RUDN Journal of MIPh
Вестник РУДН. Серия МИФ
2018 Vol. 26 No. 1 28–38
http://journals.rudn.ru/miph

УДК 519.21;519.872
DOI: 10.22363/2312-9735-2018-26-1-28-38

Система обслуживания с делением и слиянием требований, в которой требование занимает все свободные обслуживающие приборы

О. А. Осипов

Кафедра системного анализа и автоматического управления
Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского
ул. Астраханская, д. 83, г. Саратов, Россия, 410012

В работе рассматривается многоприборная система массового обслуживания с ожиданием, требования в которой делятся в момент начала обслуживания на фрагменты так, что они одновременно занимают все свободные обслуживающие приборы. Предполагается, что обслуживающие приборы имеют различные интенсивности обслуживания. Фрагменты обслуживаются независимо друг от друга, интенсивность обслуживания фрагмента зависит от его величины. Фрагмент, завершивший своё обслуживание, освобождает обслуживающий его прибор. Требование будет считаться обслуженным только после того, как будет завершено обслуживание всех его фрагментов, сразу после чего фрагменты требования объединяются, и полученное исходное требование покидает систему обслуживания.

В предположении о пуассоновском входящем потоке и экспоненциальных длительностях обслуживания фрагментов на приборах для описанной системы обслуживания с использованием матрично-геометрического метода получены точные выражения для основных стационарных характеристик. Особое внимание уделено длительности времени пребывания требований в системе обслуживания. Приводится численный пример анализа системы рассматриваемого типа, обсуждаются результаты работы и перспективы дальнейших исследований.

Представленная в работе система обслуживания может применяться в качестве модели современных многопроцессорных вычислительных систем, а также других систем с параллельным и распределённым принципом функционирования.

Ключевые слова: системы массового обслуживания с делением и слиянием требований, неоднородные приборы, матрично-геометрический метод, многопроцессорные системы, распределённое и параллельное выполнение

1. Введение

Реальные системы, в которых имеет место параллельная обработка [1] (многопроцессорные системы, GRID-системы, распределённые базы данных), получают все большее распространение. В таких системах поступающие для обработки задачи делятся на более простые для выполнения подзадачи, которые распределяются по системе, занимая выделенные для них ресурсы. После завершения своего выполнения подзадачи освобождают выделенные им ресурсы, однако исходная задача будет считаться выполненной только после выполнения всех её подзадач.

Для анализа производительности указанных выше систем используются модели теории массового обслуживания, такие как [2]: с центральным делением без синхронизирующей очереди (centralized splitting model without synchronization queue, split-merge model), с центральным делением и синхронизирующей очередью (centralized splitting model with synchronization queue), с распределённым делением и синхронизирующей очередью (distributed splitting model with synchronization queue), которые обычно относят к классу систем массового обслуживания с делением и слиянием требований (fork-join queueing systems) [3–8], а также такие родственные

Статья поступила в редакцию 20 июля 2017 г.

модели как, например, модель независимых приборов (independent server model) [9,10], модель командного обслуживания (team service model) [11].

Общим во всех перечисленных моделях является деление поступающих требований на части — фрагменты, которые обслуживаются параллельно на приборах системы, и последующее объединение обслуженных фрагментов в исходные требования. Требование считается обслуженным и покидает систему только после окончания обслуживания всех его фрагментов. Распределение фрагментов поступающих требований по приборам системы с делением и слиянием требований, как и число фрагментов, получаемых при делении одного требования, задаётся некоторой детерминированной или вероятностной [12, 13] стратегией. Так, в работах [3, 4, 6–8] требования каждый раз делятся на одинаковое число фрагментов, по одному фрагменту для каждого обслуживающего прибора системы. С другой стороны, в [12] рассматривалась модель, в которой каждое требование может быть поделено на случайное число фрагментов, которые поступают на обслуживающие приборы в соответствии с некоторым распределением вероятностей. Обзор всех основных теоретических и прикладных результатов связанных с моделями данного класса можно найти в [14, 15].

