МРНТИ 28.29.07

А. Мустафин[1], А. Кантарбаева[2]

[1]Казахский национальный исследовательский технический университет имени К.И. Сатпаева,Алматы, Казахстан
[2]Казахский национальный университет имени аль-Фараби, Алматы, Казахстан
butsuri123@gmail.com, ratnakka@gmail.com

ЖИДКОСТНАЯ МОДЕЛЬ ЦИКЛИЧЕСКОЙ ОЧЕРЕДИ

Аннотация. Построена детерминистическая модель системы массового обслуживания с открытой на вход и выход циклической очередью с отказами. Модель представляет собой систему нелинейных обыкновенных дифференциальных уравнений для зависящих от времени средних численностей популяций: клиентов (требований, заявок), ожидающих обслуживания (обработки); клиентов, получивших услугу;занятых серверов (каналов обслуживания) и свободных серверов. Особенность подхода состоит в представлении о сервере как о своего рода ферменте, который осуществляет трансформацию клиентов из одной категории в другую, возвращаясь в исходное состояние после каждого акта конверсии. Эмпирический факт, что среднее время обслуживания намного короче характерного времени ожидания, делает систему уравнений сингулярно возмущённой. Методом многих масштабов система расщепляется на медленную и быструю подсистемы, описывающие соответственно динамику клиентов и серверов. Показано, что в адиабатическом приближении количество занятых серверов мгновенно отслеживает текущий спрос в соответствии с известными соотношениями для квазистационарных концентраций ферментативной кинетики. Найдено физически допустимое стационарное решение медленной подсистемы и доказана его асимптотическая устойчивость. Проведён параметрический анализ стационарного состояния. Получен важный практический вывод о том, что стационарная очередь остаётся короткой независимо от частоты отказов при входящем потоке клиентов ниже определённого критического уровня.

Ключевые слова: очередь, жидкостная модель, отказ, повторное требование.

Введение

Быстрое развитие сферы услуг в современной экономике оживило интерес к моделям многоканальных систем массового обслуживания. Применение таких моделей особенно актуально в крупных системах обслуживания, занимающихся обработкой обращений и информированием по телефону (колл-центры), электронной и обычной почте, в интернет-чате (контакт-центры), в инфокоммуникационных сетях, а также в клиниках и других учреждениях здравоохранения [1–4].

Исследование операций, разделом которого является теория массового обслуживания, рассматривает услугу как операцию «вход-выход», выполняемую некоторым «чёрным ящиком»-преобразователем [5]. При оказании услуги трансформации подвергается не физическая сущность объекта на входе, но его состояние. Согласно получившему широкое распространение определению, предложенному Т. Хиллом [6], услуга – это изменение состояния человека или имущества, принадлежащего некоторой экономической единице, которое осуществляется в результате деятельности другой экономической единицы по предварительному согласию ранее упомянутых человека или экономической единицы. В отличие от продукта, который сразу же после его выпуска приобретает обособленное положение по отношению как к производителю, так и к потребителю, услуга единосущна и с провайдером, и с клиентом, её невозможно хранить, она представляет собой не какой-то результат, а, скорее, акт или процесс [7].

Цель настоящей работы состоит в попытке дать правдоподобную реконструкцию событий, происходящих в «чёрном ящике» услуги, понимаемой как трансформация состояния клиента в результате действий провайдера.Мы предполагаем, что изменение состояния потребителя услуги в результате его взаимодействия с провайдером происходит примерно так же, как превращение молекулы субстрата в молекулу другого вещества в биохимической реакции, катализируемой ферментом. В самом деле, в живой клетке молекула субстрата S связывается с молекулой фермента E, образуя короткоживущий фермент-субстратный комплекс $[ES]$. Этот комплекс затем распадается на молекулу продукта P и исходную молекулу фермента, которая способна катализировать новую реакцию (например, [8]):

$$S + E \rightleftharpoons [ES] \rightarrow P + E.$$

Трактуя услугу как разновидность процесса превращения, можно отождествить входящего (потенциального) клиента с субстратом, выходящего (обслуженного) клиента – с продуктом, а провайдера услуги, представленного каналом обслуживания (здесь и далее называемого сервером) – с ферментом.

