



УДК 519.21

DOI: 10.22363/2312-9735-2018-26-1-13-27

## Обзор систем параллельной обработки заявок. Часть II

А. В. Горбунова\*, И. С. Зарядов\*<sup>†</sup>, К. Е. Самуйлов\*

\* *Кафедра прикладной информатики и теории вероятностей  
Российский университет дружбы народов  
ул. Миклуто-Маклая, д. 6, Москва, Россия, 117198*

<sup>†</sup> *Институт проблем информатики  
Федеральный исследовательский центр «Информатика и управление» РАН  
ул. Вавилова, д. 44, кор. 2, Москва, Россия, 119333*

Данная работа является продолжением обзора методов исследования системы массового обслуживания вида «fork-join» (в западной классификации) или системы с расщеплением запросов. Интерес к рассматриваемой системе объясняется широким спектром задач, которые могут быть решены с её помощью, поскольку фактически речь идёт о параллельной обработке данных и их приложениях. К примеру, это может касаться анализа работы дисковых массивов, облачных вычислений, высокопроизводительных сервисов и даже процесса комплектации заказов на складе. Если в первой части обзора были описаны особенности построения данной модели и родственных ей систем, а также приведено подробное описание подхода к получению точного выражения среднего времени отклика в случае двух приборов и представлено несколько методов приближенного анализа данной характеристики в случае, когда число приборов больше двух, то во второй части обзора представлено описание других существующих методов аппроксимации среднего времени отклика. В частности, к рассматриваемым подходам приближенного анализа времени отклика относятся: матрично-геометрический метод, анализ с помощью порядковых статистик для различных типов распределения времени пребывания подзапросов.

**Ключевые слова:** система массового обслуживания, расщепление заявок, параллельное обслуживание, параллельная обработка, время отклика, время синхронизации, матрично-геометрический метод, порядковые статистики

### 1. Введение

Наряду с такими методами аппроксимации среднего времени отклика системы с расщеплением запросов, как эмпирический подход и интерполяция с помощью предельных значений загрузки системы [1], отдельного внимания заслуживают матрично-геометрический метод и анализ с помощью порядковых статистик, которому посвящена большая часть статьи, поскольку известные результаты теории порядковых статистик позволяют расширить ряд изучаемых распределений для входящего потока и времени обслуживания, а также оценить не менее важные характеристики для рассматриваемой системы: математическое ожидание и дисперсию времени синхронизации. Под временем синхронизации здесь фактически подразумевается её задержка, т.е. время между окончанием обслуживания первого и последнего подзапросов одного запроса [2].

Статья организована следующим образом: в разделе 2 описываются особенности и результаты применения матрично-геометрического метода к анализу системы с расщеплением запросов, в разделе 3 представлен подход к анализу среднего времени отклика, основанный на теории порядковых статистик, в заключении кратко подведены итоги работы.

---

Статья поступила в редакцию 5 ноября 2017 г.

Публикация подготовлена при поддержке Программы РУДН «5-100» и при финансовой поддержке РФФИ в рамках научных проектов № 18-07-00576, 16-07-00766.

## 2. Анализ с помощью матрично-геометрического метода

Идея использования матрично-геометрического подхода при анализе fork-join систем в большинстве работ сводится к рассмотрению моделей, аппроксимирующих исходную, непосредственный анализ которых позволяет сделать вывод, что изучаемый процесс, описывающий поведение системы, является квазипроцессом «размножения и гибели» со всеми вытекающими из этого следствиями [3–5]. То же касается и fork-join системы с конечными очередями к серверам в условиях пуассоновского входящего потока и экспоненциальных времён обслуживания. Так, в [6, 7] был предложен алгоритм, позволяющий свести матрицу вероятностей переходов для марковского процесса, описывающего поведение системы, к блочно-диагональному виду, что даёт возможность компактно записать систему уравнений равновесия (СУР) и решить её одним из известных численных методов и, как следствие, вычислить среднее время отклика.

Матрично-геометрический подход для оценки времени отклика fork-join системы используется в серии статей [8–10]. В первой работе [8] были предложены две модели, аппроксимирующие исходную fork-join систему с  $K$  ветвями  $M|M|1$  и различными интенсивностями обслуживания на разных серверах (приборах):

- 1) поступающий запрос принимается в систему тогда и только тогда, когда число подзапросов  $k$ -го типа, находящихся в системе, меньше некоторого фиксированного положительного целого числа  $U_k$ :  $n_k < U_k$ ,  $k = \overline{2, K}$ , в противном случае он теряется;
- 2) отказ в обслуживании на  $j$ -м сервере происходит тогда и только тогда, когда нарушается условие:  $n_i - n_j \neq U_{i,j}$  для некоторого  $i$  и для фиксированных положительных целых чисел  $U_{i,j}$ ,  $\forall i, j, i \neq j$ , причём сервер  $j$  блокируется до тех пор, пока подзапрос не закончит обслуживаться на сервере  $i$ .

Здесь вектор  $\vec{n} = (n_1, \dots, n_k, \dots, n_K)$  отражает состояние системы,  $n_k$  — число подзапросов  $k$ -го типа в системе,  $k = \overline{1, K}$ .

В результате анализа обеих систем с помощью матрично-геометрического метода удаётся определить стационарное распределение числа подзапросов в системах  $\pi(\vec{n})$ . Далее полученное распределение используется при алгоритмическом решении задачи нахождения функции распределения времени отклика:

$$F_{W_{K, \max}}(x) = \sum_{\vec{n}} \pi(\vec{n}) F_{W_{K, \max}^{\vec{n}}}(x), \quad x > 0.$$

Таким образом, задача сводится к поиску условного распределения времени отклика  $F_{W_{K, \max}^{\vec{n}}}(x)$  при условии, что система находится в состоянии  $\vec{n}$ . Для первой модифицированной модели искомое условное распределение может быть выражено с помощью распределения Эрланга, а во втором случае условное распределение получается при анализе процесса поглощения для подзапросов. Кроме того, авторы доказали, что решения, полученные при анализе этих двух систем, являются верхней и нижней границами для функции распределения времени отклика соответственно.

В следующей статье авторов [9] изучается fork-join система с пуассоновским входящим потоком, но при этом время обслуживания на однородных серверах имеет распределение Кокса. Отметим, что для стационарных вероятностей состояний в СМО  $M|C_k|1$  известно точное решение [11]. Как и в предыдущей работе рассматриваются две аналогичные, аппроксимирующие исходную систему модели:

- 1) модель, в которой все очереди кроме очереди к первому серверу имеют конечную ёмкость;
- 2) модель, в которой вводится ограничение на то, что разница между каждой парой длин очередей к серверам не должна превышать заданный порог.