В статье рассматривается многоприборная система массового обслуживания с очередью бесконечной вместимости и пуассоновским входящим потоком. Основным отличием от известных работ является то, что деление требования в момент начала обслуживания на фрагменты происходит так, что они занимают все свободные обслуживающие приборы, то есть число фрагментов, получаемых при делении одного требования, зависит от состояния системы обслуживания, а длительность обслуживания фрагмента зависит от его величины. Фрагменты обслуживаются независимо друг от друга. Фрагмент, обслуживание которого завершено, освобождает обслуживающий его прибор. Требование будет считаться обслуженным только после того, как будет завершено обслуживание всех его фрагментов, сразу после чего фрагменты требования объединяются, а полученное исходное требование покидает систему обслуживания. Также отметим, что во многих работах [4, 7, 12, 13] рассматривается случай идентичных обслуживающих приборов, в представленной же системе обслуживания приборы имеют различные интенсивности обслуживания.

Статья организована следующим образом. В разделе 2 подробно описывается изучаемая система массового обслуживания. Стационарное распределение состояний системы обслуживания, а также выражения для основных стационарных характеристик приводятся в разделах 3 и 4 соответственно. Раздел 5 содержит численные результаты анализа системы обслуживания.

2. Описание системы обслуживания

Рассматривается система обслуживания, состоящая из M обслуживающих приборов S_i, $i = 1, \ldots, M$, и очереди бесконечной вместимости с дисциплиной обслуживания FCFS. В систему обслуживания поступает пуассоновский поток требований с интенсивностью Λ. Обозначим через w класс требования и положим, что для всех требований, поступающих из источника, $w = 1$.

Требование, поступающее на обслуживание, делится на фрагменты так, что они занимают все свободные в данный момент времени обслуживающие приборы: если имеется $1 \leqslant d \leqslant M$ свободных приборов, то поступающее требование разделяется на d идентичных фрагментов с классом $w = d$. Каждый из фрагментов мгновенно занимает свободный прибор и начинает обслуживаться. То есть деление требования на $d > 1$ фрагментов происходит в том случае, когда в момент поступления этого требования очередь системы была пуста, при этом имелось более одного свободного прибора. В других случаях требование не делится, однако для удобства будем называть и такое требование фрагментом.

Длительность обслуживания любого фрагмента с классом w на приборе S_i, $i = 1, \ldots, M$, есть экспоненциально распределённая случайная величина с параметром $w\mu_i$, здесь μ_i — интенсивность обслуживания фрагмента (требования) класса $w = 1$.

Фрагмент, обслуживание которого завершено, 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 требование, которое покидает систему обслуживания.

3. Стационарное распределение системы обслуживания

Состояние системы обслуживания определим как вектор $\boldsymbol{n} = (n_0, n_1, \ldots, n_M)$, где n_0 — число требований в очереди,

$$n_i = \begin{cases} 0, & \text{если прибор } S_i \text{ свободен,} \\ w_i, & \text{если прибор } S_i \text{ обслуживает фрагмент класса } w_i; \end{cases}$$

где $i = 1, \ldots, M$.

Очевидно, что $(M+1)$-мерный процесс $\{\boldsymbol{n}(t), t \geqslant 0\}$ есть цепь Маркова с непрерывным временем, определённая на пространстве состояний $\mathcal{S}$,

$$\mathcal{S} = \left\{ \begin{matrix} (0, n_1, \ldots, n_M) : n_i \in \{0, \ldots, M\}, \\ i = 1, \ldots, M \end{matrix} \right\} \bigcup \left\{ \begin{matrix} (n_0, n_1, \ldots, n_M) : n_0 > 0, \\ n_i \in \{1, \ldots, M\},\ i = 1, \ldots, M \end{matrix} \right\}.$$

Обозначим через $q(\boldsymbol{n}, \boldsymbol{n}')$ интенсивность перехода цепи из состояния $\boldsymbol{n} \in \mathcal{S}$ в состояние $\boldsymbol{n}' \in \mathcal{S}$. Справедливо

1. если $n_i > 0$, $i = 0, 1, \ldots, M$,

$$q\left((n_0, n_1, \ldots, n_M), (n_0 + 1, n_1, \ldots, n_M)\right) = \Lambda, \tag{1}$$

$$q\left((n_0, \ldots, n_{j-1}, n_j, n_{j+1}, \ldots, n_M), (n_0 - 1, n_1, \ldots, n_{j-1}, 1, n_{j+1}, \ldots, n_M)\right) = \mu_j n_j, \tag{2}$$

   где $j \in \{1, \ldots, M\}$;
2. если $n_i > 0, i = 1, \ldots, M$,

$$q\left((0, n_1, \ldots, n_M), (1, n_1, \ldots, n_M)\right) = \Lambda; \tag{3}$$

