA} = 0, x_O = 1 - x_A - x_B\}$, and the flows are: $AB \to O, BA \to O, A \to AB, B \to BA$ all equal to $\frac{10}{6}$, $O \to A$, $O \to B$ equal to $\frac{3}{2}$. In fact, from every point in U two more trajectories start. The values of the flow rates for one of them for small initial segment of time is given by: $y_A = y_B = \frac{3}{2}$, $y_{AB} = 10$, $y_{BA} = 0$, $y_O = 2$. It is describing the situation when the light fluid A flows to the node $\bar{B}$, and the resulting heavy fluid AB in the node $\bar{B}$ blocks the fluid B in the node $\bar{B}$ from exiting. The second solution is obtained by interchanging A and B. There is a bigger set $\bar{U} \supset U$, having dimension 2, starting from where one can get to $*$, making on the way some choice of the trajectory; $\bar{U} = \{\bar{x} : x_A = x_B, x_{AB} = x_{BA}, x_O = 1 - x_A - x_B - x_{AB} - x_{BA}\}$.
- There is a cycle $\mathcal{C} \subset (\mathbb{R}^5)^+$, such that if $\bar{x}(0) \in \mathcal{C}$, then $\bar{x}(t) \in \mathcal{C}$ for all $t > 0$. For example, the point $\{\bar{x} : x_A = 1, x_B = x_{AB} = x_{BA} = x_O = 0\}$ belongs to it. All the points of the cycle $\mathcal{C}$ are uniqueness points. By $T_{\mathcal{C}}$ we will denote the time to go around the cycle $\mathcal{C}$ once. We will now describe, just applying the definitions in a straightforward way, the cycle $\mathcal{C}$, started from $\bar{x}(0) = \{x_A = 1, x_B = x_{AB} = x_{BA} = x_O = 0\}$. The first part of it happens for $t \in [0, \frac{1}{9}]$, during which time the component $x_A(t)$ decays linearly from the value 1 to 0. The component $x_B(t)$ grows linearly, and $x_B(\frac{1}{9}) = \frac{1}{9}$, the component $x_{AB}(t)$ grows linearly, and $x_{AB}(\frac{1}{9}) = \frac{8}{9}$, while x_{BA} and x_O stay zero. On the next segment, $t \in [\frac{1}{9}, 1]$, the component $x_{AB}(t)$ decays linearly, with rate 1, and so it vanishes at $t = 1$. The component $x_B(t)$ continues to grow linearly, with the same speed, so $x_B(1) = 1$. Other three components remain empty. So at time 1 we find ourselves accomplishing one half of the cycle. Therefore $T_{\mathcal{C}} = 2$.
- If $\bar{x}(0) \notin \bar{U}$, then $\bar{x}(t) \in \mathcal{C}$ once $t \ge T$, where $T = T(\gamma_O, \gamma_A, \gamma_B, \gamma_{AB}, \gamma_{BA})$. We will prove this claim under additional assumption that $\mathrm{dist}(\bar{x}(0), \mathcal{C}) \equiv \inf_{y \in \mathcal{C}} \sum_i |x_i - y_i| < \varepsilon$ for some small ε, since this will be sufficient for our purposes.

Lemma 7 (Local attractor) *There exists an $\varepsilon > 0$ such that for any initial point $\bar{x}(0)$ with $\mathrm{dist}(\bar{x}(0), \mathcal{C}) < \varepsilon$ we have $\bar{x}(t) \in \mathcal{C}$ once $t > T = T(\gamma_O, \gamma_A, \gamma_B, \gamma_{AB}, \gamma_{BA})$.*

Proof Since for every point $\bar{c}$ on $\mathcal{C}$ we have either $c_{AB} = 0$ or $c_{BA} = 0$, we see that at least one of the coordinates $x_{AB}(0)$ or $x_{BA}(0)$ has to be less than ε. Let us start with the case that both of them are positive, and we can then assume that $x_{AB}(0) < \varepsilon$, $x_{BA}(0) \geq x_{AB}(0)$. Then for a short initial segment of time, $[0, t_1]$, both coordinates $x_{AB}(0)$ and $x_{BA}(0)$ decay with the same constant rate 2, until x_{AB} will vanish, which will happen at the moment $t_1 = \frac{x_{AB}(0)}{2} < \frac{\varepsilon}{2}$. Also, for every point $\bar{c}$ on $\mathcal{C}$ we have $c_O = 0$. Therefore $x_O(0) < \varepsilon$, hence $x_O(t_1) < \varepsilon + \frac{\varepsilon}{2} = \frac{3\varepsilon}{2}$, since the total arrival rate to $\bar{O}$ does not exceed 4, while its service rate equals 3.

Now consider the case when $x_{BA}(t_1) > 0$. While $x_{BA}(t)$ keeps being positive, there is no flow from $\bar{A}$ to $\bar{B}$ and, as a result, no flow from $\bar{B}$ to $\bar{O}$. So if $x_{BA}(t)$ stays positive for $t \in [t_1, t_2 = t_1 + x_O(t_1)]$, then $x_O(t)$ becomes 0 at $t = t_2$—since it decays with rate 1 when $x_{AB} = 0$ and $x_{BA} > 0$—and will stay 0 while $x_{BA}(t)$ is positive. The coordinate $x_{AB}(t)$ stays zero as well. When finally $x_{BA}(t)$ vanishes for the first time, at $t_3 \geq t_2$, the value $x_B(t_3)$ has to be already zero. Which means that $x_A(t_3) = 1$, so $\bar{x}(t_3) \in \mathcal{C}$.

In the remaining case, the first point, t_3, where the coordinate $x_{BA}(t)$ vanishes, belongs to the segment $[t_1, t_2 = t_1 + x_O(t_1)]$. (That case contains the situation when $x_{BA}(t_1) = 0$.) Since the time $t_3 \leq 2\varepsilon$, and since initially the point $\bar{x}(0)$ was close to $\mathcal{C}$, the same is true for the point $\bar{x}(t_3)$. Since $x_{AB}(t_3) = x_{BA}(t_3) = 0$, we conclude that one of the two coordinates—either $x_A(t_3)$, or $x_B(t_3)$—should be ε-close to 1, while the remaining one, as well as $x_O(t_3)$, should be ε-close to 0. Suppose $x_A(t_3) \sim 1$. Note that the evolution of the point $\bar{x}(t_3)$ is not uniquely defined if both $x_A(t_3)$ and $x_B(t_3)$ are positive. As was explained above, there are, in fact, three options to choose from:

(i) If the choice is that the fluid A "goes first", then after a small time the coordinate x_O will vanish, and after the time of order $\frac{1}{9}$ the coordinate x_A will vanish as well, and we find ourselves on $\mathcal{C}$.

(ii) If the fluid B "goes first", then after a small time (of order ε) first x_B, and then x_{BA} will vanish, and the fluid x_A—which is in the amount of order 1, will start to decay, so we find ourselves in the situation just considered.

(iii) The remaining option is when both fluids A and B "go simultaneously" into, resp., AB and BA, at the rate $\frac{10}{6}$. As was explained above, the result will be that the levels of the fluids A and B will decay with the rate $\frac{10}{6} - \frac{3}{2} = \frac{1}{6}$, while the level x_O will be correspondingly raising. But no later than the time 6ε the level x_B will vanish, and since x_{AB} and x_{BA} both were already zero, we again are in the situation considered above, with x_O being of order ε. In fact, at any time when the system is in phase (iii), it has three options: to pass (forever!) to the phase (i) or (ii), or to stay in phase (iii). The above arguments in (i)–(iii) stays valid in this case and we get the required assertion. □

Warning. Our analysis shows that the time T needed to reach the cycle does not vanish with ε.

- *Continuity.* The above proof implies the following (weaker) substitute for the property of the continuous dependence of the trajectory on the initial condition. Let $\bar{x}(0)$ and $\bar{c}(0)$ be two initial points, and suppose that $|\bar{x}(0) - \bar{c}(0)| < \varepsilon$, and $\bar{c}(0) \in \mathcal{C}$. Then there exists a constant C, such that for any t and for any version of the x-trajectory we have $|\bar{x}(t) - \bar{c}(t)| < C\varepsilon$. The condition $\bar{c}(0) \in \mathcal{C}$ is, evidently, crucial; without it our statement fails.
- The cycle $\mathcal{C}$ depends on $\gamma_O, \gamma_A, \gamma_B, \gamma_{AB}, \gamma_{BA}$, and is non-trivial for our choice of these parameters. For some other values of $\gamma_O, \gamma_A, \gamma_B, \gamma_{AB}, \gamma_{BA}$ it is reduced to the point $*$, which then is a stable fixed point.

- All the above properties of our system would still be valid if we perturb slightly the vector $\bar{\gamma} = \{\gamma_O, \gamma_A, \gamma_B, \gamma_{AB}, \gamma_{BA}\}$ of the parameters around the point $\{3, 10, 10, 2, 2t\}$ of our choice (even if the perturbation does not respect the symmetry $\gamma_A = \gamma_B$, $\gamma_{AB} = \gamma_{BA}$).
- Let $\bar{x}(t) \subset \mathcal{C}$ be a cyclic trajectory. Let us denote by $\bar{\lambda}^{\mathcal{C}}(t)$ the corresponding (periodic) function of the inflows. The stability property just formulated implies immediately that our system in the cyclic regime is underloaded. In other words, for $\{\gamma_O, \gamma_A, \gamma_B, \gamma_{AB}, \gamma_{BA}\} = \{3, 10, 10, 2, 2\}$ there exists $\delta > 0$ such that we have

$$\frac{1}{\gamma_O} \int_0^{T_{\mathcal{C}}} \lambda_O^{\mathcal{C}}(t)\, dt < (1-\delta)\, T_{\mathcal{C}}, \tag{22}$$

$$\frac{1}{\gamma_A} \int_0^{T_{\mathcal{C}}} \lambda_A^{\mathcal{C}}(t)\, dt + \frac{1}{\gamma_{BA}} \int_0^{T_{\mathcal{C}}} \lambda_{BA}^{\mathcal{C}}(t) dt < (1-\delta)\, T_{\mathcal{C}}, \tag{23}$$

and the same relation for the $\bar{B}$ node.