Матрично-геометрический анализ указанных систем, предложенный в [12], благодаря тому, что матрицы вероятностей перехода сводятся к блочно-треугольному

виду, позволяет получить стационарные вероятности, которые используются для вывода распределения времени отклика, а, соответственно, и моментов этого распределения. Полученные выражения являются верхними и нижними границами для моментов времени отклика. Следующая работа [10] продолжает начатые исследования. Авторы предлагают алгоритмический метод для вычисления верхних и нижних оценок производительности системы, который, по их мнению, может быть расширен для анализа систем с отличными от пуассоновского входящим потоком и отличным от экспоненциального временем обслуживания на разнородных серверах.

Отметим, что матрично-геометрический метод (МГМ) использовался и при анализе систем, родственных fork-join. Так, МГМ был применён для анализа модели независимых серверов (ISM), в [13] с помощью МГМ анализируется split-merge система [1] с двумя ветвями с наложением ограничения на ёмкость накопителя одной из очередей. Были получены выражения для математического ожидания, дисперсии и функции распределения времени отклика. Также был разработан метод аппроксимации для систем с более чем с двумя ветвями. В работах [14, 15] также исследуется fork-join система, но уже с распределением фазового типа для входящего потока и с потерями. В статье [16] рассматривается ещё одна вариация классической fork-join системы, для анализа которой используется МГМ: обслуживание каждого подзапроса происходит в несколько этапов, включая фазу ожидания и фазу объединения родственных подзапросов; существует динамическая политика планирования нагрузки системы для поддержания эффективного использования сервера.

### 3. Анализ с помощью порядковых статистик

Поскольку время отклика fork-join системы классически определяется как максимум, а в некоторых случаях и как минимум из  $K$  случайных величин времён пребывания подзапросов в системе [17], то естественно, что одним из альтернативных способов оценки среднего времени отклика является использование теории порядковых статистик [18–20]. По определению [18–20], если  $\xi_1, \dots, \xi_K$  — конечная выборка, определённая на некотором вероятностном пространстве  $(\Omega, F, P)$ , и для  $\omega \in \Omega$  :  $x_i = \xi_i(\omega)$ ,  $i = 1, \dots, K$ , и далее, если перенумеровать последовательность  $\{x_i\}_{i=1}^K$  в порядке неубывания таким образом, что  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(K-1)} \leq x_{(K)}$ , то такая последовательность будет называться вариационным рядом, а его члены — порядковыми статистиками. Случайная величина же  $\xi_{(k)} : \xi_{(k)}(\omega) = x_{(k)}$  называется  $k$ -й порядковой статистикой исходной выборки. Из определения ясно, что

$$\xi_{(1)} = \min(\xi_1, \dots, \xi_K), \quad \xi_{(K)} = \max(\xi_1, \dots, \xi_K). \quad (1)$$

Таким образом, время отклика  $W_{K,\max} = \xi_{(K)}$  или  $W_{K,\min} = \xi_{(1)}$  в терминах теории порядковых статистик, где  $\xi_k$ ,  $k = \overline{1, K}$  — положительные случайные величины времён пребывания подзапросов в системе до попадания в буфер синхронизации. Следовательно, математическое ожидание времени отклика можно вычислить, зная распределения экстремальных значений  $\xi_{(1)}$  и  $\xi_{(K)}$ , причём функция распределения второй случайной величины фактически является совместной функцией распределения случайных величин  $\xi_1, \dots, \xi_K$ :

$$F_{\xi_{(K)}}(x) = P(\max(\xi_1, \dots, \xi_K) < x) = P(\xi_1 < x, \dots, \xi_K < x),$$

$$F_{\xi_{(1)}}(x) = P(\min(\xi_1, \dots, \xi_K) \leq x) = 1 - P(\xi_1 > x, \dots, \xi_K > x).$$

Обозначим через  $F_k(x)$  — функцию распределения, а через  $f_k(x)$  — плотность распределения случайной величины  $\xi_k$ ,  $k = \overline{1, K}$ . Если сделать допущение о том, что  $\xi_1, \dots, \xi_K$  — независимые, что в нашем случае, как уже говорилось выше, является упрощающим предположением, то:

$$F_{\xi_{(K)}}(x) = P(\xi_1 < x, \dots, \xi_K < x) = \prod_{i=1}^K F_i(x),$$

$$F_{\xi_{(1)}}(x) = 1 - P(\xi_1 > x, \dots, \xi_K > x) = 1 - \prod_{i=1}^K [1 - F_i(x)],$$

а  $n$ -й момент случайной величины времени отклика можно определить, вычислив интеграл:

$$E[W_{K,\max}^n] \approx E[\xi_{(K)}^n] = \int_0^{\infty} x^n f_{\xi_{(K)}}(x) dx, \quad E[W_{K,\min}^n] \approx E[\xi_{(1)}^n] = \int_0^{\infty} x^n f_{\xi_{(1)}}(x) dx,$$

где

$$f_{\xi_{(K)}}(x) = \sum_{j=1}^K \frac{f_j(x)}{F_j(x)} \prod_{i=1}^K F_i(x), \quad f_{\xi_{(1)}}(x) = \sum_{j=1}^K f_j(x) \prod_{i=1}^K \frac{1 - F_i(x)}{1 - F_j(x)}.$$

Далее, если предположить, что серверы являются однородными, т.е. случайные величины  $\xi_1, \dots, \xi_K$  не только независимы, но и одинаково распределены, а, следовательно, их функции и плотности распределения равны между собой:  $F_k(x) = F(x)$ ,  $f_k(x) = f(x)$ ,  $k = \overline{1, K}$ , то

$$P(\xi_1 < x, \dots, \xi_K < x) = F^K(x), \quad 1 - P(\xi_1 > x, \dots, \xi_K > x) = 1 - (1 - F(x))^K,$$

и, соответственно,

$$E[\xi_{(K)}^n] = K \int_0^{\infty} x^n f(x) F^{K-1}(x) dx, \quad E[\xi_{(1)}^n] = K \int_0^{\infty} x^n f(x) (1 - F(x))^{K-1} dx. \quad (2)$$

Также заметим, что математическое ожидание максимума  $K$  н.о.р.с.в. можно представить в виде [21]:

$$E[\xi_{(K)}] = \int_0^{\infty} x dF^K(x) = \int_0^{\infty} [1 - F^K(x)] dx. \quad (3)$$

Для того чтобы проанализировать время отклика, необходимо вычислить указанные интегралы. В зависимости от типа распределения для оценки интегралов могут быть применены численные методы, однако для таких распределений, как экспоненциальное, гиперэкспоненциальное и распределение Эрланга 2-го порядка, распределение Кокса, результат может быть получен в символьном виде [21–23]. Вычислительные затраты увеличиваются с ростом ветвей ( $K$ ) и увеличением порядка для распределений Эрланга или Кокса. Для уменьшения вычислительных затрат при вычислении максимума может быть использован характеристический максимум [24].

