
Mathematical Modelling and Analysis
Volume 25, Issue 4, 505–521, 2020
https://doi.org/10.3846/mma.2020.10801

http://mma.vgtu.lt
ISSN: 1392-6292
eISSN: 1648-3510

A Biologically Inspired Fluid Model of the Cyclic Service System

Aliya Kantarbayeva[a] **and Almaz Mustafin**[b]

[a] *Al-Farabi Kazakh National University*
Al-Farabi Ave. 71, 050040 Almaty, Kazakhstan
[b] *Satbayev University*
Satbayev St. 22, 050013 Almaty, Kazakhstan
E-mail(*corresp.*): a.mustafin@satbayev.university
E-mail: kantarbayeva.aliya@kaznu.kz

Received July 30, 2019; revised June 15, 2020; accepted June 16, 2020

Abstract. A deterministic fluid model in the form of nonlinear ordinary differential equations is developed to provide the description for a multichannel service system with service-in-random-order queue discipline, abandonment and re-entry, where servers are treated like enzyme molecules. The parametric analysis of the model's fixed point is given, particularly, how the arrival rate of new customers affects the steady-state demand. It is also shown that the model implies a saturating clearing function (yield vs. demand) of the Karmarkar type providing the mean service time is much shorter than the characteristic waiting time.

Keywords: fluid queues, multiple server, abandonment, re-entry, random order service, clearing function.

AMS Subject Classification: 90B22.

1 Introduction

Historically, the service sector was long considered a sort of marginal economic activity that did not fit into agriculture and manufacturing categories. However from 1980s on its share in most economies has steadily expanded. By 2020, the service sector contributes to around two thirds of the total global GDP, whereas for the most industrialized economies this indicator exceeds three fourth [34].

The rapid development of the service sector in many economies has stimulated a renewed interest in multi-server queueing models. These models are particularly important in large-scale service systems, such as customer contact

centers and health-care centers [4, 28, 35, 36]. Operations research, of which queueing theory is a branch, treats act of service and production (manufacturing) uniformly, from cybernetics perspective, as a kind of an input–output transformation process going on in an imaginary converter box [8]:

$$\text{Inputs} \longrightarrow \boxed{\text{Converter}} \longrightarrow \text{Outputs} \tag{1.1}$$

In the case of industrial production, the inputs are physical substances, such as raw materials or parts, and they are transformed by workers and machines in the factory shop into a finished commodity (output).

In the case of service, the inputs and outputs involved may not necessarily be unanimate material objects—they may appear to be people or information, and the nature of conversion may be less obvious. For service, it is not the physical nature of the input that is being transformed, but the condition (or state). Indeed, according to the widely adopted definition of service set forth by Hill [15], "A service may be defined as a change in the condition of a person, or a good belonging to some economic unit, which is brought about as a result of the activity of some other economic unit, with the prior agreement of the former person or economic unit." As opposed to a good that, once produced, acquires exteriority with regard to both manufacturer and potential consumer, services are consubstantial with provider and with customer: they cannot be stored, they are not a given outcome, but rather an act or a process [10].

Following Hill, Gadrey [9] coined the metaphorical term "service triangle", in terms of which he proposed the following wording of service: "A service activity is an operation intended to bring about a change of state in a reality C that is owned or used by consumer B, the change being effected by service provider A at the request of B, and in many cases in collaboration with him or her, but without leading to the production of a good that can circulate in the economy independently of medium C."

Figure 1, adapted from [8], illustrates different service scenarios according to the varying involvement of persons and their belongings. Generally speak-

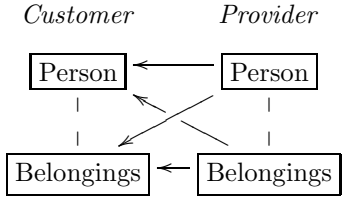

Figure 1. To the definition of service by Hill and his followers. After Fromm and Cardoso [8].

ing, a service provider has properly qualified personnel (employees with skills, competencies, and knowledge) and resources (like facilities, tools, and materials) to render a service. Five different situations are covered by the diagram on Figure 1:

(i) A person (provider) delivers a service to another person (customer) directly. No belongings are involved on either side. For example, a tutor gives private lessons to a student. In so doing, the tutor transforms less educated student (the input) into better educated student (the output);
(ii) A person equipped with belongings (provider) acts on another person (customer). For example, doctors use facilities (ward, bed, operating room), instruments (stethoscope, syringe, lancet), and materials (medicaments, plasters, dressings, disinfectants) to deliver their service to a patient. Thus, the clinic transforms ill patients (the input) into healthier patients (the output);
(iii) A person equipped with belongings (provider) acts on the belonging of a customer. The customer is not involved in the process. This situation is representative for dry cleaning, maintenance and repair services, for mail delivery and transportation of cargo. For example, the car mechanic transforms a disabled vehicle (the input) into a functional vehicle (the output);
(iv) The case of self-service, when an unmanned service is rendered by the provider. The provider typically sets up facilities or equipment that the customer can use to transform his/her condition. Examples are vending machines, phone-booths, and rental cars;
(v) Provider's belongings act on customer's belongings. For example, a program application installed on the customer's computer can request and get an update service from a remote computer program on the side of the provider.

When dealing with modeling or simulation of production and service systems, operations research mainly adheres to a black box approach to the internal structure of the "converter" shown in diagram (1.1). A black box approach allows describing the act of transformation with interface information that is externally visible to observers. Nonetheless, the internal details of the transformation are hidden. As a consequence, any original mechanism of conversion can be replaced with a proper phenomenological transfer function as long as it mimics an equivalent interface. Black box approach is often justified since the robust applied model should be able to describe a production or service without being overloaded with the full complexity of internal structures, processes, and implementations. The most well-known example of black box converter in operations research is the production function.

