ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2018 Управление, вычислительная техника и информатика № 43

УДК 519.21, 519.87
DOI: 10.17223/19988605/43/6

**О.А. Осипов**

# АНАЛИЗ RQ-СЕТИ МАССОВОГО ОБСЛУЖИВАНИЯ С ДЕЛЕНИЕМ И СЛИЯНИЕМ ТРЕБОВАНИЙ

Рассматривается экспоненциальная сеть массового обслуживания, состоящая из параллельных систем обслуживания с конечным числом мест для ожидания в очереди. Ключевой особенностью является деление поступающих требований на фрагменты, которые обслуживаются параллельно и независимо. Требование будет считаться обслуженным только после того, как завершится обслуживание всех его фрагментов. Ограниченность числа мест для ожидания в системах сети приводит к возникновению повторных вызовов. Для сети обслуживания с использованием матричных методов получены основные стационарные характеристики.
**Ключевые слова:** сети обслуживания с делением и слиянием требований; матрично-геометрическое решение; источник повторных вызовов.

Сети массового обслуживания (СеМО) с делением и слиянием требований (fork-join queueing networks) [1] являются математическими моделями, используемыми для анализа стохастических систем с параллельным или распределенным принципом функционирования (телекоммуникационные системы, распределенные базы данных, многопроцессорные системы). В таких реальных системах поступающие для обработки задачи делятся на более простые для выполнения подзадачи, которые распределяются по системе, занимая выделенные им ресурсы, однако исходная задача будет считаться выполненной только после выполнения всех ее подзадач.

В большинстве работ рассматриваются сети массового обслуживания с делением и слиянием требований, состоящие из параллельных систем массового обслуживания (СМО). Для сети обслуживания, состоящей из двух одноприборных параллельных СМО, в [2] получено выражение для производящей функции стационарного распределения вероятностей состояний сети. Анализ же сетей обслуживания большей размерности с делением и слиянием требований проводится только приближенными методами [3–5]. Обзор основных результатов за тридцатилетний период изучения СеМО с делением и слиянием требований можно найти в [6].

Характерной особенностью моделируемых реальных систем является наличие повторных обращений от поступающих для выполнения задач спустя некоторое время, они возникают в случае отказа в выполнении, что имеет место в связи с ограниченностью ресурсов системы. Для моделирования процессов, возникающих при повторных обращениях, используют модели массового обслуживания с повторными вызовами – RQ-системы и сети массового обслуживания (retrial queues). В таких моделях требование, получившее отказ в обслуживании, поступает в источник повторных вызовов (ИПВ), из которого снова пытается получить обслуживание.

Для исследования моделей обслуживания с повторными вызовами применяются матричные методы [7, 8], а также методы асимптотического анализа [9, 10]. Основные результаты можно найти в монографиях [11, 12].

В данной работе будет рассмотрена сеть массового обслуживания с делением и слиянием требований, состоящая из параллельных СМО, в которой имеют место повторные вызовы. Возникновение повторных вызовов связано с ограниченностью числа мест для ожидания в каждой системе обслуживания сети.

Статья организована следующим образом. В разд. 1 описывается изучаемая сеть массового обслуживания. Стационарное распределение вероятностей состояний сети, а также выражения для

основных стационарных характеристик приводятся в разд. 2 и 3 соответственно. Раздел 4 содержит численные примеры.

## 1. Описание сети обслуживания

Рассматривается RQ-сеть массового обслуживания с делением и слиянием требований, состоящая из $M$ параллельных одноприборных систем обслуживания, каждая с конечным числом $B$ мест для ожидания в очереди (рис. 1).

В сеть обслуживания из внешнего источника поступает пуассоновский поток требований с интенсивностью $\Lambda$. Вновь поступающее требование, застающее сеть в состоянии, когда в каждой очереди имеются свободные места для ожидания, делится на $M$ фрагментов, которые распределяются по системам сети и ожидают своего обслуживания в соответствии с дисциплиной FCFS в очередях. Длительность обслуживания фрагментов на приборе системы $i$ имеет экспоненциальное распределение с параметром $\mu_i$.

Фрагмент, завершивший свое обслуживание, освобождает обслуживающий его прибор. Требование будет считаться обслуженным только после того, как будет завершено обслуживание всех его фрагментов. Сразу после этого фрагменты требования мгновенно объединяются в исходное обслуженное требование.