3. если $n_i \geqslant 0, i = 1, \ldots, M$, и существует $j \in \{1, \ldots, M\}$ такое, что $n_j = 0$,

$$q\left((0, n_1, \ldots, n_M), (0, n'_1, \ldots, n'_M)\right) = \Lambda, \tag{4}$$

   где

$$n'_i = \begin{cases} n_i, & n_i > 0, \\ d(\boldsymbol{n}), & n_i = 0; \end{cases}$$

   $d(\boldsymbol{n})$ — число свободных приборов при нахождении системы обслуживания в состоянии $\boldsymbol{n}$;
4. если $n_i \geqslant 0$, $i = 1, \ldots, M$,

$$q\left((0, n_1, \ldots, n_{j-1}, n_j, n_{j+1}, \ldots, n_M), (0, n_1, \ldots, n_{j-1}, 0, n_{j+1}, \ldots, n_M)\right) = \mu_j n_j, \tag{5}$$

   где $j \in \{1, \ldots, M\}$.

Упорядочим состояния цепи Маркова в лексикографическом порядке. Под макросостоянием с номером i, будем понимать множество состояний $\mathcal{S}_i$, определяемое как

$$\mathcal{S}_i = \{(n_0, n_1, \ldots, n_M) \in \mathcal{S} : n_0 = i\}, \quad i = 0, 1, \ldots.$$

Мощности множеств $\mathcal{S}_0$ и $\mathcal{S}_i, i > 0$, равны $(M+1)^M$ и M^M соответственно. Заметим, что из (1) и (2) следует возможность перехода из макросостояния $\mathcal{S}_i$, $i > 0$, только в макросостояния $\mathcal{S}_{i+1}$ и $\mathcal{S}_{i-1}$, интенсивности этих переходов не зависят от i.

Цепь Маркова $\{\boldsymbol{n}(t), t \geqslant 0\}$ является квазипроцессом размножения и гибели [16], а инфинитезимальный оператор $\boldsymbol{Q} = (q(\boldsymbol{n}, \boldsymbol{n}'))$, $\boldsymbol{n}, \boldsymbol{n}' \in \mathcal{S}$, цепи имеет блочнодиагональный вид:

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

Матрицы $\boldsymbol{B}_{00}$, $\boldsymbol{B}_{01}$, $\boldsymbol{B}_{10}$ имеют размерность $(M+1)^M \times (M+1)^M$, $(M+1)^M \times M^M$, $M^M \times (M+1)^M$ соответственно. Матрицы $\boldsymbol{A}_0$, $\boldsymbol{A}_1$, $\boldsymbol{A}_2$ суть квадратные матрицы порядка M^M. Выражения (2) полностью определяют элементы матриц $\boldsymbol{A}_0$ и $\boldsymbol{B}_{10}$.

Отметим, что $\boldsymbol{A}_2 = \Lambda\mathbf{I}$, где $\mathbf{I}$— единичная матрица; $\boldsymbol{A}_1$— диагональная матрица, $\boldsymbol{A}_1 = -\operatorname{diag}(\boldsymbol{A}_0\mathbf{1} + \boldsymbol{A}_2\mathbf{1})$, где $\operatorname{diag}(\boldsymbol{a})$ задаёт диагональную матрицу с вектором $\boldsymbol{a}$ на главной диагонали, $\mathbf{1}$ — единичный вектор-столбец.

Матрицы $\boldsymbol{B}_{00}$, $\boldsymbol{B}_{01}$ определяются выражениями (3), (4), (5).

Для вычисления стационарного распределения $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)$ воспользуемся аппаратом матрично-аналитических решений [17], а именно матрично-геометрическим методом. Здесь $\boldsymbol{\pi}_i$, $i = 0, 1, \ldots$, есть вектор-строка, каждая компонента которого задаёт вероятность нахождения системы обслуживания в некотором состоянии из макросостояния $\mathcal{S}_i$ в соответствии с введённым лексикографическим порядком. Будем использовать следующие обозначения: $\boldsymbol{\pi}(\boldsymbol{n}) = \boldsymbol{\pi}(n_0, n_1, \ldots, n_M)$ — стационарная вероятность нахождения системы в состоянии $\boldsymbol{n}$; $\boldsymbol{\pi}(\mathcal{S}_i)$— стационарная вероятность нахождения системы в макросостоянии $\mathcal{S}_i$,

$$\boldsymbol{\pi}(\mathcal{S}_i) = \sum_{\boldsymbol{n} \in \mathcal{S}_i} \boldsymbol{\pi}(\boldsymbol{n}) = \boldsymbol{\pi}_i\mathbf{1}.$$

Для системы обслуживания стационарный режим будет существовать тогда и только тогда, когда выполнено условие

$$\psi = \frac{\Lambda}{\mu_1 + \ldots + \mu_M} < 1, \tag{6}$$

где ψ — коэффициент использования системы обслуживания.