The production function is an econometric statement relating the rate of production of a certain finished, or partly finished, intermediate, commodity (output), Y, to a set of the involved factors of production, ϕ_1, $\phi_2, \ldots, \phi_n$ (e.g., [31]):

$$Y = F(\phi_1, \phi_2, \ldots, \phi_n). \tag{1.2}$$

In its broad meaning, factor of production is any entity which can lead to increased output as its availability is increased. Factors may be of material, energy, human and financial nature. They may be of two types: consumable and nonconsumable. By convention, consumable factors are referred to as inputs. We use term "resource" interchangeably with "input". By consumption we understand irreversible conversion, physical embodiment, of a resource into

a material product. As the production process takes place, the resource is consumed, used up. Consuming the resource means tending to reduce its availability. The nonconsumable factors of production, commonly known as primary factors, to which belong land, capital and labor, are often lumped together as "funds". They are not resources by the definition in use. Funds are not materially transformed into an output they produce. They are transforming tools that turn the involved inputs into a product, but are not themselves embodied physically in the product. Although funds are not used up, their amount can change and they are subject to wear-and-tear. In terms of dimensions, output Y in formula (1.2), being the quantity of the commodity produced in a unit of time, is a flow variable. Inputs, most commonly are flows, although in some cases they may be stocks. Funds always are stock variables.

A special variety of production function is the clearing function introduced in the supply chain management (e.g., [1]). It is widely used for production planning. The clearing function is a deterministic relation between the throughput of a steady-state production process, Y, and the work in progress (i.e. the total volume of partially finished goods in the manufacturing system awaiting completion), W: $Y = f(W)$. The input W in this formula is readily seen to be a stock variable.

The first aim of this work is to propose a plausible reconstruction of events taking place in the black box of service, understood as a transformation of customer's state by the service provider. In other words, we propose a minimal non-phenomenological model of the service process, that is not derived from a regression analysis, but rather appeals to some first principles.

We assume that change in the condition of a customer as a result of the interaction between the customer and the service provider proceeds in much the same way as does the conversion of the substrate into a different molecule in an enzyme-catalyzed biochemical reaction. Indeed, in a living cell, a substrate molecule, S, binds to an enzyme molecule, E, to form a short-lived substrate-enzyme complex, $[ES]$. The complex then breaks up into a product, P, and the original enzyme, which can then catalyze a new reaction (e.g., [7]):

$$\xrightarrow{r} S + E \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} [ES] \xrightarrow{k_2} P + E \tag{1.3}$$

In the scheme (1.3), as is customary, an arrow points the direction of the respective reaction. The double arrow in the first step denotes a reversible reaction. Arrow labels are the respective reaction rate constants.

Treating service as a kind of transformation process, one may identify incoming (potential) customer with substrate, outgoing (served) customer with product, and server (provider) with enzyme. As applied to production process, in just the similar way we associate resource with substrate, finished good with product, and fund (e.g., worker or machine) with enzyme. All aforesaid is summarized in Table 1.

Georgescu-Roegen was the first to suggest that a production factor such as labor is like a catalyst: "The assumption is that labor, while an indispensable factor in the production of any commodity, can be substituted by other factors beyond any limit. Or in other words, an output of any size can be obtained

Table 1. Biochemical catalysis vs. production and service.

	Agents with similar functions		
Transformation process	Input	Output	Converter
Enzymatic reaction	Substrate	Product	Enzyme
Manufacturing	Resource	Finished goods	Funds
Service	Potential customer	Served customer	Provider

by any industry with as little labor as we may wish, provided that nonlabor inputs are available in unlimited amounts. Since this property presents an obvious analogy with that of a catalyst in a chemical reaction, we propose to say that labor is catalytic in industry G_k if it has the mentioned property in that industry" [12, p. 319].

A few years later, Poletaev stated that "The typical prototypes of the process of production-consumption type are chemical reaction in the presence of a catalyst, and process of industrial production. In both cases, three kinds of components are involved: inputs (substrate, raw material), funds (catalyst, equipment), and outputs (product of reaction, commodity)" [22, p. 72, own translation].

Independently, Chernavskii accentuated the analogy between biosynthesis and industrial production: "Both by spirit and by methods of research, modeling the economic and production processes is closely related to the subject we expounded above. There is nothing surprising in that biological systems with their basic variables—concentrations of substances—are similar to economic ones, where variables are the quantities of products or commodities, and the role of enzyme concentration is played by the number of machines in a shop or automatic line. In this regard, both the kinetic models of biophysics and biochemistry, and the economic models belong to the common branch of cybernetics, the so-called theory of complex systems" [29, p. 134, own translation].

Regrettably, the above mentioned inspiring insights have gone barely noticed by contemporaries. The situation started to improve after recognition of the newly emerged field of evolutionary economics [27] and the concept of "industrial metabolism" [5]. Since then, the trend has been toward the growing awareness that chain of biochemical synthesis in a living cell resembles production line in an industrial factory, where products of one machine are used by other machines for manufacturing of their own products [2]. In recent years, advanced tools of the queueing theory become more and more relevant in the studies of intracellular metabolic networks [16, 25], research work in this direction being spoken of as "biologistics" [14].

However so far only sporadic attempts have been undertaken to unveil the black box of service process with the help of interdisciplinary approach. The existing rare works lack the attempts to recognize one essential common feature of enzyme and server: they both interact with the input being transformed to form a short-lived intermediate that subsequently releases the output. For example, Niyirora and Zhuang [28] do introduce the transient server–patient

complex in their continuous model of queue in an emergency department, but do not compare time dynamics of the servers with that of the patients whatsoever. However, the coexistence of strongly varying timescales in the process of conversion may have important implications as is known from biochemistry. This is what lies behind our motives to apply more consistently the "enzymatic" approach to service. The present research is partially based on the earlier work of the authors [26] on production networks.