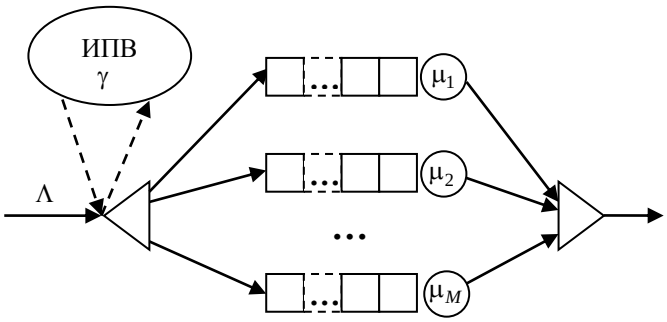


Рис. 1. RQ-сеть обслуживания с делением и слиянием требований

В том случае, когда хотя бы одна очередь заполнена полностью, поступившее требование не может быть поделено и переходит в источник повторных вызовов, где после случайной задержки вновь пытается обратиться к системам сети для получения обслуживания; если требование снова не может получить обслуживание, то оно вновь возвращается в ИПВ. Предполагается, что если в ИПВ есть требования, то длительность интервала времени между повторными вызовами имеет экспоненциальное распределение с параметром $\gamma$.

## 2. Стационарное распределение вероятностей состояний сети

Состояние рассматриваемой сети обслуживания в момент времени $t$ определим как вектор $\boldsymbol{x}(t) = (r(t), n_1(t), \ldots, n_M(t))$, где $r(t)$ – число требований в ИПВ, $n_i(t)$ – число фрагментов в системе обслуживания $i$, $i = 1, \ldots, M$.

Очевидно, что $(M + 1)$-мерный процесс $\{\boldsymbol{x}(t), t \geq 0\}$ есть цепь Маркова с непрерывным временем, определенная на пространстве состояний $\boldsymbol{X}$,

$$\boldsymbol{X} = \{(r, n_1, \ldots, n_M) : r \geq 0, 0 \leq n_i \leq B+1, i = 1, \ldots, M\}.$$

Обозначим через $q(\boldsymbol{x}, \boldsymbol{x}')$ интенсивность перехода цепи из состояния $\boldsymbol{x}$ в состояние $\boldsymbol{x}'$.

Справедливо:

1) если $n_i < B + 1$, $i = 1, \ldots, M$,

$$q\big((r, n_1, \ldots, n_M), (r, n_1 + 1, \ldots, n_M + 1)\big) = \Lambda; \qquad (1)$$

2) если $n_i < B + 1$, $i = 1, \ldots, M$, $r > 0$,

$$q\left(\left(r, n_1, ..., n_M\right), \left(r-1, n_1+1, ..., n_M+1\right)\right) = \gamma; \quad (2)$$

3) если существует $j \in \{1, ..., M\}$, такое, что $n_j = B + 1$,

$$q\left(\left(r, n_1, ..., n_{j-1}, B+1, n_{j+1}, ..., n_M\right), \left(r+1, n_1, ..., n_{j-1}, B+1, n_{j+1}, ..., n_M\right)\right) = \Lambda; \quad (3)$$

4) если существует $j \in \{1, ..., M\}$, такое, что $n_j > 0$,

$$q\left(\left(r, n_1, ..., n_{j-1}, n_j, n_{j+1}, ..., n_M\right), \left(r, n_1, ..., n_{j-1}, n_j-1, n_{j+1}, ..., n_M\right)\right) = \mu_j. \quad (4)$$

Упорядочим состояния цепи Маркова в лексикографическом порядке, под макросостоянием с номером $i$ будем понимать множество состояний $\boldsymbol{X}_i$ мощности $(B + 2)^M$, определяемое как

$$\boldsymbol{X_i} = \left\{(r, n_1, ..., n_M) \in \boldsymbol{X} : r = i\right\}.$$

Цепь Маркова $\{\boldsymbol{x}(t), t \geq 0\}$ является квазипроцессом размножения и гибели [7, 13], инфинитезимальный оператор $\boldsymbol{Q}$ цепи имеет блочно-диагональный вид:

$$\boldsymbol{Q} = \begin{pmatrix} \mathbf{B}_{00} & \mathbf{B}_{01} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{B}_{10} & \mathbf{A}_1 & \mathbf{A}_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{A}_2 & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{A}_2 & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \ddots & \ddots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

Матрицы $\mathbf{A}_0$, $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{B}_{00}$, $\mathbf{B}_{01}$, $\mathbf{B}_{10}$ суть квадратные матрицы порядка $(B + 2)^M$.

