Int. J. Production Economics 120 (2009) 378–388

Contents lists available at ScienceDirect

Int. J. Production Economics

journal homepage: www.elsevier.com/locate/ijpe

Circulation of autonomous agents in production and service networks

Olivier Gallay *, Max-Olivier Hongler

Ecole Polytechnique Fédérale de Lausanne (EPFL), STI-IMT-LPM, Station 17, 1015 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 3 January 2007
Accepted 21 January 2008
Available online 31 January 2009

Keywords:
Queueing networks
Production and service systems
Autonomous agents
History-based routing decisions
Self-organization

ABSTRACT

The inherent complexity characterizing production and/or service networks strongly favors decentralized and self-organizing mechanisms to regulate the flows of matter and information in circulation. This basic observation motivates us to study the flow dynamics in queueing networks roamed by autonomous agents which, at a given time and at a given vertex location, select their routing according to (individual) historical data (such as waiting times) collected during their past progression in the network. For several simple network configurations and despite the intrinsically non-Markovian character of the dynamics, we are able to discuss analytically the emerging collective dynamics that such a circulation of autonomous agents generates. Feedback loops in the network topology coupled with the presence of delays in the routing selection mechanisms produce a wealth of dynamical phenomena like self-sustained generically stable oscillations, spatio-temporal patterns, stabilization by noise phenomena and oscillator synchronization that are explicitly discussed in this paper.

1. Introduction

The optimal control of the flow dynamics of matter, information and money feeding complex network structures is a classical topic in operational research. This general problem arises naturally in several strategic areas such as production/supply chains, passengers and cargo transport and computerized communication systems. The flow dynamics depend jointly on the routing rules defining the ways the flows are dispatched at the network vertices and on the server dynamics which process the various items in circulation. Due to stochastic customer demands, to fluctuations in the raw material supply chains, to failures arising in the production devices, to uncertainties in operator availability and to ubiquitous financial volatility steadily affecting optimization objectives, the flow dynamics are always affected by random fluctuations. The need to model, to study and to quantify the characteristics of such complex stochastic dynamics has strongly stimulated development of queueing networks (QNs) theory. Nowadays, the QNs theory offers a wealth of reliable mathematical tools for calculating most relevant performance measures of such dynamics. The basic hypothesis behind any QN modeling is the possibility to describe the underlying dynamics by general Markovian processes. When this is realized, very general results (pioneered by the widely known Jackson factorization theorem) are available to characterize stationary flow regimes (Chen and Yao, 2001; Serfozo, 1999; Walrand, 1988), and, hence, stationary performance measures. Imposing such a Markovian character obviously limits not only the dynamics of the servers but also lays down strong restrictions on the allowable routing rules followed by the circulating items. In the present paper, we will study networks for which the Markovian character of the dynamics has to be abandoned. More precisely, the non-Markovian features will originate from the routing decisions which will be based on the items' individual

* Corresponding author. Tel.: +41 21 6935817; fax: +41 21 6933891.
E-mail address: olivier.gallay@epfl.ch (O. Gallay).

0925-5273/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijpe.2008.01.012

experiences collected during their journey through the QNs (i.e. we will speak of *history-based* (HB) *routing laws*). The presence of memory mechanisms in routing decision rules explicitly precludes the Markovian character of the underlying dynamics and this opens wide the door for the emergence of entirely new dynamical features. More particularly, the very existence of stationary regimes can no longer be taken for granted when HB rules are implemented. As we will see, the joint presence of feedback loop topology in the QNs (i.e. possibilities of flow re-injections) and HB routing decisions is responsible for the emergence of spatio-temporal patterns in the flow dynamics.

In general, numerous HB routing rules are possible. In this paper, we will focus on situations where the time spent in specific sections of the QNs will be the criterion used to select a specific route on which to engage at a bifurcating node of the network. Implicitly, such waiting time criteria require a real-time capability for each circulating agent[1] to monitor, to memorize and then to process data to ultimately form a routing decision. As a direct consequence, one realizes that we actually deal with QNs roamed by autonomous, decision making (i.e. *intelligent*) agents, with *stigmergic*[2] mutual interactions.

Networks in which HB routing decisions are present can be easily identified in various important contexts such as transportation (pedestrian, car, train and air traffic issues), production and supply chains, leisure and hospitality (theme parks, ski resorts, hotels and food industry management) and health care management. Despite this wealth of applicability, the literature devoted to formal models studying the flow dynamics involving HB routing rules remains so far rather scarce. Obviously, this is partly due to the technical difficulties inherent in the non-Markovian and non-linear features of the underlying dynamics. Nevertheless, several recent illustrations where directly related (yet mostly experimental) situations can be pointed out.

(a) *Leisure and hospitality*. In Bielen and Demoulin (2007), the influence of waiting time on the satisfaction and loyalty of customers using recurrent services is exhibited and explicitly studied. More particularly, following a survey conducted in the medical care industry (which is in many aspects in close connection with the hospitality business sector), contribution (Bielen and Demoulin, 2007) aims to investigate how customers use their waiting time satisfaction in order to determine whether to remain loyal (i.e. they will come back for future services) or alternatively to change their service provider. As discussed in Law et al. (2004), the waiting time also influences significantly the satisfaction and return decision of customers visiting fast-food facilities. General models of recurrent services where customer short-term satisfaction is driven solely by the perceived waiting time while queueing are presented in Haxholdt et al. (2003), van Ackere and Larsen (2004) and van Ackere et al. (2006). In complete analogy with the present modeling framework, customers base their service quality measure on the waiting time and will adapt their visit frequency to that particular measure. The models introduced in Haxholdt et al. (2003), van Ackere and Larsen (2004) and van Ackere et al. (2006) find natural applications in sports clubs, supermarkets and internet access management. The theme park industry is another sector to which the models presented in this paper are closely related (Kataoka et al., 2005; Kawamura et al., 2004; Kawamura et al., 2004). Roughly speaking, customers will decide to line up again for a new run at the same attraction after an exciting roller-coaster ride providing their waiting time remains acceptable for them. Ski traffic management offers another world-wide illustration where customer satisfaction and, hence, future (HB) routing decisions are directly related to suffered waiting time (Pullman and Thompson, 2002). Indeed, the waiting time spent at a ski lift clearly affects the customers' future decision on change of slope or not. Note that, throughout the present paper, only the last recorded waiting time will be used as a routing decision criterion. Relaxing this assumption by, for instance, allowing the customers to use the information collected during their successive visits to a server is also relevant. A first study including this feature is proposed in Gallay and Hongler (2008).

(b) *Supply chains and production management*. The need, in supply chain management, for coordination strategies leading to adaptive, flexible and collective behaviors motivated a recent contribution proposed by Surana et al. (2005), in which the authors show how a coherent global behavior can be generated by using only elementary components with local interactions. This paper explains how basic concepts and operational tools of Complex Adaptive Systems (CAS) fit naturally and efficiently for characterizing the supply chains dynamics. In this context, contributions (Kumara et al., 2003; Surana et al., 2005) expose the dynamics of simple QNs for which feedback loops and delays coexist and yield temporal oscillations of queue contents. Originally introduced in the framework of telephone switching systems (Erramilli and Forys, 1991), these particular QNs have been further applied in the context of supply chains in Kumara et al. (2003) and Surana et al. (2005). Note, however, that, contrary to the waiting time criterion to be studied in the present paper and which mostly characterizes leisure and hospitality systems, the HB features in Erramilli and Forys (1991), Kumara et al. (2003) and Surana et al. (2005) are due to HB changes of the agents' priority status in re-entrant queueing systems. The relevance and legitimacy of CAS in logistics, as well as their practical implications, have been recently strongly emphasized in Hülsmann et al. (2008) and

[1] From now on, we will speak of agents, items and customers interchangeably.