We verify workability of the enzymatic model of service by the example of a cyclic service system. Within the steadily growing service center literature, the object of our research fits into the single customer class, single server type category of queueing systems with parallel multiple servers, queue discipline of service in random order, unlimited-size buffer, abandonment from queue, and re-entry. At present, queueing theory is held to be the principal mathematical tool to describe processes in waiting lines using the apparatus of probability theory (e.g., [17]). For a broad range of purposes, however, deterministic fluid models, originally intended to analyze the stability of an underlying queueing system [6], prove to be an effective alternative to discrete stochastic queueing models [3, 11]. They are an approximation technique for studying important dynamic characteristics of queueing networks. When the deterministic variability in the arrival and departure flows seems to dominate over the stochastic volatility, it may be safe to neglect the stochastic part of the model altogether. And if the number of customers in a system varies over a wide range, then the discrete nature of individual customers also might be ignored. This is similar to a passage from multiparticle system to the hydrodynamical limit of continuous medium. Discrete jobs are replaced by continuous fluid mass with scalar density field and vector velocity field, which follows the same routing as before. The resulting set of ordinary differential equations (ODEs) of mass balance is referred to as a fluid model. The solutions of fluid models are often handy to analyze as opposed to the corresponding queueing network equations.

The model of a service system suggested in the present paper likens servers to enzymes, which shift customers from the category of "potential" to the category of "served". This enables us to state the model in fluid form in terms of nonlinear ODEs of chemical kinetics. Assuming that the mean service time is much shorter than the typical waiting time makes possible to apply a quasi-steady-state approximation to the number of busy servers. This variable is shown to hastily adapt to the momentary demand following the famous Michaelis–Menten saturation curve of enzyme kinetics. As a result, we easily derive the clearing function of the system, which is shown to be the Karmarkar function [18] widely used in production planning. The proposed model allows for an explicit steady-state solution and complete analysis of its stability. We also analyze how the steady-state demand is affected by the arrival rate of new customers.

2 The model

Consider an open service system with a group of identical servers helping the customers (processing the incoming orders). Each of the servers can process

any of the arriving orders and we assume here that they do so one at a time. Potential customers arrive at the service system from the outside with a constant rate, and if all servers are busy processing jobs, the arriving customer has to join the queue. Would-be customers may be impatient and leave after a while, should they have to wait too long before a server is available. Those customers are lost (they will go to another company).

Potential customers waiting for service gather in a common buffer which feeds all servers. The waiting space is assumed unlimited. When a server finishes the processing of its current job, it randomly picks another customer to enter service from the queue. The probability for a waiting customer to obtain service depends on the number of customers waiting at the same time and is independent of how long the customer has already waited. This type of the service discipline is known as service in random order (SIRO). One example of SIRO is a system where the service discipline is processor sharing and the service times are independent and identically exponentially distributed. Processor sharing corresponds to the server equally dividing its service amongst all the customers in the queue. This is relevant in many computational settings. Another examples where SIRO may be encountered are call centers, or an air defense system. SIRO is not as popular scheduling discipline as the well-known first-come-first-served (FCFS) rule, however this fact rather reflects considerations of social justice than of economic performance. Indeed, in case of memoryless arrivals, independent and identically distributed service times and a single server, SIRO and FCFS both provide the same mean waiting time, given by the Pollaczek–Khintchine formula [32, pp. 256–259].

In our model the terms "inventory", "work in progress" (WIP) and "buffer stock" are regarded as synonyms and mean the current number of potential customers in the queue. In fact, this is the instantaneous demand. The population of potential customers does not include those in service (current customers).

Customer-server contact may not necessarily culminate in satisfaction of customer's individual demand, in other words completion of service. Occasionally, current customer may evince hesitation and decide to postpone the service, or may be turned down by a server (e.g., in the event that the order is submitted with irregular documentation). In both cases such a customer returns to the queue.

Once completely served, current customer turns to a past customer. After a while, past customers either leave the system, or replenish the ranks of potential customers over again. The latter happens when the change in the condition of a customer is temporary. For example, the obtained commodity is perishable, or represents durable goods which need replacement/repair because of their continuous use, or is attributable to obsolescence, or falls in the category of style-and-fashion goods, etc. Thus the queue actually consists of the mixture of external and internal arrivals.

The systematic study of cyclic queues has been initiated by Taylor and Jackson [33] and Koenigsberg [20,21]. Shortle and coauthors [32, ch. 5, pp. 213–254] review the most notable results obtained by now in this field of research with probabilistic approach. We opt for the fluid formalism. The issue of the accuracy of fluid approximation as applied to our "enzymatic" model of a

service system is touched upon in Section 4.

We can represent the totality of the events listed above in the form of pseudochemical equations presented by Figure 2. In the figure referred above,

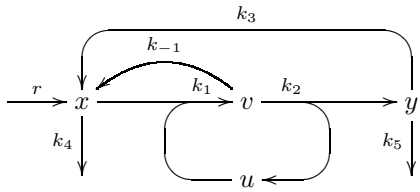

Figure 2. The modeled open cyclic service system.

x is the population of potential customers (in other words, queue length or demand), u is the population of free servers, v is the population of busy (operative) servers, and y is the population of successfully served (past) customers. The constants $r, k_{-1}, k_1, \ldots, k_5$ depict the various rates with which these processes proceed.

A flowchart such as the one given by Figure 2 encodes both the sequence of steps and the rates with which these steps occur. It can be translated into a set of ODEs that describe rates of change of stock quantities of the participating agents. To write corresponding equations, we naturally can choose to use what chemists call the "law of mass action" (e.g., [24, sect. 102, pp. 306–310]), which states that when two or more agents are involved in a conversion step, the rate of conversion is proportional to the product of their quantities. By convention, the mass-action rate constants are the proportionality constants. They are indicated in Figure (2) as arrow labels. Namely:

- r is the arrival rate of the customers;
- k_1 is the capture rate of a would-be customer by a server;
- k_{-1} is the frequency with which orders are being withdrawn or denied after beginning of service;
- k_2 is server's turnover number, or the maximum number of fulfilled orders that a single server is capable to deliver at one unit of time. Thereby, $1/(k_{-1} + k_2)$ is the mean lifetime of the complex "server–customer", or the mean processing time of a server, or simply service time;
- k_3 is the frequency with which customers rejoin the queue upon completion of service. Thus, $1/k_3$ is how long lasts "aftereffect" of a service;
- k_4 is the frequency of abandonments;
- k_5 is the out-migration rate of the past customers.