Мы апробируем работоспособность ферментативной модели оказания услуги на примере циклической системы массового обслуживания. Используется формализм непрерывного приближения, в котором дискретный поток требований заменяется сплошным потоком «жидкости».Детерминистические жидкостные модели в виде систем обыкновенных дифференциальных уравнений (ОДУ) для средних величин оказываются удобнее для исследованиявопросов устойчивости по сравнению с исходными стохастическими моделями теории массового обслуживания[9, 10].

Модель

Рассмотрим открытую систему обслуживания с группой параллельных одинаковых серверов (приборов, каналов обслуживания), обслуживающих входящих клиентов (заявки, требования). Постоянный поток потенциальных потребителей услуги поступает в систему извне, и если все имеющиеся серверы заняты, то прибывший становится в очередь. Потенциальный клиент может проявить нетерпение и через некоторое время покинуть очередь совсем. Такие клиенты считаются потерянными для системы – они могут направиться в другую компанию того же профиля.

Клиенты, ожидающие обслуживания, собираются в общем буфере, снабжающем все серверы. Размер буфера предполагается неограниченным. В нашей модели буферный запас, длина очереди, текущий спрос и популяция потенциальных клиентов являются синонимами. Каждый сервер может принять к обработке любое из поступивших требований, причём в состоянии обслуживать одновременно только одного клиента. Закончив обслуживание, сервер принимает следующего клиента из буфера. Популяция потенциальных потребителей не включает в себя клиентов, обслуживаемых в данный момент.

Контакт клиента с сервером не обязательно заканчивается получением желаемой услуги. Клиент может внезапно засомневаться и отложить обслуживание. Провайдер также может отклонить требование, если, например, заявка составлена не по форме. В обоих случаях клиент возвращается обратно в очередь.

После успешного обслуживания потребитель пополняет популяцию бывших клиентов. Спустя некоторое время он может либо совсем утратить интерес к данной услуге, либо опять присоединиться к очереди на обслуживание. Последнее имеет место, если изменение состояния клиента носит временный характер. Например, приобретённый покупателем товар может быть скоропортящимся или подлежащим замене (ремонту), вследствие постоянной эксплуатации, или подверженным моральному износу, или входить в категорию модных товарови тому подобное. Таким образом, очередь в буфере всегда состоит из смеси новых и бывших клиентов.

Перечисленную выше совокупность процессов, происходящих в изучаемой циклической очереди, можно представить в виде схемы псевдохимических реакций, как показано на Рисунке 1.

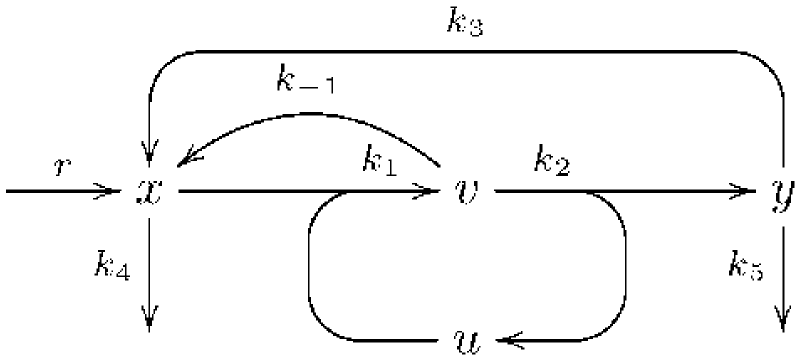

Рисунок 1 – Открытая циклическая система массового обслуживания.Здесь x – потенциальные клиенты, u –свободные серверы, v – серверы, занятые обслуживанием, y –потребители, получившие услугу. Константы r, k_{-1}, k_1, …, k_5 характеризуют скорости соответствующих процессов.