Матрицы $\mathbf{A}_0$, $\mathbf{A}_2$ задаются выражениями (2) и (3) соответственно, $\mathbf{B}_{10} = \mathbf{A}_0$, $\mathbf{B}_{01} = \mathbf{A}_2$. Выражения (1) и (4) определяют внедиагональные элементы матриц $\mathbf{A}_1$, $\mathbf{B}_{00}$; диагональные элементы матриц определяются из условий:

$$\mathbf{A}_0\mathbf{1} + \mathbf{A}_1\mathbf{1} + \mathbf{A}_2\mathbf{1} = \mathbf{0},$$

$$\mathbf{B}_{00}\mathbf{1} + \mathbf{B}_{01}\mathbf{1} = \mathbf{0},$$

где $\mathbf{1}$ обозначает единичный вектор-столбец.

Для вычисления стационарного распределения $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)$ воспользуемся аппаратом матрично-аналитических решений [7, 13], а именно матрично-геометрическим методом. Здесь $\boldsymbol{\pi}_i$, $i = 0, 1, \ldots$ есть вектор-строка, каждая компонента которого задает вероятность нахождения в некотором состоянии из макросостояния $\boldsymbol{X}_i$ в соответствии с введенным лексикографическим порядком.

Будем использовать следующие обозначения: $\boldsymbol{\pi}(\boldsymbol{x}) = \boldsymbol{\pi}(r, n_1, \ldots, n_M)$ – стационарная вероятность нахождения сети обслуживания в состоянии $\boldsymbol{x}$; $\boldsymbol{\pi}(\boldsymbol{X}_i)$ – стационарная вероятность нахождения сети в макросостоянии $\boldsymbol{X}_i$,

$$\boldsymbol{\pi}(\boldsymbol{X_i}) = \sum_{\boldsymbol{x} \in \boldsymbol{X_i}} \boldsymbol{\pi}(\boldsymbol{x}) = \boldsymbol{\pi}_i\mathbf{1}.$$

Для сети обслуживания стационарный режим будет существовать тогда и только тогда, когда выполнено условие [13]

$$\boldsymbol{\alpha}\mathbf{A}_0\mathbf{1} > \boldsymbol{\alpha}\mathbf{A}_2\mathbf{1},$$

где $\boldsymbol{\alpha}$ есть решение уравнения $\boldsymbol{\alpha}(\mathbf{A}_0 + \mathbf{A}_1 + \mathbf{A}_2) = \mathbf{0}$ с условием нормировки $\boldsymbol{\alpha}\mathbf{1} = 1$.

Известно, что тогда стационарное распределение имеет следующий вид:

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_1\mathbf{R}^{i-1}, i = 1, 2, ...,$$

где $\mathbf{R}$ есть решение уравнения $\mathbf{A_2} + \mathbf{RA_1} + \mathbf{R^2A_0} = \mathbf{0}$, векторы $\boldsymbol{\pi}_0$ и $\boldsymbol{\pi}_1$ находятся как решение уравнения

$$\begin{pmatrix} \boldsymbol{\pi}_0 & \boldsymbol{\pi}_1 \end{pmatrix} \begin{pmatrix} \mathbf{B}_{00} & \mathbf{B}_{01} \\ \mathbf{B}_{10} & \mathbf{A}_1 + \mathbf{RA}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \end{pmatrix},$$

с условием нормировки $\boldsymbol{\pi_0}\mathbf{1} + \boldsymbol{\pi_1}\mathbf{(I-R)^{-1}1} = 1$, здесь $\mathbf{I}$ – единичная матрица.