### 3.1. Экспоненциальное распределение

Допустим, что времена пребывания подзапросов в подсистемах являются независимыми экспоненциально распределёнными случайными величинами с плотностями распределения  $f_k(x) = \lambda_k e^{-\lambda_k x}$ ,  $x > 0$ ,  $k = \overline{1, K}$ .

Для того, чтобы определить моменты высшего порядка максимума  $K$  независимых экспоненциальных случайных величин, эффективнее с вычислительной точки зрения может быть дифференцирование соответствующее число раз преобразования Лапласа–Стилтьеса (ПЛС) совместной функции распределения максимума  $\pi_{\xi_{(K)}}(s)$  по  $s$  и в последующем приравнивание  $s$  к нулю. В частности, для случая  $K = 2$  функция распределения, плотность распределения и ПЛС [25, 26] равны:

$$F_{\xi_{(2)}}(x) = F_1(x)F_2(x) = 1 - e^{-\lambda_1 x} - e^{-\lambda_2 x} + e^{-(\lambda_1 + \lambda_2)x},$$

$$\begin{aligned} f_{\xi_{(2)}}(x) &= \frac{dF_2(x)}{dx} = f_1(x)F_2(x) + f_2(x)F_1(x) = \\ &= \lambda_1 e^{-\lambda_1 x} + \lambda_2 e^{-\lambda_2 x} - (\lambda_1 + \lambda_2) e^{-(\lambda_1 + \lambda_2)x}, \end{aligned}$$

$$\pi_{\xi_{(2)}}(s) = \int_0^{\infty} f_2(x)e^{-sx} dx = \frac{\lambda_1}{s + \lambda_1} + \frac{\lambda_2}{s + \lambda_2} - \frac{\lambda_1 + \lambda_2}{s + \lambda_1 + \lambda_2}.$$

Тогда математическое ожидаемое максимума равно [21, 23, 25, 26]:

$$E\xi_{(2)} = - \left. \frac{d\pi_2(s)}{ds} \right|_{s=0} = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} - \frac{1}{\lambda_1 + \lambda_2}.$$

Общее выражение для ПЛС максимума  $K$  экспоненциально распределённых случайных величин может быть записано в следующем виде [27]:

$$\pi_{\xi_{(K)}}(s) = \sum_{i=1}^K (-1)^{i-1} \sum_{j=1}^i \sum_{m=1}^j \frac{\sum_{l=1}^i \lambda_{(m+l-1)_K}}{s + \sum_{l=1}^i \lambda_{(m+l-1)_K}},$$

где  $(m+l-1)_K$  обозначает сложение по модулю  $K$ . Кроме того, ПЛС максимума  $K$  независимых экспоненциально распределённых случайных величин с параметрами  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$  и плотностью распределения  $f_{\xi_{(K)}}(\boldsymbol{\lambda}, x)$  с ПЛС  $\pi_{\xi_{(K)}}(\boldsymbol{\lambda}, s)$  можно представить с помощью рекуррентной формулы [28]:

$$\left( s + \sum_{j=1}^k \lambda_j \right) \pi_{\xi_{(k)}}(\boldsymbol{\lambda}, s) = \sum_{j=1}^k \lambda_j \pi_{\xi_{(k-1)}}(\boldsymbol{\lambda}_{/j}, s), \quad 1 \leq k \leq K,$$

где « $/j$ » означает исключение  $\lambda_j$ , т.е.  $\boldsymbol{\lambda}_{/j} = (\lambda_1, \dots, \lambda_{j-1}, \lambda_{j+1}, \dots, \lambda_m)$  и  $\pi_0(\mathbf{0}, s) = 1$ . Тогда момент  $n$ -го порядка максимума  $K$  независимых экспоненциально распределённых случайных величин равен:

$$M_K(\boldsymbol{\lambda}, n) = \left( nM_K(\boldsymbol{\lambda}, n-1) + \sum_{j=1}^K M_{K-1}(\boldsymbol{\lambda}_{/j}, n) \right) \left( \sum_{j=1}^K \lambda_j \right)^{-1}, \quad (4)$$

$K \geq 1$ ,  $M_0(\mathbf{0}, n) = 0$  для всех  $n \geq 1$  и  $M_K(\boldsymbol{\lambda}, 0) = 1$  для всех  $K \geq 0$ .

Рассмотрим ещё одну важную характеристику производительности системы облачных вычислений — время синхронизации. Время синхронизации определяется как

время между поступлением первого и последнего подзапросов одного запроса в буфер синхронизации, иными словами, это разность между максимумом и минимумом из времён пребывания подзапросов в системе:

$$E[W_K] = E[W_{K,\max}] - E[W_{K,\min}].$$

В теории порядковых статистик данная величина называется размахом [18–20]. Одни из первых результатов анализа этого показателя были приведены в работе [2], а именно численное решение оптимизационной задачи по минимизации среднего времени синхронизации. В статье [25] были представлены оценки математического ожидания в случае неоднородных приборов, а в случае однородных приборов ( $\lambda_k = \lambda, k = \overline{1, K}$ ) получена оценка и для дисперсии времени синхронизации:

$$D[W_K] \approx \frac{1}{\lambda^2} \sum_{i=1}^{K-1} \frac{1}{i^2}.$$

Вычислить дисперсию исследуемой величины, для определения которой недостаточно знать значения дисперсий случайных величин  $W_{K,\max}$  и  $W_{K,\min}$ , можно и в неэкспоненциальном случае, действуя аналогичным образом, а именно пользуясь известными в теории порядковых статистик формулой для функции распределения размаха положительных н.о.р.с.в.:

$$F_{W_K}(x) \approx F_{\xi_{(K)} - \xi_{(1)}}(x) = \sum_{i=1}^K \int_0^{\infty} f_i(y) \prod_{j=1, j \neq i}^K [F_j(y+x) - F_j(y)] dy, \quad (5)$$

либо непосредственно формулой для вычисления дисперсии размаха выборки объёма  $K$  н.о.р.с.в. с функцией распределения  $F(x)$  [18]:

$$\begin{aligned} D[W_K] &\approx D[\xi_{(K)} - \xi_{(1)}] = \\ &= 2 \int_{-\infty}^{\infty} (1 - F^K(y) - [1 - F(x)]^K + [F(y) - F(x)]^K) dx dy - (E[W_K])^2. \end{aligned} \quad (6)$$