Keeping track of each participant allows us to derive the following set of equations:

$$\dot{x} = r + k_{-1}v + k_3 y - k_1 ux - k_4 x, \tag{2.1a}$$

$$\dot{y} = k_2 v - (k_3 + k_5)y, \tag{2.1b}$$

$$\dot{u} = (k_{-1} + k_2)v - k_1 ux, \tag{2.1c}$$

$$\dot{v} = k_1 ux - (k_{-1} + k_2)v, \tag{2.1d}$$

where overdots mean differentiation with respect to time t. All parameters in the model are nonnegative. (Notice that the model suggested in [28] is a special case of the system (2.1) with $k_{-1} = k_3 = k_5 = 0$.)

Adding equations (2.1c) and (2.1d) reveals a conserved quantity u_0, the total number of servers, free and busy:

$$u + v = u_0. \tag{2.2}$$

With the aid of (2.2) the system (2.1) can be reduced by eliminating either u, or v. We arbitrarily choose to eliminate u. Next we nondimensionalize the remaining three equations by changing over to the scaled variables $\tau = t/T$, $\xi = x/K$, $\eta = y/K$, and $\zeta = v/u_0$, where

$$T = 1/(k_1 u_0), \quad K = (k_{-1} + k_2)/k_1. \tag{2.3}$$

According to (2.3), T is chosen to be a new unit of time. It is a characteristic time a potential customer spends waiting (in the queue) before beginning service. This time is seen to be inversely proportional to the total number of installed servers, u_0. The motivation behind normalizing x and y by K will be provided in Section 3. As to v, replacing it with a quantity ζ scaled relative to u_0, so that $0 < \zeta < 1$, is quite natural.

The dimensionless fluid equations then become

$$\dot{\xi} = \rho + \alpha\eta + (1 - \mu)\zeta + \zeta\xi - (1 + \beta)\xi, \tag{2.4a}$$

$$\dot{\eta} = \mu\zeta - (\alpha + \gamma)\eta, \tag{2.4b}$$

$$\varepsilon\,\dot{\zeta} = \xi - \zeta\xi - \zeta, \tag{2.4c}$$

where overdots now mean differentiation with respect to τ. Quantities

$$\begin{aligned} &\alpha = k_3/(k_1 u_0), \ \ \beta = k_4/(k_1 u_0), \ \ \gamma = k_5/(k_1 u_0), \\ &\varepsilon = k_1 u_0/(k_{-1} + k_2), \ \mu = k_2/(k_{-1} + k_2), \ \rho = r/\left((k_{-1} + k_2)u_0\right) \end{aligned} \tag{2.5}$$

all are new dimensionless parameters. It is worth noting that $\mu < 1$. According to (2.5), ε is the ratio of the frequency with which a server and a customer associate to the frequency with which the complex server–customer dissociates. In other words, it is the ratio of the mean service (holding) time to the mean waiting time. From physical considerations, only nonnegative solutions of the model equations (2.4) make sense.

3 Results

When deriving the equations (2.4) no approximations have been made. However it is known [35] that the waiting times are often longer than the service times meaning our parameter ε should be assumed small. Inasmuch as $\varepsilon \ll 1$, the system (2.4) is singularly perturbed. The slow variables ξ and η are respective populations of potential and past customers, while the fast variable, ζ, is the number of busy servers. The standard practice of reducing such systems is multiple-scale analysis whereby fast variable is adiabatically eliminated. One

has to establish the validity of the adiabatic elimination in each specific case. As applied to (2.4), Fenichel–Tikhonov theorem requires, among other things, (i) fixed point $\bar{\zeta}(\xi, \eta)$ of the fast equation (2.4c) to be an isolated root of the algebraic equation $\zeta = 0$ and to retain stability at all allowed values of the slow variables ξ and η, and (ii) initial condition ζ_0 to fall within the domain of influence of $\bar{\zeta}$ for all initial values ξ_0 and η_0 [23, ch. 3, p. 53–70]. The influential works to clarify the applicability of the technique to enzymatic reactions have been carried out by Klonowski [19], and Segel and Slemrod [30].

To decompose system (2.4) into fast and slow parts, introduce fast time variable $\vartheta = \tau/\varepsilon$. Now rescale (2.4) by replacing τ with $\vartheta\varepsilon$ and, after taking $\varepsilon = 0$, it becomes

$$\xi' = 0, \quad \eta' = 0, \quad \zeta' = \xi - \zeta\xi - \zeta, \tag{3.1}$$

where 'primes' stand for differentiation with respect to ϑ. This is the fast subsystem, where ξ and η are replaced by their initial values and treated as parameters. It yields the inner solution, valid for $\tau = \mathcal{O}(\varepsilon)$.

Setting $\varepsilon = 0$ in (2.4) leads to the slow subsystem

$$\dot{\xi} = \rho + \alpha\eta + (1 - \mu)\zeta + \zeta\xi - (1 + \beta)\xi, \tag{3.2a}$$

$$\dot{\eta} = \mu\zeta - (\alpha + \gamma)\eta, \tag{3.2b}$$

$$0 = \xi - \zeta\xi - \zeta, \tag{3.2c}$$

which produces the outer solution, valid for $\tau = \mathcal{O}(1)$. In this singular limit, as $\varepsilon \to 0$, the subsystem defines a slow flow over the slow manifold given by algebraic equation (3.2c). Outer solution is valid for those values of ξ and η, for which the quasi-steady states of the fast subsystem (3.1) are stable.