We now want to consider the "open" system Δ^o, which is obtained from Δ by the following construction: every exiting fluid now goes not to the corresponding node, but leaves the system. On the other hand, there is some inflow, $\bar{\lambda}(t)$, entering the system from the outside.

For the future use we introduce now the following compact subset $\mathcal{K} \equiv \mathcal{K}(\bar{\gamma}) \subset (\mathbb{R}^5)^+$:

$$\mathcal{K} = \left\{ \bar{x} \in \left(\mathbb{R}^5\right)^+ : \max\left\{ \frac{1}{2}x_{AB} + \frac{1}{10}x_B, \frac{1}{2}x_{BA} + \frac{1}{10}x_A, \frac{1}{3}x_O \right\} < 10 \right\}.$$

Lemma 8 *Let the open system Δ^o be in the state $\bar{x}(0) \in (\mathbb{R}^5)^+$. Consider the functional L on $\mathbb{R}^5$, given by*

$$L(\bar{x}) = \begin{cases} 0 & \text{if } \bar{x} \in \mathcal{K}, \\ \max\{\frac{1}{2}x_{AB} + \frac{1}{10}x_B, \frac{1}{2}x_{BA} + \frac{1}{10}x_A, \frac{1}{3}x_O\} & \text{otherwise.} \end{cases} \tag{24}$$

Suppose that $L(\bar{x}(0)) > 10$. Then there exists a constant $C > 0$, such that for all external flow rates $\{\bar{\lambda}(t), t \in [0, T_{\mathcal{C}}]\}$ which are close enough to $\{\bar{\lambda}^{\mathcal{C}}(t), t \in [0, T_{\mathcal{C}}]\}$ in the L^1 distance, we have

$$L(\bar{x}(0)) - L(\bar{x}(T_{\mathcal{C}})) > C.$$

Proof For the case of the inflows with rates $\bar{\lambda}^{\mathcal{C}}(t)$ our statement follows from the underload property, due to the relations (22)–(23). Therefore it holds for inflows that are close enough, by continuity. □

4.3 M Coupled Fluid Networks

Let Δ_M be the dynamical system on $(\mathbb{R}^{5M})^+$, obtained from M copies of Δ, interconnected in the mean-field manner, as follows. Each node $\bar{O}_i$, $i = 1, \ldots, M$, is connected to *all* of the nodes $\bar{A}_j$, $\bar{B}_j$, $j = 1, \ldots, M$, and its fluid, of the amount $x_{O,i}$, flows into the nodes $\bar{A}_j$, $\bar{B}_j$ in equal amounts. The rate of each of these flows is now $M^{-1}\frac{\gamma_O}{2} = \frac{3}{2M}$, which means, as before, that three units of the fluid $x_{O,i}$ leave $\bar{O}_i$ per unit time, so each of the set $\{\bar{A}_j\}$ and $\{\bar{B}_j\}$ gets $\frac{3}{2}$ units of incoming fluids, A and B, per unit time. In a similar way, the fluid A from every node $\bar{A}_i$ is splitted among all nodes $\{\bar{B}_j\}$, so the rate of every individual flow $\bar{A}_i \to \bar{B}_j$ is $M^{-1}\gamma_A = \frac{10}{M}$, and so on. The priorities are kept the same: if the node $\bar{A}_i$, say, is in the state with both amounts $x_{A,i}$ and $x_{BA,i}$ positive, then *the fluid BA goes first*.

Again, we first describe the open network, i.e. the Non-Homogeneous Dynamical System. We will not need the general case here; it is enough for us to consider the net inflow defined by the same function $\bar{Y}(t) \in (\mathbb{R}^5)^+$ as in the previous subsection; every node then gets $\frac{1}{M}$-th part of the inflow, so, for example, for each $i = 1, \ldots, M$ the net inflow function of the node $\bar{O}_i$ equals to $\frac{1}{M} Y_O(t)$. Once we are also given the initial values $\bar{x} \in (\mathbb{R}^{5M})^+$ of the fluid levels, the net outflows $Z_{a,i}^{\bar{x},\bar{Y}}(t)$, $a \in \{O, A, B, AB, BA\}$, $i = 1, \ldots, M$, are defined as above, see (18).

Passing to the closed system, instead of relations (19)–(21), we impose the relations

$$Y_O(t) = \sum_{i=1}^{M} (Z_{AB,i}^{\bar{x},\bar{Y}}(t) + Z_{BA,i}^{\bar{x},\bar{Y}}(t)), \tag{25}$$

$$Y_{A,j}(t) = \sum_{i=1}^{M} \frac{1}{2} Z_{O,i}^{\bar{x},\bar{Y}}(t), \tag{26}$$

$$Y_{AB,j}(t) = \sum_{i=1}^{M} Z_{A,i}^{\bar{x},\bar{Y}}(t). \tag{27}$$

Again, all the functions $Y_*(t)$ and $Z_*(t)$ are Lipschitz continuous and, hence, differentiable almost everywhere. These derivatives will be denoted, again, by $y_*(t)$ and $z_*(t)$. For each node (say, $\bar{A}_i$) of Δ_M, the evolution of the amount of fluids $x_{A,i}(t)$ and $x_{BA,i}(t)$ is found from the corresponding inflows $\frac{1}{M} Y_A(t)$ and $\frac{1}{M} Y_{BA}(t)$ and initial states $x_{A,i}(0)$ and $x_{BA,i}(0)$ in the same manner as for the network Δ.

We will consider trajectories $\bar{x}^M(t) \in (\mathbb{R}^{5M})^+$ with $|\bar{x}^M(t)| = M$, i.e. we have the unit amount of fluid per elementary system $\Delta \subset \Delta_M$.

For every M we will define now another dynamical system, acting on $(\mathbb{R}^5)^+$. This one, also denoted by Δ_M, will be of central importance for the present paper. However, it will be not the usual dynamical system. It will be defined not as a group of transformations of $(\mathbb{R}^5)^+$, but directly on the (sub)set $\mathcal{M}_M$ of some atomic probability measures on $(\mathbb{R}^5)^+$. This transformation will not be linear on $\mathcal{M}$, and for that reason we will call it *non-linear dynamical system.* In fact, it is just a convenient representation of our initial dynamical system on $(\mathbb{R}^{5M})^+$, suitable for passing to the limit $M \to \infty$. The construction is very simple:

To every point $\bar{x}^M = \{(x_O^i, x_{\bar{A}}^i = (x_A^i, x_{BA}^i), x_{\bar{B}}^i = (x_B^i, x_{AB}^i)), i = 1, \ldots, M\} \in (\mathbb{R}^{5M})^+$ we can assign a probability measure on $(\mathbb{R}^5)^+$ in the following way: we put $\mu_O = \frac{1}{M} \sum_{i=1}^{M} \delta_{x_O^i}$, $\mu_{\bar{A}} = \frac{1}{M} \sum_{i=1}^{M} \delta_{x_{\bar{A}}^i}$, $\mu_{\bar{B}} = \frac{1}{M} \sum_{i=1}^{M} \delta_{x_{\bar{B}}^i}$, and we define $\mu \equiv \mu_{\bar{x}^M} = \mu_O \times \mu_{\bar{A}} \times \mu_{\bar{B}} \in \mathcal{M}_M$. Now, if the point $\bar{x}^M$ evolves according to Δ_M, so is the measure μ; moreover, the evolution $\mu(t)$ of μ is well defined and does not depend on the choice of the preimage, so if $\mu_{\bar{x}_1^M} = \mu_{\bar{x}_2^M}$, then $\mu_{\bar{x}_1^M}(t) = \mu_{\bar{x}_2^M}(t)$.

The set of measures $\mathcal{M}_M$ on $(\mathbb{R}^5)^+$consists of all measures μ, having the properties

1. μ is a product,

$$\mu \equiv \left(\mu_O, \mu_{\bar{A}}, \mu_{\bar{B}}\right) \equiv \mu_O \times \mu_{\bar{A}} \times \mu_{\bar{B}} \equiv \Pi_{\bar{o}}[\mu] \times \Pi_{\bar{A}}[\mu] \times \Pi_{\bar{B}}[\mu], \tag{28}$$

   of probability measures on $\mathbb{R}^1 = \{x_O\}$, resp. $\mathbb{R}^2 = \{x_A, x_{BA}\}$ and $\mathbb{R}^2 = \{x_B, x_{AB}\}$. Here we denote by Π_*-s the various projections (or marginals),
2. we have $\int |\bar{x}| d\mu = \int x_O d\mu_O + \int (x_A + x_{BA}) d\mu_{\bar{A}} + \int (x_B + x_{AB}) d\mu_{\bar{B}} = 1$,

3. we have $\mu_O = \frac{1}{M}\sum_{i=1}^{M}\delta_{\bar{x}_i}$ for some (not necessarily distinct) $\bar{x}_i \in \mathbb{R}^1$, $i = 1, \ldots, M$, likewise $\mu_{\bar{A}} = \frac{1}{M}\sum_{i=1}^{M}\delta_{\bar{x}'_i}$, $\mu_{\bar{B}} = \frac{1}{M}\sum_{i=1}^{M}\delta_{\bar{x}''_i}$, $\bar{x}'_i, \bar{x}''_i \in \mathbb{R}^2$.