### 3.2. Распределение Эрланга

Формула для аппроксимации среднего времени отклика fork-join системы с *распределением Эрланга*  $l_k$ -го порядка ( $\varepsilon_{l_k}$ ), которое представляет собой сумму  $l_k$  независимых случайных величин, распределённых по одному и тому же экспоненциальному закону с параметром  $\lambda_k$  для времени пребывания в  $k$ -й ветви системы,  $k = \overline{1, K}$  с функцией и плотностью распределения:

$$F(x) = 1 - e^{-\lambda_k x} \sum_{i=0}^{l_k-1} \frac{(\lambda_k x)^i}{i!}, \quad f(x) = \frac{\lambda_k (\lambda_k x)^{l_k-1}}{(l_k-1)!} e^{-\lambda_k x},$$

где  $\lambda_k \geq 0; l_k = 1, 2, \dots; x \geq 0$ , а математическое ожидание равно  $l_k/\lambda_k$ , представлена в [29]:

$$E[W_{K,\max}] \approx \int_0^{\infty} \left[ 1 - \prod_{k=1}^K \left( 1 - e^{-\lambda_k x} \sum_{i=0}^{l_k-1} \frac{(\lambda_k x)^i}{i!} \right) \right] dx.$$

Для случая  $K = 2$ :

$$E[W_{2,\max}] \approx \sum_{k=1}^2 \frac{l_k}{\lambda_k} - \sum_{m=0}^{l_1-1} \sum_{n=0}^{l_2-1} \binom{m+n}{m} \frac{\lambda_1^m \lambda_2^n}{(\lambda_1 + \lambda_2)^{m+n+1}}.$$

### 3.3. Распределение экстремальных значений

Ещё один метод вычисления максимума н.о.р.с.в. — это его аппроксимация распределением экстремальных значений [30]. Распределение экстремальных значений представляет собой предельное распределение наибольшего  $\xi_{(K)}$  (наименьшего  $\xi_{(1)}$ ) из значений н.о.р. непрерывных с.в. при бесконечном увеличении их числа  $K \rightarrow \infty$ , т.е., говоря иными словами, это предельное распределение наибольшего (наименьшего) выборочного значения при бесконечном увеличении объёма выборки из непрерывного распределения [30], хотя, конечно, нельзя сказать, что распределение экстремального значения — это распределение  $\xi_{(K)}$  при  $K \rightarrow \infty$ , поскольку фактически оно будет вырожденным, поэтому, чтобы получить невырожденное предельное распределение, рассматривают линейное преобразование  $\xi_{(K)}$  с коэффициентами, зависящими от объёма выборки, что в действительности аналогично нормировке, но по большому счету не исчерпывается только последовательностями линейных преобразований [30].

Существует три типа семейств экстремальных распределений [30]:

- 1) тип I — распределение типа Гумбеля, которое ещё иногда называют дважды экспоненциальным:

$$F_{\xi_{(K)}}(x) = \exp\left(-e^{-\frac{x-\alpha}{\theta}}\right), \quad (7)$$

- 2) тип II — распределение типа Фреше:

$$F_{\xi_{(K)}}(x) = \begin{cases} 0, & x < \alpha, \\ \exp\left[-\left(\frac{x-\alpha}{\theta}\right)^{-n}\right], & x \geq \alpha, \end{cases}$$

- 3) тип III — распределение типа Вейбулла:

$$F_{\xi_{(K)}}(x) = \begin{cases} \exp\left[-\left(\frac{x-\alpha}{\theta}\right)^n\right], & x < \alpha, \\ 0, & x \geq \alpha, \end{cases}$$

где  $\alpha, \theta > 0$  и  $n > 0$  — параметры. Распределения типа II и III приводятся к распределению типа I с помощью преобразований  $(\ln(\xi_{(K)} - \alpha))$  и  $(-\ln(\xi_{(K)} - \alpha))$ , соответственно, поэтому в основном акцент делается на распределении типа Гумбеля, плотность распределения которого имеет вид:

$$f_{\xi_{(K)}}(x) = \theta^{-1} e^{-\frac{x-\alpha}{\theta}} \exp\left(-e^{-\frac{x-\alpha}{\theta}}\right),$$

а математическое ожидание и дисперсия равны:

$$E[\xi_{(K)}] = \alpha + \gamma\theta, \quad D[\xi_{(K)}] = \frac{\pi^2\theta^2}{6},$$

где  $\gamma \approx 0,577721$  — это постоянная Эйлера–Маскерони. При  $\alpha = 0$  и  $\theta = 1$  получается стандартная форма распределения типа I — распределение Гумбеля:

$$F_{\xi_{(K)}}(x) = \exp(-e^{-x}), \quad f_{\xi_{(K)}}(x) = \exp(-x - e^{-x}). \quad (8)$$

В [30] упоминается об установленной связи между свойствами генерального распределения  $F(x)$  и типом предельного распределения. Полученные условия, которым, вообще говоря, удовлетворяют далеко не все распределения (например, распределения с «тяжёлыми хвостами»), являются необходимыми и достаточными для сходимости к одному из трёх типов семейств экстремальных значений и касаются поведения  $F(x)$  при больших (малых) значениях  $x$ , если речь идёт о наибольших (наименьших) значениях случайных величин. Причём при одном и том же исходном распределении наибольшее и наименьшее значения могут иметь предельные распределения, относящиеся к разным типам. К распределениям, которые удовлетворяют условию сходимости к типу I можно отнести *нормальное, экспоненциальное и логистическое*, к типу II — *распределение Коши*, а к типу III — распределения, сосредоточенные на ограниченной сверху части числовой оси [30]. Так, в [20] было показано, что распределение максимального значения  $\xi_{(K)}$  в выборке из случайных величин с экспоненциальным распределением с параметром  $\lambda = 1$  приближается к *распределению Гумбеля* в его стандартной форме с ростом объёма выборки, т.е.

$$\xi_{(K)} \approx \ln K + \zeta, \quad E[\xi_{(K)}] \approx \ln K + \gamma, \quad D[\xi_{(K)}] \approx D[\zeta] \approx \pi^2/6,$$

где  $\zeta$  — это случайная величина с распределением Гумбеля с функцией распределения из (8), математическим ожиданием  $E[\zeta] = \gamma$  и дисперсией  $D[\zeta] = \pi^2/6$ .

Отметим, что для случая, когда каждая порядковая статистика имеет стандартное распределение экстремальных значений типа I из (8), математическое ожидание максимума вычисляется по формуле [30]:  $E[\xi_{(K)}] = \alpha + \gamma\theta + \theta \ln K$ .

