УДК 519.21
DOI: 10.22363/2312-9735-2017-25-4-350-362

Обзор систем параллельной обработки заявок

А. В. Горбунова[*], **И. С. Зарядов**[*†], **К. Е. Самуйлов**[*], **Э. С. Сопин**[*†]

[*] *Кафедра прикладной информатики и теории вероятностей*
Российский университет дружбы народов
ул. Миклухо-Маклая, д. 6, Москва, Россия, 117198
[†] *Институт проблем информатики*
Федеральный исследовательский центр «Информатика и управление» РАН
ул. Вавилова, д. 44, кор. 2, Москва, Россия, 119333

Данная работа является первой в серии из двух статей, посвящённых обзору систем массового обслуживания вида «fork-join» (в западной классификации) или системам с расщеплением запросов. Указанная система является естественной моделью для многих других реальных систем. В статье описаны особенности построения этой модели и родственных ей систем, основные их характеристики. Отдельное внимание уделяется методам анализа времени отклика системы. Поскольку точное выражение для среднего времени отклика известно только для случая двух приборов, в статье приведено подробное описание подхода к получению точного выражения этой характеристики. Для случая, когда число приборов больше двух, различными методами получены аппроксимации среднего времени отклика, что объясняется сложностью исследований из-за существующей зависимости между очередями подзапросов в силу общих моментов поступления. В работе представлено несколько методов приближенного анализа: различные варианты эмпирической аппроксимации, т.е. методы, уточняющие полученные характеристики благодаря использованию результатов имитационного моделирования; интерполяция с помощью предельных значений загрузки системы в случаях с отличными от экспоненциального распределениями для входящего потока и времени обслуживания.

Ключевые слова: система массового обслуживания, расщепление заявок, параллельное обслуживание заявок, параллельная обработка, время отклика

1. Введение

Название исследуемой системы с параллельной обработкой запросов или с расщеплением запросов не имеет утвердившегося в русском языке терминологического аналога, в англоязычной же литературе используется термин «fork-join queueing system» [1–3]. Интерес к указанной системе объясняется тем, что она является естественной моделью для многих других реальных систем. В качестве примеров можно привести сферу обслуживания, медицинские приложения, производственные системы, системы облачных вычислений и др. [4]. Если же говорить об информационных технологиях, то прежде всего стоит упомянуть распределённые вычисления, технология которых оказала значительное влияние и на концепцию облачных вычислений, а также параллельные вычисления, например обработку пакетов данных, телефонных вызовов и др. Однако, несмотря на широкий спектр задач, которые решаются с помощью систем массового обслуживания с параллельной обработкой запросов, и их популярность среди зарубежных авторов, в нашей стране данная система исследовалась значительно меньше [4–9]. Статья организована следующим образом: в разделе 2 описываются особенности построения системы с расщеплением запросов и родственных ей систем, в разделе 3 подробно описан метод точного анализа среднего времени отклика для случая расщепления на два подзапроса, в разделе 4 представлены два метода аппроксимации среднего времени отклика, в

Статья поступила в редакцию 22 июня 2017 г.
Исследование выполнено при финансовой поддержке Минобрнауки России (соглашение № 02.A03.21.0008) и при финансовой поддержке РФФИ в рамках научных проектов № 15-07-03051, 15-07-03608.

заключении кратко подведены итоги и анонсировано содержание второй части работы.

2. Системы параллельного обслуживания заявок

Общая идея функционирования системы fork-join (рис. 1) заключается в следующем: на вход поступает поток заявок (запросов) $T_1, T_2, \ldots$, в момент поступления в точке «расщепления» (англ. fork point) заявка T_i, $i = \overline{1, \infty}$ разделяется на K родственных заявок (подзапросов): $ST_1^i \stackrel{\text{def}}{=} ST_1, ST_2^i \stackrel{\text{def}}{=} ST_2, \ldots, ST_K^i \stackrel{\text{def}}{=} ST_K$, каждая из которых отправляется в очередь на обслуживание к приборам или серверам с номерами $1, 2, \ldots, K$ соответственно. Предполагается, что запросы имеют одну и ту же структуру и каждый сервер предназначен для выполнения конкретных задач: подзапрос ST_k может обслуживаться только на k-м сервере, $k = \overline{1, K}$. Обработанные подзапросы попадают в буфер синхронизации, где дожидаются родственных им подзапросов. В момент поступления последнего родственного подзапроса одного из запросов происходит их синхронизация, т.е. объединение, после чего обработанный запрос покидает систему. Здесь и далее будем считать,что синхронизация происходит мгновенно.

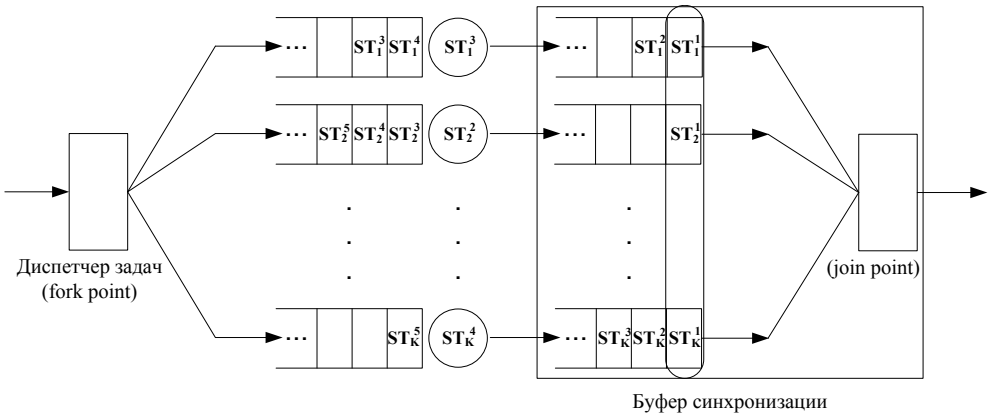

Рис. 1. Система массового обслуживания типа fork-join

Стоит отметить, что описанную модель рассматривают и как сеть массового обслуживания (СеМО): после расщепления подзапросы одновременно поступают на вход ветвей, которые представляют собой одноканальные системы массового обслуживания (СМО) с очередями. Однако данный подход предполагает дальнейший переход к рассмотрению независимых СМО, что является упрощающим допущением в силу существующей зависимости между очередями подзапросов вследствие общих моментов поступления. Тем не менее таких сложностей не возникает для случая, когда каждая подсистема состоит из бесконечного числа приборов [6, 7].