Схема кодирует как последовательность превращений, так и скорости их протекания. Она может быть формализована в виде системы кинетических уравнений для скоростей изменения количеств участвующих агентов. Для записи соответствующих уравнений воспользуемся известным в химии законом действующих масс, согласно которому скорость образования продукта в реакции взаимодействия пропорциональна количеству каждого из участвующих реагентов. Коэффициент пропорциональности в законе действующих масс – это константа скорости соответствующего акта превращения, которая предполагается известной. Константы каждого процесса обозначены на Рисунке 1 метками у стрелок. Они имеют следующий смысл: r – скорость притока клиентов в систему извне; k_1 – частота приёма требований индивидуальным сервером; k_{-1} – частота отзывов (или отклонений) требований на обслуживание после установления контакта между клиентом и сервером; k_2 – число оборотов сервера, то есть максимальное количество удовлетворённых им заявок за единицу времени. Таким образом, $(k_{-1}+k_2)^{-1}$ представляет собой среднее время занятости сервера; k_3 – частота возвращения клиентов в очередь на повторное обслуживание. Тем самым k_3^{-1} характеризует время, в течение которого сохраняется ценность полученной услуги, или время жизни изменённого состояния клиента; k_4 – частота, с которой нетерпеливые потребители покидают очередь; k_5 – частота, с которой провайдер услуги по разным причинам утрачивает лояльность бывших клиентов.

Записывая балансовое соотношение для каждого из участников схемы на Рисунке 1, мы получаем систему ОДУ:

$$\begin{aligned}
\dot{x} &= r + k_{-1}v + k_3 y - k_1 ux - k_4 x,\\
\dot{y} &= k_2 v - (k_3 + k_5)y,\\
\dot{u} &= (k_{-1} + k_2)v - k_1 ux,\\
\dot{v} &= k_1 ux - (k_{-1} + k_2)v,
\end{aligned} \tag{1}$$

где точки сверху означают дифференцирование по времени t. Все параметры модели неотрицательны. (Заметим, что модель, рассмотренная в работе \cite{Niyirora:2017}, является частным случаем системы (1) при $k_{-1} = k_3 = k_5 = 0$.)

Складывая третье и четвёртое уравнения системы (1), выявляем первый интеграл, смысл которого состоит в сохранении общего числа имеющихся серверов $u_0 : u + v = u_0$. С его помощью систему (1) можно упростить, исключая u либо v. Произвольно выбирая для исключения u, мы следующим шагом приводим три оставшихся уравнения к безразмерному виду путём перехода к новым переменным

$$\tau = t / T, \quad \xi = x / K, \quad \eta = y / K, \quad \zeta = v / u_0, \tag{2}$$
$$\text{где} T = (k_1 u_0)^{-1} \text{ и } K = (k_{-1} + k_2) / k_1.$$

Согласно формулам (2) безразмерное время τ отсчитывается в единицах T. Величина T имеет смысл характерного времени, проводимого потенциальным клиентом в очереди перед началом обслуживания. Это время обратно пропорционально общему количеству серверов u_0. Обоснование нормировки x и y на величину K будет дано в следующем разделе. Что касается переменной v, то её нормировка на u_0 представляется очевидной.