[2] Stigmergic qualifies indirect communication in a self-organizing system where individual parts (here the circulating items) interact with one another by modifying their local environment (here the environment is basically the queue length).

Wycisk et al. (2008). The actual impact of decentralized management and of the resulting self-organization on process and product quality is addressed in Massotte and Bataille (2000).

In close connection to actual production issues, let us also mention the recent contributions devoted to *Real-Time Queueing Systems* (RTQS) (Baldwin et al., 2000; Doytchinov et al., 2001; Lehoczky, 1997). Unlike standard queueing theory, RTQS focus on the ability of a queue discipline to meet production task timing requirements, for instance the distribution of lateness. In the RTQS described in Lehoczky (1997), each incoming task in a queueing system is endowed with a *due date* before which it has to be completed. To reduce the potential lateness, a *dynamic scheduling policy* in which waiting tasks are placed by decreasing *leadtime* (i.e. the more urgent being the closest to the server) is implemented. This dynamical scheduling rule can be viewed as a special illustration of HB routing. Indeed, this *earliest-deadline-first priority rule* implies that each incoming task triggers a rearrangement (i.e. a real-time routing) dependent on the waiting history of all tasks present in the queue.

Finally, contributions (Klein et al., 2005; Whitby et al., 2001) discuss network configurations where the joint effect of non-linearities and delays affects the underlying dynamics, thus showing similar features with our present study. Both in Klein et al. (2005) and Whitby et al. (2001), the dynamics exhibit, not surprisingly, an oscillatory behavior. However, while the delay is, in our work, self-induced by the agents themselves, the delay is produced by an external source in Klein et al. (2005) and Whitby et al. (2001).

Our paper is organized as follows. In Section 2, we introduce the constitutive elements, namely a feedback loop and HB routing mechanism, to show how stable temporal oscillations can already be generated in such a simple QN. In Sections 3 and 4, we analyze more complex networks involving two servers with feedback loop topology. The open network considered in Section 3 studies the traffic load partition between two parallel servers. The closed network of Section 4 idealizes the quality of service competition which arises between two service providers.

2. Single feedback loop — siphon dynamics

The simplest possible network composed of a single queue with the presence of a feedback routing node is sketched in Fig. 1. An incoming flow of customers, described by a renewal process with mean inter-arrival time $1/\lambda$ and probability distribution $A(x)$ with density $dA(x)$, is served by a processing unit whose service times are i.i.d. random variables with mean $1/\mu$, probability distribution $B(x)$ and density $dB(x)$. Accordingly, the parameters λ and μ are, respectively, the incoming and service rates of the renewal processes. We assume that the distributions $A(x)$ and $B(x)$ have finite moments. Here, we suppose that the traffic intensity $\rho=(\lambda/\mu)<1\Leftrightarrow\lambda<\mu$, which ensures the stability of the queueing system in absence of feedback loop. Assume also that the waiting room capacity is unlimited and that the service discipline is first-in-first-out (FIFO). After being served, the routing of each customer at the QN decision node n will be either:

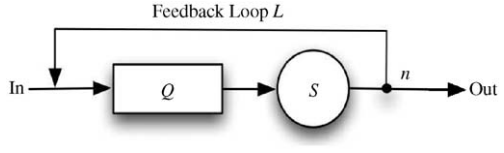

Fig. 1. A single stage queueing system with feedback loop.

(i) to leave the system definitively,
(ii) to follow the feedback loop and line up again to be served once more.

Several well-known contributions (D'Avignon and Disney, 1976; Peköz and Joglekar, 2002; Takàcs, 1963) consider the situation arising when the decision between the choices (i) and (ii) is taken randomly. When this is the case, by imposing a stationary flow balance (i.e. incoming equals outgoing flow), the system is driven into a self-consistent stationary regime. As we will now see, such purely stationary flows strongly differ from the queue dynamics observed when "intelligent" agents, able to base their routing on historical data (here the time spent while queueing and being served), circulate in the network. Specifically, following (Filliger and Hongler, 2005), assume now that each customer is able to record the total waiting time W he spent to receive service (i.e. W is the sum of queueing and processing times; in queueing theory, W is commonly known as the sojourn time). Assume further that W controls the routing decision between the alternatives (i) and (ii), namely the following HB routing rule $\mathscr{R}$ is implemented:

$$\mathscr{R}=\begin{cases}\text{follow alternative (i)} & \text{if } W>P,\\ \text{follow alternative (ii)} & \text{if } W\leqslant P,\end{cases} \tag{1}$$

where P is called here the *patience parameter* of the agents in circulation. When alternative (ii) is chosen, we will speak of loyal customers, as the agents are pleased with the server and then return to it for another service. Note that the flow of customers taking the feedback loop is added to the flow of fresh customers entering the system (with rate λ). We assume that, when joining the queue, a loyal customer behaves as a fresh customer (i.e. only the last recorded sojourn time will be determining for its routing at node n). Note that the routing is now clearly HB — it is determined by the sojourn time that each customer spent to be served. Later in this paper, we focus on homogenous agents for which P is a common value. The underlying idea behind this type of model is to study how, by modifying the quality of service (here the sojourn time), it is possible to enhance the circulation of loyal customers (i.e. those taking the feedback loop).

As discussed in Filliger and Hongler (2005), let us now show that when the circulating items apply the HB routing rule $\mathscr{R}$ stated in Eq. (1) and when P is large

enough, quasi-deterministic cyclo-stationary regimes emerge, i.e. *stable temporal oscillations* of the queue level $Q(t)$ are observed and this is independent of the detailed nature of the probability laws $A(x)$ and $B(x)$. Indeed, despite the presence of the fluctuations, this robust and quasi-deterministic behavior is directly reminiscent from the *law of large numbers*. The importance of the relative fluctuations around the associated average sojourn time $\langle W\rangle$ (which is the sum of individual processing times) decreases for large queue content $Q(t)$ (a quantitative characterization is given in Filliger and Hongler, 2005). Accordingly, for large P, the dynamics can be discussed via a deterministic approach (involving a constant service time $1/\mu$) (Filliger and Hongler, 2005; Hongler et al., 2004). Hence, for a given queue length N_c and a given corresponding patience parameter $P = N_c/\mu$, an incoming tagged customer (called $\mathscr{C}$ from now on) lining behind N_c other customers, will, when reaching node n, choose the alternative (i) (i.e. leave the system). Indeed, for such a deterministic regime, the measured total waiting time $W = (N_c/\mu) + (1/\mu) > P$. However, before $\mathscr{C}$ makes its way through the queue and reaches the node n, the queue content $Q(t)$ still increases at the (deterministic) rate λ (as nobody leaves the system during this time interval), implying a delay mechanism in the draining of the queue content. As soon as $\mathscr{C}$ reaches n, and thus leaves the system, a second dynamical phase is triggered. During this second phase, the customers arriving immediately after $\mathscr{C}$ do also experience a waiting time exceeding P and, hence, will also leave the system. As $\lambda < \mu$, the queue population $Q(t)$ decreases during this second dynamical phase and the depletion lasts until a satisfied customer (and hence his immediate successors) reach the node n. When this happens, the first dynamical phase starts again and $Q(t)$ fills up at rate λ. The alternation between these two dynamical phases produces a *cyclo-stationary behavior whose very existence is entirely due to the elementary "intelligence" attached to the circulating agents* — elementary "intelligence" being here due to the capability to monitor and to memorize the time while queueing and to take an individual routing decision accordingly. It is enlightening to visualize the queue dynamics by using the hydrodynamic analogy sketched in Fig. 2. Indeed, one can convince oneself that the time-dependent queue content level is fully analogous to the liquid oscillations arising in a self-siphoning "Tantalus glass" (sketched in the right of Fig. 2). In addition, for large P, the purely deterministic context ensuing from the law of large numbers enables an elementary derivation of both the amplitude Δ and the period Π of the queue population. Following Filliger and Hongler (2005) for further analytical details, we obtain (see also Figs. 2a and b):