Анализ *SPM (splitting and matching queueing system) системы массового обслуживания* в [10] является одной из самых ранних работ на эту тему. В ней рассматриваются авиапассажиры и их багаж, которые прибывают вместе на самолёте в аэропорт, но разделены до тех пор, пока пассажиры не получат свои вещи в зоне выдачи багажа. В этой статье анализируется система с двумя ветвями $D|M|1$, т.е. с детерминированным входящим потоком и экспоненциальным временем обслуживания на каждом из серверов.

SM (split-merge queueing system) система является одним из вариантов fork-join системы массового обслуживания (рис. 2). Её отличие заключается в том, что на

свободные серверы не поступит ни один подзапрос следующего запроса до того момента, пока не обслужатся все подзапросы текущего запроса [11], т.е. происходит блокировка каждого освободившегося сервера до окончания обработки всех задач запроса.

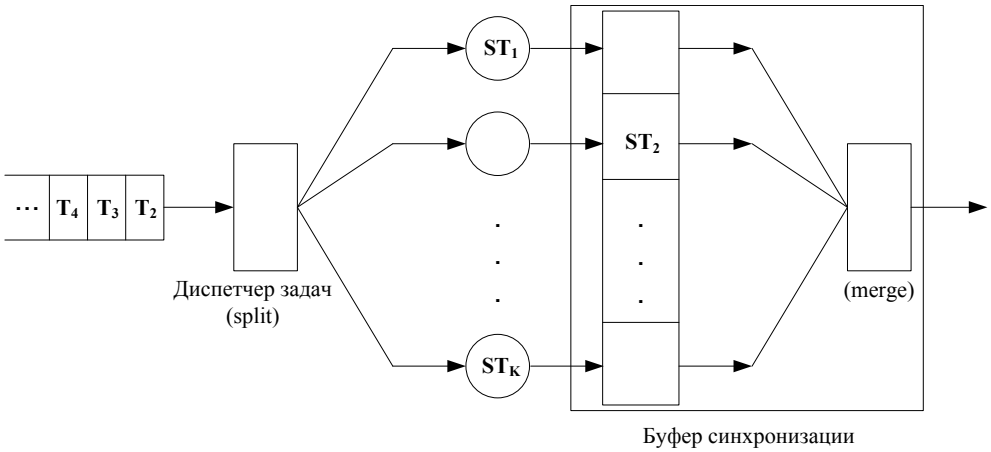

Рис. 2. Система массового обслуживания типа split-merge

FF (fission-fusion queueing system) система — это ещё одна разновидность fork-join системы (рис. 3). Если допустить, что все подзапросы $ST_1, ST_2, \ldots, ST_K$ являются идентичными и неразличимыми: $ST_k = ST, k = \overline{1, K}$, то запрос может покинуть систему после обработки любых K подзапросов, иначе говоря, подзапросы могут принадлежать разным запросам [11].

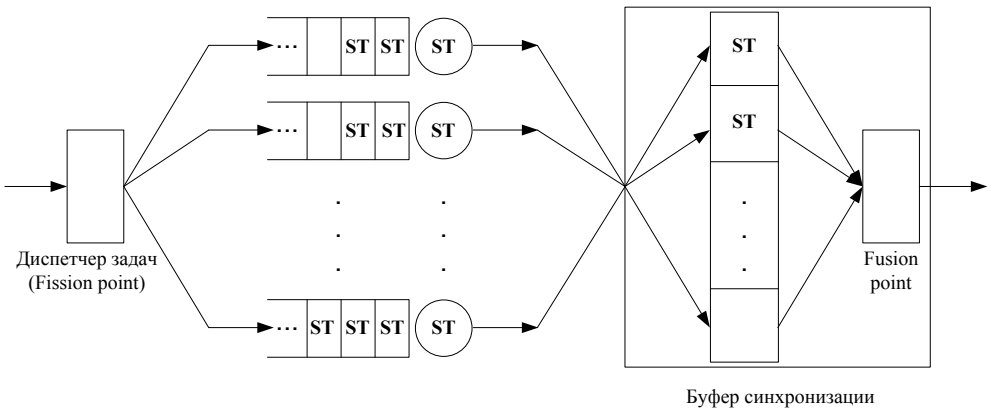

Рис. 3. Система массового обслуживания типа fission-fusion

Модель независимых серверов (independent server model, ISM) представляет собой многоканальную систему массового обслуживания. Для обслуживания поступившего запроса требуется случайное число серверов, не превышающее их общее количество, причём обслуживание очередного запроса не может начаться до тех пор, пока число свободных серверов не станет по крайней мере равно числу серверов, требуемых для обслуживания. В частности, в работе [12] рассматриваемая система состоит из однородных серверов, времена обслуживания являются независимыми и имеют экспоненциальное распределение. Входящий поток является пуассоновским,

а число серверов, требуемых для обработки запроса, определяется некоторой вероятностью. Запросы обслуживаются в порядке поступления (First Come First Served, FCFS) и могут покинуть систему только после освобождения всех затребованных серверов. Поскольку времена обслуживания на отдельных серверах являются независимыми, то сервер, закончивший обслуживание запроса, становится доступным для других задач. Таким образом, время обслуживания запроса является максимумом случайного числа экспоненциально распределённых случайных величин. Для модели были получены такие характеристики, как распределение времени обслуживания, распределение числа занятых серверов, а также достаточные условия для существования стационарного распределения.