Безразмерные жидкостные уравнения принимают вид

$$\begin{aligned} \dot{\xi} &= \rho + \alpha\eta + (1-\mu)\zeta + \zeta\xi - (1+\beta)\xi, \\ \dot{\eta} &= \mu\zeta - (\alpha+\gamma)\eta, \\ \varepsilon\dot{\zeta} &= \xi - \zeta\xi - \zeta, \end{aligned} \tag{3}$$

где точки сверху теперь подразумевают дифференцирование по τ. Величины

$$\begin{aligned} &\alpha = k_3 / (k_1 u_0), \quad \beta = k_4 / (k_1 u_0), \quad \gamma = k_5 / (k_1 u_0), \\ &\varepsilon = k_1 u_0 / (k_{-1} + k_2), \quad \mu = k_2 / (k_{-1} + k_2), \quad \rho = r / \left((k_{-1} + k_2) u_0\right) \end{aligned} \tag{4}$$

– новые безразмерные параметры. Следует заметить, что $\mu < 1$. Согласно определению (4) ε – это отношение времени обслуживания $(k_{-1} + k_2)^{-1}$к времени ожидания$(k_1 u_0)^{-1}$.Из физических соображений имеют смысл только неотрицательные решения уравнений (3).

Результаты

При выводе уравнений (3) мы не делали никаких специальных предположений относительно параметров. Однако известно [4], что время ожидания часто длиннее времени обслуживания, поэтому следует считать параметр ε малым. При $\varepsilon \ll 1$система (3) становится сингулярно-возмущённой с медленными переменнымиξ, ηи быстрой переменнойζ. Стандартной практикой редукции таких систем служит метод многих масштабов, посредством которого быстрая переменная адиабатически исключается. Законность процедуры исключения устанавливается теоремойФенихеля–Тихонова [11].

В сингулярном пределе $\varepsilon \to 0$ третье уравнение системы (3) заменяется алгебраическим $\xi - \zeta\xi - \zeta = 0$.Алгебраическое уравнение задаёт медленное многообразие, по которому движется изображающая точка системы на масштабах времени $\tau = O(1)$. Оно же определяет и квазистационарное состояние быстрой переменной: $\zeta = \xi / (1+\xi)$. Напомним, что ζ – это доля занятых серверов. Покуда нормированный спрос мал, то есть $\xi \ll 1$, эта доля тоже адекватно мала, $\zeta \approx \xi$, означая низкую загруженность серверов.

Однако при высоких уровнях спроса, когда $\xi \gg 1$, практически все серверы заняты обслуживанием: $\zeta \to 1$. Когда $\xi = 1$, доля занятых серверов составляет половину от их общего числа в системе. В этом состоит причина выбора величины K в качестве масштабной единицы для размеров популяций потенциальных и бывших клиентов. В биохимии K носит название константы Михаэлиса.

Отсюда следует, что на временах порядка $\tau = O(1)$ динамика системы массового обслуживания (3) управляется двумя уравнениями:

$$\begin{aligned} \dot{\xi} &= \rho + \alpha\eta - \mu\xi / (1+\xi) - \beta\xi, \\ \dot{\eta} &= \mu\xi / (1+\xi) - (\alpha+\gamma)\eta. \end{aligned} \tag{5}$$

Первый член в правой части второго уравнения (5) представляет собой выпуск рассматриваемой системы, то есть количество клиентов, получающих услугу в единицу времени: $Y = \mu\xi / (1+\xi)$. В этой формуле можно узнать знаменитое уравнение Михаэлиса–Ментен из ферментативной кинетики \cite[гл.2]{Cornish:1979}. Её отличительной чертой является насыщение выпуска в ответ на увеличение спроса. При низком спросе выпуск приблизительно пропорционален спросу. Однако при высоком спросе поток получивших услугу стремится к постоянной величине.

Система (5) имеет два стационарных состояния. Из них лишь одно всегда лежит в первом квадранте фазовой плоскости $\xi\eta$, то есть физически осмысленно:

$$\begin{aligned} \bar{\xi} &= \left((B^2 + 4\beta(\alpha+\gamma)^2\rho)^{1/2} - B \right) \Big/ \left(2\beta(\alpha+\gamma) \right), \\ \bar{\eta} &= \left(B + 2(\alpha+\gamma)\rho - (B^2 + 4\beta(\alpha+\gamma)^2\rho)^{1/2} \right) \Big/ \left(2\gamma(\alpha+\gamma) \right), \end{aligned} \tag{6}$$

где ради краткости обозначено $B = (\alpha+\gamma)(\beta-\rho) + \gamma\mu$. Анализ показывает, что неподвижная точка (6) всегда устойчива по типу узла. Фазовые траектории входят в первый квадрант слева и снизу, делая тем самым стационарную точку (6) глобально устойчивой для всех мыслимых начальных условий.