Properties of Δ_M:

- The set of the fixed points of Δ_M consists of measures $\mu = \Pi_{\bar{o}}[\mu] \times \delta_{x_{\bar{A}}=0} \times \delta_{x_{\bar{B}}=0}$, where $\delta_{x_{\bar{A}}=0}$ and $\delta_{x_{\bar{B}}=0}$ are unit atoms at the origin, while $\Pi_{\bar{o}}[\mu]$ is the projection on the coordinate x_O. In words, that means that all the fluid stays permanently in the O-nodes (in arbitrary amounts, adding up to M). This fact follows from Proposition 5.
- If $\mu(0) = \delta_{\bar{x}} \in \mathcal{M}_M$, then $\mu(t) = \delta_{\bar{x}(t)}$, where $\bar{x}(t)$ is the trajectory of Δ with $\bar{x}(0) = \bar{x}$. In particular if $\bar{x} \in \mathcal{C}$, then $\bar{x}(t) \in \mathcal{C}$.
- Note that the dynamics Δ_M on $\mathcal{M}_M$ is "non-linear", in the sense that in general in the situation when $\mu(0) = \alpha\mu'(0) + (1-\alpha)\mu''(0)$, $\mu(0), \mu'(0), \mu''(0) \in \mathcal{M}_M$, $0 < \alpha < 1$, we have $\mu(t) \neq \alpha\mu'(t) + (1-\alpha)\mu''(t)$ for $t > 0$. (There is nothing strange or unusual in this relation, since the dynamical system in question is itself defined on the space of measures as its state space, and not on the $(\mathbb{R}^5)^+$.)
- Let ρ_{KROV} be the Kantorovich-Rubinstein-Ornstein-Vaserstein distance on the probability measures on $(\mathbb{R}^5)^+$, corresponding to the metric $\rho(\bar{x}, \bar{y}) = \sum_{i=1}^{5}|x_i - y_i|$ on $(\mathbb{R}^5)^+$. (We recall briefly, that if μ, μ' are two probability measures on a metric space (X, ρ), then $\rho_{KROV}(\mu, \mu') = \inf_\kappa \int \rho(x, x')d\kappa(x, x')$, where the inf is taken over all probability measures κ on $X \times X$, such that $\kappa(A \times X) = \mu(A)$, $\kappa(X \times A) = \mu'(A)$. Suppose that the initial measure $\mu(0)$ is close enough to the cycle $\mathcal{C}$, which means that for some $x \in \mathcal{C}$ we have

$$\rho_{KROV}\left(\mu\left(0\right), \delta_x\right) < \varepsilon. \tag{29}$$

Let $Y(t) \in \mathcal{Y}(\mu(0))$ be one of the possible net inflows, corresponding to the initial state $\mu(0)$, and $\mu(t)$ be the corresponding evolution. Then there exists the time $T = T(M, \gamma_O, \gamma_A, \gamma_B, \gamma_{AB}, \gamma_{BA})$, such that for all $t \geq T$ we have $\mu(t) = \delta_{\bar{x}(t)}$ with $\bar{x}(t) \in \mathcal{C}$. Moreover, let us define the compact Λ_K by

$$\Lambda_K = \{\bar{x} = (x_O, x_A, \ldots, x_B) : x_O < K, x_A < K, \ldots, x_B < K\},$$

and let $\mu|_K(t)$ be the evolution of the restriction $\mu(0)|_K$ under the same evolution, defined by the flow rates $Y(t) \in \mathcal{Y}(\mu(0))$. (Once the inflows $Y(t)$ are fixed, the evolution becomes the usual (non-autonomous) *linear* dynamical system, so we can apply the dynamics to the summand $\mu(0)|_K$ of the measure $\mu(0)$.) We choose K to be large enough, so that from $\rho_{KROV}(\mu(0), \delta_x) < \varepsilon$ for some $x \in \mathcal{C}$ it follows that

$$\mu\left(0\right)[K] > 1 - \varepsilon. \tag{30}$$

We now claim the following:

Lemma 9 *Under conditions* (29) *and* (30), *there exists the time moment* $T' = T'(K, \gamma_O, \gamma_A, \gamma_B, \gamma_{AB}, \gamma_{BA})$, *such that for every* $t \geq T'$ (*uniformly in M!*) *there exists a point* $\bar{x}(t)$, *such that* $\mu|_K(t)$ *is just the atom at that point*: $\mu|_K(t) = c\delta_{\bar{x}(t)}$. *Moreover*, $\mathrm{dist}(\bar{x}(t), \mathcal{C}) < \tilde{c}$, *and* $c = \mu(0)[K] \to 1$ *while* $\tilde{c} \to 0$ *as* $K \to \infty$.

Notes

- We will prove only the statement about the existence of the time moment T', since below we will not use the time moment $T(M, \gamma_O, \gamma_A, \gamma_B, \gamma_{AB}, \gamma_{BA})$.
- Our proof can be extended *literally* to the case $M = \infty$ of the next subsection.

Proof Note first that due to the mean-field nature of our graph, the flows to all the $\bar{A}$-nodes are equal at every time moment (as well as to all the $\bar{B}$-nodes or $\bar{O}$-nodes). Consider now any subset Q of the $\bar{A}$-nodes, $|Q| \le M$, and let I_{BA} be the index, for which $(x_{BA})_{I_{BA}}(0) \ge (x_{BA})_i(0)$ for all $i \in Q$. Then, clearly, this relation holds at later moments, i.e. $(x_{BA})_{I_{BA}}(t) \ge (x_{BA})_i(t)$. In the same way, define the index I_A as the one, for which $(x_A)_{I_A}(0) \ge (x_A)_i(0)$. Then, if all the variables $(x_{BA})_i(0)$ are equal for $i \in Q$, the relation $(x_A)_{I_A}(t) \ge (x_A)_i(t)$ holds at all later moments. We will use this property for the set $Q = \{i\}$ of indices, which satisfy $(x_{BA})_i(0) < K$, $(x_A)_i(0) < K$.

Let us show that there exists the time moment, such that before it every node in Q will be empty for some time duration. In view of what was said before, it means that after that time moment all the nodes in Q will be synchronized. To see this depletion, note that the initial supply of the fluid BA at any node $i \in Q$ is not exceeding K, so it will be over before $t' = \frac{K}{2}$. Next, there exists a moment t'', after which $\frac{99}{100}$ (say) of "atoms" of the heavy fluid BA, passing through our node $i \in Q$ were at earlier moments at some $\bar{O}$-node. Indeed, another option would be that such an atom was staying at some $\bar{B}$ node for all the time duration t''. That means that the initial total amount of fluid at this node was very high, once t'' is chosen to be large. However, the proportion of such nodes has to be small, due to the simple fact that the total amount of fluid per node is of the order of one. But the rate, at which the fluid goes from the $\bar{O}$ node to the $\bar{B}$ node is never higher than $\frac{3}{2}$. Let K'' be the amount of fluid at our node i at the moment t''. Clearly it is at most $12t''$. Suppose the node is not empty during the time interval $[t'', t'' + T]$. Than it works all the time at full capacity. Let $0 \le k(t) \le 1$ be the fraction at moment t of the capacity of the node, used by the heavy fluid, while the remaining fraction $1 - k(t)$ is used by the light fluid. Then the amount of heavy fluid, which left the server during this time interval, is $2\int_{t''}^{t''+T} k(t)dt$, while the corresponding amount of the light fluid is $10\int_{t''}^{t''+T}(1 - k(t))dt$. Since the light fluid flows into the node with the rate at most $\frac{3}{2}$, we have that the relation

$$10\int_{t''}^{t''+T}(1 - k(t))\,dt \le \frac{3}{2}T + K''$$

has to hold, since the amount of light fluid, leaving the node, can not exceed the initial amount present at the node plus the amount which came to the node during the time interval T. For the heavy fluid we similarly have

$$2\int_{t''}^{t''+T} k(t)dt \le \frac{3}{2}T + 10\left(\frac{1}{100}T\right) + K''.$$

The two relations imply that

$$10T \le \left(9\frac{1}{2}\right)T + 6K''.$$

So $T \le 12K''$, which establishes our depletion claim for the $\bar{A}$ (as well as for $\bar{B}$) nodes at some moment t''', independent of M and ε, provided only that ε is small.

Thus far we were not using the condition $\rho_{KROV}(\mu(0), \delta_x) < \varepsilon$, without which our claim about the existence of the time moment $T'(K, \gamma_O, \gamma_A, \gamma_B, \gamma_{AB}, \gamma_{BA})$ is not valid—see, for example, the first property of the dynamics Δ_M. We will use it now, in dealing with the $\bar{O}$ nodes. Due to the above discussion and the continuity property, our statement is reduced to the following one: consider the initial measure $\mu(0)$, having the properties: $\rho_{KROV}(\mu(0), \delta_{\bar{x}}) < \varepsilon'$ $(= C\varepsilon$, see Continuity Property of the previous section), while in

the decomposition $\mu(0) = \mu(0)_{\bar{O}} \times \mu(0)_{\bar{A}} \times \mu(0)_{\bar{B}}$ we have $\mu(0)_{\bar{A}} = (1-\varepsilon)\delta_{\tilde{x}_{\bar{A}}} + \varkappa_{\bar{A}}$, $\mu(0)_{\bar{B}} = (1-\varepsilon)\delta_{\tilde{x}_{\bar{B}}} + \varkappa_{\bar{B}}$, with the vectors $\tilde{x}_{\bar{A}}, \tilde{x}_{\bar{B}} \in \mathbb{R}^2$ close to the corresponding projections (x_A, x_{BA}), resp. (x_B, x_{AB}) of the vector $\bar{x}$. But that means that the flows into the $\bar{O}$ nodes will be almost always almost equal to these on the cycle $\mathcal{C}$, so in finite time all of the $\bar{O}$-nodes which initially have their levels $\leq K$ will become empty. Indeed, on the cycle the flow to the $\bar{O}$ node has rate 2, while the capacity of these nodes equals 3.