В случае *модели группового обслуживания (team service model, TSM)*, так же как и в случае ISM системы, для обработки задач требуется наличие нескольких свободных серверов одновременно, но при этом серверы начинают обслуживание этих задач и освобождаются одновременно [13, 14].

Также существует такой вариант fork-join системы, когда поступающие на вход в систему несколько потоков запросов делятся на подзапросы и отправляются на обслуживание на серверы в соответствии с заданной матрицей вероятностей, т.е. запросы поступают как K независимых потоков, затем делятся на части, которые распределяются по серверам в соответствии с некоторой стратегией (политикой), которая и определяется матрицей [15, 16].

Первые результаты исследования fork-join системы массового обслуживания были получены для случая расщепления заявки на два подзапроса ST_1, ST_2 с пуассоновским входящим потоком и экспоненциальным временем обслуживания на обоих серверах [17]. Однако несмотря на марковость предположений анализ fork-join систем массового обслуживания даже для случая $K = 2$ затруднителен по следующим причинам: процессы поступления подзапросов на серверы коррелируют; модель описывается двумерной цепью Маркова с бесконечным числом состояний для каждой размерности.

В силу указанных выше факторов точного решения для распределения числа подзапросов ST_k, $k = \overline{1, K}$ в fork-join системе не было найдено. Но при этом для случая пуассоновского входящего потока и экспоненциальных времён обслуживания можно получить явные выражения для маргинальных вероятностей [4].

Приближенное же решение для распределения числа подзапросов в случае очередей неограниченных размеров может быть получено для достаточно большой ёмкости накопителя — такой, что вероятность потери запросов в связи с этой большой ёмкостью будет пренебрежимо мала, и в данном случае будут применимы итерационные численные методы решения систем линейных уравнений большой размерности [18].

Большинство авторов в своих работах исследуют один из основных показателей качества функционирования подобных систем массового обслуживания — время отклика $W_{K,\max} = \max(\xi_1, \ldots, \xi_K)$, где ξ_k — случайная величина времени пребывания подзапроса ST_k в системе до его попадания в буфер синхронизации, $k = \overline{1, K}$. Стоит сразу отметить, что точное выражение для среднего времени отклика было получено только для fork-join системы с двумя ветвями $M|M|1$ ($K = 2$) в [1] с помощью результатов работы [17]. Для случая $K > 2$ с помощью различных методов были получены аппроксимации среднего времени отклика.

3. Анализ системы с двумя ветвями M|M|1

Модель цепи Маркова с непрерывным временем для fork-join системы с двумя ветвями $M|M|1$ и накопителем неограниченной ёмкости рассматривается в [17]. Этот анализ был расширен для ветвей $M|G|2$ в [19]. Для изучения стационарного и нестационарного режимов основной краевой задачи (дифференциальное уравнение

вместе с набором дополнительных ограничений) в работе [19] использовались методы исследования функций комплексной переменной, представленные, например, в [20]. В ходе последующего обсуждения более подробно остановимся на анализе статьи [17], поскольку она ведёт к точному решению задачи для fork-join системы с двумя ветвями $M|M|1$.

Итак, рассмотрим fork-join систему, попадая в которую запрос расщепляется на $K = 2$ подзапроса, входящий поток является пуассоновским с интенсивностью $\lambda = 1$, а интенсивности обслуживания на двух серверах удовлетворяют условию: $1 < \mu_1 \leqslant \mu_2$, гарантируя, что со временем система не переполнится, т.е. число подзапросов не устремится к бесконечности. Поведение системы описывается непрерывным марковским процессом $X(t) = \{x_1(t), x_2(t),\, t \geqslant 0\}$, где $x_k(t)$ — число подзапросов k-го типа, соответственно $k = 1, 2$. Указанные состояния интерпретируются следующим образом: если $X(t) = \{(i, j), i, j \geqslant 0\}$ в некоторый момент времени t, то в системе находится i заявок 1-го типа, т.е. ST_1, и j заявок 2-го типа, т.е. ST_2, а $p_{i,j}$ — стационарная вероятность указанного состояния. Далее записывается система уравнений равновесия в терминах двойной производящей функции $P(z, w) = \sum_i \sum_j p_{i,j} z^i w^j$ [17, 21, 22]:

$$Q(z, w)P(z, w) = N(z, w), \quad |z|, |w| \leqslant 1,$$

где

$$Q(z, w) = 1 + \mu_1\mu_2 zw - \mu_1 w - \mu_2 z - z^2 w^2,$$
$$N(z, w) = \mu_2 z(w - 1)P(z, 0) + \mu_1 w(z - 1)P(0, w).$$

Анализ производящей функции позволяет получить выражения для $P(z, 0)$ и $P(0, w)$, которые для симметричного случая $\mu_1 = \mu_2 = \mu$ равны между собой:

$$P(z, 0) = P(0, z) = \frac{(\mu - 1)^{3/2}}{\mu(\mu - z)^{1/2}}. \tag{1}$$

Также с помощью преобразований производящей функции в граничных точках: $P(z, 0)$ и $P(0, w)$, в [17] были получены асимптотические выражения для стационарных вероятностей состояний $p_{i,0}$ и $p_{0,j}$ при $i \to \infty, j \to \infty$ соответственно. Это ещё раз свидетельствует о том, что данная задача требует более совершенных математических методов решения.