$$\Delta = P\mu, \tag{2}$$

$$\Pi = P\left[2 + \lambda\mu - \lambda + \frac{\mu - \lambda}{\lambda}\right], \tag{3}$$

and provided $P \gg \max(1/\lambda, 1/\mu)$ both Eqs. (2) and (3) are in perfect agreement with simulation experiments (see Fig. 3), as discussed in Filliger and Hongler (2005) and Hongler et al. (2004), and this for any possible choice of the probability distributions $A(x)$ and $B(x)$. As shown in Fig. 3, the influence of the law of large numbers explicitly grows as P increases and causes the curves to become smoother and (quasi-)deterministically periodic with growing P.

3. Parallel feedback loops

Let us now increase the complexity of the QN and pay attention to the configuration $\mathscr{D}$, formed by a dipole of feedback queues, as sketched in Fig. 4. Two feedback queueing systems of the type introduced in Section 2 are placed in parallel. The total incoming external customers feeding this system is a renewal process with rate Λ. At a first decision node (DN) n_e (where e stands for *entry*), the agents face two routing possibilities: to either join server S_u, or to join S_d. In front of S_u and S_d, the agents wait in queues whose respective contents will be denoted by $Q_u(t)$ and $Q_d(t)$ (from Fig. 4, the indices u and d stand for *up* and *down*, respectively). We will write by μ_u and μ_d the respective service rates of S_u and S_d. The capacities of both queues $Q_u(t)$ and $Q_d(t)$ are assumed to be unlimited and the service policy is FIFO. The presence of the feedback loops introduces two DN's n_u and n_d. As in Section 2, at n_u

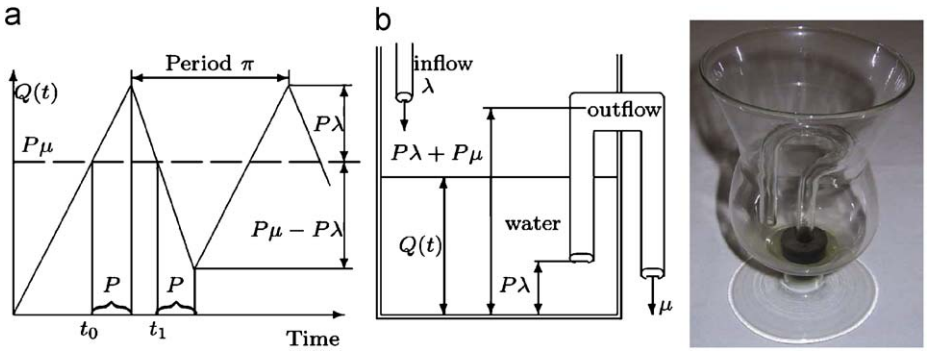

Fig. 2. Hydrodynamic analogy. (A) The agent entering at t_0 is the first one of a whole cluster U of unsatisfied customers and triggers the alternation of $Q(t)$ from the increasing to the decreasing state at $t_0 + P$. The last agent belonging to the cluster of unsatisfied customers U is the one entering just before t_1 and triggers the switch of $Q(t)$ from the decreasing to the increasing state at $t_1 + P$. This simple delay dynamic repeats and creates stable oscillations. (b) The "Tantalus glass" siphon model. The queue length corresponds to the water level $Q(t)$. The inflow and outflow rates are λ, respectively μ. The siphon leaves a water residue of height $P\lambda$ due to the constant inflow during P. The effective siphon length is $P\mu$.

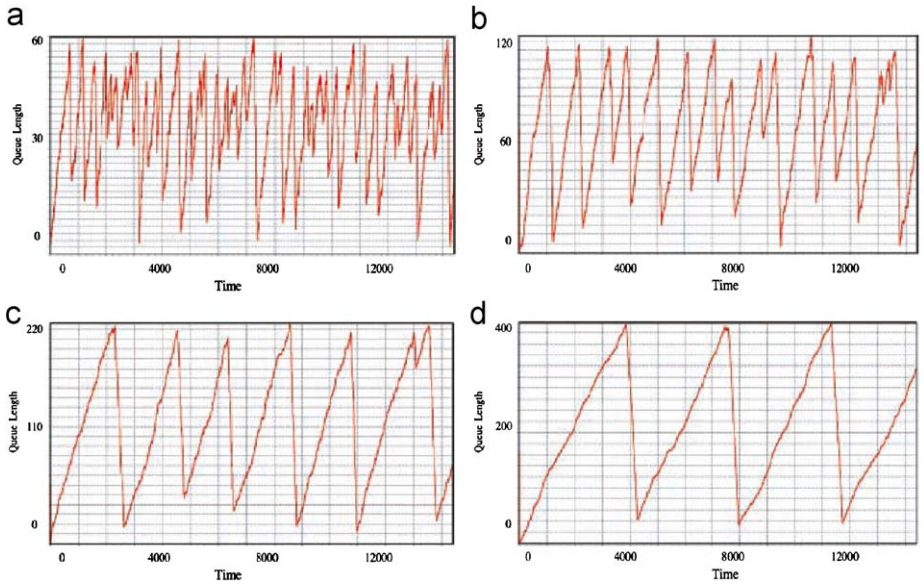

Fig. 3. Queue length oscillations obtained by simulation for exponentially distributed inter-arrival and service times with $\lambda = 0.1$, $\mu = 1$ and (a) $P = 50$, (b) $P = 100$, (c) $P = 200$, (d) $P = 350$. These simulated behaviors (and more precisely the influence of the law of large numbers, which is increasing with P) remain qualitatively the same for any other possible inter-arrival and service times probability distributions.

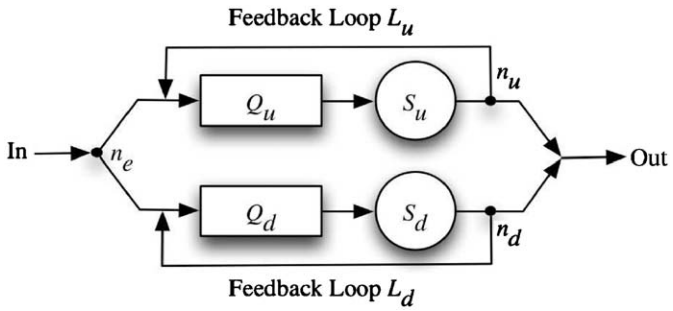

Fig. 4. A bifurcation of queueing systems with feedback loop.

and n_d the decision to enter into the feedback loop depends on the sojourn time W individually measured by each customer. In the following, we will separately consider three typical scenarios depending on the agents' ability to gather information.