Известно [16, 17], что при выполнении условия (6) стационарное распределение имеет вид $\boldsymbol{\pi}_i = \boldsymbol{\pi}_1\boldsymbol{R}^{i-1}$, $i = 1, 2, \ldots$, где $\boldsymbol{R}$ есть решение уравнения $\boldsymbol{A}_2 + \boldsymbol{R}\boldsymbol{A}_1 + \boldsymbol{R}^2\boldsymbol{A}_0 = \mathbf{0}$, векторы $\boldsymbol{\pi}_0$ и $\boldsymbol{\pi}_1$ находятся как решение уравнения

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

с условием нормировки $\boldsymbol{\pi}_0\mathbf{1} + \boldsymbol{\pi}_1(\mathbf{I} - \boldsymbol{R})^{-1}\mathbf{1} = 1$.

4. Вычисление стационарных характеристик

Используя стационарное распределение $\boldsymbol{\pi}$, определим математическое ожидание $\bar{B}$ числа требований в очереди системы обслуживания и математическое ожидание $\overline{W}$ длительности пребывания требований в очереди,

$$\bar{B} = \sum_{i=1}^{\infty} i\boldsymbol{\pi}(\mathcal{S}_i) = \boldsymbol{\pi}_1 \sum_{i=1}^{\infty} i\boldsymbol{R}^{i-1}\mathbf{1} = \boldsymbol{\pi}_1(\mathbf{I} - \boldsymbol{R})^{-2}\mathbf{1},$$

$$\overline{W} = \bar{B}/\Lambda.$$

Одной из основных характеристик, представляющих интерес в системах данного класса, является длительность времени пребывания требования [4, 6, 7] в системе обслуживания, которая определяется как длительность интервала времени от поступления требования в систему до возвращения требования в источник. Под длительностью обслуживания требования в системе обслуживания будем понимать максимум из длительностей интервалов времени, затраченных на обслуживание приборами всех фрагментов этого требования. Обозначим через $\overline{V}$ математическое ожидание длительности обслуживания требований в системе обслуживания, тогда математическое ожидание $\overline{T}$ длительности времени пребывания требований в системе обслуживания $\overline{T} = \overline{W} + \overline{V}$.

Рассмотрим два случая для поступающего из источника требования.

1. Требование застаёт все приборы системы обслуживания занятыми и встаёт в очередь. Тогда длительность обслуживания требования есть случайная величина с экспоненциальным распределением.
2. Требование застаёт в системе обслуживания $d > 0$ свободных обслуживающих приборов. В этом случае длительность обслуживания требования есть максимум из d независимых случайных величин с экспоненциальными распределениями.

Пусть система обслуживания находится в состоянии $\boldsymbol{n} = (n_0, n_1, \ldots, n_M)$, $n_0 > 0$, тогда вероятность $\alpha_i(n_1, \ldots, n_M)$ завершения обслуживания фрагмента на приборе S_i, $i = 1, \ldots, M$, определяется как

$$\alpha_i(n_1, \ldots, n_M) = \frac{\mu_i n_i}{\mu_1 n_1 + \ldots + \mu_M n_M};$$

м.о. $\boldsymbol{v}(n_1, \ldots, n_M)$ длительности обслуживания для требования, занявшего освободившийся прибор, определяется выражением

$$\boldsymbol{v}(n_1, \ldots, n_M) = \sum_{i=1}^{M} \frac{\alpha_i(n_1, \ldots, n_M)}{\mu_i}.$$