Обсуждение

Потребовав, чтобы стационарный нормированный спрос, даваемый первой формулой из (6), не превышал единицы, находим, при каком соотношении параметров модели очередь на обслуживание остаётся относительно короткой:

$$\bar{\xi} < 1 \quad \forall(\rho \le \rho^* \wedge \beta > 0) \vee (\rho > \rho^* \wedge \beta > \rho - \rho^*), \tag{7}$$

где $\rho^* = \gamma\mu / \left(2(\alpha+\gamma)\right)$– критическая скорость притока клиентов в систему. Примечательно, что при докритической скорости притока новых клиентов, такой что $\rho \le \rho^*$, стационарный спрос остаётся низким независимо от частоты уходов из очереди (отказов), как показано на Рисунке 2.

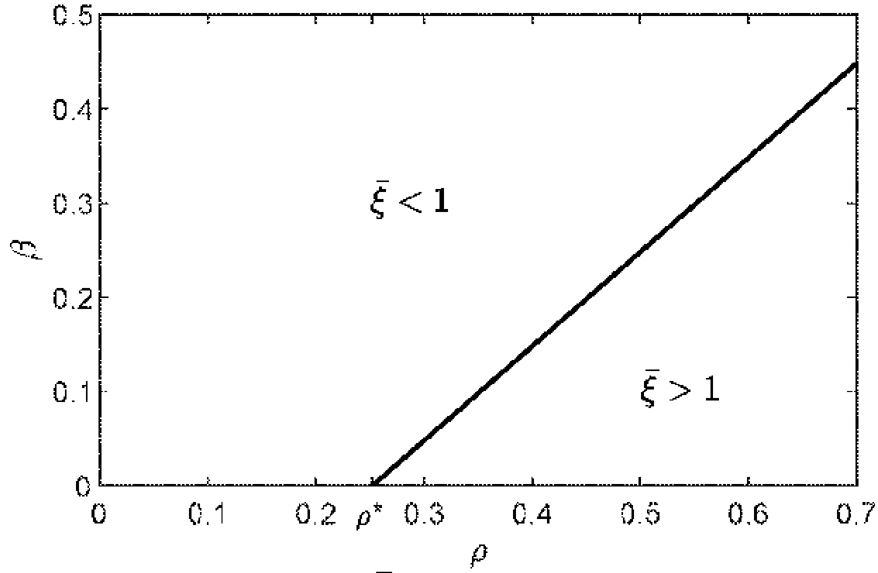

Рисунок 2 – Проекция линии уровня $\bar{\xi} = 1$ стационарного решения (6) системы (5) на плоскость параметров $\rho\beta$.Прямая $\beta = \rho - \rho^*$ делит плоскость на две области: выше прямой – стационарный спрос низкий, ниже – высокий. Параметры расчёта: $\alpha = 1{,}1$; $\gamma = 1{,}4$; $\mu = 0{,}9$; $\rho^* = 0{,}252$.

В первом приближении для стационарного спроса в окрестности параметров $\rho = \rho^*$ и $\beta = 0$ получаем: $\bar{\xi} = 1 + 2(\rho - \rho^*)/\rho^* - 2\beta/\rho^*$. Как и следовало ожидать, увеличение входящего потока требований удлиняет очередь, а повышение частоты отказов– сокращает.