3.1. Fixed entrance dispatching rule

Let us start with blind agents only being able to record the total waiting time spent to receive service (i.e. queueing + processing times) but unable to observe the contents $Q_u(t)$ and $Q_d(t)$. Hence, in this case, the routing decision at node n_e does not depend on agents' "intelligence" and an incoming customer selects between S_u and S_d by using either a deterministic or a random rule, independent of the content of Q_u and Q_d. Typical dispatching rules can be:

(i) *Deterministic polling*. In this case, the time horizon is divided into deterministic intervals T_u and T_d during which S_u and S_d, respectively, are alternatively fed with the total incoming traffic Λ. The conditions $\rho_u = T_u/(T_u + T_d) \cdot \Lambda/\mu_u < 1$ and $\rho_d = T_d/(T_u + T_d) \cdot \Lambda/\mu_d < 1$ ensure the stability of the system. In view of Section 2, it is not surprising that stable oscillations of the queue contents will, here again, be observed. However, instead of being smooth, the alternative feeding of the servers creates indentations in the time evolutions of Q_u and Q_d. The frequency of the alternations, given by T_u and T_d, determines the indentation structure. Qualitatively, increasing the frequency of the alternations does decrease the roughness of the curve. For large P, the amplitudes and frequencies of the two decoupled oscillations can be determined using Eqs. (2) and (3) with the parameters μ_u, $\lambda_u = T_u\Lambda/(T_u + T_d)$ on one hand and $\mu_d, \lambda_d = T_d\Lambda/(T_u + T_d)$ on the other hand. This is in perfect agreement with the simulation experiments given in Fig. 5.

(ii) *Random dispatching rule*. Here, we typically consider a Bernoulli sampling of the incoming flow, where

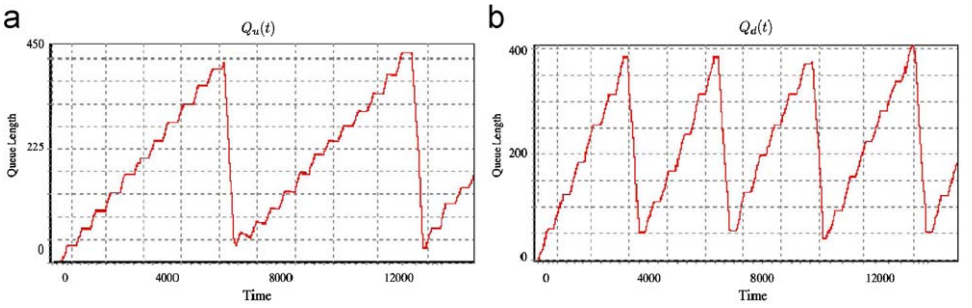

Fig. 5. Deterministic polling entrance rule: queue length indented oscillations obtained by simulation for exponentially distributed inter-arrival and service times with $\Lambda = 0.2$, $\mu_u = 1.25$, $\mu_d = 1$, $T_u = 200$, $T_d = 300$ and $P = 350$.

the Bernoulli random variable is determined by a parameter r, $(0 \leqslant r \leqslant 1)$. A partial traffic with rate $r\Lambda$ enters into server S_u while a traffic with rate $(1-r)\Lambda$ enters into S_d. The Bernoulli sampling implies that both systems S_u and S_d evolve independently and individually follow the dynamics exposed in Section 2. For large P, two decoupled, quasi-deterministic cyclo-stationary oscillations whose amplitudes and frequencies were determined by using Eqs. (2) and (3) with parameters μ_u, $r\Lambda$ and μ_d, $(1-r)\Lambda$ respectively.

3.2. Entrance dispatching based on partial observation of the queue contents — noise induced stabilization

Besides chronometers to record W, each customer is now endowed with a "visual system" enabling him to observe, in real-time, the instantaneous queue content $Q_u(t)$. Assume however that $Q_d(t)$ always remains hidden to the incoming agents, although they do know the average service rate μ_d. At time t, an incoming agent at node n_e first observes the queue content $Q_u(t)$ and, based on his observation, decides either to enter S_u or to join S_d. Once entered into a queue, neither reneging nor jockeying (i.e. jumping between S_u and S_d) is allowed. Note that except for the presence of feedback loops, this network configuration is fully similar to the two gas stations network studied in Hassin (1996). Recall that in this contribution, two gas stations are located one after the other on a main road. A driver who needs to refuel is only able to observe the queue length $Q_u(t)$ at the first station (which would be here S_u). Then, he compares $Q_u(t)$ to the conditional expected queue content at the second station (here S_d) and decides either to enter into the first station or to postpone his refueling and enter into the second one.

Returning to our present model, we assume from now on that an incoming agent decides:

(a) either to enter S_u whenever $Q_u(t)$ strictly stays below a threshold value N^* (i.e. when $Q_u(t) < N^*$),
(b) or to enter into S_d otherwise.

At the DN's n_u and n_d, the routing rules depend, as in Section 2, on a patience parameter P which is again assumed to be common to all agents. Assume that the patience P and the threshold control parameter N^* are adjusted as

$$P \geqslant \frac{N^* + 1 + \delta}{\mu_u}, \quad \delta \in \mathbb{N}^+, \tag{4}$$

where δ denotes a tolerance level above the expected sojourn time. We can interpret (further details are given below) the routing decision at n_u to be a formal illustration of the Maister (1985) first principle of the psychology of waiting lines, namely: "*satisfaction equals perception minus expectation*". Indeed, at the DN n_e, the level N^* defines, via P as given by Eq. (4), an expected admissible sojourn time. Later, when reaching n_u, each agent compares his measured sojourn time (playing the role of the *perceived* sojourn time) with P (playing the role of the *expected* sojourn time) and then takes his routing decision.

Consider first the deterministic dynamics where S_u operates with a fixed service time $1/\mu_u$. When, at a given time t, $Q_u(t) = N^*$ agents are waiting in front of S_u, they will remain loyal to S_u forever (i.e. these agents will loop forever and ever). Indeed, their measured sojourn time W never exceeds P and, the dynamics being deterministic, no perturbation will alter this dynamically "frozen" situation. As a result, once $Q_u(t) \equiv N^*$, the server S_u remains definitively unavailable for any external incomer and the global incoming traffic with rate Λ is entirely dispatched to S_d. Whenever $\Lambda/\mu_d > 1$, the queueing system will thus be unstable (i.e. $\lim_{t\to\infty} Q_d(t) = \infty$). Assume now that random fluctuations affect the service times of S_u. While Eq. (4) is still satisfied on average, service time noise triggers, at node n_u, a random flow of unsatisfied customers, who will definitively leave the system. Hence, with the presence of noise (in the service time), the availability (for external traffic) of S_u effectively increases — remember that this availability is null in absence of noise. Consequently, part of the global incoming traffic is now processed by S_u. For a selected range of control parameters, we may simultaneously have

$$\rho_u = \frac{\alpha\Lambda}{\mu_u} < 1 \quad \text{and} \quad \rho_d = \frac{(1-\alpha)\Lambda}{\mu_d} < 1, \quad 0 \leqslant \alpha \leqslant 1, \tag{5}$$

where $\alpha\Lambda$ and $(1-\alpha)\Lambda$ stand for the stationary average rates of the partial traffic flows feeding S_u and S_d, ($\alpha \equiv 0$ corresponds to the purely deterministic case considered before). Whenever Eq. (5) holds, both queueing branches are dynamically stable. The previous qualitative reasoning

suggests that it exists a critical variance $\sigma^2_{u,c}$ of the service times of S_u (and hence a critical value α_c) such that:

(a) for $\sigma^2_u \geqslant \sigma^2_{u,c}$, the queueing system is stable,
(b) for $\sigma^2_u < \sigma^2_{u,c}$, the queueing system is unstable.

Hence, we can speak here of a *noise-induced stabilization of the dynamics*, which is studied below in more details both experimentally and analytically.