Будем рассматривать систему обслуживания при условии, что в очереди есть требования, тогда множество $\mathcal{S}^* = \{(n_1, \ldots, n_M) : n_i \in \{1, \ldots, M\}, i = 1, \ldots, M\}$, определяет множество состояний обслуживающих приборов. При данном предположении вероятность $\alpha_i(n_1, \ldots, n_{i-1}, n_i, n_{i+1}, \ldots, n_M)$ является вероятностью перехода из состояния $(n_1, \ldots, n_{i-1}, n_i, n_{i+1}, \ldots, n_M)$ в состояние $(n_1, \ldots, n_{i-1}, 1, n_{i+1}, \ldots, n_M)$, обусловленного завершением обслуживания некоторого фрагмента на приборе S_i. Случайный процесс, вложенный по моментам ухода фрагментов, является цепью Маркова с дискретным временем и множеством состояний $\mathcal{S}^*$. Пусть состояния в $\mathcal{S}^*$ упорядочены в лексикографическом порядке, а $\boldsymbol{P}$ задаёт матрицу переходов для цепи Маркова.

Предположим, что требование поступает в систему, когда в очереди находится $i > 0$ требований. Рассмотрим эволюцию цепи Маркова, описывающей переходы между состояниями приборов. В этом случае $\boldsymbol{\pi}(i, n_1, \ldots, n_M)/\boldsymbol{\pi}(\mathcal{S}_i)$ задаёт вероятность нахождения приборов в состоянии $(n_1, \ldots, n_M)$ при условии нахождения в очереди i требований. Вектор $\boldsymbol{\pi}_i/\boldsymbol{\pi}(\mathcal{S}_i)$ будет являться вектором начального распределения для рассматриваемого процесса. Через i завершений обслуживания поступившее в систему обслуживания требование будет первым в очереди на обслуживание, следовательно, математическое ожидание $\overline{V}_i$ длительности обслуживания требования, заставшего i требований в очереди, определяется как

$$\overline{V}_i = \frac{\boldsymbol{\pi}_i}{\boldsymbol{\pi}(\mathcal{S}_i)} \boldsymbol{P}^i \boldsymbol{v}, \quad i = 1, 2, \ldots,$$

где $\boldsymbol{v}$ есть вектор-столбец, составленный из элементов множества $\{\boldsymbol{v}(n_1, \ldots, n_M) : (n_1, \ldots, n_M) \in \mathcal{S}^*\}$ в соответствии с введённым лексикографическим порядком на множестве $\mathcal{S}^*$.

Рассмотрим теперь случай, когда поступающее в систему обслуживания требование не застаёт в очереди требований. Если все обслуживание приборы заняты, то математическое ожидание длительности обслуживания определяется аналогично предыдущему случаю. Пусть требование застаёт некоторые обслуживающие приборы свободными; обозначим через $D(\boldsymbol{n})$ множество номеров свободных приборов для состояния $\boldsymbol{n} \in \mathcal{S}$. Положим для некоторого $\boldsymbol{n}$, $D(\boldsymbol{n}) \neq \varnothing$, тогда требование разделится на $d(\boldsymbol{n}) > 0$ фрагментов, длительность обслуживания этого требования будет максимумом из независимых экспоненциально распределённых случайных величин. Таким образом, для требований, которые застают очередь системы пустой, математическое ожидание $\overline{V}_0$ длительности обслуживания требования определяется как

$$\overline{V}_0 = \frac{1}{\boldsymbol{\pi}(\mathcal{S}_0)} \left(\sum_{(n_1,\ldots,n_M) \in \mathcal{S}^*} \boldsymbol{\pi}(0, n_1, \ldots, n_M) \boldsymbol{v}(n_1, \ldots, n_M) + \right.$$

$$\left. + \sum_{\boldsymbol{n} \in \mathcal{S},\, d(\boldsymbol{n}) > 0} \boldsymbol{\pi}(\boldsymbol{n}) \mathsf{E} \left[\max_{i \in D(\boldsymbol{n})} \{ \exp\left(d(\boldsymbol{n}) \mu_i\right) \} \right] \right),$$

здесь E обозначает оператор математического ожидания, $\exp(a)$ — случайная величина с экспоненциальным распределением с параметром a. Обсуждение нахождения распределения максимума случайных величин можно найти, например, в [16, 18].