Summarizing, we have thus far that after a finite time, independent of M, all the nodes in Q are synchronized, and moreover all the $\bar{O}$-nodes in Q are empty. The application of the Continuity Property and the Attraction Lemma 7 finishes the proof. □

4.4 $M \to \infty$ Fluid Network

This is again a dynamical system, Δ_∞, acting on probability measures $\mathcal{M} = \mathcal{M}((\mathbb{R}^5)^+)$ on $(\mathbb{R}^5)^+$. One way of defining it is to say that the family $\mu(t) \in \mathcal{M}$ is a trajectory of Δ_∞, iff for any M there is a trajectory $\mu_M(t) \in \mathcal{M}_M$ of Δ_M, so that for every t we have $\mu_M(t) \to \mu(t)$ weakly.

Now we will give another description of Δ_∞, which does not make use of the limit $M \to \infty$. As was the case with the dynamics Δ and Δ_M, we will define first the NHDS version of Δ_∞. Let the measure $\mu_0 \in \mathcal{M}((\mathbb{R}^5)^+)$ and let the function $\bar{Y}(t) \in (\mathbb{R}^5)^+$ be given, which is the net inflow of our fluids. Then the corresponding evolution $\mu = \{\mu_t \in \mathcal{M}((\mathbb{R}^5)^+)\}$ of the state μ_0 is defined to be just the evolution of the measure μ_0 under the dynamics $\Delta(\bar{Y})$. Again, the net outflows $\bar{Z}^{\bar{x},\bar{Y}}(t)$, $\bar{x} \in (\mathbb{R}^5)^+$ are given by (18).

Now we can define the evolution $\mu = \{\mu_t \in \mathcal{M}((\mathbb{R}^5)^+)\}$ of the initial measure μ_0 under Non-Linear Dynamical System Δ_∞ (NLDS) as the NHDS evolution of it under any dynamics $\Delta(\bar{Y})$, with $\bar{Y}$ satisfying the equations:

$$\begin{aligned} Y_O(t) &= \int (Z^{\bar{x},\bar{Y}}_{AB}(t) + Z^{\bar{x},\bar{Y}}_{BA}(t)) d\mu_0(\bar{x}), \\ Y_A(t) &= \frac{1}{2}\int Z^{\bar{x},\bar{Y}}_{O}(t)\, d\mu_0(\bar{x}), \\ Y_{AB}(t) &= \int Z^{\bar{x},\bar{Y}}_{A}(t) d\mu_0(\bar{x}), \end{aligned} \tag{31}$$

and symmetric relations for B and BA variables. The set of all such flows $\bar{Y}(t)$ will be denoted by $\mathcal{Y}(\mu_0)$.

We are calling the dynamical system Δ_∞ non-linear, since in general we will have for $\mu_0 = \frac{1}{2}(\mu_0' + \mu_0'')$ that $\mu_t \neq \frac{1}{2}(\mu_t' + \mu_t'')$ when $t > 0$. Note also that if $\mu_0 = \delta_{\bar{x}(0)}$ for some point $\bar{x}(0)$, then the family μ_t is a Δ_∞ trajectory iff $\mu_t = \delta_{\bar{x}(t)}$, with $\bar{x}(t)$ being some Δ-evolution of $\bar{x}(0)$.

Properties of Δ_∞:

1. the set of the fixed points of Δ_∞ consists of measures $\mu = \Pi_{\bar{o}}[\mu] \times \delta_{x_{\bar{A}}=0} \times \delta_{x_{\bar{B}}=0}$, where $\Pi_{\bar{o}}[\mu]$ is the projection on the coordinate x_O.
2. if $\mu(0) \in \mathcal{M}_M$ for some M, then the Δ_M-dynamics and Δ_∞-dynamics with that initial data coincide. In particular, if $\mu(0) = \delta_{\bar{x}} \in \mathcal{M}$ with $\bar{x} \in \mathcal{C}$, then $\mu(t) = \delta_{\bar{x}(t)}$, where $\bar{x}(t)$ is the trajectory of Δ with $\bar{x}(0) = \bar{x}$.

For all our purposes it is sufficient to prove the following property of our NLDS.

Proposition 10 *Let the measure $\mu = \mu(0)$ on $(\mathbb{R}^5)^+$ have the following properties*:

(i) *Unit mass*:

$$\int_{(\mathbb{R}^5)^+} (x_A + x_B + x_{BA} + x_{AB} + x_O)\, d\mu = 1, \tag{32}$$

(ii) *Exponential moment condition*: *for some* $\alpha > 0$, $A < \infty$ *we have*

$$\langle \exp\{\alpha L(\bar{x})\}\rangle_{\mu(0)} < 3A \tag{33}$$

(*see* (24)),

(iii) *For some* $x \in \mathcal{C}$ *we have*

$$\rho_{KROV}(\mu(0), \delta_x) < \varepsilon, \tag{34}$$

with ε *small enough* (*depending on* α *and* A).

Consider now some Non-Linear Dynamical System Δ_∞ (*NLDS*), *defined by the initial state* $\mu(0)$. *In other words*, $\Delta_\infty = \Delta_\infty(\bar{Y}(\cdot))$, *for some* $\bar{Y}(\cdot) \in \mathcal{Y}(\mu(0))$. *Then for* $t \to \infty$ *the evolving measure* $\mu(t)$ *satisfies*:

(I)

$$\rho_{KROV}\left(\mu(t), \delta_{z(t)}\right) \to 0 \tag{35}$$

for appropriate $z(t) \in \mathcal{C}$,

(II)

$$\langle \exp\{\alpha L(\bar{x})\}\rangle_{\mu(t)} \to 0. \tag{36}$$

Moreover, the convergence in (35), (36) *is uniform over all* $\bar{Y}(\cdot) \in \mathcal{Y}(\mu(0))$ *and all initial measures* $\mu(0)$ *satisfying* (32)–(34).

To establish it we first prove the following simpler fact.

Proposition 11 *For every* T, ε *there exists a value* $\bar{\varepsilon}(T, \varepsilon)$, *such that the following holds*:

(i) *For every* T *we have* $\bar{\varepsilon}(T, \varepsilon) \to 0$ *as* $\varepsilon \to 0$.
(ii) *Let* μ *be any measure satisfying* (i)–(iii) *above. Then for every* $t \in [0, T]$ *we have* $\rho_{KROV}(\mu(t), \delta_{x(t)}) < \bar{\varepsilon}(T, \varepsilon)$.

Proof Let the sequence of measures $\mu_n(0)$ converge to δ_x in the *KROV* metrics, and let it satisfy the conditions of the Proposition 10. Let us consider the set of trajectories $\mu_n(t)$, $0 \le t \le T$. The family of these trajectories, viewed as functions of $t \in [0, T]$, is a family of uniformly bounded functions (due to the compactness of the set of measures with the properties (i)–(iii)), which also are equicontinuous. Indeed, every version of the vector field, along which any of the measures $\mu_n(t)$ has to evolve, is continuous and bounded in norm by the constant $\gamma_A = 10$. Therefore it is compact, and so has a limit point, the function $\mu(t)$, with $\mu(0) = \delta_x$. But $\mu(t)$ has to be a trajectory of NLDS, and since there is just one such trajectory starting from δ_x, we conclude that $\mu(t) = \delta_{x(t)}$.

Therefore $\mu_n(t) \to \delta_{x(t)}$ for every $t \in [0, T]$, and this convergence is uniform in t. The existence of the function $\bar{\varepsilon}(T, \varepsilon)$ follows from the compactness of the balls $\{\mu : \rho_{KROV}(\mu, \delta_x) < a\}$ of measures satisfying the moment condition. □

Proof of the Proposition 10 1. To begin with, we prove that once ε is small enough, there exists the time moment T_1, at which for any initial $\mu(0)$, satisfying conditions of our proposition,

$$\langle \exp\{\alpha L(\bar{x})\}\rangle_{\mu(T_1)} < \varepsilon. \tag{37}$$

Indeed, suppose first that the value of the exponential moment of the initial measure $\tilde{\mu}(0)$ is fixed, $\langle \exp\{\alpha L(\bar{x})\}\rangle_{\tilde{\mu}(0)} = M$, and also that

$$\rho_{KROV}\left(\tilde{\mu}(t), \delta_{z(t)}\right) < \tilde{\varepsilon} \quad \text{for all } t \in [0, T_{\mathcal{C}}], \text{ with } \tilde{\varepsilon} \text{ small enough,} \tag{38}$$

where $T_{\mathcal{C}}$ is the time it takes to go once around the cycle $\mathcal{C}$. We claim that at the moment $T_{\mathcal{C}}$ we have $\langle \exp\{\alpha L(\bar{x})\}\rangle_{\tilde{\mu}(T_{\mathcal{C}})} < cM$, where $c < 1$ is some constant, which depends only on the parameters of our model. In particular, c is independent of $\tilde{\mu}(0)$.

To see that we note first that for any point $\bar{x}_0$ with $L(\bar{x}_0) > 10$ (see definition (24)) and under *Cyclic Dynamics* $\Delta_\infty = \Delta_\infty(\bar{Y}(\cdot))$, for $\bar{Y}(\cdot) \in \mathcal{Y}(\delta_z)$, $z \in \mathcal{C}$, we have after the time shift by $T_{\mathcal{C}}$ that $L(\bar{x}_{T_{\mathcal{C}}}) < L(\bar{x}_0) - C(\gamma_*)$, where the constant $C(\gamma_*) > 0$ depends only on the service rates γ_*. This follows directly from Lemma 8 of Sect. 4.2. From the same lemma and due to the condition (38) we have that under any dynamics $\Delta_\infty(\bar{Y}(\cdot))$ with $\bar{Y}(\cdot) \in \mathcal{Y}(\tilde{\mu}(0))$ we have for the same point that $L(\bar{x}_{T_{\mathcal{C}}}) < L(\bar{x}_0) - \frac{1}{2}C(\gamma_*)$. So we have proven our claim, with $c = \exp\{-\frac{1}{2}C(\gamma_*)\}$.