Используя результаты анализа в [17], среднее время отклика для fork-join системы с двумя симметричными ветвями выводится в Приложении В в [1]. Интенсивность пуассоновского входящего потока имеет значение λ, интенсивности обслуживания на двух серверах раны μ, т.е. $\mu_1 = \mu_2 = \mu$, причём $\lambda < \mu$, так, что нагрузка на каждый из серверов $\rho = \lambda/\mu$ меньше единицы:

$$E\left[W_{2,\max}\right] = \left[H_2 - \frac{\rho}{8}\right] E\left[W_{1,\max}\right] = \frac{12 - \rho}{8} E\left[W_{1,\max}\right],$$

где $H_2 = 1,5, E\left[W_{1,\max}\right] = (\mu - \lambda)^{-1}$ — среднее время отклика для СМО $M|M|1$ [22–24].

Вывод $E\left[W_{2,\max}\right]$ основан на наблюдении, что эта величина есть сумма времени пребывания в системе $M|M|1$, т.е. $W_{1,\max}$, и времени, проведённого подзапросом в системе после обслуживания на сервере в ожидании второго, и до ухода из неё (время синхронизации) S: $E\left[W_{2,\max}\right] = E\left[W_{1,\max}\right] + E[S]$.

Согласно формуле Литтла, $N = 2\lambda E[S]$, где N — это среднее число подзапросов, которые обслужились на приборе и ждут свою пару [22–24]. Через q_k обозначим

вероятность того, что число подзапросов во второй очереди превышает число подзапросов в первой очереди на k. Благодаря симметрии $q_k = q_{-k}$, $k \geqslant 1$ и, таким образом, $N = 2\sum\limits_{k=1}^{\infty} kq_k$, следовательно $E[S] = \dfrac{1}{\lambda}\sum\limits_{k=1}^{\infty} kq_k$.

Из уравнений, представленных в [1], можно получить:

$$q_k = \sum_{i=k}^{\infty} p_{i,0}. \tag{2}$$

Тогда $q_k = q_{k+1} + p_{k,0}$, где $p_{k,0}$ — это вероятность того, что в первой очереди находится k подзапросов, а во второй — ноль. Далее с учётом формулы (1) можем записать:

$$P(z,0) = P(0,z) \stackrel{\text{def}}{=} P(z) = \sum_{i=0}^{\infty} p_{i,0} z^i = \frac{(1-\rho)^{3/2}}{(1-\rho z)^{1/2}}. \tag{3}$$

Затем подставляем в формулу для времени синхронизации выражение для q_k из (2) и меняем порядок суммирования:

$$E[S] = \frac{1}{\lambda}\sum_{k=1}^{\infty} k \sum_{i=k}^{\infty} p_{i,0} = \frac{1}{\lambda}\sum_{i=1}^{\infty} p_{i,0} \sum_{k=1}^{i} k = \frac{1}{\lambda}\sum_{i=1}^{\infty} \frac{(i+1)i}{2} p_{i,0}. \tag{4}$$

Потом вычисляем значения первой и второй производной при $z = 1$ от выражения для производящей функции $P(z,0)$, полученном в (3) и от выражения, которое записывается исходя из определения этой функции:

$$\left.\frac{\mathrm{d}P(z)}{\mathrm{d}z}\right|_{z=1} = \frac{1}{2}\rho = \sum_{i=1}^{\infty} i p_{i,0}, \quad \left.\frac{\mathrm{d}^2P(z)}{\mathrm{d}z^2}\right|_{z=1} = \frac{3\rho^2}{4(1-\rho)}\rho = \sum_{i=1}^{\infty} i^2 p_{i,0} - \sum_{i=1}^{\infty} i p_{i,0}.$$

Теперь можем подставить полученные формулы в (4):

$$E[S] = \frac{1}{2\lambda}\left(\sum_{i=1}^{\infty} i^2 p_{i,0} + \sum_{i=1}^{\infty} i p_{i,0}\right) = \frac{1}{2\lambda}\left(\frac{\rho^2 + 2\rho}{4(1-\rho)} + \frac{\rho}{2}\right) = \frac{\rho(4-\rho)}{8\lambda(1-\rho)}, \tag{5}$$

следовательно, среднее время отклика определяется следующим выражением:

$$E\left[W_{2,\max}\right] = E\left[W_{1,\max}\right] + E[S] = \frac{12-\rho}{8}\frac{1}{\mu-\lambda} = \frac{12-\rho}{8}E\left[W_{1,\max}\right].$$

4. Аппроксимация времени отклика

4.1. Эмпирическая формула аппроксимации

Один из способов аппроксимации времени отклика для fork-join системы с K ветвями $M|M|1$ и одинаковыми интенсивностями обслуживания на серверах был предложен в [1]. Идея предложенного метода аппроксимации появилась из наблюдения за поведением времени отклика при проведении численных экспериментов.

Верхнюю границу для среднего времени отклика можно получить, сделав допущение о независимости случайных величин времён пребывания подзапросов в системе, или, что тоже самое, рассмотреть CeMO, состоящую из K независимых параллельно функционирующих систем $M|M|1$. Данное предположение подтверждается приведённым в [1] доказательством того факта, что случайные величины времён

пребывания подзапросов в системе являются положительно ассоциированными. По определению [25] две случайные величины ξ и η являются положительно ассоциированными, если $E[f(\xi)g(\eta)] \geqslant E[f(\xi)]E[g(\eta)]$, т.е. ковариация любых пар неубывающих функций f и g является неотрицательной, а одно из свойств положительно ассоциированных случайных величин $\xi_1, \xi_2, \ldots \xi_K$ заключается в том, что [25]

$$P\left(\max_{1\leqslant i\leqslant K} \xi_i > x\right) \leqslant 1 - \prod_{i=1}^{K} P\left(\xi_i < x\right).$$

Нижняя граница для среднего времени отклика получается, если «пренебречь» очередями (т.е. для значений ρ близких к нулю): в этом случае время отклика будет являться максимумом K независимых одинаково распределённых случайных величин (н.о.р.с.в.) со средним $1/\mu$.