Тогда для математического ожидания $\overline{V}$ длительности обслуживания требований справедливо

$$\overline{V} = \sum_{i=0}^{\infty} \boldsymbol{\pi}(\mathcal{S}_i) \overline{V}_i = \boldsymbol{\pi}(\mathcal{S}_0) \overline{V}_0 + \sum_{i=1}^{\infty} \boldsymbol{\pi}_i \boldsymbol{P}^i \boldsymbol{v} = \boldsymbol{\pi}(\mathcal{S}_0) \overline{V}_0 + \boldsymbol{\pi}_1 \sum_{i=0}^{\infty} \boldsymbol{R}^i \boldsymbol{P}^i \boldsymbol{P} \boldsymbol{v}. \tag{7}$$

Рассмотрим подробнее ряд $\sum\limits_{i=0}^{\infty} \boldsymbol{R}^i \boldsymbol{P}^i$, обозначив через $\boldsymbol{X}$ его сумму,

$$\sum_{i=0}^{\infty} \boldsymbol{R}^i \boldsymbol{P}^i = \mathbf{I} + \boldsymbol{R}(\mathbf{I} + \boldsymbol{R}\boldsymbol{P} + \boldsymbol{R}^2 \boldsymbol{P}^2 + \ldots)\boldsymbol{P} = \mathbf{I} + \boldsymbol{R}\boldsymbol{X}\boldsymbol{P} = \boldsymbol{X},$$

$$\boldsymbol{X} = \mathbf{I} + \boldsymbol{R}\boldsymbol{X}\boldsymbol{P}. \tag{8}$$

Матрица $\boldsymbol{X}$ есть решение уравнения (8), которое может быть переписано в привычном для уравнения Сильвестра [19, 20] виде (9):

$$\boldsymbol{R}^{-1} = \boldsymbol{R}^{-1}\boldsymbol{X} - \boldsymbol{X}\boldsymbol{P}. \tag{9}$$

Из (7) после преобразований получаем

$$\overline{V} = \boldsymbol{\pi}(\mathcal{S}_0)\overline{V}_0 + \boldsymbol{\pi}_1\boldsymbol{X}\boldsymbol{P}\boldsymbol{v}.$$

Математическое ожидание числа требований в системе $\overline{N} = \Lambda\overline{T}$.

5. Пример

На основании полученных выражений рассмотрим изменение математического ожидания длительности пребывания требований в системе обслуживания в зависимости от числа обслуживающих приборов ($M = 2, 3, 4$) и интенсивности входящего потока. Предполагается, что $\mu_1 = \ldots = \mu_M = 1$, коэффициент использования меняется в следующих границах: $0,1 \leqslant \psi \leqslant 0,9$.

Графики зависимости представлены на рис. 1. Отметим, что полученные с использованием выражений численные результаты были подтверждены также результатами имитационного моделирования.

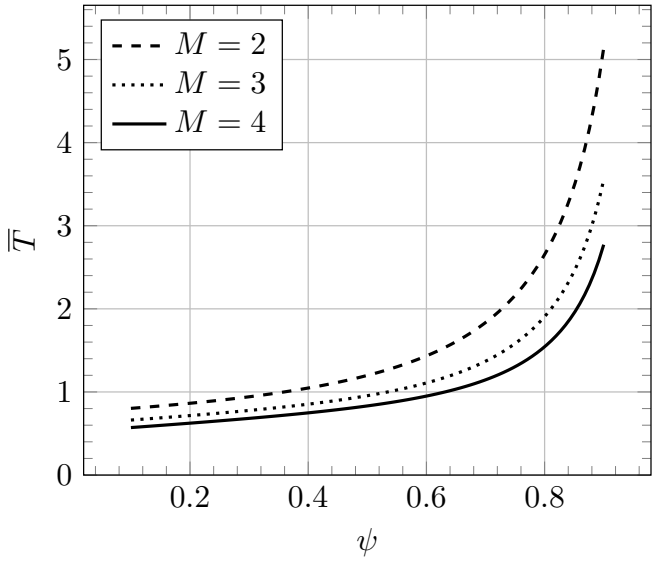

Рис. 1. Зависимость математического ожидания длительности пребывания требований в системе обслуживания

Отметим, что когда коэффициент использования приближается к единице, рассматриваемая система ведёт себя аналогично системе с M обслуживающими приборами без деления требований, тогда в случае, когда все обслуживающие приборы имеют одинаковую интенсивность обслуживания μ, математическое ожидание $\overline{T}$ длительности времени пребывания требований в системе может быть приближённо найдено как

$$\overline{T} \approx \frac{1}{\mu}\left[p_0\frac{M^{M-1}\psi^M}{M!(1-\psi)^2} + 1\right], \quad p_0 = \left[\frac{(M\psi)^M}{M!(1-\psi)} + \sum_{i=0}^{M-1}\frac{(M\psi)^i}{i!}\right]^{-1}.$$

.