3.2.1. Experimental observations

The above dynamical behavior can be explicitly observed in simulation experiments where the incoming flow of customers is an exponential process with parameter Λ and the S_u service times are drawn from a probability density $dB_u(x)$ being:

(i) uniform with support $[(1/\mu_u)-\xi,(1/\mu_u)+\xi]$ with $\xi \geqslant 0$ (thus $\sigma^2_u = \xi^2/3$). The following numerical values were used: $\Lambda = 1.11$, $1/\mu_u = 1/\mu_d = 1$, $N^* = 28$ and $P = 30$ (i.e. $\delta = 1$ in Eq. (4)). We observe that for $\xi \geqslant 0.118 \Rightarrow \sigma^2_{u,c} \geqslant 0.0046$, the queueing system remains stable, while it becomes unstable (i.e. $\lim_{t\to\infty} Q_d(t) = \infty$) for smaller values of ξ.
(ii) a normal law $\mathcal{N}(1/\mu_u, \sigma^2_u)$. For the same numerical values as above, we observe that for $\sigma^2_u \geqslant \sigma^2_{u,c} = 0.0046$, the queueing system remains stable, while it becomes unstable for $\sigma^2_u < \sigma^2_{u,c}$.

3.2.2. Analytical approach

To analytically discuss the stability issue, let us consider the situation where the service times of S_u are independent Bernoulli random variables with values $\{1/\mu_u, 1/\mu^+\}$ and corresponding probabilities $(1-q)$ and q respectively, $0 \leqslant q \ll 1$. We assume that $\mu^+ < \mu_u$ and interpret $1/\mu^+$ (with $1/\mu^+ > 1/\mu_u$) as the effective service time occurring when a failure alters the ordinary behavior of the server S_u. Remember that the agents follow the FIFO rule and are homogenous in their patience parameter P, chosen here to fulfill

$$P < \frac{N^*}{\mu_u} + \frac{1}{\mu^+} \quad \text{and} \quad P > \frac{N^*+1}{\mu_u}, \tag{6}$$

where the second expression is actually Eq. (4) with $\delta = 1$. When, at a given time t, $Q_u(t) \equiv N^* - 1$, an incoming tagged customer $\mathscr{C}$ at DN n_e will decide to enter S_u. Later on, when $\mathscr{C}$ reaches n_u, he will, according to Eq. (6), choose:

(a) either to follow the feedback loop, whenever no failure occurred during the service of the N^* customers who were directly in front of him (including the customer who was served when $\mathscr{C}$ joined $Q_u(t)$) and during his own service,
(b) or to leave the system, whenever one or more failures occurred during the service of the N^* customers who were directly in front of him and during his own service.

Hence, in absence of failures and when $Q_u(t) \equiv N^*$, the agents will remain in the feedback loop forever and, at DN's n_e and n_u, neither an externally new incomer nor a leaving customer will be observed. However, as soon as failures occur in S_u, Eq. (6) implies that one or more customers will definitively leave the system after the decision at n_u. Hence, this implies that the global incoming traffic will now be shared between S_u and S_d. Assume that

$$\mu_d < \Lambda \iff \rho_d = \frac{\Lambda}{\mu_d} > 1. \tag{7}$$

Thus, S_d cannot sustain alone the full traffic load without being in an unstable regime ($\rho_d > 1 \Rightarrow \lim_{t\to\infty} Q_d(t) = \infty$). Remember that $\alpha\Lambda$ and $(1-\alpha)\Lambda$ denote the rates of the average partial traffics processed by S_u and S_d, respectively. It exists a critical incoming flow, defined by $(1-\alpha_c)\Lambda$, above which the queue $Q_d(t)$ becomes unstable. For the associated traffic intensities, this implies that

$$\rho_u = \frac{\alpha\Lambda}{\mu_u} < 1 \quad \text{and} \quad \rho_d = \frac{(1-\alpha)\Lambda}{\mu_d} < 1, \quad \forall \alpha > \alpha_c, \tag{8}$$

$$\rho_{d,c} = \frac{(1-\alpha_c)\Lambda}{\mu_d} = 1, \tag{9}$$

where $\rho_{d,c}$ is the critical traffic load driving the queue $Q_d(t)$ to its marginal stability regime.

To proceed further with analytical considerations, let us now focus on rare events regimes (RER), for which more than a single failure during N^*+1 consecutive ordinary services is a highly improbable event. As N^* is the threshold value governing the decision at node n_e and P fulfills Eq. (6), the RER is expected when $N^*+1 \ll (1/q)$. Under the RER, each failure triggers the drainage of the queue $Q_u(t)$. Indeed, due to the FIFO scheduling rule, when a failure occurs at time t, the last agent in $Q_u(t)$ will experience a sojourn time larger than P when arriving at n_u. So will also do the N^*-1 agents directly lining behind him (i.e. these are the loyal customers traveling in the loop and feeding $Q_u(t')$ for $t' > t$). As it has been discussed in Section 2, this produces a *siphon avalanche*, here of size N^*. In the RER, the succession of these siphon events will be approximately uncorrelated. Hence, in the stationary regime, we can simply estimate the outgoing flow rate λ_u at DN n_u as being given by

$$\lambda_u = \text{Prob}\{\textit{a single failure occurs}\} N^* \mu_u = q N^* \mu_u. \tag{10}$$

When Eq. (10) holds, the partial traffic on S_d is given by

$$\rho_d = \frac{\lambda_d}{\mu_d} = \frac{\Lambda - \lambda_u}{\mu_d} = \frac{\Lambda - qN^*\mu_u}{\mu_d}. \tag{11}$$

The marginal stability of queue $Q_d(t)$ is attained at the critical traffic $\rho_d = \rho_{d,c} = 1$, which implies

$$q \geqslant q_c := \frac{\Lambda - \mu_d}{N^*\mu_u}. \tag{12}$$

In terms of α_c, we can write

$$\alpha_c = 1 - \frac{\mu_d}{\Lambda}. \tag{13}$$

Finally, we can also express the stability condition given by Eq. (12) in terms of the critical variance $\sigma^2_{u,c}$ of the

Table 1
Stability conditions obtained when using a discrete events simulator with the following parameters: $N^* = 28$, $1/\mu_d = 1/\mu_u = 1$, $1/\mu_+ = 3$ and $P = 30$.

Global incoming traffic Λ	Simulated stability condition on q	Simulated stability condition on σ_u^2
1.05	0.0017	0.00075
1.1	0.0034	0.0015

No discrepancy between simulated and theoretical results have been observed up to the shown precision.

underlying Bernoulli random variable. We obtain

$$\sigma_u^2 \geqslant \sigma_{u,c}^2 = q_c(1-q_c)\left(\frac{1}{\mu^+} - \frac{1}{\mu_u}\right)^2. \tag{14}$$

The numerical experiments reported in Table 1 are in perfect agreement with Eqs. (12) to (14).

While the concept of stabilization by noise is currently discussed in the context of stochastic differential equations (Arnold et al., 1983; Has'minskiĭ, 1980; Ruszczynski and Kish, 2000), the present class of models exemplifies clearly that such a random stabilization can be encountered in multi-agent systems where a non-linearity (in our case, the feedback loop) is present.