Таким образом, имеем:

$$H_K(1/\mu) \leqslant E\left[W_{K,\max}\right] \leqslant H_K E[W],$$

где $H_K = \sum\limits_{i=1}^{K} 1/i$ — это частичная сумма гармонического ряда. Верхняя и нижняя границы растут с одной и той же скоростью H_K, иными словами, для больших значений K границы имеют порядок $O(\ln K)$. Поэтому, зная значение $E\left[W_{2,\max}\right]$, можем записать: $E\left[W_{K,\max}\right] \approx S_K(\rho) E\left[W_{2,\max}\right]$, где $S_K(\rho)$ — это коэффициент масштабирования, растущий со скоростью $O(\ln K)$. Следовательно, аппроксимация времени отклика может быть представлена в следующем виде: $S_K = \alpha(\rho) + \beta(\rho) H_K$. Поскольку по определению $S_2(\rho) = 1$, то, подставив это значение в предыдущее выражение, получим, что $\beta(\rho) = (1 - \alpha(\rho))/H_2$, таким образом,

$$S_K(\rho) = \alpha(\rho) + \frac{1 - \alpha(\rho)}{H_2} H_K, \quad K \geqslant 2.$$

Посредством имитационного моделирования в [1] определяется значение $\alpha(\rho) \approx 4\rho/11$. Окончательно имеем

$$E\left[W_{K,\max}\right] \approx \left[\frac{H_K}{H_2} + \frac{4}{11}\left(1 - \frac{H_K}{H_2}\right)\rho\right] \frac{12 - \rho}{8} \frac{1}{\mu - \lambda}, \quad K \leqslant 2. \tag{6}$$

Полученное приближение, исходя из численного анализа, проведённого авторами, имеет погрешность аппроксимации, не превышающую 5% для значений $2 \leqslant K \leqslant 32$.

В статье [26] исследуется аналогичная fork-join система с K ветвями $M|M|1$. Полученная аппроксимация является средним арифметическим верхней и нижней оценок. В качестве верхней оценки используется видоизменённое выражение для верхней оценки из вышеупомянутой работы [1]. Нижняя оценка получается при анализе эквивалентной величины времени отклика, но для системы с непараллельной организацией очереди [27]. Итак,

$$\frac{1}{\mu}\left(H_K + \rho \sum_{k=1}^{K} \frac{1}{k(k-\rho)}\right) \leqslant E\left[W_{K,\max}\right] \leqslant \frac{H_K}{\mu}\left(1 + \frac{\rho}{1-\rho}\right) = \\ = \frac{1}{\mu}\left(H_K + \rho \sum_{k=1}^{K} \frac{1}{k(1-\rho)}\right).$$

Далее вычисляется среднее арифметическое полученных оценок, в результате чего имеем:

$$E\left[W_{K,\max}\right] \approx \frac{1}{\mu}\left[H_K + \frac{\rho}{2(1-\rho)}\left(\sum_{k=1}^{K}\frac{1}{k-\rho} + (1-2\rho)\sum_{k=1}^{K}\frac{1}{k(k-\rho)}\right)\right]. \qquad (7)$$

4.2. Интерполяция с помощью предельных значений загрузки системы

В [28] предложен другой метод аппроксимации времени отклика — комбинация методов интерполяции высокой и слабой входных нагрузок (heavy and light traffic interpolation approximations). Указанные техники в отличие от описанного выше метода не используют результаты имитационного моделирования, однако их применение может быть расширено и до анализа fork-join систем не только с экспоненциальным временем обслуживания или с пуассоновским входящим потоком.

Для рассматриваемой задачи найти решение в аналитическом виде очень затруднительно, однако возможно получить асимптотические формулы для искомых характеристик. Интерполяция слабой загрузкой системы является результатом её работы в режиме слабой нагрузки, т.е. когда интенсивность входящего потока λ очень мала. В этом случае целесообразно обратиться к разложению в ряд Тэйлора характеристик производительности системы (в частности, функции распределения времени отклика — как функции от λ — в окрестности нуля), с помощью которого можно определить неизвестные величины в представлении исследуемой функции в виде полинома от λ порядка n. Рассматриваются только полиномы нулевой и первой степени. Если же говорить об интерполяции с помощью высокой загрузки, то для fork-join системы речь идёт об анализе такого режима, в котором значение λ очень близко к значению μ. Ключевым параметром при интерполяции с помощью метода высокой нагрузки является параметр β, два крайних значения которого интерпретируются как два предельных случая: если $\beta = 0$, то это означает, что входящий поток является детерминированным, если $\beta = 1$, то время обслуживания — детерминированное, а ветви fork-join системы являются K независимыми $D|GI|1$ и $GI|D|1$ СМО, соответственно.

Таким образом, благодаря исследованию поведения функции времени отклика в граничных значениях загрузки системы удаётся, в отличие от метода, описанного в [1], не прибегая к численным экспериментам для определения констант в интерполяционной формуле, определить их точные выражения в замкнутой форме.