### 3. Вычисление стационарных характеристик

Используя стационарное распределение $\boldsymbol{\pi}$, определим математическое ожидание (м.о.) $\bar{N}_R$ числа требований в ИПВ:

$$\bar{N}_R = \sum_{i=1}^{\infty} i\boldsymbol{\pi}(\boldsymbol{X}_i) = \boldsymbol{\pi}_1 \sum_{i=1}^{\infty} i\mathbf{R}^{i-1}\mathbf{1} = \boldsymbol{\pi}_1(\mathbf{I}-\mathbf{R})^{-2}\mathbf{1}.$$

Обозначим через $(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_{(B+2)^M})$ вектор-столбец, составленный в соответствии с введенным лексикографическим порядком из всех элементов множества $\{(n_1, \ldots, n_M) : 0 \le n_i \le B+1\}$, который отображает все возможные способы распределения фрагментов по $M$ системам сети. Пусть $\boldsymbol{n} = (n_1, \ldots, n_M)$, через $\boldsymbol{\pi}(\boldsymbol{n})$ обозначим следующую стационарную вероятность:

$$\boldsymbol{\pi}(\boldsymbol{n}) = \sum_{r=0}^{\infty} \boldsymbol{\pi}(r, n_1, \ldots, n_M).$$

Положим $\mathbf{d} = (\boldsymbol{\pi}(\boldsymbol{n}_1), \ldots, \boldsymbol{\pi}(\boldsymbol{n}_{(B+2)^M}))$, тогда справедливо

$$\mathbf{d} = \sum_{i=0}^{\infty} \boldsymbol{\pi}_i = \boldsymbol{\pi}_0 + \boldsymbol{\pi}_1(\mathbf{I}-\mathbf{R})^{-1}.$$

Предположим, что сеть обслуживания находится в состоянии $(r, n_1, \ldots, n_M)$, тогда число требований, разделившихся на фрагменты, будет, очевидно, равно $\max\{n_1, \ldots, n_M\}$. Математическое ожидание $\bar{N}_S$ числа таких требований

$$\bar{N}_S = \sum_{i=0}^{\infty} \sum_{\boldsymbol{x} \in \boldsymbol{X}_i} \max\{n_1, \ldots, n_M\}\boldsymbol{\pi}(\boldsymbol{x}) = \mathbf{d}\left(\max\{\boldsymbol{n}_1\}, \ldots, \max\left\{\boldsymbol{n}_{(B+2)^M}\right\}\right).$$

Под длительностью пребывания требования в сети обслуживания будем понимать длительность интервала времени от момента разделения требования на фрагменты и распределения по системам до завершения обслуживания последнего из этих фрагментов. Тогда м.о. $\bar{T}_S$ длительности времени пребывания требований в сети обслуживания будет равно

$$\bar{T}_S = \bar{N}_S / \Lambda.$$

С учетом возможного пребывания в ИПВ м.о. $\bar{T}$ длительности интервала от поступления требования из источника до завершения его обслуживания будет равно

$$\bar{T} = \frac{\bar{N}_S + \bar{N}_R}{\Lambda}.$$

Обозначим через $b$ вероятность того, что требование, поступающее из источника, перейдет в ИПВ, тогда интенсивность $\Lambda_R$ потока требований из источника в ИПВ составит

$$\Lambda_R = b\Lambda,$$

$$b = \sum_{\substack{i=1 \\ \max\{\boldsymbol{n}_i\}=B+1}}^{(B+2)^M} \boldsymbol{\pi}(\boldsymbol{n}_i).$$

Требование, находящееся в ИПВ, осуществляет попытки занятия систем обслуживания сети, математическое ожидание $\bar{T}_R$ суммарной длительности времени пребывания требований в ИПВ, будет равно

$$\bar{T}_R = \frac{\bar{N}_R}{\Lambda_R}.$$

Для требования, находящегося в ИПВ, можно рассмотреть число повторных обращений к системам обслуживания. Обозначим через $\bar{K}$ математическое ожидание числа попыток захвата систем обслуживания. Будем исходить из следующих соображений: интенсивность входящего потока из источника в ИПВ равна $\Lambda_R$, требования из ИПВ обращаются к системам сети, интенсивность обращений к системам равна $\gamma(1-\boldsymbol{\pi}(\mathbf{X}_0))$. Тогда

$$\bar{K} = \frac{\gamma(1-\boldsymbol{\pi}(\boldsymbol{X}_0))}{\Lambda_R}.$$

## 4. Примеры

Рассмотрим изменение м.о. $\bar{T}$ длительности интервала времени от поступления требования из источника до завершения его обслуживания в зависимости от числа мест для ожидания в очереди ($B$ = 0, 1, 2) и интенсивности $\Lambda$ входящего потока. Результаты численных экспериментов изображены на рис. 2, здесь $M$ = 2, $\mu_1 = \mu_2 = 6$, $\gamma = 10$.