3.3. Flow dispatching based on fully observable queues — synchronization of oscillations

Here, we assume that both queues $Q_u(t)$ and $Q_d(t)$ can be observed simultaneously by the incoming agents. Thus, compared with Section 3.2, the information gathering process has been further increased. Based on the queue contents, several dispatching policies at the DN n_e can be constructed. Among the simplest and most natural rules, let us here focus on the policy sending a new externally incoming customer to the shortest observed queue. This *Shortest-Queue-First* (SQF) rule implies the natural emergence, for large common patience parameter P, of *synchronized stable temporal oscillations* of the queue contents $Q_u(t)$ and $Q_d(t)$. This happens for any initial conditions of the queue populations. As before, when P is large and common to all agents, a purely deterministic approach is perfectly suitable. We assume that $(\Lambda/(\mu_u + \mu_d)) < 1$ to ensure the stability of the system. Let us consider, without loss of generality, that $(1/\mu_u) \geqslant (1/\mu_d)$. The two following cases may arise:

(1) *Non-generic case*: two identical servers (i.e. $1/\mu_u = 1/\mu_d$).
The total incoming traffic is evenly divided between the two servers, both receiving a partial traffic with rate $\Lambda/2$. The amplitude and period of the common synchronized stable temporal oscillations of the queue contents $Q_u(t)$ and $Q_d(t)$ are given by Eqs. (2) and (3) with parameters $\Lambda/2$ and μ.
(2) *Generic case*: two servers with service rate ratio $(1/\mu_u) > (1/\mu_d)$.
Even though the servers do not work at the same speed, the queue contents $Q_u(t)$ and $Q_d(t)$ are equal at any time, provided $\Lambda/\mu_d > 1$ (i.e. provided S_d is not able to handle alone the total incoming flow). The greater speed of S_d implies that the customers joining this server will remain satisfied for a longer queue length than with S_u. As a consequence of the SQF rule, there will be more unsatisfied customers with server S_u and this server will thus process a greater part of the global incoming traffic than S_d (i.e. S_u will absorb more fresh customers, but these customers will stay less time in the system than those joining S_d). As shown in Fig. 6, two distinct dynamics may emerge depending on the arrival and service rates.

4. Competing services in closed market contexts

Consider the closed network sketched in Fig. 7, formed by two feedback queue models as discussed in Section 2. The servers S_k $(k = 1,2)$ composing the network have an average service time $1/\mu_k$, where μ_k $(k = 1,2)$ stand for the service rates. Without loss of generality, we assume that $\mu_1 \leqslant \mu_2$. The total number $N \in \mathbb{N}^+$ of agents circulating in the network is fixed and we allow the capacities C_k $(k = 1,2)$ of both queues to be large enough to accommodate the entire population (i.e. one assumes that $C_k \geqslant N$ for $k = 1,2$). Directly inspired from Section 2, each circulating customer is equipped with a clock and monitors its total waiting time spent to receive service (i.e. its sojourn time) — the clock is reset to $t = 0$ each time a customer enters into a queue, the clock time value T_k is obtained when reaching a node n_k $(k = 1,2)$. The measured value T_k of each customer is then compared with a fixed and common to all customers patience parameter P. Thanks to the time monitoring, the following HB routing rule $\mathscr{R}$ can be implemented:

$$\mathscr{R} = \begin{cases} \text{Go to the feedback loop and} \\ \quad \text{hence go to server } S_k & \text{if } T_k \leqslant P, \\ \text{Avoid the feedback loop and} \\ \quad \text{hence go to server } S_{\lceil k+1 \rceil} & \text{if } T_k > P, \end{cases} \tag{15}$$

with the notation

$$S_{\lceil k+1 \rceil} = \begin{cases} S_1 & \text{if } k = 2, \\ S_2 & \text{if } k = 1. \end{cases}$$

Writing $N_1(t)$ and $N_2(t) = N - N_1(t)$ for the number of customers (including the one being served) waiting respectively in Q_1 and Q_2 and using, as in Section 2, a fluid queueing picture to describe the population dynamics, we can write

$$N_1(t) = N_1(0) + \int_0^t [\mu_2 \mathbb{I}(W_2(s) \geqslant P)\mathbb{I}(N_1(s) < N) - \mu_1 \mathbb{I}(W_1(s) \geqslant P)\mathbb{I}(N_1(s) > 0)]\, ds, \tag{16}$$

where $W_k(s)$ is the sojourn time of customers lining in queue Q_k. The function $\mathbb{I}\{E\}$ is the indicator of the event E (i.e. $\mathbb{I}\{E\} \equiv 1$ when the event $\{E\}$ is realized and 0 otherwise). We will assume, from now on, that $P \in [P_{min}, P_{max}]$ with P_{min} being large enough to safely allow,

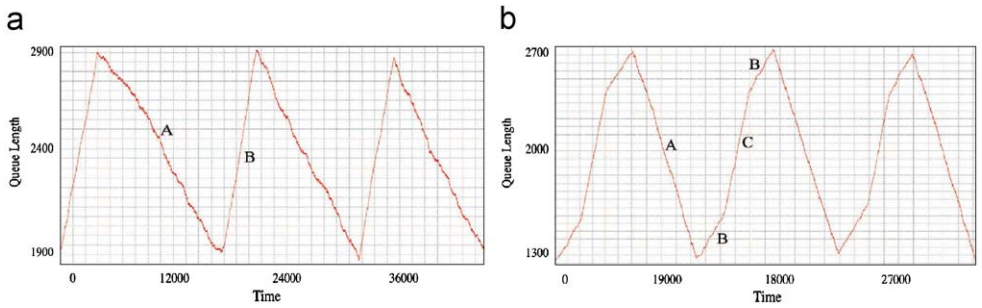

Fig. 6. The SQF policy implies that $0 \leqslant |Q_u(t) - Q_d(t)| \leqslant 1$, $\forall t$. Here, we only show the state of $Q_u(t)$ in the figures. (a) Queue content $Q_u(t)$ when $P = 2500$ and the inter-arrival and service processes are exponential with parameters $\Lambda = 1.25$, $1/\mu_u = 1.6$ and $1/\mu_d = 1.2$. The amplitude and period of the common synchronized stable temporal oscillations are given by $\Delta = P\mu_d/2$ and $\Pi = P((\mu_d/(\mu_d + \mu_u - \Lambda)) + (\mu_d/(\Lambda - \mu_u)))$, respectively. The two different slopes are given by $\mathbf{A} = ((\Lambda - \mu_u - \mu_d)/2)$ and $\mathbf{B} = ((\Lambda - \mu_u)/2)$. (b) Queue content $Q_u(t)$ when $P = 2500$ and the inter-arrival and service processes are exponential with parameters $\Lambda = 0.9$, $1/\mu_u = 1.6$ and $1/\mu_d = 1.2$. The dynamics differs from case (a) by the presence of a time interval with slope $\mathbf{C} = \Lambda/2$. During this interval, customers in S_u and S_d are all satisfied. On the other hand, during the time intervals with slope **A** and **B**, the customers in S_u are unsatisfied (the customers in S_d being unsatisfied only during the interval with slope **A**). For instance, in the configuration (a), all the customers joining S_u are unsatisfied, because $Q_u(t)$ always remains above the critical threshold. The complexity of the dynamics in case (b) requires more involved computations, which precludes to give simple and compact expressions for the amplitude and the period of the synchronized oscillations. However, due to the deterministic nature of the dynamics (when P is large), an analytical characterization is still feasible.