### 3.4. Произвольное распределение

Рассмотрим обобщение вывода в [28] для аппроксимации моментов максимума случайных величин с произвольным распределением [4]. Рассмотрим случайную величину  $\xi = \max(\xi_1, \xi_2)$ , где  $\xi_k$  — неотрицательные и независимые с функциями распределения  $F_k(x)$ , а  $\pi_k(s)$  — это ПЛС этих функций распределения,  $k = 1, 2$ . Тогда математическое ожидание  $\xi$  можно выразить следующим образом:

$$\begin{aligned} E[\xi] &= E[\xi_1] + P(\xi_2 > \xi_1) \cdot E[\xi_2 - \xi_1 | \xi_2 > \xi_1] = \\ &= E[\xi_1] + \int_0^{\infty} F_1(x) dF_2(x) \cdot E[\xi_2 - \xi_1 | \xi_2 > \xi_1]. \end{aligned}$$

Пусть  $\mu_1$  и  $\sigma_1$  — первые два момента случайной величины  $\xi_1$ , а  $\mu_2$  и  $\sigma_2$  — случайной величины  $\xi_2$ . Тогда, если  $\xi_1$  — экспоненциальная, то

$$E[\xi_2 - \xi_1 | \xi_2 > \xi_1] = \frac{\sigma_2}{2\mu_2}.$$

Далее окончательно получаем следующую аппроксимацию:

$$E[T] = \mu_1 + \frac{\sigma_2 \pi_1(\mu_2^{-1})}{2\mu_2},$$



которая фактически является точным результатом для случая, когда обе случайные величины  $\xi_1$  и  $\xi_2$  имеют экспоненциальное распределение. Таким образом, в общем случае, когда речь идёт о  $K$  случайных величинах, необходимо аппроксимировать ПЛС функции распределения максимума  $(K - 1)$  случайных величин. Поэтому допустим, что уравнение (4) применимо не только к экспоненциальным случайным величинам. Обозначим  $M(K, \lambda, \sigma)$  среднее значение максимума  $K$  случайных величин с математическими ожиданиями  $\mu = (\mu_1, \dots, \mu_K)$ ,  $\lambda = (\mu_1^{-1}, \dots, \mu_K^{-1})$  и вторыми моментами  $\sigma = (\sigma_1, \dots, \sigma_K)$ . Тогда  $M(K, \lambda, \sigma)$  можно выразить рекуррентно:

$$M(k, \lambda, \sigma) = \frac{1}{k} \sum_{i=1}^k M(k-1, \lambda/i, \sigma/i) + \frac{1}{2} \lambda_i \sigma_i \pi_{\xi_{(k-1)}}(\lambda/i, \lambda_i), \quad (9)$$

где  $k = 2, \dots, K$  и  $M(1, \lambda_1, \sigma_1) = 1/\lambda_1$ . Если же все случайные величины имеют одинаковые распределения со средним  $\lambda_i = \lambda$  и  $\sigma_i = \sigma$  для  $1 \leq i \leq K$ , то

$$M(k, \lambda, \sigma) = \frac{1}{\lambda} + \frac{\sigma \lambda}{2} (H_k - 1).$$

Результаты моделирования для оценки точности приближения приведены в [28], а также в [31] с использованием уравнения из (2). Рассматриваются следующие распределения: распределение Эрланга 2-го, 3-го и 4-го порядков ( $\varepsilon_2, \varepsilon_3, \varepsilon_4$ ) и распределение Парето с функцией распределения вида  $F_p(x) = 1 - \alpha(x + \gamma)^{-\beta}$ ,  $\beta > 2$ , для  $\beta = 4$  и  $\beta = 5$ . Условие  $\beta > 2$  необходимо для того, чтобы первые два момента были конечными,  $\alpha = \gamma^\beta$  и  $\gamma = \beta - 1$ , так что  $m = 1$  и  $M = 2 + 2/(\beta - 2)$ .

Погрешность аппроксимации выражений (9) и (2) мала и для  $\varepsilon_2$ , но быстро увеличивается начиная с  $K = 16$ , для  $\varepsilon_3$  и  $\varepsilon_4$ . Интересно, что оба уравнения дают точные результаты для распределения Парето, а последнее уравнение даёт более точные результаты в обоих случаях.

### 3.5. Аппроксимация на основе границ моментов порядковых статистик

Если математическое ожидание и дисперсию генерального распределения обозначить  $\mu$  и  $\sigma^2$  соответственно, то справедливо [18, 19]:

$$E[\xi_{(K)}] \leq \mu + \sigma \frac{K-1}{\sqrt{2K-1}}, \quad E[\xi_{(1)}] \geq \mu - \sigma \frac{K-1}{\sqrt{2K-1}}.$$

Поэтому в [18, 19] для максимума  $K$  н.о.р.с.в.  $\xi_{(K)}$  была предложена аппроксимация следующего вида:

$$E[\xi_{(K)}] \approx \mu + \sigma G(K), \quad (10)$$

где

$$G(K) \leq \frac{K-1}{\sqrt{2K-1}}. \quad (11)$$

Заметим, для экспоненциального распределения справедливо, что  $G(K) = H_K - 1$ , а для равномерного распределения выполняется:  $G(K) = \sqrt{3}(K-1)/(K+1)$  [27]. Если же математическое ожидание равно нулю, а стандартное квадратическое отклонение равно единице, то, очевидно,  $E[\xi_{(K)}] \approx G(K)$ . Знак равенства в выражении (11) имеет место для н.о.р.с.в.  $(\xi_k, k = 1, K)$  с функцией и плотностью распределения

следующего вида:

$$F(x) = \left(\frac{1+bx}{K}\right)^{1/(K-1)}, \quad f(x) = \frac{b}{K} \left(\frac{1+bx}{K}\right)^{1/(K-1)-1},$$

где

$$-\frac{\sqrt{2K-1}}{K-1} \leq x \leq \sqrt{2K-1}, \quad b = \frac{K-1}{\sqrt{2K-1}}.$$

Для симметричного распределения  $F(x) = 1 - F(-x)$  справедливо:

$$E[\xi_{K,\max}] \leq \frac{K}{2} \left[ \frac{2(1 - 1/\binom{2K-2}{K-1})}{2K-1} \right]^{1/2}.$$

Для аппроксимации ожидаемого значения максимума н.о.р.с.в. с *нормальным распределением* в работе [32] используется  $G(K) = \sqrt{2 \ln(K)}$  в формуле (10), т.е.:

$$E[\xi_{(K)}] \approx \mu + \sigma \sqrt{2 \ln(K)}.$$

Более точная аппроксимация дана в [30]:

$$E[\xi_{(K)}] \approx \mu + \sigma \left[ \sqrt{2 \ln(K)} - \frac{\ln(\ln(K)) - \ln(4\pi) - 2\gamma}{2\sqrt{2 \ln(K)}} \right]. \quad (12)$$