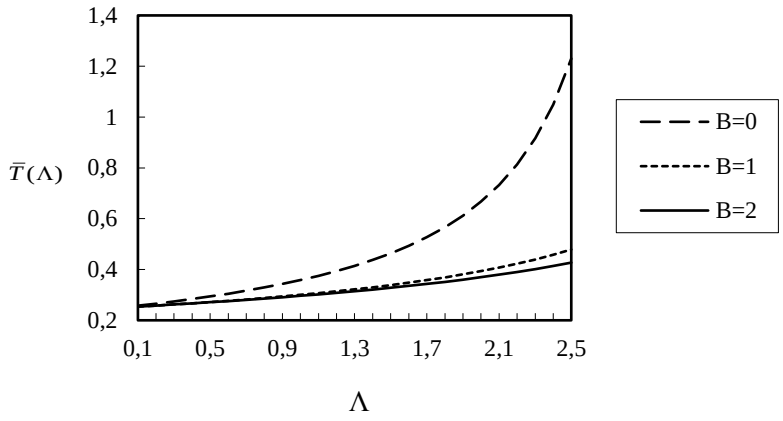


Рис. 2. Графики зависимости $\bar{T}(\Lambda)$ при 0,1 < $\Lambda$ < 2,5

В данном примере численно решим задачу оптимального выбора параметра $\gamma$ – интенсивности поступления требований из ИПВ, исходя из следующего: будем предполагать, что каждая неудачная попытка поступления из ИПВ приводит к выплате $C_K$; с другой стороны, $C_T$ есть плата за нахождение одного требования в ИПВ в течение единицы времени.

Тогда целевая функция $C_{KT}(\gamma)$ имеет вид:

$$C_{KT}(\gamma) = C_K \bar{K}(\gamma) + C_T \bar{T}_R(\gamma).$$

На рис. 3 представлен график целевой функции $C_{KT}(\gamma)$ для следующих параметров сети обслуживания: $\Lambda$ = 1, $M$ = 5, $\mu_1 = \ldots = \mu_5 = 2$, $B$ = 2, $C_K$ = 1, $C_T$ = 0,5.

При интенсивности поступления из ИПВ $\gamma_{KT}$ = 1,98, целевая функция достигает минимума и принимает значение $C_{KT}(\gamma_{KT})$ = 4,15.

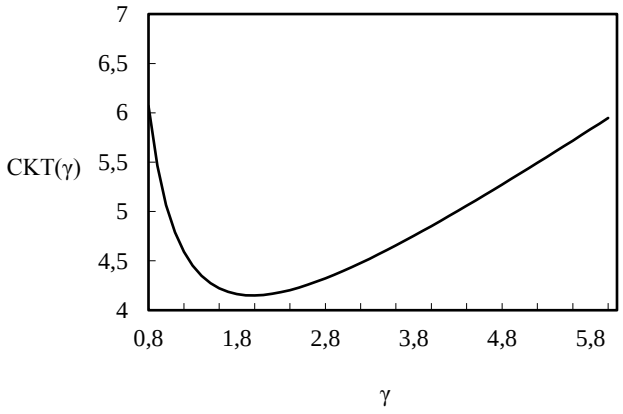


Рис. 3. График целевой функции $C_{KT}(\gamma)$ при 0,8 < $\gamma$ < 6

Пусть теперь целевая функция определена следующим образом:

$$C_{Kb}(\gamma) = C_K \bar{K}(\gamma) + C_b(1 - b(\gamma)),$$

т.е. каждая неудачная попытка поступления из ИПВ приводит к выплате $C_K$; с другой стороны, $C_b$ есть плата за нахождение всех очередей систем в незаполненном состоянии.

На рис. 4 представлен график целевой функции $C_{Kb}(\gamma)$ для следующих параметров сети обслуживания: $\Lambda = 1$, $M = 5$, $\mu_1 = \ldots = \mu_5 = 2$, $B = 2$, $C_K = 1$, $C_b = 50$.