Let now k be the smallest integer, such that $3Ac^k < \varepsilon$. We want to repeat k times the procedure of the previous paragraph. Its duration is, evidently, $kT_{\mathcal{C}}$. Suppose that the parameter ε is so small that the function $\bar{\varepsilon}(kT_{\mathcal{C}}, \varepsilon)$, defined in Proposition 11 satisfies $\bar{\varepsilon}(kT_{\mathcal{C}}, \varepsilon) < \tilde{\varepsilon}$ (see (38)). Then the desired repetition is possible, and so (37) indeed holds, with $T_1 = kT_{\mathcal{C}}$.

2. Summarizing, at the moment T_1 we have:

(i)

$$\langle \exp\{\alpha L(\bar{x})\}\rangle_{\mu(T_1)} < \varepsilon, \tag{39}$$

(ii) $\rho_{KROV}(\mu(T_1), \delta_z) < \bar{\varepsilon}(T_1, \varepsilon)$ for some $z \in \mathcal{C}$.

Let us use now the compact $\mathcal{K}$. From (39) immediately follows that $\mu(T_1)[\mathcal{K}] > 1 - \varepsilon$. Hence, due to Lemma 9 of Sect. 4.3, there is the time moment $T_2 = T_2(\mathcal{K}, \varepsilon)$, after which the restricted measure $\mu(T_1)|_{\mathcal{K}}$ would evolve under the dynamics $\Delta_\infty(\bar{Y}(\cdot))$ with $\bar{Y}(\cdot) \in \mathcal{Y}(\mu(0))$ to a δ-atom, of mass at least $1 - \varepsilon$. According to Lemma 11—or rather its $M = \infty$ version—for some $z \in \mathcal{C}$ the distance $\rho_{KROV}(\mu(T_1 + T_2), \delta_z t)$ at that moment would be at most $\bar{\varepsilon}(T_2, \bar{\varepsilon}(T_1, \varepsilon))$, which is small. If it would have been the case that this atom has unit mass, than after time $T_{\mathcal{C}}$ it would already be on the cycle $\mathcal{C}$. Since, however, we know only that its mass is above $1 - \varepsilon$, we can claim that at the time moment $T_1 + T_2 + T_{\mathcal{C}}$ its distance from $\mathcal{C}$ is at most $K_1\varepsilon$, where K_1 is some universal constant. Therefore for the total measure $\mu(T_1 + T_2 + T_{\mathcal{C}})$ we can claim that

(ii′) $\rho_{KROV}(\mu(T_1 + T_2 + T_{\mathcal{C}}), \delta_z) < K_2\varepsilon$ for some $z \in \mathcal{C}$ and some universal K_2, while

(i′) $\langle \exp\{\alpha L(\bar{x})\}\rangle_{\mu(T_1+T_2+T_{\mathcal{C}})} < c\varepsilon$, with $c < 1$ the same as in the first step of the proof.

So what happened is that the estimator ε of the exponential moment, appearing in (39), multiplied by K_2, is the estimator of the *KROV* distance on the next step, while the estimate of the exponential moment is improved by a constant c. This argument can be iterated; on the next step we will have

(i″) $\langle \exp\{\alpha L(\bar{x})\}\rangle_{\mu(T_1+2(T_2+T_{\mathcal{C}}))} < c^2\varepsilon$,

(ii″) $\rho_{KROV}(\mu(T_1 + 2(T_2 + T_{\mathcal{C}})), \delta_z) < K_2 c\varepsilon$ for some $z \in \mathcal{C}$, and so on, which completes the proof. □

5 Euler Scaling Limit of the NLMP

In this section we will show that in the limit of high load our Non-Linear Markov process tends to the fluid model, introduced above.

Theorem 12 *Suppose that for every N we are given the initial state ν^N of the (scaled—see (1)) NLMP ∇_∞^N, and the sequence ν^N converges to the measure μ in the KROV metric, i.e. $\rho_{KROV}(\nu^N, \mu) \to 0$. Consider all the KROV-limit points $\mu(t)$ of the set of trajectories $\{\tilde{\nu}^N(t) = \nu^N(Nt), t \in [0, T]\}$, $N = 1, 2, \ldots$. Every such limit point, called Euler fluid limit of the NLMP, is necessarily a trajectory of the fluid model Δ_∞.*

(*Any trajectory of Δ_∞, obtained via this limit, will be called a fluid solution.*)

Note As was mentioned above, the system Δ_∞ does not possess the uniqueness property. The uniqueness property for the subclass of the trajectories of Δ_∞—the fluid solution trajectories—does not hold as well. That means that the above set of all limit trajectories $\mu(t)$ might contain more than one element, and so the trajectory $\mu_{\nu^N}(t)$, which satisfies $\mu_{\nu^N}(0) = \mu$, does depend on the sequence $\nu^N \to \mu$.

Proof 1. We start by considering the open system. Then we can consider every node separately. Let us take the node $\bar{A}$, say. In the open system case its evolution is defined by prescribing the initial state, $\nu_{\bar{A}}^N$—the distribution of the quantity $q_{\bar{A}}(0) \in \mathbb{R}^2$, which is the initial queue, scaled by the factor $\frac{1}{N}$ —together with the rate function $\lambda_{\bar{A}}^N(t) \equiv \{\lambda_A^N(t), \lambda_{BA}^N(t)\} \in \mathbb{R}^2$, $t \geq 0$, which defines the Poisson flows of incoming clients. For our applications it is enough to consider the case when all our rate functions λ-s are uniformly bounded:

$$\lambda_*^N(*) \leq CN, \tag{40}$$

since this is definitely the case for our closed system. The service times of the clients are exponential, with respective rates $N\gamma_A$, $N\gamma_{BA}$ (this scaling is due to the Euler limit we are going to study). As above, the BA clients have priority. That means that if, while a user of class A is being served, a class BA user arrives, the service of A user is interrupted until the moment when there will be no BA users in queue (the preemptive priority service discipline).

For the future use we will introduce the functions

$$\Lambda_i^N(t) = \frac{1}{N} \int_0^t \lambda_i^N(s) ds, \quad i = A, BA,$$

which together form a 2D vector $\Lambda^N(t)$. We will also denote by $Nq_{\bar{A}}^N(t) \in \mathbb{R}^2$ the pair of queues at the moment t. By $NQ_i^N(t)$, $i = A, BA$ we denote the number of clients which have arrived to the server $\bar{A}$ during the time interval $[0, t]$. We can as well assume that to every client the service time is assigned at the moment of its arrival. The sum of these required service times for clients arrived during the time interval $[0, t]$ will be denoted by $NW_i^N(t)$; this function has a stair-like graph. By $Nw_i^N(t)$ we denote the remaining required times for clients queuing *or being served* at the moment t; the graph of these functions are saw-like.

We also consider the fluid model at this node. So $q_{\bar{A}}(t)$ will denote the evolution of the initial measure $q_{\bar{A}}(0)$ on $\mathbb{R}^2$ under fluid dynamics governed by the inflow with rate $\lambda_{\bar{A}}(t)$. By $Q_i(t) = \Lambda_i(t)$ we denote the amounts of fluid arriving to our node during the time interval

$[0, t]$. Again, the fluid BA has priority over the fluid A, that is, it goes out first at rate γ_1 whenever present.

Our goal is to prove convergence to the fluid limit. Let us assume that the scaled inflows and initial states of the node converge to those of the fluid model as $N \to \infty$, that is

$$\lim_{N\to\infty} \sup_{t\in[0,T]} \|\Lambda^N(t) - \Lambda(t)\| = 0 \tag{41}$$

and

$$\lim_{N\to\infty} \rho_{KROV}\left(q^N(0), q(0)\right) = 0. \tag{42}$$

In what follows, we will need the following three bounds. The first one is the statement that the process $W^N(t)$ is very close to the function $\gamma^{-1}\Lambda^N(t)$—namely,

$$\lim_{N\to\infty} \mathbb{E}\left(\sup_{t\in[0,T]} \|W^N(t) - \gamma^{-1}\Lambda^N(t)\| \right) = 0. \tag{43}$$

A simpler claim concerns the sum $Nw^N(0)$ of the service times of all the users present in the queue $Nq^N(0)$ at the initial moment $t = 0$. Namely, for the conditional distribution of $w^N(0)$ under the condition $q^N(0) = q$ we have

$$\mathbb{E}\left\| \left(w^N(0) \,\middle|\, q^N(0) = q\right) - \gamma^{-1}q \right\| \le \psi(N)\|q\| \tag{44}$$

for some $\psi(N) \to 0$ as $N \to \infty$. Moreover,

$$\sup_{t\in[0,T]} \rho_{KROV}\left(w^N(t), \gamma^{-1}q^N(t)\right) \le \psi(N) \sup_{t\in[0,T]} \rho_{KROV}\left(q^N(t), \mathbf{0}\right). \tag{45}$$

Of course, for every fixed t the convergence in (43) follows from the Central Limit Theorem. The problem is that we need the convergence at all moments t. We will obtain (43) by constructing the finite-point event, which contain the one we are interested. Indeed, consider the event $E(t_0, c)$, $c > 0$, which consists of all trajectories such that

$$W_A^N(t_0) > \gamma_A^{-1}\left(\Lambda_A^N(t_0) + c\right). \tag{46}$$

Note that the function $\Lambda_A^N(t)$ has its derivative $\le C$ (see (40), so if the event $E(t_0, c)$ happens, then all the events $E(t_0 + \tau, c - C\tau)$ happen as well, since the function $W_A^N(t)$ is non-decreasing on every trajectory. Therefore we can replace the infinite union by the finite one:

$$\bigcup_{0\le t_0\le T} E(t_0, c) \subset \bigcup_{k=0}^{2CT/c} E\left(k\frac{c}{2C}, \frac{c}{2}\right),$$

and use the fact that for any fixed $t \le T$ and c the probability of the event $E(t, \frac{c}{2})$ is exponentially small in N. The remaining cases (corresponding to the second coordinate, AB, and to the lower estimate in (46)) are immediate.