В размерном виде условие (7) выглядит так:

$$\bar{x} < K \quad \forall (0 < r \le r^* \wedge k_4 > 0) \vee (r > r^* \wedge k_4 > (r - r^*)/K),$$
$$\text{где } K = (k_{-1} + k_2)/k_1 \text{ и } r^* = k_2 u_0 k_5 / \left(2(k_3 + k_5)\right).$$

При ограниченной вместимости буфера снизить угрозу его переполнения можно прежде всего путём повышения критической скорости притока потенциальных клиентов. В выражении для r^* множитель $k_5/(k_3 + k_5)$ заключён в пределах от 0 до 1: нижнее значение соответствует случаю, когда среднее время сохранения лояльности бывшего клиента к провайдеру, k_5^{-1}, намного продолжительнее времени жизни результата обслуживания, k_3^{-1}; верхнее – случаю, когда услуга долговременная, а бывшие клиенты быстро утрачивают интерес или потребность в получении повторной услуги у данного провайдера. Поэтому более действенным средством отодвигания критического порога может послужить увеличение фактора $k_2 u_0$– максимально возможной скорости обслуживания клиентов всеми серверами, имеющимися в системе. Это достигается укомплектованием системы обслуживания дополнительным персоналом или оборудованием либо повышением числа оборотов, k_2, каждого канала обслуживания.

На ходу, однако, такую реорганизацию провести затруднительно. Более практичной мерой представляется увеличение частоты отказов k_4 сразу по достижении притоком клиентов критической скорости r^*. Учёт отказов от обслуживания является важной особенностью рассматриваемой модели. И не только потому что реальной очереди без отказов не бывает – таково поведение потребителя. Без отказов стационарный режим в модели становится невозможным при превышении входящего трафика клиентов над максимальной скоростью обслуживания, в результате чего длина очереди неограниченно растёт. Включение в модель отказов делает её более робастной и тем самым – более реалистичной. При необходимости частотой отказов в системе массового обслуживания намного проще манипулировать в динамике по сравнению с другими параметрами. Можно, например, постоянно следить за скоростью прибытия клиентов, и в случае достижения ею критической величины побуждать потенциальных потребителей отложить получение услуги на более поздний срок путём объявлений о задержке обслуживания.

Заключение

Нами построена жидкостная модель системы массового обслуживания, представляющая собой открытую циклическую очередь. Мы обратились к биологической аналогии провайдера услуги с ферментом-катализатором, составив балансовые дифференциальные уравнения для потоков клиентов, подобные уравнениям химической кинетики. Анализ модели позволил сделать ряд содержательных выводов о свойствах и поведении системы в стационарном режиме и в динамике, которые затруднительно получить в рамках вероятностной постановки. Важно подчеркнуть, что мы использовали язык химических реакций лишь для удобства описания процессов взаимодействия экономических агентов, когда переход агента из одного состояния в другое осуществляется через образование и дезинтеграцию короткоживущего промежуточного комплекса. Этот формализм может с успехом быть применён и для моделирования других явлений, где имеют место акты трансформации вход-выход.

СПИСОК ЛИТЕРАТУРЫ

[1]. ArmonyM., ShimkinN., WhittW. The impact of delay announcements in many-server queues with abandonment// Operations Research. 2009. Vol. 57, no. 1. P. 66–81.

[2]. Yom-TovG.B., MandelbaumA. Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing// Manufacturing and Service Operations Management. 2014. Vol. 16, no. 2. P. 283–299.

[3]. NiyiroraJ., ZhuangJ. Fluid approximations and control of queues in emergency departments//European Journal of Operational Research. 2017. Vol. 261, no. 3. P. 1110–1124.

[4]. WhittW. Time-varying queues// Queueing Models and Service Management. 2018. Vol. 1, no. 2. P. 79–164.

[5]. FrommH., CardosoJ. Foundations// Fundamentals of Service Systems/ Ed. by J. Cardoso, H. Fromm, S. Nickel etal. New York, NY: Springer, 2015. Service Science: Research and Innovations in the Service Economy. P. 1–32.