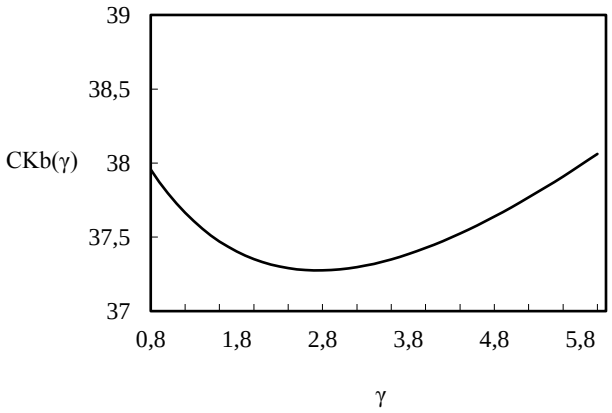


Рис. 4. График целевой функции $C_{Kb}(\gamma)$ при $0{,}8 < \gamma < 6$

Минимальное значение целевой функции достигается при интенсивности поступления из ИПВ $\gamma_{Kb} = 2{,}74$, $C_{Kb}(\gamma_{Kb}) = 37{,}27$.

## Заключение

В статье рассмотрена RQ-сеть массового обслуживания с делением и слиянием требований. Получены выражения для основных стационарных характеристик сети обслуживания. Приведены примеры и рассмотрены задачи численной оптимизации.

Представленная в работе сеть массового обслуживания может применяться в качестве модели многопроцессорных вычислительных систем, а также других систем с параллельным и распределенным принципами функционирования.

ЛИТЕРАТУРА

1. Narahari Y., Sundarrajan P. Performability analysis of fork-join queueing systems // Journal of the Operational Research Society. 1995. V. 6, No. 10. P. 1237–1249.
2. Flatto L., Hahn S. Two parallel queues created by arrivals with two demands I // SIAM Journal of Applied Mathematics. 1984. V. 4, No. 5. P. 1041–1053.
3. Nelson R., Tantawi A.N. Approximate analysis of fork/join synchronization in parallel queues // IEEE Trans. Comp. 1988. V. 37, No. 6. P. 739–743.
4. Ko S.-S., Serfozo R.F. Sojourn times in G/M/1 fork-join networks // Naval Research Logistics (NRL). 2008. V. 55, No.5. P. 432–443.
5. Ko S.-S., Serfozo R.F. Response times in M/M/s fork-join networks // Adv. Appl. Prob. 2004. V. 36, No. 3. P. 432–443.
6. Thomassian A. Analysis of fork/join and related queueing systems // ACM Computing Surveys. 2014. V. 47, No. 2. P. 17:1–17:71.
7. Neuts M. Matrix-Geometric Solutions in Stochastic Models: an Algorithmic Approach. Baltimore : The Johns Hopkins University Press, 1981. 352 p.
8. Klimenok V.I., Savko R.Ch. Tandem system with retrials and impatient customers // Automation and Remote Control. 2015. V. 76, No. 8. P. 1387–1399.
9. Назаров А.А., Моисеева С.П. Метод асимптотического анализа в теории массового обслуживания. Томск : Изд-во НТЛ, 2006. 112 с.
10. Nazarov A.A., Semenova I.A. Asymptotic analysis of retrial queueing systems // Optoelectronics, Instrumentation and Data Processing. 2011. V. 47, No. 4. P. 406–413.
11. Falin G.I., Tempelton J.G.C. Retrial queues. London : Chapman & Hall, 1997. 328 p.
12. Artalejo J.R., Gomez-Corral A. Retrial Queueing Systems. A Computational Approach. Springer, 2008. 332 p.
13. He Q.-M. Fundamentals of Matrix-Analytic Methods. New York : Springer, 2014. 349 p.

Поступила в редакцию 29 декабря 2017 г.

Osipov O.A. (2018) ANALYSIS OF FORK/JOIN QUEUEING NETWORKS WITH RETRIALS. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie vychislitelnaja tehnika i informatika* [Tomsk State University Journal of Control and Computer Science]. 43. pp. 49–55

DOI: 10.17223/19988605/43/6

In this paper, we consider a fork/join queueing network with retrials. Jobs arrive at the network according to a Poisson process with rate $\Lambda$. Each node operates like a /M/1/B queueing system under a FCFS discipline.