The proof of (45) proceeds in a similar way. First of all, there is a natural coupling between the processes $w^N(t)$ and $q^N(t)$, corresponding to the fact that we can assume that the service time of every client is known at its arrival moment. Consider the event $E'(t_0, c)$, consisting of the trajectories where

$$w_A^N(t_0) \ge \gamma_A^{-1}\left(q_A^N(t_0) + c\right).$$

We would like to use the argument similar to the above, saying that if the bad event $E'(t_0, c)$ happens at t_0, then on a whole segment around t_0 something unlikely has to happen as well. Partially it can be done, since for every trajectory we have $w_A^N(t_0+\tau) \geq w_A^N(t_0) - \tau$, and so we have for every trajectory in $E'(t_0, c)$ that $w_A^N(t_0+\tau) \geq \gamma_A^{-1}(q_A^N(t_0) + c - \gamma_A \tau)$. If we can claim that $q_A^N(t_0) + c - \gamma_A \tau > q_A^N(t_0+\tau) + c/2$ for all τ small enough, then we would be done. However, the outcome that $q_A^N(t_0+\tau) > q_A^N(t_0) + c/2 - \gamma_A \tau$ is not excluded, even if τ is very small. Yet, the probability of the increase of the queue by $c/2 - \gamma_A \tau$ during the time τ is exponentially small in τ as $\tau \to 0$. Therefore, the event $\bigcup_{0 \leq t_0 \leq T} E'(t_0, c)$ is contained in the union $[\bigcup_{k=0}^{2C'T/c} E'(k\frac{c}{2C'}, \frac{c}{2})] \cup [\bigcup_{k=0}^{2C'T/c} E''(k\frac{c}{2C'}, \frac{c}{4})]$ for some suitably chosen C', where $E''(k\frac{c}{2C'}, \frac{c}{4})$ is the event that on the segment $[k\frac{c}{2C'}, (k+1)\frac{c}{2C'}]$ the increment $q_A^N((k+1)\frac{c}{2C'}) - q_A^N(k\frac{c}{2C'}) > c/4$, and we are done.

From (41) and (43), we get

$$\lim_{N \to \infty} \mathbb{E}\left(\sup_{t \in [0,T]} \left\| W^N(t) - \gamma^{-1} \Lambda(t) \right\| \right) = 0. \tag{47}$$

Note that

$$\rho_{KROV}\left(w^N(0), \gamma^{-1} q(0)\right) \leq \mathbb{E}\big(\mathbb{E}\left\|\left(w^N(0) \,\middle|\, q^N(0) = q\right) - \gamma^{-1} q\right\|\big) + \rho_{KROV}\left(\gamma^{-1} q^N(0), \gamma^{-1} q(0)\right).$$

(Here we treat $w^N(0)$ in the l.h.s. as the probability distribution.) So we get from (42) and (44) that

$$\rho_{KROV}\left(w^N(0), \gamma^{-1} q(0)\right) \leq \varphi(N) \rho_{KROV}\left(q(0), \delta_{\mathbf{0}}\right), \tag{48}$$

where $\varphi(N) \to 0$ as $N \to \infty$, and $\mathbf{0} \in \mathbb{R}^2$ is the origin.

Next, we need the following estimate (see for instance [9]):

Lemma 13 *Let $\Lambda(t)$ and $\Lambda'(t)$ be two inflows to the fluid priority node with initial (non-random) fluid levels $q(0)$ and $q'(0)$. Then*

$$\sup_{t \in [0,T]} \|q(t) - q'(t)\| \leq L(\|q(0) - q'(0)\| + \sup_{t \in [0,T]} \|\Lambda(t) - \Lambda'(t)\|),$$

where $L = L(\gamma_1, \gamma_2)$.

Proof Let us consider a fluid single-class node with variable capacity. Namely, let $q_1(0)$ be the (scalar) initial fluid level, let $\Lambda_1(t)$ be the inflow, and introduce $S_1(t)$ to be the server capacity, which is the amount of work the server can do during the time interval $[0, t]$. (For example, in our situation $S_1(t) = \gamma_1 t$, but we will consider more general case, with $S_1(t)$ not necessarily linear.) Introduce the *virtual level*

$$V(t) = q_1(0) + \Lambda_1(t) - S_1(t) \tag{49}$$

and the *unused service capacity*

$$U(t) = \max\left\{0, -\inf_{s \in [0,t]} V(s)\right\}. \tag{50}$$

Then

$$q_1(t) = V(t) + U(t), \quad t \geq 0. \tag{51}$$

Let us introduce the sup norm on the space of functions. Then the functionals $\{q_1(0), \Lambda_1(\cdot), S_1(\cdot)\} \to V(\cdot)$ and $\{V(\cdot), U(\cdot)\} \to q_1(\cdot)$, given by (49) and (51), have finite norms, since they are linear. The non-linear functional $V(\cdot) \to U(\cdot)$, given by (50), has finite norm as well. Indeed, the functional $V(\cdot) \to \inf_{[0,\cdot]} V(\cdot)$ has norm ≤ 1, since for any pair $x(\cdot), y(\cdot)$ of scalar functions

$$\left| \inf_{s\in[0,t]} x(s) - \inf_{s\in[0,t]} y(s) \right| \leq \sup_{s\in[0,t]} |x(s) - y(s)|,$$

and so

$$\sup_{t\in[0,T]} \left| \inf_{s\in[0,t]} x(s) - \inf_{s\in[0,t]} y(s) \right| \leq \sup_{t\in[0,T]} |x(t) - y(t)|;$$

the same holds for the functional $\{x(t)\} \to \max\{0, x(t)\}$, since

$$\sup_{t\in[0,T]} |\max\{0, x(t)\} - \max\{0, y(t)\}| \leq \sup_{t\in[0,T]} |x(t) - y(t)|.$$

Therefore the composed functional, taking the triplet $\{q_1(0), \Lambda_1(\cdot), S_1(\cdot)\}$ to the pair $\{q_1(\cdot), U(\cdot)\}$, defined by (49)–(51), has finite norm.

That proves the desired statement for the first component of q. Now, to finish the proof, we note that the capacity of the server for the users of the second class is given by $S_2(t) = \gamma_2 U(t)$, where $U(t)$ is the unused server capacity for the high-priority class (with $S_1(t) = \gamma_1 t$, $t \geq 0$). Then we repeat the argument above. □

Next we formulate as a separate statement the obvious remark that the evolution of the current remaining service time variable, $w(t)$, coincides with the evolution of the level of some evidently constructed fluid system.

Lemma 14 *Let users u_j of two possible types $i = A, BA$, with service times h_i^j arrive at the initially empty server at times t_i^j, $j = 1, 2, \ldots$, and let $w_i(t)$ be the evolution of the remaining service times. Consider also the fluid model with two classes of fluids, which starts in the empty state and is governed by the fluid inflows*

$$\Lambda_i(t) = \gamma_i \sum_{j: t_i^j \leq t} h_i^j.$$

(I.e., our fluids have "viscosities" γ_1^{-1} and γ_2^{-1}.) Then at every moment $t \geq 0$ the current levels of fluids at the server equal to $\gamma_i w_i(t)$, $i = 1, 2$.

Another auxiliary result is needed:

Lemma 15 *Let $q \in \mathbb{R}^2$ be (random) queue to our server, and w be the corresponding (random) amount of total work (=service time needed). The service times of the users are independent, and within the i-th class identically distributed with mean γ_i^{-1}. Then, for any $v \in \mathbb{R}^2$,*

$$\rho_{KROV}\left(\gamma^{-1} q, \delta_v\right) \leq \rho_{KROV}\left(w, \delta_v\right). \tag{52}$$

Proof Since the norm $\|\cdot\|$ is convex, we have for the conditional random variable $w|q=\bar{q}$ that $\|\gamma^{-1}\bar{q}-v\| \le \mathbb{E}\|(w|q=\bar{q})-v\|$. Averaging over $\bar{q}$ gives (52). □

Now, we derive the following result

Proposition 16 *Under the assumptions made above*,

$$\lim_{N\to\infty} \sup_{t\in[0,T]} \rho_{KROV}\left(q^N(t), q(t)\right) = 0. \tag{53}$$

Proof First,

$$\sup_{t\in[0,T]} \rho_{KROV}\left(q^N(t), q(t)\right) \le \max\{\gamma_1,\gamma_2\} \sup_{t\in[0,T]} \rho_{KROV}\left(\gamma^{-1}q^N(t), \gamma^{-1}q(t)\right).$$

Then we write a chain of inequalities. First of all we have

$$\begin{aligned}&\sup_{t\in[0,T]} \rho_{KROV}\left(\gamma^{-1}q^N(t), \gamma^{-1}q(t)\right)\\ &\quad\le \sup_{t\in[0,T]} \rho_{KROV}\left(w^N(t), \gamma^{-1}q^N(t)\right) + \sup_{t\in[0,T]} \rho_{KROV}\left(w^N(t), \gamma^{-1}q(t)\right),\end{aligned}$$

and the first summand can be bounded by (45). To estimate the distance $\rho_{KROV}(w^N(t), \gamma^{-1}q(t))$ we have to exhibit some joint distribution of $\gamma^{-1}q(t)$ and $w^N(t)$. We take the following one: first, we choose the coupling between $\gamma^{-1}q(0)$ and $w^N(0)$, using (44) and (42), getting

$$\rho_{KROV}\left(w^N(0), \gamma^{-1}q(0)\right) \le 2\psi(N)\rho_{KROV}\left(q(0), \mathbf{0}\right).$$