Сравнение результатов для случая, когда $M = 4$, $\mu = 1$, приведено в табл. 1.

Таблица 1
Математическое ожидание длительности пребывания требований в системе для двух различных моделей

ψ	0.1	0.3	0.5	0.7	0.8	0.9	0.95
с делением требований	0,5714	0,6821	0,8324	1,1540	1,5604	2,8001	5,2955
без деления требований	1,0002	1,0132	1,0870	1,3572	1,7455	2,9694	5,4571

Видно, что во всех случаях система без деления требований на фрагменты имеет большее математическое ожидание длительности пребывания требований в системе, то есть может быть использована для нахождения верхней границы для $\overline{T}$.

6. Заключение

В статье рассмотрена многоприборная система массового обслуживания с очередью бесконечной вместимости, в которой требование делится на фрагменты так, что они занимают все свободные обслуживающие приборы. С использованием матрично-геометрического метода получено стационарное распределение вероятностей состояний системы, а также точные выражения для основных стационарных характеристик системы (математическое ожидание длительности пребывания требований в системе обслуживания, математическое ожидание длительности пребывания требований в очереди и т.д.).

В качестве направления дальнейших исследований могут быть рассмотрены различные дисциплины разделения требований, например, с ограничением на максимальное число фрагментов, порождаемых одним требованием.

Литература

1. Models for Parallel and Distributed Computation / Ed. by R. Corrêa, I. Dutra, M. Fiallos, F. Gomes. — Springer US, 2002.
2. *Narahari Y., Sundarrajan P.* Performability Analysis of Fork-join Queueing Systems // Journal of the Operational Research Society. — 1995. — Vol. 46, No 10. — Pp. 1237–1249.
3. *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.
4. *Nelson R., Tantawi A. N.* Approximate Analysis of Fork/Join Synchronization in Parallel Queues // IEEE Transactions on Computers. — 1988. — Vol. 37, No 6. — Pp. 739–743.
5. *Ko S.-S., Serfozo R. F.* Response Times in $M|M|s$ Fork-Join Networks // Advances in Applied Probability. — 2004. — Vol. 36, No 3. — Pp. 854–871.
6. 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. 31–38.
7. 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.

8. *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.
9. *Green L.* A Queueing System in Which Customers Require a Random Number of Servers // Operations Research. — 1980. — Vol. 28, No 6. — Pp. 1335–1346.
10. *Rumyantsev A., Morozov E.* Stability Criterion of a Multiserver Model with Simultaneous Service // Annals of Operations Research. — 2015. — Vol. 252, No 1. — Pp. 29–39.
11. *Omahen K., Schrage L.* A Queueing Analysis of a Multiprocessor System with Shared Memory // Proceedings of the Symposium on Computer Communication Networks and Teletraffic. — 1972. — Pp. 77–88.
12. *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.
13. *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.
14. *Thomasian A.* Analysis of Fork/Join and Related Queueing Systems // ACM Computing Surveys. — 2014. — Vol. 47, No 2. — Pp. 17:1–17:71.
15. 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, вып. 4. — С. 350–362.
16. *He Q.-M.* Fundamentals of Matrix-Analytic Methods. — New York: Springer, 2014.
17. *Neuts M. F.* Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. — Baltimore: The Johns Hopkins University Press, 1981.
18. *David H. A., Nagaraja H. N.* Order Statistics. — John Wiley & Sons, Inc., 2003.
19. *Gantmacher F. R.* The Theory of Matrices. — Chelsea Publishing Company, 1959.
20. *Ланкастер П.* Теория матриц. — М.: Наука, 1973.

UDC 519.21;519.872
DOI: 10.22363/2312-9735-2018-26-1-28-38

A Heterogeneous Fork-Join Queueing System in Which Each Job Occupy All Free Servers

O. A. Osipov

Department of System Analysis and Automatic Control
Saratov State University (SSU)
83 Astrahanskaya St., Saratov, 410012, Russian Federation

In this paper, we consider a multiserver queueing system with heterogeneous servers in which each job is split to be serviced into a number of tasks, one for each free server. The tasks are serviced independently, but service time depends on weight of the tasks. A job is considered to be complete only when all the tasks associated with the job have been executed to completion.