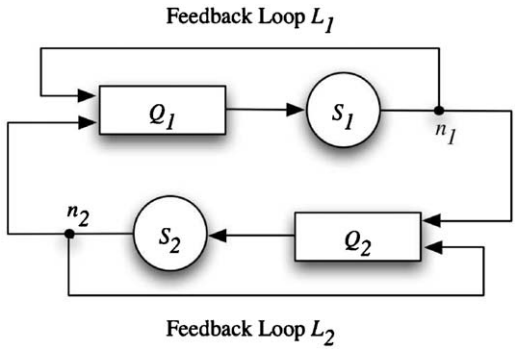

Fig. 7. Closed network of two servers with feedback loop.

as in Section 2, a deterministic analysis (thanks to the law of large numbers). In addition we impose that

$$\mu_1 P_{max} < N \quad \Rightarrow \quad N_1(0) \leqslant N \leqslant \mu_1 P_{max}, \tag{17}$$

which, in view of the rule $\mathscr{R}$ stated in Eq. (15), rules out the trivial situations occurring when all the customers are systematically satisfied and therefore stay loyal to their initial server.

The total number of customers being fixed, the dynamical state of the system is entirely determined by the single variable $N_1(t)$. While, for the deterministic evolution, a complete analytical characterization of the dynamics is possible (Labouchère, 2005), we present here only the most relevant qualitative features exhibited by this dynamical system. Four separated regimes, summarized in Fig. 8, can be characterized by the initial queue content $N_1(0)$

- $\mathscr{N}$*-regime*: $N_1(0) \geqslant \mu_1 P$ and $N_1(0) > (N - \mu_2 P)$.
  Starting at time $t = 0$ with $N_1(0) \geqslant \mu_1 P$ customers lining in Q_1, we conclude that after time $t = P$, customers reaching the node n_1 will be unsatisfied and therefore leave to populate Q_2. To study the role played by the second condition: $N_1(0) > (N - \mu_2 P) \Leftrightarrow (N - N_1(0)) = N_2 < \mu_2 P$, two sub-cases have to be examined:

  (1) *Generic case*: $\mu_1 < \mu_2$.
    (a) When $\mu_2 P > N$, the server S_2 alone is able to accommodate, with satisfaction, all customers. Consider the situation where $N_1(0)$ customers initially populate Q_1. As $N_1(0) \geqslant \mu_1 P$, the first $\mu_1 P$ customers reaching the node n_1 will be satisfied and therefore return to line again in Q_1. The remaining $N_1(0) - \mu_1 P$ customers, being unsatisfied, go to line in Q_2. At their second visit to S_1, the $\mu_1 P$ customers initially satisfied will, when reaching n_1 for the second time, be unsatisfied. Indeed, they did effectively wait $P + (N_1(0)/\mu_1)$ in Q_1 to receive their second service. This mechanism implies that ultimately all customers leave the first node and populate Q_2 and stay there forever. This behavior can be used to mimic how a performant service can ultimately monopolize an entire market sector.
    (b) In the case where $\mu_2 P \leqslant N$, the server S_2 alone is not able to accommodate, with satisfaction, all customers. Hence, temporal oscillations of the queue contents will be sustained.

  (2) *Non-generic case*: $\mu_1 = \mu_2$.
    In this case, none of the servers is able to accommodate, with satisfaction, the entire population. Hence, only oscillating regimes are generated.
- $\mathscr{W}$*-regime*: $N_1(0) \geqslant \mu_1 P$ and $N_1(0) \leqslant (N - \mu_2 P)$.
  Starting with $N_1(0) \geqslant \mu_1 P$ customers in Q_1 implies that at time $t = P$, customers leave Q_1 to enter into Q_2. Again, to study the role played by the second condition: $N_1(0) \leqslant (N - \mu_2 P) \Leftrightarrow (N - N_1(0)) = N_2 \geqslant \mu_2 P$, two sub-cases have to be examined:

  (1) *Generic case*: $\mu_1 < \mu_2$.
    The second condition stated above implies similarly that, at time $t = P$, customers leave Q_2 to enter into Q_1. Hence, unsatisfied customers are systematically generated.

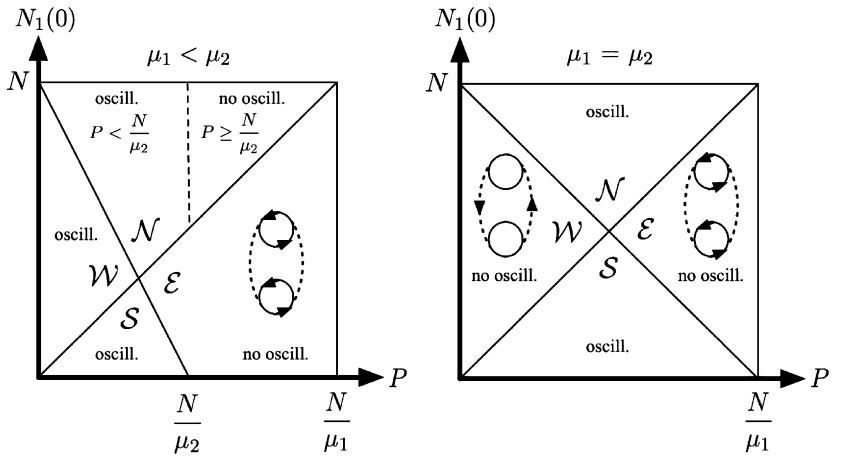

Fig. 8. Summary of the different regimes obtained in function of the initial condition and the patience parameter.

(2) *Non-generic case*: $\mu_1 = \mu_2$.
Here, one can show that the queue contents remain constant, as customers travel from Q_1 to Q_2 in a similar way, exactly as they would do in a closed tandem fluid queue without feedback.

- $\mathcal{S}$*-regime*: $N_1(0) < \mu_1 P$ and $N_1(0) \leqslant (N - \mu_2 P)$.
Starting with $N_1(0) < \mu_1 P$ implies that customers initially in Q_1 are satisfied and stay in that queue. The second condition: $N_1(0) \leqslant (N - \mu_2 P) \Leftrightarrow (N - N_1(0)) = N_2 \geqslant \mu_2 P$ implies that, for $t = P$, customers in Q_2 are unsatisfied and therefore leave to populate Q_1. The discussion of this case is very similar to the $\mathcal{N}$-regime. Here, however, due to the fact that $\mu_1 < \mu_2$, the server S_2 will never be able to attract the entire market and therefore, only oscillating behaviors of the queue contents are observable.
- $\mathcal{E}$*-regime*: $N_1(0) < \mu_1 P$ and $N_1(0) > (N - \mu_2 P)$.
Starting with $N_1(0) < \mu_1 P$ implies that customers initially in Q_1 are satisfied and stay in Q_1. The second condition: $N_1(0) > (N - \mu_2 P) \Leftrightarrow (N - N_1(0)) = N_2 < \mu_2 P$ implies similarly that customers are satisfied and therefore stay in Q_2. In this regime, no exchange of customers is observed and the system effectively behaves as if it were formed by two independent servers.

Refined analytical results and the rigorous proofs of the above results rely on taking into account the ability for both servers to be able to accommodate, with satisfaction, the entire customers population. This can be discussed by introducing the parameter $\Delta = N - (\mu_1 + \mu_2)P$. For $\Delta > 0$, a systematic flow of unsatisfied customers will always be generated. Conversely, when $\Delta < 0$, distinct regimes from those previously discussed are triggered, depending on the initial conditions (this explicit ergodicity violation is due to the intrinsic non-Markovian property of the underlying dynamics). This competing servers dynamics illustrates explicitly how rich spatio-temporal structures can be generated by HB routing in QNs.