The quasi-equilibrium for the fast subsystem (3.1) is given by $\zeta = \xi/(1 + \xi)$, and it is asymptotically stable for any positive ξ. Recall that ζ is the simultaneous fraction of busy servers. As long as the normalized demand keeps small, i.e. $\xi \ll 1$, this fraction remains adequately small, meaning the servers are strongly underloaded. However at high levels of demand, for $\xi \gg 1$, all the servers become busy. When $\xi = 1$, the number of busy servers will be one-half their total number. Therein lies the reason to consider the quantity $K = (k_{-1} + k_2)/k_1$ an apt normalization constant for the populations of customers. K is called Michaelis constant in biochemistry.

Hence, it follows from (3.2) that for time scales on the order of $\tau = \mathcal{O}(1)$ the dynamics of the service system (2.4) is governed by a pair of equations

$$\dot{\xi} = \rho + \alpha\eta - \mu\xi/(1 + \xi) - \beta\xi, \tag{3.3a}$$

$$\dot{\eta} = \mu\xi/(1 + \xi) - (\alpha + \gamma)\eta. \tag{3.3b}$$

The first term in the right-hand side of (3.3b) is the output of our service system, or how many customers are being successfully served by all available servers in a unit of time:

$$\begin{aligned} &\tilde{Y} = \mu\xi/(1 + \xi) && \text{(nondimensional form)}, \\ &Y = k_2 u_0 x/(K + x) && \text{(dimensional form)}. \end{aligned} \tag{3.4}$$

This is a clearing function since it depends on the number of potential customers waiting in the queue, i.e. on WIP. More precisely, this is a clearing function in the form originally introduced by Karmarkar [18]. However, as opposed to the formula empirically obtained by Karmarkar for supply chains, the argument ξ appearing in our formula does not have to be an equilibrium value of WIP. In deriving the clearing function we did not require the service system to operate in a steady-state mode. Nonetheless, the number of busy servers, ζ, being the fast variable, after a short transient of order $\mathcal{O}(\varepsilon)$ keeps in a quasi-steady state with respect to the current demand, ξ. In (3.3a), the arrival rate of potential customers, ρ, may not be necessarily constant, but if the timescale of its typical variations is much longer than the service time, then the expression (3.4) for the clearing function will remain valid.

One can recognize (3.4) as another version of the Michaelis–Menten equation of enzyme kinetics [7, ch. 2]. The hyperbolic input–output relationship of the type (3.4) is not uncommon in natural sciences: it describes a relationship between the number of active sites of the surface undergoing adsorption and pressure (Langmuir equation), the sigmoidal oxygen-binding curve of haemoglobin and the fraction of a macromolecule saturated by ligand as a function of the ligand concentration (Hill equation), growth rate of microorganisms in a nutrient solution (Monod equation), numerical response of predator to prey population density (Holling type II response), and the like.

A distinguishing characteristic of the equation (3.4) is saturated response of the output to the current demand. For low levels of demand, the output is roughly proportional to the demand. At high levels of the demand, though, the release of customers having received their service approaches a constant value.

System (3.3) has two steady states, of which only one always lies in the first quadrant of the $\xi\eta$ phase plane, in other words is physically meaningful:

$$\bar{\xi} = \left(\sqrt{B^2 + 4\beta(\alpha+\gamma)^2\rho} - B\right) / \left(2\beta(\alpha+\gamma)\right), \tag{3.5a}$$

$$\bar{\eta} = \left(B + 2(\alpha+\gamma)\rho - \sqrt{B^2 + 4\beta(\alpha+\gamma)^2\rho}\right) / \left(2\gamma(\alpha+\gamma)\right), \tag{3.5b}$$

where $B = (\alpha+\gamma)(\beta-\rho) + \gamma\mu$ is an auxiliary quantity introduced for the sake of simplicity. This steady state is always a stable node, because $\operatorname{Tr}\mathbf{J} = -\alpha - \beta - \gamma - \mu/(1+\bar{\xi})^2 < 0$, $\det\mathbf{J} = \beta(\alpha+\gamma) + \gamma\mu/(1+\bar{\xi})^2 > 0$, and $(\operatorname{Tr}\mathbf{J})^2 - 4\det\mathbf{J} > 0$, where $\mathbf{J}$ is the Jacobian matrix of the system (3.3) evaluated at fixed point (3.5). Equations (3.3) suggest that phase trajectories enter the first quadrant from the left and from below, thus making the nodal point (3.5) globally stable for all conceivable initial conditions.

4 Discussion and conclusions

Two populations of customers coexist in the service system under consideration. They are interdependent, though do not interact directly. One, with the normalized head count ξ, is the queue for service and consists of both newcomers and re-entrants. Another, with the normalized head count η, consists of those whose demand has been satisfied by the moment. In steady state,

as follows from the equations (3.3), the ratio of the two populations adheres to the formula $\bar{\xi}/\bar{\eta} = (\alpha+\gamma)(1+\bar{\xi})/\mu$. As will readily be observed, in the case of a low steady-state demand, $\bar{\xi} < 1$, this ratio takes the minimum value $(\alpha+\gamma)/\mu = (k_1u_0)^{-1}(1+k_{-1}/k_2)(k_3+k_5)$. Factor $1+k_{-1}/k_2$ is thought to be of order $\mathcal{O}(1)$ under the realistic assumption that $k_{-1} < k_2$ (the rate of rejected requests is lower than that of the fulfilled ones). Thus the expression for $\min(\bar{\xi}/\bar{\eta})$ can be interpreted simply as the ratio of the characteristic waiting time, $(k_1u_0)^{-1}$, to the mean time of customer's sojourn in the category of past clients, $(k_3+k_5)^{-1}$.

Requiring the steady-state normalized demand given by (3.5a) to not exceed unity, we find a combination of model parameters for which the queue length remains relatively short:

$$\bar{\xi} < 1 \quad \forall(\rho \leqslant \rho^* \wedge \beta > 0) \vee (\rho > \rho^* \wedge \beta > \rho - \rho^*), \tag{4.1}$$

where $\rho^* = \frac{1}{2}\gamma\mu/(\alpha+\gamma)$ is the critical arrival rate of potential customers to the system.

At subcritical arrival rates, such that $\rho \leqslant \rho^*$, the steady-state demand will keep low irrespective of the frequency of abandonments, as illustrated in Figure 3.