Стоит отметить, что в уравнении (12) существует некоторый сдвиг (смещение), который корректируется путём вычитания  $0,1727K^{-0,2750}$  из выражения в скобках в (12) [33]. Таким образом, имеем:

$$D[\xi_{(K)}] \approx \frac{1,64492\sigma^2}{2 \ln(K)}.$$

На основе формулы (10) в [34] для случайной величины времени отклика было представлено приближение:

$$E[W_{K,\max}] \approx E[W_{1,\max}] + F_K \sigma_{W_{1,\max}}(\rho) \alpha_K(\rho), \quad (13)$$

где  $E[W_{1,\max}]$  и  $\sigma_{W_{1,\max}}$  обозначают математическое ожидание и среднее квадратическое отклонение времени отклика СМО  $M|G|1$  с коэффициентом загрузки системы  $\rho$ . Напомним, что если времена обслуживания подзапросов распределены по произвольному закону  $B(x)$  [4], то

$$E[W_{1,\max}] = b + \frac{\lambda b^{(2)}}{2(1-\rho)},$$

где  $b = \int_0^\infty x dB(x)$ ,  $b^{(2)} = \int_0^\infty x^2 dB(x)$ ,  $\rho = \lambda b$ .  $F_K$  определяется из (13) при  $\rho = 0$  и  $\alpha_K(\rho) = 1$  следующим выражением:

$$F_K = \frac{b_{K,\max} - b}{\sqrt{b^{(2)} - b^2}},$$

где  $b_{K,\max} = \int_0^{\infty} [1 - B^K(x)] dx$  — математическое ожидание максимума  $K$  н.о.р.с.в. с функцией распределения  $B(x)$ . Легко показать, что для экспоненциального распределения  $F_K = H_K - 1$ , а для распределения экстремальных значений  $F_K = \sqrt{6} \ln(K)/\pi$  [30]. Сравнительный анализ полученного приближения и результатов имитационного моделирования при заданном распределении времени отклика может быть использован для получения выражения  $\alpha_K(\rho)$  [34]:

$$\alpha_K(\rho) = \frac{E[W_{K,\max}] - E[W_{1,\max}]}{F_K \sigma_{W_{1,\max}}}.$$

Так, например, для fork-join системы с двумя ветвями  $M|M|1$  из выражения (6) следует, что  $\alpha_2(\rho) = 1 - \rho/4$ .

#### 4. Заключение

В статье рассмотрены подходы к приближенному анализу времени отклика: матрично-геометрический метод, анализ с помощью порядковых статистик для различных типов распределения времени пребывания подзапросов в системе. Также приводятся результаты для математического ожидания и дисперсии времени синхронизации. Стоит отметить, что существуют исследования, посвящённые анализу работы сетей массового обслуживания, в качестве узла или узлов которой выступает система с расщеплением, но обзор подобных работ выходит за рамки поставленной перед авторами задачи.

#### Литература

1. Обзор систем параллельной обработки заявок / А. В. Горбунова, И. С. Зарядов, К. Е. Самуйлов, Э. С. Сопин // Вестник РУДН. Серия: Математика, информатика, физика. — 2017. — Т. 25, № 4. — С. 350—362.
2. *Tsimashenka I., Knottenbelt W. J.* Reduction of Subtask Dispersion in Fork-Join Systems // Computer Performance Engineering. — Springer Berlin Heidelberg, 2013. — Pp. 325–336.
3. *Башарин Г. П.* Лекции по математической теории телетрафика. — Москва: РУДН, 2009. — 342 с.
4. *Бочаров П. П., Печинкин А. В.* Теория массового обслуживания. — Москва: Изд-во РУДН, 1995. — 529 с.
5. Queueing Theory / P. P. Bocharov, C. D'Apice, A. V. Pechinkin, S. Salerno. — Brill Academic Publishers, 2004. — 457 p.
6. *Мокров Е. В., Самуйлов К. Е.* Модель системы облачных вычислений в виде системы массового обслуживания с несколькими очередями и с групповым поступлением заявок // Т-Comm — Телекоммуникации и Транспорт. — 2013. — Т. 11, № 7. — С. 139–141.
7. *Мокров Е. В., Чукарин А. В.* Анализ показателей эффективности системы облачных вычислений с миграцией серверов // Т-Comm — Телекоммуникации и Транспорт. — 2014. — Т. 8, № 8. — С. 64–67.
8. *Balsamo S., Mura I.* Approximate Response Time Distribution in Fork and Join Systems // SIGMETRICS Performance Evaluation Review. — 1995. — Vol. 23, No 1. — Pp. 305–306.
9. *Balsamo S., Mura I.* On Queue Length Moments in Fork and Join Queueing Networks with General Service Times // Computer Performance Evaluation Modelling Techniques and Tools. LNCS. — 1997. — Vol. 1245. — Pp. 218–231.