Для случая fork-join системы с K ветвями $M|M|1$ аппроксимация времени отклика имеет вид:

$$E\left[W_{K,\max}\right] \approx \left[H_K + (V_K - H_K)\frac{\lambda}{\mu}\right]\frac{1}{\mu-\lambda}, \quad 0 \leqslant \mu, \quad K \geqslant 2, \qquad (8)$$

где

$$V_K = \sum_{i=1}^{K}\binom{K}{i}(-1)^{i-1}\sum_{m=1}^{i}\binom{i}{m}\frac{(m-1)!}{i^{m+1}}.$$

В случае распределения Эрланга второго порядка для входящего потока с функцией распределения $A(x) = 1 - (1+2\lambda x)e^{-2\lambda x}$, $x \geqslant 0$ и для времени обслуживания с функцией распределения $B(x) = 1 - (1+2\mu x)e^{-2\mu x}$, $x \geqslant 0$:

$$E\left[W_{K,\max}\right] \approx \left[F_K + \left(\frac{V_K}{2} - F_K\right)\frac{\lambda}{\mu}\right]\frac{1}{\mu-\lambda}, \quad 0 \leqslant \lambda < \mu, \quad K = 2,3,\ldots, \qquad (9)$$

где

$$F_K \equiv \frac{1}{\mu}\sum_{r=1}^{K}\binom{K}{r}(-1)^{r-1}\sum_{m=0}^{r}\binom{r}{m}\frac{m!}{2r^{m+1}}, \quad K = 2, 3, \ldots$$

Также в работе [28] для того, чтобы получить оценку среднего времени отклика в случаях с другими распределениями для входящего потока и времени обслуживания, метод интерполяции высокой нагрузки был модифицирован, и анализировались три значения ключевой константы $\beta = 0, 1/2, 1$, что привело к квадратичной интерполяции, благодаря которой, в сочетании с методом низкой нагрузки, были получены формулы аппроксимации среднего времени отклика для следующих немарковских случаев:

– распределение Эрланга второго порядка для входящего потока с функцией распределения $A(x) = 1 - (1 + 2\lambda x)e^{-2\lambda x}$, $x \geqslant 0$ и экспоненциальное время обслуживания с функцией распределения $B(x) = 1 - e^{-\mu x}$, $x \geqslant 0$:

$$E\left[W_{K,\max}\right] \approx \left[H_K + \left(\frac{2}{3}V_K - \frac{5}{6}H_K - \frac{1}{12}\right)\frac{\lambda}{\mu}\right]\frac{1}{\mu - \lambda}, \quad 0 \leqslant \lambda < \mu, \quad K = 2, 3, \ldots;$$

– пуассоновский входящий поток с функцией распределения $A(x) = 1 - e^{-\lambda x}$, $x \geqslant 0$ и распределение Эрланга второго порядка времени обслуживания с функцией распределения $B(x) = 1 - (1 + 2\mu x)e^{-2\mu x}$, $x \geqslant 0$:

$$E\left[W_{K,\max}\right] \approx \left[F_K + \left(\frac{1}{6} - \frac{H_K}{12} + \frac{2}{3}V_K - F_K\right)\frac{\lambda}{\mu}\right]\frac{1}{\mu - \lambda}, \quad 0 \leqslant \lambda < \mu, \quad K = 2, 3, \ldots,$$

а также для гиперэкспоненциального входящего потока и экспоненциального времени обслуживания, пуассоновского входящего потока и гиперэкспоненциального времени обслуживания, формулы для которых в силу их громоздкости приводить не будем.

5. Заключение

В статье рассмотрены особенности построения систем массового обслуживания с расщеплением запросов. Представлен метод точного анализа среднего времени отклика ($K = 2$), а также его аппроксимации. Во второй части работы будут рассмотрены другие подходы к приближенному анализу времени отклика: матрично-геометрический метод, анализ с помощью порядковых статистик для различных типов распределения времени пребывания подзапросов в системе.

Литература

1. *Nelson R., Tantawi A. N.* Approximate Analysis of Fork/Join Synchronization in Parallel Queues // IEEE Transactions on Computers. — 1988. — Vol. 37. — Pp. 739–743.
2. *Thomasian A.* Analysis of Fork/Join and Related Queueing Systems // ACM Computing Surveys (CSUR). — 2014. — Vol. 47, No 2. — Pp. 17:1–17:71.
3. *Tsimashenka I., Knottenbelt W. J.* Reduction of Subtask Dispersion in Fork-Join Systems // Computer Performance Engineering. — Springer Berlin Heidelberg, 2013. — Pp. 325–336.
4. Аппроксимация времени отклика системы облачных вычислений / А. В. Горбунова, И. С. Зарядов, С. И. Матюшенко, К. Е. Самуйлов, С. Я. Шоргин // Информатика и её применения. — 2015. — Т. 9, вып. 3. — С. 32–38.