5. Conclusion and perspectives

The main stream in queueing networks (QNs) is to consider the circulation of classes of items sharing all the same attributes. In this context, a single item is fully representative of all the members belonging to its class. Our present paper definitely differs from such a classical point of view and considers the flow dynamics in networks where each circulating item is an autonomous agent able to adapt its routing according to historical data monitored during its past journey through the network. Accordingly, a single circulating item is not a copy of the others and the resulting dynamics is not covered by the ordinary tools of QNs theory. Global dynamics of interacting autonomous agents explicitly belong to the vast realm of complex systems, for which collective properties emerge from the individual "intelligence" endowed to each agent. The underlying history-based routing decision mechanism, considered in our paper, violates the basic hypothesis of classical queueing models. Hence, QNs theory involving autonomous agents can be viewed as a new topic in itself. While the generic character and the synergetic modeling potential offered by agent systems dynamics have already been abundantly explored in basic sciences (physics, chemistry, biology) and in social sciences (economics, finance, psychology, car traffic), it presently remains largely open for investigations in production and service QNs. With the sophistication of production and supply chain facilities, decentralized management methods relying on autonomous entities imposes itself as a natural way to explore. In this context, the increasing availability of RFID (Radio Frequency Identification Devices) technology offers the possibility for a wide implementation of such local intelligence to circulating items in production and service networks. In particular, flow control of flexible production systems with complex structures is gradually decentralized to "intelligent" pallets (i.e. carrying units). Ideally, such devices should be able, in real-time, to select autonomously, according to ad-hoc historical production data

and real-time observations, the best possible routing alternatives. Obviously, the ultimate goal of production and services managers will be to determine the efficient compromise between pure interventionism (due to centralized controls) and self-organization (due to the swarm intelligence of the agents).

Acknowledgment

This work is partially supported by the FNS (Fonds National Suisse pour la Recherche) under Grant no. 200020-117608/1.

References

Arnold, L., Crauel, H., Wihstutz, V., 1983. Stabilization of linear systems by noise. SIAM Journal on Control and Optimization 21 (3), 451–461.
D'Avignon, G.R., Disney, R.L., 1976. Single-server queues with state-dependent feedback. INFOR Journal 14 (1), 71–85.
Baldwin, R.O., Davis IV, N.J., Kobza, J.E., Midkiff, S.F., 2000. Real-time queueing theory: a tutorial presentation with an admission control application. Queueing Systems 35, 1–21.
Bielen, F., Demoulin, N., 2007. Waiting time influence on the satisfaction-loyalty relationship in services. Managing Service Quality 17 (2), 174–193.
Chen, H., Yao, D.D., 2001. Fundamentals of Queueing Networks. Springer, Berlin.
Doytchinov, B., Lehoczky, J., Shreve, S., 2001. Real-time queues in heavy traffic with earliest-deadline-first queue discipline. The Annals of Applied Probability 11 (2), 332–378.
Erramilli, A., Forys, L.J., 1991. Oscillations and chaos in a flow model of switching system. IEEE Journal on Selected Area in Communications 9 (2), 171–178.
Filliger, R., Hongler, M.-O., 2005. Syphon dynamics—a soluble model of multi-agents cooperative behavior. Europhysics Letters 70 (3), 285–291.
Gallay, O., Hongler, M.-O., 2008. Weariness and loyalty loss in recurrent service models. In: Proceedings of MOSIM'08, Paris.
Hassin, R., 1996. On the advantage of being the first server. Management Science 42 (4), 618–623.
Has'minskiĭ, R.Z., 1980. Stochastic Stability of Differential Equations. Sijthoff & Noordhoff, Alphen aan den Rijn.
Haxholdt, C., Larsen, E.R., van Ackere, A., 2003. Mode locking and chaos in a deterministic queueing model with feedback. Management Science 49 (6), 816–830.
Hongler, M.-O., Cheihkrouhou, N., Glardon, R., 2004. An elementary model for customer fidelity. In: Proceedings of MOSIM-04, Nantes (France), Lavoisier Editions, 2, pp. 899–906.
Hülsmann, M., Grapp, J., Li, Y., 2008. Strategic adaptivity in global supply chains—competitive advantage by autonomous cooperation. International Journal of Production Economics 114, 14–26.
Kataoka, T., Kawamura, H., Kurumatani, K., Ohuchi, A., 2005. Distributed visitors coordination system in Theme Park problem. In: MMAS 2004, Lecture Notes in Artificial Intelligence, vol. 3446, pp. 335–348.
Kawamura, H., Kataoka, T., Kurumatani, K., Ohuchi, A., 2004. Investigation of global performance affected by congestion avoiding behavior in Theme Park problem. IEEJ Transactions EIS 124 (10), 1922–1929.
Kawamura, H., Kurumatani, K., Ohuchi, A., 2004. Modeling of Theme Park problem with multiagent for mass user support. In: Lecture Notes in Computer Science, vol. 3012, pp. 48–69.
Klein, M., Metzler, R., Bar-Yam, Y., 2005. Handling emergent resource use oscillations. IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans 35 (3), 327–336.
Kumara, S.R.T., Ranjan, P., Surana, A., Narayanan, V., 2003. Decision making in logistics: a chaos theory based analysis. Annals of the CIRP 52 (1), 381–384.
Labouchère, G., 2005. Comportement d'auto-organisation de la dynamique des clients en marché clos induit par des décisions basées sur leur historique. Diploma Work, EPFL.
Law, A.K.Y., Hui, Y.V., Zhao, X., 2004. Modeling repurchase frequency and customer satisfaction for fast food outlets. International Journal of Quality & Reliability Management 21 (5), 545–563.
Lehoczky, J.P., 1997. Using real-time queueing theory to control lateness in real-time systems. Performance Evaluation Review 25 (1), 158–168.
Maister, D.H., 1985. The psychology of waiting lines. In: Czepiel, J.A., Solomon, M.R., Surprenant, D.F. (Eds.), The Service Encounter. D.C. Heath, Lexington, Mass (Chapter 8).
Massotte, P., Bataille, R., 2000. Future production systems: influence of self-organization on approaches to quality engineering. International Journal of Production Economics 64, 359–377.
Peköz, E.A., Joglekar, N., 2002. Poisson traffic flow in a general feedback queue. Journal of Applied Probability 39 (3), 630–636.
Pullman, M., Thompson, G.M., 2002. Evaluating capacity—and demand—management decisions at a ski resort. Cornell Hotel & Restaurant Administration Quarterly 43 (6), 25–36.
Ruszczynski, P.S., Kish, L.B., 2000. Noise enhanced efficiency of ordered traffic. Physics Letters A 276, 187–190.
Serfozo, R., 1999. Introduction to Stochastic Networks. Springer, Berlin.
Surana, A., Kumara, S., Greaves, M., Raghavan, U.N., 2005. Supply-chain networks: a complex adaptive systems perspective. International Journal of Production Research 43 (20), 4235–4265.
Takàcs, L., 1963. A single-server queue with feedback. The Bell System Technical Journal 42, 505–519.
van Ackere, A., Larsen, E.R., 2004. Self-organising behaviour in the presence of negative externalities: a conceptual model of commuter choice. European Journal of Operational Research 157, 501–513.
van Ackere, A., Haxholdt, C., Larsen, E.R., 2006. Long-term and short-term customer reaction: a two-stage queueing approach. System Dynamics Review 22 (4), 349–369.
Walrand, J., 1988. An Introduction to Queueing Networks. Englewood Cliffs.
Whitby, S., Parker, D., Tobias, A., 2001. Non-linear dynamics of duopolistic competition: a R and D model and simulation. Journal of Business Research 51, 179–191.
Wycisk, C., McKelvey, B., Hülsmann, M., 2008. "Smart parts" supply networks as complex adaptive systems: analysis and implications. International Journal of Physical Distribution & Logistics Management 38 (2), 108–125.