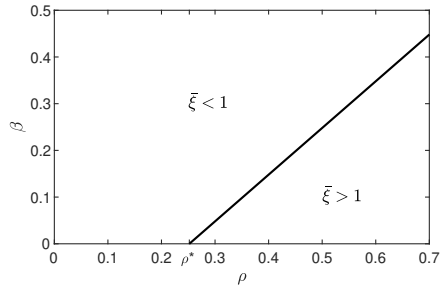

Figure 3. Contour line $\bar{\xi} = 1$ of the steady-state solution (3.5a) to the system (3.3) projected onto the plane $\rho\beta$ ("customers arrival rate–frequency of abandonments"). Line $\beta = \rho - \rho^*$, where $\rho^\star = \frac{1}{2}\gamma\mu/(\alpha+\gamma)$ is the critical arrival rate, partitions the positive quadrant into two domains: above the line the steady-state demand is low, and below is high. The parameters chosen for simulation: $\alpha = 1.1$, $\gamma = 1.4$, $\mu = 0.9$, and $\rho^* = 0.252$.

To a first approximation, the steady-state demand in the neighborhood of $\rho = \rho^*$ and $\beta = 0$ is given by $\bar{\xi} = 1+2(\rho-\rho^*)/\rho^*-2\beta/\rho^*$. As could be expected, the queue gets longer whenever the arrival rate increases, and shorter whenever abandonments become more frequent.

The dimensional form of the condition (4.1) is as follows:

$$\bar{x} < K \quad \forall(0 < r \leqslant r^* \wedge k_4 > 0) \vee (r > r^* \wedge k_4 > (r-r^*)/K),$$

where $K = (k_{-1}+k_2)/k_1$ and $r^* = \frac{1}{2}k_2u_0k_5/(k_3+k_5)$.

The threat of buffer overflow could be reduced, first and foremost, by elevating the critical arrival rate. In the formula for r^*, factor $k_5/(k_3+k_5)$ is sandwiched between 0 and 1: the lower limit may occur when the loyalty to

provider, k_5^{-1}, lasts longer than the service "aftereffect", k_3^{-1}, while the upper limit is the case when the loyalty is transient and the results of a service are long-lasting. Thus the potential of that factor is very limited. A more effective means to push the critical threshold up could be increasing the factor k_2u_0 that stands for the maximum possible yield of all servers in the system. This can be achieved by setting up more servers and/or by increasing the turnover of an individual server.

On the run, however, such a reorganization is difficult to implement. A more practical measure is to increase the frequency of abandonments, k_4, as soon as the inflow of new customers reaches the critical value r^*. Abandonment from service is an important feature of the model. Not only because there is no real queue without abandonment—such is the nature of consumer behavior. Without abandonment, a steady-state mode in our system would be impossible for the arrival rates higher than the maximum service rate. As a result, the queue length would blow up. The inclusion of abandonment in the model makes it more robust and thereby—more realistic. In case of necessity, it is much easier to manipulate the frequency of abandonments compared to other model parameters. For example, the provider may constantly monitor the arrival rate of new customers, and encourage those in the queue to postpone their visit, should the inflow reach a prescribed critical level.

The discussion of the properties of the open system would be incomplete without mentioning the distinct features of its closed counterpart. Closed version of the service system is obtained by putting $r = k_4 = k_5 = 0$ in equations (2.1). The resulting system of kinetic equations has two conserved quantities: u_0, given by (2.2), and x_0, given by $x + v + y = x_0$, where x_0 is the overall number of customers, potential, current and past. Thus the initial system of four equations can be reduced to a system of just two, which, upon a rescaling $\tau = tk_1x_0$, $\xi = x/x_0$, and $\eta = y/x_0$, takes the following nondimensional form:

$$\begin{aligned} \dot{\xi} &= c\eta - (\xi + a)(\xi + \eta - 1) - \epsilon\xi, \\ \dot{\eta} &= -b(\xi + \eta - 1) - c\eta, \end{aligned} \tag{4.2}$$

where $a = k_{-1}/(k_1x_0)$, $b = k_2/(k_1x_0)$, $c = k_3/(k_1x_0)$, $\epsilon = u_0/x_0$, and overdots denote differentiation with respect to the new time variable τ.

A realistic assumption $\epsilon \ll 1$ leads to the only physically feasible steady-state solution of the system (4.2) to within $\mathcal{O}(\epsilon)$:

$$\bar{\xi} = 1 - \frac{\epsilon(b + c)}{c(a + b + 1)}, \quad \bar{\eta} = \frac{\epsilon b}{c(a + b + 1)}. \tag{4.3}$$

This is a stable node in the phase plane $\xi\eta$ for all admissible values of the parameters, because $\det \mathbf{J} = c(a + b + 1) + \mathcal{O}(\epsilon) > 0$, $\operatorname{Tr} \mathbf{J} = -a - b - c - 1 + \mathcal{O}(\epsilon) < 0$, and $(\operatorname{Tr} \mathbf{J})^2 - 4 \det \mathbf{J} = (a + b - c + 1)^2 + \mathcal{O}(\epsilon) > 0$, where $\mathbf{J}$ is the Jacobian matrix of (4.2), evaluated at the fixed point (4.3). One can see that the steady-state population of past customers is always relatively small within the smallness of ϵ. Evidently, the number of past customers is in the direct relationship with server's turnover number and in the inverse relationship with all other parameters of the model.