Given the joint realization of the initial values $(\gamma^{-1}q(0), w^N(0))$, the evolution of the coordinate $q(t)$ is deterministic, defined by the flows $\Lambda_{\bar{A}}(t)$ of the arriving fluids. The evolution $w^N(t)$ is stochastic, governed by the Poisson process with net rates $\Lambda_i^N(t)$. Therefore for every t we have to exhibit the coupling between the distribution of the vector $w^N(t)$ and deterministic value $q(t)$. The resulting *KROV* distance is precisely what the Lemmas 13 and 14 allow us to control:

$$\begin{aligned}&\rho_{KROV}\left(w^N(t), \gamma^{-1}q(t) \;\middle|\; w^N(0), q(0)\right)\\ &\quad\equiv \mathbb{E}\left(\left\|w^N(t)-\gamma^{-1}q(t)\right\| \;\middle|\; w^N(0), q(0)\right)\\ &\quad\le L\left(\|w^N(0)-\gamma^{-1}q(0)\| + \mathbb{E}\left(\sup_{s\in[0,t]}\|W^N(s)-\gamma^{-1}\Lambda(s)\|\right)\right).\end{aligned}$$

It remains to apply bounds (47) and (48) and deduce (53). □

2. The proof of our statement for the closed system does not require any extra arguments, since the closed system is a special case of the open system, where all the flows satisfy the relations defining the closed system. □

6 Main Result

In this section we finally formulate and prove our main theorem, which claims that the NLMP started from some special initial state behaves similarly to the fluid model in its

periodic regime, at all times $t \in (0, \infty)$. Since we know already that the NLMP is in turn a limit of networks of size M, as $M \to \infty$, our theorem implies that the large size ($M \gg 1$) Markov process ∇_M^N behaves similarly to the fluid model for a very long time, which time diverges as $M \to \infty$, provided the number N of clients per node exceeds some value N_0. In particular, there are initial states for the networks ∇_M^N, which lead to a long time oscillations, before the network reaches its stationary state.

Theorem 17 *Let $\varepsilon > 0$. Then there exist the values N_0, $\varepsilon' > 0$, $\alpha > 0$ and $E < \infty$ such that for all $N > N_0$ the states $\nu^N(t)$ of the NLMP process ∇_∞^N, started at the initial state $\nu^N(0)$ with the properties*:

$$\rho_{KROV}\left(\nu^N(0), \delta_{x(0)}\right) < \varepsilon',$$

with $x(0) \in \mathcal{C}$,

$$\langle \exp\{\alpha L(\bar{x})\}\rangle_{\nu^N(0)} < E,$$

satisfies for all $t > 0$

$$\rho_{KROV}(\nu^N(t), \delta_{x_v}) < \varepsilon, \tag{54}$$

where $x_v \in \mathcal{C}$ is some moving point, depending on the process $\nu = \{\nu^N(t), t \geq 0\}$. In particular, the process $\nu^N(t)$ has no limit as $t \to \infty$.

In words, we are proving that if we start the NLMP with high load N per server, from the state close to some atomic measure δ_z with z belonging to the cycle, then it never goes to a limit.

For that, we need a general lemma, which is formulated in the Euler scaling. First, we recall the definitions. Let $\bar{\lambda}(t) = \{\lambda_i(t), i = 1, \ldots, k\}$ be the rates of Poisson inflows of the customers of k types, and γ_i, $i = 1, \ldots, k$ be their rates of service. The discipline of service will be irrelevant here; we need only that the server is not idle if the queue is not empty. We call the flow $\bar{\lambda}(t)$ to be underloaded, with parameters (T, δ), if for any t

$$\sum_{i=1}^{k} \frac{1}{\gamma_i} \int_t^{t+T} \lambda_i(t)dt < (1-\delta)\, T.$$

Below we are talking about the flow with load N. That means that we consider the situation when the input rates are given by $N\bar{\lambda}(t) = \{N\lambda_i(t), i = 1, \ldots, k\}$, while the service rates are equal to $N\gamma_i$.

Lemma 18 *Consider the Non-Homogeneous Markov Process $\mu(t)$, started from the initial state $\mu(0)$, and suppose that its generating rate function $\bar{\lambda}(t)$ is underloaded, with parameters (T, δ). Then there exist values $\alpha > 0$ and $A < \infty$, depending only on the pair (T, δ), such that if the exponential moment $\langle \exp\{\alpha L(\bar{x})\}\rangle_{\mu(0)}$ of the initial state is finite, then for times $t > t(\mu(0))$ and for any load N*

$$\langle \exp\{\alpha L(\bar{x})\}\rangle_{\mu(t)} < A.$$

Moreover, there exists the time $\mathcal{T} = \mathcal{T}(T, \delta)$, such that for any initial state $\mu(0)$, satisfying the estimate

$$\langle \exp\{\alpha L(\bar{x})\}\rangle_{\mu(0)} < 3A,$$

we have, for any load N, that

$$\langle \exp\{\alpha L(\bar{x})\}\rangle_{\mu(\mathcal{T})} < 2A.$$

Proof We consider Poisson inflow with a general distribution η of the service time, having finite exponential moment. In particular, exponential service time fits.

The users arrive to the node according to the Poisson processes with rates $\lambda_i(t)$. Their service times are i.i.d. with distribution functions $\eta_i(h)$. Let $E_i = \mathbb{E}(\eta_i)$. We study the dynamics of the remaining service time, hence, the service discipline is of no importance. The regime we are interested in is the underloaded regime; that means that for some $\delta > 0$, all T large enough and all $t \geq 0$

$$\int_t^{t+T} \left(\sum_i \lambda_i(s) E_i \right) ds \leq T(1-\delta).$$

Let $d\mu_t(u)$ be the current distribution of the remaining service time. We want to study the exponential moment

$$Q_\alpha(t) = \int_0^\infty q_\alpha(u) d\mu_t(u), \tag{55}$$

where $q_\alpha(u) = e^{\alpha u}$. We will show that the moment $Q_\alpha(t)$ satisfies the equation

$$Q_\alpha(t) \leq e^{C_1 - \beta t} Q_\alpha(0) + C_2, \quad t \geq 0.$$

The statistics of the observable L will then be easy to derive.

Note that the underload condition ensures the absolute continuity of $Q_\alpha(t)$. We will need the quantities

$$\Phi_i^\alpha = \int_0^\infty \left(e^{\alpha h} - 1\right) d\eta_i(h).$$

We assume that $\Phi_i^\alpha < +\infty$ for all $\alpha \leq \overline{\alpha}$ with $\overline{\alpha} > 0$.

Before studying the moments (55), we will consider the situation of the "broken" server, when the clients (of one type) only come, but are not served. The queue then only grows in time, as is the workload u. The corresponding exponential moment will be denoted by $Q_\alpha^{(1)}(t)$. We have:

$$\dot{Q}_\alpha^{(1)}(t) = \lambda(t) \Phi^\alpha Q_\alpha^{(1)}(t). \tag{56}$$

Indeed, the event of arrival of a user with service time h at the queue with the current workload u shifts the workload to the value $u + h$, so the value of q_α changes from $e^{\alpha u}$ to $e^{\alpha(u+h)} = e^{\alpha u} + e^{\alpha u}(e^{\alpha h} - 1)$. In order to find $\dot{Q}_\alpha^{(1)}(t)$, we have to multiply the increment $e^{\alpha u}(e^{\alpha h} - 1)$ by the rate $\lambda(t)$ of the arrival event and to integrate it with respect to $d\mu_t(u) \times d\eta(h)$, since u and h are independent. In this way we arrive to (56).

Next, let us study the case of "broken pipe", when the inflow is zero, so the server works only on the initial supply of clients. The evolution of the distribution μ_t of the workload is given by the following simple relation:

$$\mu_{t+s}[a,b] = \begin{cases} \mu_t[a+s, b+s] & \text{if } 0 < a < b, \\ \mu_t(-\infty, b+s] & \text{if } a \leq 0 < b. \end{cases}$$

In words, the atom at $u = 0$ grows with time. We denote the corresponding exponential moment by $Q_\alpha^{(2)}(t)$. The straightforward computation shows that

$$\dot{Q}_\alpha^{(2)}(t) = \alpha p_0(t) - \alpha Q_\alpha^{(2)}(t), \tag{57}$$

where $p_0(t)$ is the current probability of the queue to be empty. Note that $Q_\alpha^{(2t)}(t)$ is absolutely continuous and (57) holds for almost all t.

In the general case of several inflows we put the two relations together to get

$$\dot{Q}_\alpha(t) = \left[\left(\sum_i \lambda_i(t)\Phi_i^\alpha\right) - \alpha\right] Q_\alpha(t) + \alpha p_0(t). \tag{58}$$

Let us now rewrite Φ_i^α. We have

$$\begin{aligned}\Phi_i^\alpha &= \int_0^\infty \left(e^{\alpha h} - 1\right) d\eta_i(h) \\ &= \int_0^\infty \alpha h \, d\eta_i(h) + \int_0^\infty \left[e^{\alpha h} - 1 - \alpha h\right] d\eta_i(h) \equiv \alpha E_i + \alpha F_i(\alpha),\end{aligned} \tag{59}$$

where E_i is the mean service time and $F_i(\alpha)$ is continuous function of $\alpha \in [0, \overline{\alpha}]$, which satisfies $F_i(\alpha) = O(\alpha)$ as $\alpha \to 0$. From (58) and (59) we get for $\alpha < 1$ the bound

$$\dot{Q}_\alpha(t) \le \alpha \left[\left(\sum_i \lambda_i(t) E_i\right) - 1 + \sum_i \lambda_i(t) F_i(\alpha)\right] Q_\alpha(t) + 1. \tag{60}$$

The Euler scaling with parameter N changes $\lambda(t)$ to $\lambda^N(t) = N\lambda(t)$ and $\eta(h)$ to $\eta^N(h) = N\eta(Nh)$. Hence, $\lambda_i^N(t)E_i^N$ does not depend on N. Let us show that $NF_i^N(\alpha)$ is small for all N, once α is small. Indeed,

$$\begin{aligned}N\alpha F^N(\alpha) &= N^2 \int_0^\infty \left[e^{\alpha h} - 1 - \alpha h\right] d\eta(Nh) \\ &= N^2 \int_0^\infty \left[e^{\frac{\alpha}{N} Nh} - 1 - \frac{\alpha}{N} Nh\right] d\eta(Nh) \\ &= N^2 \frac{\alpha}{N} F\left(\frac{\alpha}{N}\right) \sim \alpha^2.\end{aligned}$$

Hence, we get a uniform bound for α small enough and all $N \ge 1$ simultaneously (and, by the limit, for the fluid model "$N = \infty$" as well):

$$\dot{Q}_\alpha^N(t) \le \alpha \left[\sum_i \lambda_i(t)\left(E_i + \varkappa(\alpha)\right) - 1\right] Q_\alpha^N(t) + 1, \tag{61}$$

where $\varkappa(\alpha) \sim \alpha$.