When each node has less then B+1 tasks, an arriving job is split into $M$ tasks which are simultaneously assigned to the $M$ nodes. The tasks are serviced independently, and their service times at node $i$ have an exponential distribution with rate $\mu_i$, $i = 1, \ldots, M$. When all of its $M$ tasks are finished, the job is completed and exits the network.

In other case, the job goes to the retrial orbit to retry for service after a random time. The retrial policy is assumed to be independent of the number of jobs in the orbit, i.e., a constant retrial policy.

We represent the state of the network over time by the stochastic process $\boldsymbol{x}(t) = (r(t), n_1(t), \ldots, n_M(t))$, where $r(t)$ denotes the number of jobs in orbit at time $t$, $n_i(t)$ denotes the number of tasks in node $i$ at time $t$. This is an $(M + 1)$-dimensional continuous time Markov chain on the state space $\boldsymbol{X}$,

$$\boldsymbol{X} = \{(r, n_1, \ldots, n_M) : r \geq 0, 0 \leq n_i \leq B+1, i = 1, \ldots, M\}.$$

Applying a matrix-geometric approach, we obtain the stationary distribution of the number of jobs in the network under exponential assumptions. Using the distribution, we determine performance measures. Finally, some numerical examples and a section of conclusions commenting the main research contributions of this paper are presented.

The results can be used for the performance analysis of multiprocessor systems and other modern distributed systems.

Keywords: fork/join queueing networks; retrial queues; matrix-geometric method; distributed computing systems.

*OSIPOV Oleg Alexandrovich* (National Research Saratov State University, Russian Federation).
E-mail: oleg.alex.osipov@gmail.com

REFERENCES

1. Narahari, Y. & Sundarrajan, P. (1995) Performability analysis of fork-join queueing systems. *Journal of the Operational Research Society*. 6(10). pp. 1237–1249. DOI: 10.1057/jors.1995.17
2. Flatto, L. & Hahn, S. (1984) Two parallel queues created by arrivals with two demands I. *SIAM Journal of Applied Mathematics*. 44(5). pp. 1041–1053. DOI: 10.1137/0144074
3. Nelson, R. & Tantawi, A.N. (1988) Approximate analysis of fork/join synchronization in parallel queues. *IEEE Trans. Comp*. 37(6). pp. 739–743. DOI: 10.1109/12.2213
4. Ko, S.-S. & Serfozo, R.F. (2008) Sojourn times in G/M/1 fork-join networks. *Naval Research Logistics (NRL)*. 55(5). pp. 432–443. DOI: 10.1002/nav.20294
5. Ko, S.-S. & Serfozo, R.F. (2004) Response times in M/M/s fork-join networks. *Advances in Applied Probability*. 36(3). pp. 432–443. DOI: 10.1017/S000186780001315X
6. Thomassian, A. (2014) Analysis of fork/join and related queueing systems. *ACM Computing Surveys*. 47(2). pp.17:1-17:71.
7. Neuts, M. (1981) *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Baltimore: The Johns Hopkins University Press.
8. Klimenok, V.I. & Savko, R.Ch. (2015) Tandem system with retrials and impatient customers. *Automation and Remote Control*. 76(8). pp. 1387–1399. DOI: 10.1134/S0005117915080056
9. Nazarov, A.A. & Moiseeva, S.P. (2006) *Metod asimptoticheskogo analiza v teorii massovogo obsluzhivaniya* [Method of asymptotic analysis in queuing theory]. Tomsk: NTL.
10. Nazarov, A.A. & Semenova, I.A. (2011) Asymptotic analysis of retrial queueing systems. *Optoelectronics, Instrumentation and Data Processing*. 47(4). DOI: 10.3103/S8756699011040121
11. Falin, G.I., Tempelton, J.G.C. (1997) *Retrial queues*. London: Chapman & Hall.
12. Artalejo, J.R. & Gomez-Corral, A. (2008) *Retrial Queueing Systems. A Computational Approach*. Springer.
13. He, Q.-M. (2014) *Fundamentals of Matrix-Analytic Methods*. New York: Springer.

.