10. *Balsamo S., Donatiello L., Van Dijk N. M.* Bound Performance Models of Heterogeneous Parallel Processing Systems // *IEEE Transactions on Parallel and Distributed Systems*. — 1998. — Vol. 9, No 10. — Pp. 1041–1056.
11. *Perros H. G.* On the  $M/C_k/1$  Queue // *Performance Evaluation*. — 1983. — Vol. 3, No 2. — Pp. 83–93.
12. *Neuts M. F.* Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. — Courier Corporation, 1981. — 332 p.
13. *Rao B. M., Posner M. J. M.* Algorithmic and Approximation Analyses of the Split and Match Queue // *Stochastic Models*. — 1985. — Vol. 1, No 3. — Pp. 433–456.
14. *Takahashi M., Ōsawa H., Fujisawa T.* On a Synchronization Queue with Two Finite Buffers // *Queueing Systems*. — 2000. — Vol. 36. — Pp. 107–123.
15. *Takahashi M., Takahashi Y.* Synchronization Queue with Two MAP Inputs and Finite Buffers // *Proc. of the Third International Conference on Matrix Analytical Methods in Stochastic Models*. — 2000. — Pp. 375–390.
16. Generalized Parallel-Server Fork-Join Queues with Dynamic Task Scheduling / *M. S. Squillante, Y. Zhang, A. Sivasubramaniam, N. Gautam* // *Annals of Operations Research*. — 2008. — Vol. 160, No 1. — Pp. 227–255.
17. *Joshi G., Soljanin E., Wornell G.* Efficient Redundancy Techniques for Latency Reduction in Cloud Systems // *arXiv preprint arXiv:1508.03599*. — 2015.
18. *David H. A.* Order Statistics. — Wiley, New York, 1981.
19. *David H. A., Nagaraja H. N.* Order Statistics. — John Wiley & Sons, 2003. — 458 p.
20. *Gumbel E. J.* Statistics of Extremes. — New York: Columbia University Press, 1958. — 375 p.
21. *Allen A. O.* Probability, Statistics, and Queueing Theory: With Computer Science Applications. — Gulf Professional Publishing, 1990. — 740 p.
22. *Kleinrock L.* Queueing Systems, Volume I: Theory. — Wiley Interscience, 1975. — 448 p.
23. *Trivedi K. S.* Probability and Statistics with Reliability, Queueing, and Computer Science Applications (2nd ed.). — John Wiley & Sons, 2002. — 830 p.
24. *Gravey A.* A Simple Construction of an Upper Bound for the Mean of the Maximum of  $N$  Identically Distributed Random Variables // *Journal of Applied Probability*. — 1985. — Vol. 22. — Pp. 844–851.
25. The Estimation of Probability Characteristics of Cloud Computing Systems with Splitting of Requests / *A. V. Gorbunova, I. S. Zaryadov, S. I. Matushenko, E. S. Sopin* // *Distributed Computer and Communication Networks. DCCN 2016. Communications in Computer and Information Science*. — 2016. — Vol. 678. — Pp. 418–429.
26. *Самуйлов К. Е., Зарядов И. С., Горбунова А. В.* Анализ времени отклика системы облачных вычислений // *T-Comm: Телекоммуникации и транспорт*. — 2015. — Т. 9, № 11. — С. 57–61.
27. *Thomasian A.* Analysis of Fork/Join and Related Queueing Systems // *ACM Computing Surveys (CSUR)*. — 2014. — Vol. 47, No 2. — Pp. 17:1–17:71.
28. *Harrison P., S. Z.* Queueing Models with Maxima of Service Times // *Computer Performance Evaluation. Modelling Techniques and Tools*. — Springer Berlin Heidelberg, 2003. — Pp. 152–168.
29. *Thomasian A., Menon J.* RAID5 Performance with Distributed Sparing // *IEEE Transactions on Parallel and Distributed Systems*. — 1997. — Vol. 8, No 6. — Pp. 640–657.
30. *Johnson N. L., Kotz S., Balakrishnan N.* Continuous Univariate Distributions. — Wiley Series in Probability and Statistics, 1995. — Vol. 1, 752 p.
31. *Lebrecht A. S., Knottenbelt W. J.* Response Time Approximations in Fork-Join Queues // *In Proceedings of the 23rd Annual UK Performance Engineering Workshop (UKPEW'07)*. — 2007.
32. *Arnold B. C.* Distribution-Free Bounds on the Mean of the Maximum of a Dependent Sample // *SIAM Journal on Applied Mathematics*. — 1985. — Vol. 38, No 1. — Pp. 163–167.

33. *Petzold M.* A Note on the First Moment of Extreme Order Statistics from the Normal Distribution // rapport nr.: Seminar Papers. — 2000.
34. *Thomasian A., Tantawi A. N.* Approximate Solutions for  $M/G/1$  Fork/Join Synchronization // Proceedings of the 26th conference on Winter simulation (WSC'94) / Society for Computer Simulation International. — 1994. — Pp. 361–368.

UDC 519.21

DOI: 10.22363/2312-9735-2018-26-1-13-27

## A Survey on Queuing Systems with Parallel Serving of Customers. Part II

A. V. Gorbunova\*, I. S. Zaryadov\*<sup>†</sup>, K. E. Samouylov\*

\* *Department of Applied Probability and Informatics  
Peoples' Friendship University of Russia (RUDN University)  
6 Miklukho-Maklaya St., Moscow, 117198, Russian Federation*

<sup>†</sup> *Institute of Informatics Problems  
Federal Research Center "Computer Science and Control" Russian Academy of Sciences  
44-2 Vavilova St., Moscow, 119333, Russian Federation*

This paper is a continuation of the survey of the “fork-join” queuing systems (in the western classification) or the systems with splitting of queries. Interest in such systems is explained by a wide range of problems that can be solved with their help, since in fact it is a matter of parallel processing of data and their applications. For example, this may concern the analysis of disk arrays, cloud computing, high-performance services and even the process of picking orders in a warehouse. In the first part of the survey, the main features of the described model (and related systems) and its construction were introduced. Also the detailed description of the approach to obtaining an accurate expression of the average response time in the case of two devices was presented as well as several methods of approximate analysis of this characteristic (the case when the number of devices is more than two). This part of the survey is devoted to the description of other existing methods for approximating the average response time. In particular, the approaches of the approximate analysis of the response time are as follows: the matrix-geometric method, the analysis with the help of order statistics for various types of distribution of the service time of subqueries.

**Key words and phrases:** queuing system, splitting of requests, parallel service of requests, parallel processing, response time, synchronization delay, matrix geometric method, order statistics

## References

1. A. V. Gorbunova, I. S. Zaryadov, K. E. Samouylov, E. S. Sopin, A Survey on Queuing Systems with Parallel Serving of Customers, RUDN Journal of Mathematics, Information Sciences and Physics 25 (4) (2017) 350–362, in Russian.
2. I. Tsimashenka, W. J. Knottenbelt, Reduction of Subtask Dispersion in Fork-Join Systems, in: Computer Performance Engineering, Springer Berlin Heidelberg, 2013, pp. 325–336.
3. G. P. Basharin, Lectures on the Mathematical Theory of Teletraffic, PFUR, Moscow, 2009, in Russian.
4. P. P. Bocharov, A. V. Pechinkin, Queueing Theory, PFUR, Moscow, 1995, in Russian.
5. P. Bocharov, C. D'Apice, A. Pechinkin, S. Salerno, Queueing theory, Brill Academic Publishers, 2004.
6. E. V. Mokrov, K. E. Samouylov, Modeling of Cloud Computing as a Queuing System with Batch Arrivals, T-Comm — Telecommunications and transport 11 (7) (2013) 139–141, in Russian.