As a by-product of the model analyzed in this work we derived the Karmarkar clearing function, a concept widely used nowadays for production planning. Technically, it is a well known result for steady-state mode. However we extended the notion of the clearing function to non-steady-state situations and succeeded to find out the conditions for its validity. The questions like "How good is an approximation offered by the clearing function?" and "When is it expected to hold, and under what conditions would it fail?" are among hot topics in queueing theory as evidenced by the review written by Armbruster [1]. So, there are two strongly varying timescales in our model of a service system: the longer one, corresponding to the dynamics of consumer populations, and the shorter, corresponding to the dynamics of the number of servers engaged in rendering service. Presence of this time hierarchy makes possible the quasi-steady-state approximation and, hence, the saturated response of the output to the input (buffer stock) in the form of the Karmarkar clearing function, identical with the Michaelis–Menten equation. In terms of our "enzymatic" model with two timescales, we are able to suggest a more sound justification of the clearing function: the momentary number of busy servers would be in a quasi-steady state with respect to the demand, provided the service time is much shorter than the waiting time. Consequently, the condition $\varepsilon \ll 1$ is expected to be sufficient to assure the validity of the clearing function in non-steady-state service systems.

The last (but not the least) issue concerns the accuracy of the fluid approximation in our model. Just as the deterministic approach fails to capture the discrete and stochastic nature of chemical reactions at low concentrations and small reaction volume, so does the continuous mass-action treatment of server-customer contacts at small numbers of interacting agents, whether customers or servers. As many service systems involve operation at extremely small quantities, such discrete stochastic effects are well relevant for our bio-inspired model. For some queues, large fluctuations in the demand may be destructive. The evolution of the population of customers due to interactions with servers-enzymes is adequately described by Markov processes, which can be formalized in terms of the chemical master equation (CME) for probability.

Recent years, as is evidenced by the review article [13], were marked with notable advances in the field of theoretical stochastic enzyme kinetics. In particular, there has been analyzed the Michaelis-Menten reaction mechanism with substrate inflow catalyzed by one enzyme molecule in a compartment, as encoded by equation (1.3). One may notice, that scheme (1.3) is nothing but the core of Figure 2 depicting our cyclic queue, so the main stochastic properties of the former could be extended to the latter. In mechanism (1.3), suppose Ω is the volume of the compartment in which the reaction occurs, E_0 is the total enzyme concentration (cf. u_0 from (2.2)), and $K_M = (k_{-1} + k_2)/k_1$ is the Michaelis–Menten constant (cf. K from (2.3)). By solution of the CME it follows that in the limit $K_M\Omega \gg 1$, the mean rate of product formation, $\langle \dot{P} \rangle$, reduces to the deterministic Michaelis-Menten equation $\dot{P} = k_2 E_0 S/(K_M + S)$ (cf. (3.4)) with $E_0 = 1/\Omega$. Given the definition of K_M and the fact that $k_{-1} + k_2$ represents the frequency with which complex dissociates and k_1/Ω is the frequency with which a substrate and an enzyme molecule associate, it fol-

lows that $K_M\Omega \gg 1$ implies the condition wherein bimolecular binding occurs relatively rarely compared with complex breakdown. Hence fluctuations in the substrate concentration are small. Monte Carlo simulations of reaction mechanism with enzyme molecule numbers in the range of 10–100 and with physiologically realistic parameters show that whenever the criterion $K_M\Omega \gg 1$ is satisfied, the Michaelis–Menten equation accurately describes the relationship between the rate of product formation and the mean substrate concentration.

In terms of our model, the condition $K_M\Omega \gg 1$ means $K/u_0 \gg 1$ or, what amounts to the same thing, $\varepsilon \ll 1$. The latter is precisely our key assumption as stated in the first paragraph of Section 3. Thus the deterministic fluid approximation, including the Michaelis–Menten mechanism of input-output transformation, can be safely applied to the service system depicted in Figure 2, whenever the mean service time is much shorter than the characteristic waiting time. At the same time, exploring the possibilities offered by more fundamental CME approach will constitute a future direction for work on our model.

It is important to underline that we used the notation of chemical reactions simply to describe things that combine and the things that they produce, and that this framework can be employed to model other phenomena, where acts of conversion are involved, in a similar way.

References

[1] D. Armbruster. The production planning problem: Clearing functions, variable lead times, delay equations and partial differential equations. In D. Armbruster and K. G. Kempf(Eds.), *Decision Policies for Production Networks*, pp. 289–302. Springer, London, 2012. https://doi.org/10.1007/978-0-85729-644-3_12.

[2] D. Armbruster, K. Kaneko and A. S. Mikhailov(Eds.). *Networks of Interacting Machines: Production Organization in Complex Industrial Systems and Biological Cells*, volume 3 of *World Scientific Lecture Notes in Complex Systems*. World Scientific, Singapore, 2005.

[3] D. Armbruster, D. Marthaler and C. Ringhofer. Kinetic and fluid model hierarchies for supply chains. *Multiscale Modeling & Simulation*, **2**(1):43–61, 2003. https://doi.org/10.1137/S1540345902419616.

[4] M. Armony, N. Shimkin and W. Whitt. The impact of delay announcements in many-server queues with abandonment. *Operations Research*, **57**(1):66–81, 2009. https://doi.org/10.1287/opre.1080.0533.

[5] R.U. Ayres and U.E. Simonis(Eds.). *Industrial Metabolism: Restructuring for Sustainable Development*. United Nations University Press, Tokyo, 1994.

[6] M. Bramson. *Stability of Queueing Networks*. Number 1950 in Lecture Notes in Mathematics. Springer, Berlin; Heidelberg, 2008. https://doi.org/10.1007/978-3-540-68896-9.

[7] A. Cornish-Bowden. *Fundamentals of Enzyme Kinetics*. Wiley-Blackwell, Weinheim, 4th edition, 2012.

[8] H. Fromm and J. Cardoso. Foundations. In J. Cardoso, H. Fromm, S. Nickel, G. Satzger, R. Studer and C. Weinhardt(Eds.), *Fundamentals of Service Systems*, Service Science: Research and Innovations in the Service Economy, pp. 1–32. Springer, New York, NY, 2015. https://doi.org/10.1007/978-3-319-23195-2_1.

[9] J. Gadrey. The characterization of goods and services: an alternative approach. *Review of Income and Wealth*, **46**(3):369–387, 2000. https://doi.org/10.1111/j.1475-4991.2000.tb00848.x.