5. *Вышенский С. В., Григорьев П. В., Дубенская Ю. Ю.* Идеальный синхронизатор маркированных пар в сети разветвление-объединение // Обозрение прикладной и промышленной математики. — 2008. — Т. 15, № 3. — С. 385–399.
6. *Моисеева С. П., Ивановская И. А.* Исследование математической модели параллельного обслуживания заявок смешанного типа // Известия Томского политехнического университета. Управление, вычислительная техника и информатика. — 2010. — Т. 317, № 5. — С. 32–34.
7. *Моисеева С. П., Жидкова Л. А.* Исследование системы параллельного обслуживания кратных заявок простейшего потока // Известия Томского политехнического университета. Управление, вычислительная техника и информатика. — 2011. — Т. 17, № 4. — С. 49–54.
8. 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 // Proceedings of the Nineteenth International Scientific Conference Russia: Distributed computer and communication networks: control, computation, communications (DCCN-2016). — Vol. 3. — 2016. — Pp. 467–472.
9. Оценка вероятностных характеристик системы облачных вычислений с расщеплением запросов / А. В. Горбунова, И. С. Зарядов, С. И. Матюшенко, Э. С. Сопин // Информационные технологии и математическое моделирование (ИТММ-2016): Материалы XV Международной конференции имени А. Ф. Терпугова. — 2016. — С. 167–172.
10. *Mandelbaum M., Itzhak A.-B.* Introduction to Queueing with Splitting and Matching // Israel Journal of Technology. — 1968. — Vol. 6, No 5. — Pp. 376–382.
11. *Duda A., Czachórski T.* Performance Evaluation of Fork and Join Synchronization Primitives // Acta Informatica. — 1987. — Vol. 24, No 5. — Pp. 525–533.
12. *Green L.* A Queueing System in which Customers Require a Random Number of Servers // Operations Research. — 1980. — Т. 28, № 6. — С. 1335–1346.
13. *Omahen K. J., Schrage L.* A Queueing Analysis of a Multiprocessor System with Shared Memory // Proc. of the Symposium on Computer Communication Networks and Teletraffic. — 1972. — Pp. 77–88.
14. *Thomasian A., Avizienis A.* Dynamic Scheduling of Tasks Requiring Multiple Processors // Proceedings of the 11th IEEE Computer Society International Conference (COMPCON'75 Fall). — 1975. — Pp. 77–80.
15. *Javidi T.* Cooperative and Non-Cooperative Resource Sharing in Networks: a Delay Perspective // IEEE Transactions on Automatic Control. — 2008. — Vol. 53, No 9. — Pp. 2134–2142.
16. *Kumar A., Shorey R.* Performance Analysis and Scheduling of Stochastic Fork-Join Jobs in a Multicomputer System // IEEE Transactions on Parallel and Distributed Systems. — 1993. — Vol. 10, No 4. — Pp. 1147–1164.
17. *Flatto L., Hahn S.* Two Parallel Queues Created by Arrivals with Two Demands I // SIAM Journal on Applied Mathematics. — 1984. — Vol. 44, No 5. — Pp. 1041–1053.
18. Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications / G. Bolch, S. Greiner, H. de Meer, K. S. Trivedi. — John Wiley & Sons, 2006. — P. 896.
19. *Baccelli F.* Two Parallel Queues Created by Arrivals with Two Demands: the $M|G|2$ Symmetrical Case // INRIA Rapport de Recherche. — 1985. — Vol. 426.
20. *Boyce W. E., DiPrima R. C.* Elementary Differential Equations and Boundary Value Problems. — John Wiley & Sons, 2012. — P. 809.
21. *Башарин Г. П.* Введение в теорию вероятностей. — Москва: РУДН, 1990. — 228 с.
22. *Бочаров П. П., Печинкин А. В.* Теория массового обслуживания. — Москва: Изд-во РУДН, 1995. — С. 529.
23. *Башарин Г. П.* Лекции по математической теории телетрафика. — Москва: РУДН, 2009. — 342 с.
24. Queueing Theory / P. P. Bocharov, C. D'Apice, A. V. Pechinkin, S. Salerno. — Brill

Academic Publishers, 2004. — P. 457.
25. *Barlow R. E., Proschan F.* Statistical Theory of Reliability and Life Testing: Probability Models. — John Wiley & Sons, 1981. — P. 290.
26. *Varki E., Merchant A., Chen H.* The $M|M|1$ Fork-Join Queue with Variable Subtasks . — http://www.cs.unh.edu/~varki/publication/2002-nov-open.pdf.
27. *Varki E.* Response Time Analysis of Parallel Computer and Storage Systems // IEEE Transactions on Parallel and Distributed Systems. — 2001. — Vol. 12, No 11. — Pp. 1146–1161.
28. *Varma S., Makowski A. M.* Interpolation Approximations for Symmetric Fork-Join Queues // Performance Evaluation. — 1994. — Vol. 20. — Pp. 245–265.

UDC 519.21
DOI: 10.22363/2312-9735-2017-25-4-350-362

A Survey on Queuing Systems with Parallel Serving of Customers

A. V. Gorbunova[*], I. S. Zaryadov[*†], K. E. Samouylov[*], E. S. Sopin[*†]

[*] *Department of Applied Probability and Informatics*
Peoples' Friendship University of Russia (RUDN University)
6 Miklukho-Maklaya St., Moscow, 117198, Russian Federation
[†] *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 the first in a series of two articles devoted to the review of "fork-join" (in the western classification) queuing systems or systems with the splitting of incoming queries. This system is a natural model for many other real systems. The article describes the fork-join queueing model construction and main characteristics of this model. Special attention is paid to methods of analysis of the response time of the system. Since the exact expression for the mean response time is known only for the case of two servers, the article gives a detailed description of the approach to obtaining an accurate expression of this characteristic. For the case when the number of servers is more than two, approximations of the mean response time are obtained by different methods, which is explained by the complexity of the studies due to the existing dependence between the queues of subqueries due to common arrival moments. The paper presents several methods of approximate analysis: various variants of empirical approximation, i.e. methods that refine the obtained characteristics by using the results of simulation modeling; interpolation methods using system load limit values in cases when the incoming flow and service time distributions are not exponential.

Key words and phrases: queuing system, splitting of requests, parallel service of requests, parallel processing, response time

References

1. R. Nelson, A. N. Tantawi, Approximate Analysis of Fork/Join Synchronization in Parallel Queues, IEEE Transactions on Computers 37 (1988) 739–743.
2. A. Thomasian, Analysis of Fork/Join and Related Queueing Systems, ACM Computing Surveys (CSUR) 47 (2) (2014) 17:1–17:71.
3. I. Tsimashenka, W. J. Knottenbelt, Reduction of Subtask Dispersion in Fork-Join Systems, in: Computer Performance Engineering, Springer Berlin Heidelberg, 2013, pp. 325–336.
4. A. V. Gorbunova, I. S. Zaryadov, S. I. Matyushenko, K. E. Samouylov, S. Ya. Shorgin, The Approximation of Response Time of a Cloud Computing System, Informatics and applications 9 (2015) 32–38, in Russian.