7. E. V. Mokrov, A. V. Chukarin, Performance Analysis of Cloud Computing System with Live Migration, T-Comm — Telecommunications and transport 8 (8) (2014) 64–67, in Russian.
8. S. Balsamo, I. Mura, Approximate Response Time Distribution in Fork and Join Systems, SIGMETRICS Performance Evaluation Review 23 (1) (1995) 305–306.
9. S. Balsamo, I. Mura, On Queue Length Moments in Fork and Join Queuing Networks with General Service Times, Computer Performance Evaluation Modelling Techniques and Tools. LNCS 1245 (1997) 218–231.
10. S. Balsamo, L. Donatiello, N. M. Van Dijk, Bound Performance Models of Heterogeneous Parallel Processing Systems, IEEE Transactions on Parallel and Distributed Systems 9 (10) (1998) 1041–1056.
11. H. G. Perros, On the  $M/C_k/1$  Queue, Performance Evaluation 3 (2) (1983) 83–93.
12. M. F. Neuts, Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach, Courier Corporation, 1981.
13. B. M. Rao, M. J. M. Posner, Algorithmic and Approximation Analyses of the Split and Match Queue, Stochastic Models 1 (3) (1985) 433–456.
14. M. Takahashi, H. Ōsawa, T. Fujisawa, On a Synchronization Queue with Two Finite Buffers, Queueing Systems 36 (2000) 107–123.
15. M. Takahashi, Y. Takahashi, Synchronization Queue with Two MAP Inputs and Finite Buffers, in: Proc. of the Third International Conference on Matrix Analytical Methods in Stochastic Models, 2000, pp. 375–390.
16. M. S. Squillante, Y. Zhang, A. Sivasubramaniam, N. Gautam, Generalized Parallel-Server Fork-Join Queues with Dynamic Task Scheduling, Annals of Operations Research 160 (1) (2008) 227–255.
17. G. Joshi, E. Soljanin, G. Wornell, Efficient Redundancy Techniques for Latency Reduction in Cloud Systems, arXiv preprint arXiv:1508.03599.
18. H. A. David, Order Statistics, Wiley, New York, 1981.
19. H. A. David, H. N. Nagaraja, Order Statistics, John Wiley & Sons, 2003.
20. E. J. Gumbel, Statistics of Extremes, Columbia University Press, New York, 1958.
21. A. O. Allen, Probability, Statistics, and Queueing Theory: With Computer Science Applications, Gulf Professional Publishing, 1990.
22. L. Kleinrock, Queueing Systems, Volume I: Theory, Wiley Interscience, 1975.
23. K. S. Trivedi, Probability and Statistics with Reliability, Queueing, and Computer Science Applications (2nd ed.), John Wiley & Sons, 2002.
24. A. Gravey, A Simple Construction of an Upper Bound for the Mean of the Maximum of  $N$  Identically Distributed Random Variables, Journal of Applied Probability 22 (1985) 844–851.
25. A. V. Gorbunova, I. S. Zaryadov, S. I. Matushenko, E. S. Sopin, The Estimation of Probability Characteristics of Cloud Computing Systems with Splitting of Requests, Distributed Computer and Communication Networks. DCCN 2016. Communications in Computer and Information Science 678 (2016) 418–429.
26. K. E. Samouylov, I. S. Zaryadov, A. V. Gorbunova, The Response Time Analysis of Cloud Computing System, T-Comm — Telecommunications and transport 9 (11) (2015) 57–61, in Russian.
27. A. Thomasian, Analysis of Fork/Join and Related Queueing Systems, ACM Computing Surveys (CSUR) 47 (2) (2014) 17:1–17:71.
28. P. Harrison, Z. S., Queueing Models with Maxima of Service Times, in: Computer Performance Evaluation. Modelling Techniques and Tools, Springer Berlin Heidelberg, 2003, pp. 152–168.
29. A. Thomasian, J. Menon, RAID5 Performance with Distributed Sparing, IEEE Transactions on Parallel and Distributed Systems 8 (6) (1997) 640–657.
30. N. L. Johnson, S. Kotz, N. Balakrishnan, Continuous Univariate Distributions, Vol. 1, Wiley Series in Probability and Statistics, 1995.

31. A. S. Lebrecht, W. J. Knottenbelt, Response Time Approximations in Fork-Join Queues, in: In Proceedings of the 23rd Annual UK Performance Engineering Workshop (UKPEW'07), 2007.
32. B. C. Arnold, Distribution-Free Bounds on the Mean of the Maximum of a Dependent Sample, SIAM Journal on Applied Mathematics 38 (1) (1985) 163–167.
33. M. Petzold, A Note on the First Moment of Extreme Order Statistics from the Normal Distribution, in: rapport nr.: Seminar Papers, 2000.
34. A. Thomasian, A. N. Tantawi, Approximate Solutions for  $M/G/1$  Fork/Join Synchronization, in: Proceedings of the 26th conference on Winter simulation (WSC'94), Society for Computer Simulation International, 1994, pp. 361–368.

© Горбунова А. В., Зарядов И. С., Самуйлов К. Е., 2018

**Для цитирования:**

Горбунова А. В., Зарядов И. С., Самуйлов К. Е. Обзор систем параллельной обработки заявок. Часть II // Вестник Российского университета дружбы народов. Серия: Математика. Информатика. Физика. — 2018. — Т. 26, № 1. — С. 13–27. — DOI: 10.22363/2312-9735-2018-26-1-13-27.

**For citation:**

Gorbunova A. V., Zaryadov I. S., Samouylov K. E. A Survey on Queuing Systems with Parallel Serving of Customers. Part II, RUDN Journal of Mathematics, Information Sciences and Physics 26 (1) (2018) 13–27. DOI: 10.22363/2312-9735-2018-26-1-13-27. In Russian.

**Сведения об авторах:**

**Горбунова Анастасия Владимировна** — кандидат физико-математических наук, ассистент кафедры прикладной информатики и теории вероятностей РУДН (e-mail: gorbunova\_av@rudn.university, тел.: +7(495)9550927)

**Зарядов Иван Сергеевич** — кандидат физико-математических наук, доцент кафедры прикладной информатики и теории вероятностей РУДН, старший научный сотрудник ИПИ ФИЦ ИУ РАН (e-mail: zaryadov\_is@rudn.university, тел.: +7(495)9550927)

**Самуйлов Константин Евгеньевич** — профессор, доктор технических наук, заведующий кафедрой прикладной информатики и теории вероятностей РУДН (e-mail: samuylov\_ke@rudn.university, тел.: +7(495)9550956)

**Information about the authors:**

**Gorbunova A. V.** — Candidate of Physical and Mathematical Sciences, assistant of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia (RUDN University) (e-mail: gorbunova\_av@rudn.university, phone: +7(495)9550927)

**Zaryadov I. S.** — Candidate of Physical and Mathematical Sciences, assistant professor of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia (RUDN University); Senior Researcher of Institute of Informatics Problems of Federal Research Center "Computer Science and Control" Russian Academy of Sciences (e-mail: zaryadov\_is@rudn.university, phone: +7(495)9550927)

**Samouylov K. E.** — professor, Doctor of Engineering Science, head of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia (RUDN University) (e-mail: samuylov\_ke@rudn.university, phone: +7(495)9550956)