[10] F. Gallouj. Innovation in services and the attendant old and new myths. *The Journal of Socio-Economics*, **31**(2):137–154, 2002. https://doi.org/10.1016/S1053-5357(01)00126-3.

[11] D. Gamarnik. Fluid models of queueing networks. In J.J. Cochran, L.A. Cox, P. Keskinocak, J.P. Kharoufeh and J.C. Smith(Eds.), *Wiley Encyclopedia of Operations Research and Management Science*. Wiley, Hoboken, NJ, 2011. https://doi.org/10.1002/9780470400531.eorms0329.

[12] N. Georgescu-Roegen. Some properties of a generalized Leontief model. In *Analytical Economics: Issues and Problems*, chapter 9, pp. 316–337. Harvard University Press, Cambridge, MA, 1966. https://doi.org/10.4159/harvard.9780674281639.c19.

[13] R. Grima, N.G. Walter and S. Schnell. Single-molecule enzymology à la Michaelis–Menten. *FEBS Journal*, **281**(2):518–530, 2014. https://doi.org/10.1111/febs.12663.

[14] D. Helbing, D. Armbruster, A.S. Mikhailov and E. Lefeber. Information and material flows in complex networks. *Physica A: Statistical Mechanics and its Applications*, **363**(1):xi–xvi, 2006. https://doi.org/10.1016/j.physa.2006.01.042.

[15] T.P. Hill. On goods and services. *Review of Income and Wealth*, **23**(4):315–338, 1977. https://doi.org/10.1111/j.1475-4991.1977.tb00021.x.

[16] P. Hochendoner, C. Ogle and W.H. Mather. A queueing approach to multi-site enzyme kinetics. *Interface Focus*, **4**:1–11, 2014. https://doi.org/10.1098/rsfs.2013.0077.

[17] W.J. Hopp and M.L. Spearman. *Factory Physics: Foundations of Manufacturing Management*. McGraw-Hill, New York, NY, 3rd edition, 2008.

[18] U.S. Karmarkar. Manufacturing lead times, order release and capacity loading. In S. C. Graves, A. H. G. Rinnooy Kan and P. H. Zipkin(Eds.), *Logistics of Production and Inventory*, volume 4 of *Handbook in Operations Research and Management Science*, chapter 6, pp. 287–329. North Holland, Amsterdam, 1993.

[19] W. Klonowski. Simplifying principles for chemical and enzyme reaction kinetics. *Biophysical Chemistry*, **18**(2):73–87, 1983. https://doi.org/10.1016/0301-4622(83)85001-7.

[20] E. Koenigsberg. Cyclic queues. *Operational Research Quarterly*, **9**(1):22–35, 1958. https://doi.org/10.2307/3007650.

[21] E. Koenigsberg. Twenty five years of cyclic queues and closed queue networks: A review. *Journal of the Operational Research Society*, **33**(7):605–619, 1982. https://doi.org/10.1057/jors.1982.136.

[22] G.I. Kolesova and I.A. Poletaev. Nekotorye voprosy issledovaniia sistem s limitiruiushchimi faktorami [Selected problems in research of the systems with limiting factors]. In *Upravliaemye sistemy [Controllable systems]*, number 3, pp. 71–80. Institute of Mathematics, Siberian Branch of the USSR Academy of Sciences, Novosibirsk, 1969. (In Russian)

[23] C. Kuehn. *Multiple Time Scale Dynamics*, volume 191 of *Applied Mathematical Sciences*. Springer, New York, NY, 2014.

[24] L.D. Landau and E.M. Lifshitz. *Statistical Physics. Part 1*, volume 5 of *Course of Theoretical Physics*. Elsevier Butterworth-Heinemann, Burlington, MA, 3rd edition, 2013.

[25] E. Levine and T. Hwa. Stochastic fluctuations in metabolic pathways. *Proceedings of the National Academy of Sciences*, **104**(22):9224–9229, 2007. https://doi.org/10.1073/pnas.0610987104.

[26] A. Mustafin and A. Kantarbayeva. Opening the Leontief's black box. *Heliyon*, **4**(5):e00626, 2018. https://doi.org/10.1016/j.heliyon.2018.e00626.

[27] R.R. Nelson and S.G. Winter. *An Evolutionary Theory of Economic Change*. Belknap Press of Harvard University Press, Cambridge, MA, 1982.

[28] J. Niyirora and J. Zhuang. Fluid approximations and control of queues in emergency departments. *European Journal of Operational Research*, **261**(3):1110–1124, 2017. https://doi.org/10.1016/j.ejor.2017.03.013.

[29] Yu.M. Romanovskii, N.M. Stepanova and D.S. Chernavskii. *Chto takoe matematicheskaia biofizika: Kineticheskie modeli v biofizike [What is mathematical biophysics: Kinetic models in biophysics]*. Prosveshchenie, Moscow, 1971. (In Russian)

[30] L.A. Segel and M. Slemrod. The quasi-steady-state assumption: A case study in perturbation. *SIAM Review*, **31**(3):446–477, 1989. https://doi.org/10.1137/1031091.

[31] R.W. Shephard. *Theory of Cost and Production Functions*. Princeton University Press, Princeton, NJ, 2016.

[32] J.F. Shortle, J.M. Thompson, D. Gross and C.M. Harris. *Fundamentals of Queueing Theory*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 5th edition, 2018.

[33] J. Taylor and R.R.P. Jackson. An application of the birth and death process to the provision of spare machines. *Operational Research Quarterly*, **5**(4):95–108, 1954. https://doi.org/10.2307/3007087.

[34] The World Bank. *World development indicators. Table 4.2: Structure of output*. The World Bank Group, Washington, DC, 2020. Available from Internet: `http://wdi.worldbank.org/table/4.2`.

[35] W. Whitt. Time-varying queues. *Queueing Models and Service Management*, **1**(2):79–164, 2018.

[36] G.B. Yom-Tov and A. Mandelbaum. Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, **16**(2):283–299, 2014. https://doi.org/10.1287/msom.2013.0474.