5. S. V. Vyshenski, P. V. Grigoriev, Yu. Yu. Dubenskaya, Ideal Synchronizer for Marked Pairs in Fork-Join Network, Review of applied and industrial mathematics 15 (3) (2008) 385–399, in Russian.
6. S. P. Moiseeva, I. A. Ivanovskaya, Analysis of the Mathematical Model of Parallel Service of Mixed Type Requests, Bulletin of the Tomsk Polytechnic University. Control, Computer Science and Technology 317 (5) (2010) 32–34, in Russian.
7. S. P. Moiseeva, L. A. Zhidkova, Investigation of the Parallel Service System with Multiple Claims of the Poisson Process, Bulletin of the Tomsk Polytechnic University. Control, Computer Science and Technology 17 (4) (2011) 49–54, in Russian.
8. 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, in: Proceedings of the Nineteenth International Scientific Conference Russia: Distributed computer and communication networks: control, computation, communications (DCCN-2016), Vol. 3, 2016, pp. 467–472.
9. 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, in: Proceedings of the 15th International Conference named after A. F. Terpugov: Information technologies and mathematical modelling (ITMM-2016), 2016, pp. 167–172, in Russian.
10. M. Mandelbaum, A.-B. Itzhak, Introduction to Queueing with Splitting and Matching, Israel Journal of Technology 6 (5) (1968) 376–382.
11. A. Duda, T. Czachórski, Performance Evaluation of Fork and Join Synchronization Primitives, Acta Informatica 24 (5) (1987) 525–533.
12. L. Green, A Queueing System in which Customers Require a Random Number of Servers, Operations Research 28 (6) (1980) 1335–1346.
13. K. J. Omahen, L. Schrage, A Queueing Analysis of a Multiprocessor System with Shared Memory, in: Proc. of the Symposium on Computer Communication Networks and Teletraffic, 1972, pp. 77–88.
14. A. Thomasian, A. Avizienis, Dynamic Scheduling of Tasks Requiring Multiple Processors, in: Proceedings of the 11th IEEE Computer Society International Conference (COMPCON'75 Fall), 1975, pp. 77–80.
15. T. Javidi, Cooperative and Non-Cooperative Resource Sharing in Networks: a Delay Perspective, IEEE Transactions on Automatic Control 53 (9) (2008) 2134–2142.
16. A. Kumar, R. Shorey, Performance Analysis and Scheduling of Stochastic Fork-Join Jobs in a Multicomputer System, IEEE Transactions on Parallel and Distributed Systems 10 (4) (1993) 1147–1164.
17. L. Flatto, S. Hahn, Two Parallel Queues Created by Arrivals with Two Demands I, SIAM Journal on Applied Mathematics 44 (5) (1984) 1041–1053.
18. G. Bolch, S. Greiner, H. de Meer, K. S. Trivedi, Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications, John Wiley & Sons, 2006.
19. F. Baccelli, Two Parallel Queues Created by Arrivals with Two Demands: the $M|G|2$ Symmetrical Case, INRIA Rapport de Recherche 426.
20. W. E. Boyce, R. C. DiPrima, Elementary Differential Equations and Boundary Value Problems, John Wiley & Sons, 2012.
21. G. P. Basharin, Introduction to Probability Theory, PFUR, Moscow, 1990, in Russian.
22. P. P. Bocharov, A. V. Pechinkin, Queueing Theory, PFUR, Moscow, 1995, in Russian.
23. G. P. Basharin, Lectures on the Mathematical Theory of Teletraffic, PFUR, Moscow, 2009, in Russian.
24. P. P. Bocharov, C. D'Apice, A. V. Pechinkin, S. Salerno, Queueing Theory, Brill Academic Publishers, 2004.
25. R. E. Barlow, F. Proschan, Statistical Theory of Reliability and Life Testing: Probability Models, John Wiley & Sons, 1981.
26. E. Varki, A. Merchant, H. Chen, The $M|M|1$ Fork-Join Queue with Variable

Subtasks.
URL http://www.cs.unh.edu/~varki/publication/2002-nov-open.pdf
27. E. Varki, Response Time Analysis of Parallel Computer and Storage Systems, IEEE Transactions on Parallel and Distributed Systems 12 (11) (2001) 1146–1161.
28. S. Varma, A. M. Makowski, Interpolation Approximations for Symmetric Fork-Join Queues, Performance Evaluation 20 (1994) 245–265.

Для цитирования:
Горбунова А. В., Зарядов И. С., Самуйлов К. Е., Сопин Э. С. Обзор систем параллельной обработки заявок // Вестник Российского университета дружбы народов. Серия: Математика. Информатика. Физика. — 2017. — Т. 25, № 4. — С. 350–362. — DOI: 10.22363/2312-9735-2017-25-4-350-362.

For citation:
Gorbunova A. V., Zaryadov I. S., Samouylov K. E., Sopin E. S. A Survey on Queuing Systems with Parallel Serving of Customers, RUDN Journal of Mathematics, Information Sciences and Physics 25 (4) (2017) 350–362. DOI: 10.22363/2312-9735-2017-25-4-350-362. In Russian.

Сведения об авторах:
Горбунова Анастасия Владимировна — 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)
Сопин Эдуард Сергеевич — кандидат физико-математических наук, доцент кафедры прикладной информатики и теории вероятностей РУДН, старший научный сотрудник ИПИ ФИЦ ИУ РАН (e-mail: sopin_es@rudn.university, тел.: +7(495)9550927)

Information about the authors:
Gorbunova A. V. — 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)
Sopin E. 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: sopin_es@rudn.university, phone: +7(495)9550927)