Applying a matrix-geometric approach, we obtain the exact expression for the stationary distribution of the number of jobs in the system under exponential assumptions. Using the distribution, we derive other important performance measures. Special attention is paid to the sojourn time in the queueing system (the time to complete a job). 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.

Key words and phrases: fork-join queueing systems, heterogeneous servers, matrix-geometric method, multiprocessor systems, distributed computing systems

References

1. R. Corrêa, I. Dutra, M. Fiallos, F. Gomes (Eds.), Models for Parallel and Distributed Computation, Springer US, 2002. doi:10.1007/978-1-4757-3609-0.
2. Y. Narahari, P. Sundarrajan, Performability Analysis of Fork-join Queueing Systems, Journal of the Operational Research Society 46 (10) (1995) 1237–1249. doi:10.1057/jors.1995.171.
3. L. Flatto, S. Hahn, Two Parallel Queues Created by Arrivals with Two Demands I, SIAM Journal on Applied Mathematics 44 (5) (1984) 1041–1053. doi:10.1137/0144074.
4. R. Nelson, A. N. Tantawi, Approximate Analysis of Fork/Join Synchronization in Parallel Queues, IEEE Transactions on Computers 37 (6) (1988) 739–743. doi:10.1109/12.2213.
5. S.-S. Ko, R. F. Serfozo, Response Times in $M|M|s$ Fork-Join Networks, Advances in Applied Probability 36 (3) (2004) 854–871. doi:10.1017/s000186780001315x.
6. 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. doi:10.14357/19922264150304.
7. 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. doi:10.1007/s10479-008-0312-7.
8. 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.
9. L. Green, A Queueing System in Which Customers Require a Random Number of Servers, Operations Research 28 (6) (1980) 1335–1346.
10. A. Rumyantsev, E. Morozov, Stability Criterion of a Multiserver Model with Simultaneous Service, Annals of Operations Research 252 (1) (2015) 29–39. doi:10.1007/s10479-015-1917-2.
11. K. Omahen, L. Schrage, A Queueing Analysis of a Multiprocessor System with Shared Memory, in: Proceedings of the Symposium on Computer Communication Networks and Teletraffic, 1972, pp. 77–88.
12. 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.
13. T. Javidi, Cooperative and Non-Cooperative Resource Sharing in Networks: A Delay Perspective, IEEE Transactions on Automatic Control 53 (9) (2008) 2134–2142. doi:10.1109/TAC.2008.930186.
14. A. Thomasian, Analysis of Fork/Join and Related Queueing Systems, ACM Computing Surveys 47 (2) (2014) 17:1–17:71. doi:10.1145/2628913.
15. 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 (2017) 350–362, in Russian. doi:10.22363/2312-9735-2017-25-4-350-362.
16. Q.-M. He, Fundamentals of Matrix-Analytic Methods, Springer, New York, 2014. doi:10.1007/978-1-4614-7330-5.
17. M. F. Neuts, Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach, The Johns Hopkins University Press, Baltimore, 1981.
18. H. A. David, H. N. Nagaraja, Order Statistics, John Wiley & Sons, Inc., 2003. doi:10.1002/0471722162.
19. F. R. Gantmacher, The Theory of Matrices, Chelsea Publishing Company, 1959.
20. P. Lancaster, M. Tismenetsky, The Theory of Matrices, 2d edition, Academic Press, 1985, in Russian.

Для цитирования:
Осипов О. А. Система обслуживания с делением и слиянием требований, в которой требование занимает все свободные обслуживающие приборы // Вестник Российского университета дружбы народов. Серия: Математика. Информатика. Физика. — 2018. — Т. 26, № 1. — С. 28–38. — DOI: 10.22363/2312-9735-2018-26-1-28-38.

For citation:
Osipov O. A. A Heterogeneous Fork-Join Queueing System in Which Each Job Occupy All Free Servers, RUDN Journal of Mathematics, Information Sciences and Physics 26 (1) (2018) 28–38. DOI: 10.22363/2312-9735-2018-26-1-28-38. In Russian.

Сведения об авторах:
Осипов Олег Александрович — ассистент кафедры системного анализа и автоматического управления СГУ (e-mail: oleg.alex.osipov@gmail.com, тел.: +7(8452)213620)

Information about the authors:
Osipov O. A. — assistant of Department of System Analysis and Automatic Control of Saratov State University (SSU) (e-mail: oleg.alex.osipov@gmail.com, phone: +7(8452)213620)