The solution to the linear equation

$$\dot{x}(t) = a(t)x(t) + b(t)$$

is given by the formula

$$x(t) = g(0,t)x(0) + \int_0^t g(s,t)b(s)\,ds,$$

where $g(s,t) = e^{\int_s^t a(\tau)d\tau}$. We apply it to (61), with $x(t) = Q_\alpha(t)$, $a(t) = \alpha[\sum_i \lambda_i(t)(E_i + \varkappa(\alpha)) - 1]$ and $b(t) = 1$. By the underload assumption,

$$\int_s^t a(\tau)d\tau \le C_1 - \beta(t-s)$$

for some $C_1, \beta > 0$ and for all $s < t$, once α is small. Then, $\int_0^t g(s,t)ds \le C_2$ for all $t \ge 0$.

Hence,

$$Q_\alpha(t) \le e^{C_1 - \beta t} Q_\alpha(0) + C_2, \quad t \ge 0.$$

In our case the distribution η_i is exponential with the parameter γ_i. Let us show finally that the exponential bound on the workload implies an exponential bound on the number of customers (may be, with another exponent).

Indeed, under the condition that we are in the state with n_1 and n_2 customers of two classes, the conditional distribution of the workload u is a measure $\mu_{n_1 n_2}$ on $(\mathbb{R}^1)^+$, with mean value $\bar{u} = \frac{n_1}{\gamma_1} + \frac{n_2}{\gamma_2}$. By convexity of the exponent,

$$\int e^{\alpha u} d\mu_{n_1 n_2}(u) \ge e^{\alpha \bar{u}},$$

which provides us with the upper bound

$$e^{\frac{\alpha}{\gamma_1+\gamma_2}(n_1+n_2)} \le \int e^{\alpha u} d\mu_{n_1 n_2}(u).$$

Taking expectations with respect to $n_1(t), n_2(t)$ we get

$$\mathbb{E}\left(e^{\frac{\alpha}{\gamma_1+\gamma_2}(n_1(t)+n_2(t))}\right) \le \int e^{\alpha u} d\mu_t(u) = Q_\alpha(t),$$

which is the desired estimate. □

Proof of the Main Theorem The proof proceeds by "induction" in time. We suppose inductively that at a certain (Euler) moment T the NLMP with the load N is in the state $\mu(T)$, having two properties:

$$\langle \exp\{\alpha L(\bar{x})\}\rangle_{\mu(T)} < 3A, \tag{62}$$

$$\rho_{KROV}(\mu(T), \delta_x) < \varepsilon \tag{63}$$

for some $x \in \mathcal{C}$. We will show that there exists the time T', at which the same two conditions hold for the measure $\mu(T+T')$—except for different point x on the cycle $\mathcal{C}$.

To see this we first consider the Non-Linear Dynamical System (NLDS) Δ_∞, with initial state $\mu(T)$. In other words, $\Delta_\infty = \Delta_\infty(\bar{Y}(\cdot))$, for some $\bar{Y}(\cdot) \in \mathcal{Y}(\mu(T))$. We can use the Proposition 10, which tells us that for any T' large enough $\rho_{KROV}(\Delta_\infty^{T'}\mu(T), \delta_{x(T')}) < \varepsilon/3$. Choosing one such T' (uniformly in $\mu(T)$, satisfying (62)–(63)) we can claim that for the

NLMP evolution we have $\rho_{KROV}(\mu(T+T'),\delta_{x(T')}) < 2\varepsilon/3$, provided only that N is large enough; indeed, we know from Theorem 12 that the NLMP converges in the KROV metric to NLDS on any finite time interval, as $N \to \infty$, which convergence is uniform over the set of initial measures satisfying (62). Note that we thus have reproduced the condition (63).

Proposition 11 tells us that for all $t \le T'$ $\rho_{KROV}(\Delta^t_\infty \mu(T), \delta_{x(t)}) < \bar{\varepsilon}(T', \varepsilon)$. Due to the same convergence statement, for the NLMP evolution we have $\rho_{KROV}(\mu(T+t), \delta_{x(t)}) < 2\bar{\varepsilon}(T', \varepsilon)$ for all $t \le T'$. In words, the measure $\mu(T+t)$ goes very close to the cycle trajectory. Since we want to use Lemma 18, we can as well assume that $T' > \mathcal{T}$, where $\mathcal{T}$ is the time introduced in this lemma. Now all its conditions are satisfied, so Lemma 18 tells us that $\langle \exp\{\alpha L(\bar{x})\}\rangle_{\mu(T+T')} < 2A$, thus the condition (62) is reproduced as well. □

7 Conclusions

Our main result indicates that there is an important similarity between large queuing networks and large systems of statistical mechanics. Namely, we have shown that the load per server plays for some networks the same role as the inverse temperature in statistical mechanics. At high load the network can lose the property of uniqueness of the stationary state and start to behave in the oscillatory manner. This phenomenon looks similar to the fact that some 3D systems with continuous symmetry are not ergodic under Glauber dynamics, when the temperature is low enough.

It is very interesting to understand how general this phenomenon is; our expectations are that such non-ergodic behavior is a characteristic feature of the high load regime.

In the forthcoming publications we will show that the behavior in the low load regime is always ergodic, which corresponds to the high temperature uniqueness of statistical mechanics.

Acknowledgements We would like to thank our colleagues—in particular, J. Chayes, F. Kelly, M. Biskup, O. Ogievetsky, G. Olshansky, Yu. Peres, S. Pirogov,—for valuable discussions and remarks, concerning the topic of this paper. A. Rybko would like to acknowledge the financial support and hospitality of CPT, Luminy, Marseille, where part of the work was done. This work was also partially supported by RFBR Grants 06-01-72556, 07-01-92215, and 07-01-92216.

References

1. Bramson, M.: Instability of FIFO queueing networks. Ann. Appl. Probab. **4**, 414–431 (1994)
2. Bramson, M.: Instability of FIFO queueing networks with quick service times. Ann. Appl. Probab. **4**, 693–718 (1994)
3. Dai, J.G.: On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. Ann. Appl. Probab. **5**, 49–77 (1995)
4. Dobrushin, R.L., Karpelevich, F.I., Vvedenskaya, N.D.: Queuing systems with choice of shortest queue—asymptotic approach. Probl. Pereda. Inf. **32**(1), 20–36 (1996)
5. Ethier, S.N., Kurtz, T.G.: Markov Processes. Characterization and Convergence. Wiley, New York (1986)
6. Hepp, K., Lieb, E.H.: Phase transitions in reservoir-driven open systems with applications to lasers and superconductors. Helv. Phys. Acta **46**, 574–603 (1973)
7. Kleinrock, L.: Communication Nets, Stochastic Message Flow and Delay. McGraw-Hill, New York (1964)
8. Kumar, P., Seidman, T.: Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. IEEE Trans. Automat. Contr. **35**, 289–298 (1990)
9. LeBoudec, J.Y., Thiran, P.: Network Calculus—A Theory of Deterministic Queuing Systems. Lecture Notes in Computer Science, vol. 2050. Springer, Berlin (2001)

10. Liggett, T.M.: Interacting particle systems. Grundlehren der Mathematischen Wissenschaften, vol. 276. Springer, New York (1985)
11. McKean, H.P. Jr.: A class of Markov processes associated with nonlinear parabolic equations. Proc. Natl. Acad. Sci. USA **56**, 1907–1911 (1966)
12. McKean, H.P. Jr.: An exponential formula for solving Boltzmann's equation for a Maxwellian gas. J. Comb. Theory **2**, 358–382 (1967)
13. Puhalsky, A., Rybko, A.: Non-ergodicity of queueing networks when their fluid model is unstable. Probl. Inf. Transm. **36**, 26–46 (2000)
14. Rybko, A.N., Shlosman, S.B.: Poisson hypothesis for information networks, http://fr.arxiv.org/PS_cache/math/pdf/0406/0406110.pdf. Sinai's Festschrift, Moscow Math. J. **5**, 679–704 (2005). Tsfasman's Festschrift, Moscow Math. J. **5**, 927–959 (2005)
15. Rybko, A.N., Shlosman, S.B., Vladimirov, A.: Self-averaging property of queuing systems, http://fr.arxiv.org/abs/math.PR/0510046. Probl. Inf. Transm. (4) (2006)
16. Rybko, A.N., Stolyar, A.L.: Ergodicity of stochastic processes describing the operation of open queuing networks. Probl. Inf. Transm. **28**, 199–220 (1992)
17. Stolyar, A.L.: The asymptotics of stationary distribution for a closed queueing system. Probl. Pereda. Inf. **25**(4), 80–92 (1989) (in Russian). Translation in Probl. Inf. Transm. **25**(4), 321–331 (1990)
18. Stolyar, A.L.: On the stability of multiclass queueing networks: a relaxed sufficient condition via limiting fluid processes. Markov Process. Relat. Fields **1**, 491–512 (1995)