[6]. HillT.P. On goods and services// Review of Income and Wealth. 1977. Vol. 23, no. 4. P. 315–338.

[7]. GalloujF. Innovation in services and the attendant old and new myths// The Journal of Socio-Economics. 2002. Vol. 31, no. 2. P. 137–154.

[8]. Cornish-BowdenA. Fundamentals of Enzyme Kinetics. 4th ed. Weinheim: Wiley-Blackwell, 2012. 498 p.

[9]. ArmbrusterD., MarthalerD., RinghoferC. Kinetic and fluid model hierarchies for supply chains// Multiscale Modeling and Simulation. 2003. Vol. 2, no. 1. P. 43–61.

[10]. GamarnikD. Fluid models of queueing networks// Wiley Encyclopedia of Operations Research and Management Science/ Ed. by J.J. Cochran, L.A. Cox, P. Keskinocak etal. Hoboken, NJ : Wiley, 2011.

[11]. KuehnC. Multiple Time Scale Dynamics. New York, NY: Springer, 2014. Vol. 191 of Applied Mathematical Sciences. 814 p.

ЦИКЛДІК КЕЗЕК ҮШІН СҰЙЫҚТЫҚ МОДЕЛІ

А. Мұстафин[1], Ә. Қаңтарбаева[2]

[1]Сәтбаев университеті, Алматы, Қазақстан
[2]әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан
butsuri123@gmail.com, ratnakka@gmail.com

Аңдатпа. Ашық циклдік жаппай қызмет көрсету жүйесінің сұйық моделі істен шығумен және қайталама талаптармен ұсынылған. Модель келесі айнымалы: кіріс клиенттер (талаптар); қызмет алған клиенттер; бос серверлер (қызмет көрсету арналары); және бос серверлер үшін сызықты емес дифференциалдық теңдеулер жүйесінің түрі бар. Тәсіл сервердің ферментке ұқсас әрекет ететіні туралы идеяға негізделген. Сингулярлық ауытқулар теориясы әдістерінің көмегімен бастапқы жүйе екі санаттың әрбір клиенттерінің динамикасына арналған екі теңдеуге келтіріледі. Жұмыс істейтін серверлердің квазистационарлық саны белгілі қисық Михаэлис-Ментен ферментативті кинетиканың болуы көрсетілген. Модельдің физикалық тұрғыдан ойластырылған қозғалмайтын нүктесі табылды және оның асимптотикалық тұрақтылығы дәлелденді. Стационарлық шешімге параметрлік талдау жүргізілді. Модель клиенттердің түсу жылдамдығына байланысты сұранысты басқару тәсілдерін болжайды.

Тірек сөздер:кезек, сұйықтық модель, істен шығу, қайта талап.

A FLUID MODEL FOR THE CYCLIC QUEUE

A. Mustafin[1], A. Kantarbayeva[2]

[1]Satbayev University, Almaty, Kazakhstan
[2]al-Farabi Kazakh National University, Almaty, Kazakhstan
butsuri123@gmail.com, rantakka@gmail.com

Abstract. A fluid model for the open cyclical service system with abandonment and re-entry is proposed. The model has the form of a system of nonlinear ordinary differential equations for the following variables: incoming customers;outgoing customers;busy servers; and free servers. The key feature of our approach is the ideathatserver acts as an enzyme transforming the condition of a customer.Using the methods of the singular perturbation theory the original system is reduced to a couple of equations for the respective dynamics of customers of both categories. The quasi-steady-state number of busy servers is shown to follow the well-known Michaelis-Menten saturation curve of the enzyme kinetics. The physically feasible fixed point of the model is found and proven to be asymptotically stable. The parametric analysis of the steady-state solution is performed. The model predicts how the demand could be controlled depending on the arrival rateof the customers.

Key words*:* queue, fluid model, abandonment, re-entry.