УДК 004.032
doi:10.31799/1684-8853-2025-6-42-50
EDN: EGLAUQ

Научные статьи
Articles

Оценка характеристик модели распределенных транзакционных приложений с микросервисной архитектурой и параллельными узлами

А. В. Горбунова[a], *канд. физ.-мат. наук, старший научный сотрудник, orcid.org/0000-0002-9183-0426, avgorbunova@list.ru*
[a]*Институт проблем управления им. В. А. Трапезникова РАН, Профсоюзная ул., 65, Москва, 117997, РФ*

Введение: *микросервисная архитектура, позволяющая создавать приложения как набор независимых микросервисов для совместной работы над выполнением некоторого общего клиентского запроса, стала в последнее время основой, или даже стандартом, для развертывания сложных систем, затрагивающих множество физических структур и устройств. Кроме того, внедрение в подобные системы, особенно системы с высокой загрузкой, параллельные сценарии обслуживания, позволяет повысить их эффективность и производительность.* ***Цель:*** *разработать математическую модель распределенных транзакционных приложений с микросервисной архитектурой и параллельными узлами и оценить такой показатель ее функционирования, как среднее время отклика.* ***Результаты:*** *представлена математическая модель распределенных транзакционных приложений с микросервисной архитектурой в виде сети массового обслуживания с последовательными узлами, один из которых имеет параллельную структуру с несколькими подузлами, число которых больше двух. На основе метода декомпозиции для анализа сетей массового обслуживания предлагается подход к оценке среднего времени отклика рассматриваемой системы с использованием известных результатов для оценки отдельных узлов сети типа G/G/1, а также узлов с разделением и параллельным обслуживанием. Результаты вычислительных экспериментов позволяют сделать выводы о допустимости использования предложенного подхода, а также получить рекомендации относительно применимости формул для различных уровней загруженности системы, в частности тех, для которых средняя погрешность аппроксимации не превышает 10 %.* ***Практическая значимость:*** *предложенная в работе модель и метод ее исследования могут быть использованы для первичной оценки и прогнозирования среднего времени отклика транзакционных приложений с параллельными узлами при разных уровнях загруженности системы и, как следствие, способствовать поддержанию необходимого качества обслуживания пользователей транзакционных приложений.*

Ключевые слова *— транзакционные приложения, микросервисная архитектура, распределенные системы, параллельные операции, сеть массового обслуживания, среднее время отклика, fork-join, G/G/1.*

Для цитирования: Горбунова А. В. Оценка характеристик модели распределенных транзакционных приложений с микросервисной архитектурой и параллельными узлами. *Информационно-управляющие системы*, 2025, № 6, с. 42–50. doi:10.31799/1684-8853-2025-6-42-50, EDN: EGLAUQ
For citation: Gorbunova A. V. Evaluation of the characteristics of a distributed transactional application model with microservice architecture and fork-join structures. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2025, no. 6, pp. 42–50 (In Russian). doi:10.31799/1684-8853-2025-6-42-50, EDN: EGLAUQ

Введение

Транзакционные приложения представляют собой системы, обслуживающие (оказывающие) транзакционные услуги, под которыми в большинстве случаев подразумевается обработка финансовых или коммерческих операций. Примерами транзакционных приложений могут являться системы 1) онлайн-банкинга, 2) электронной коммерции, 3) бронирования билетов и другие сервисы, связанные с транзакциями. В целом, речь идет о системах, управляющих большим потоком транзакций и имеющих, как правило, распределенную базу данных, поскольку традиционные базы данных с реляционной архитектурой в такой ситуации оказываются недостаточно продуктивными [1–3]. При этом под транзакционными услугами, соответственно, понимаются в первом случае операции по управлению пользователем своими счетами, а именно перевод денежных средств, оплата различного рода счетов, обмен валюты и т. п., во втором случае — операции, связанные с онлайн-торговлей товарами или услугами, а именно прием и обработка заказов, проверка статуса заказа и т. п., в третьем случае — это операции, связанные с продажей билетов, оплатой их стоимости и т. д. То есть это операции с базами данных, которые обслуживают рабочий процесс и могут включать в себя создание, удаление или изменение данных.

Естественно, что для систем транзакционных услуг важную роль играет именно производительность и, возможно, даже большую по сравнению, например, с интерактивностью (активным взаимодействием пользователя с системой), как

в случае с платформами, организующими видеоконференции, онлайн-игры или онлайн-чаты, позволяющими пользователям общаться в реальном времени через интернет. Особенно это касается высоконагруженных систем.

В таких условиях предпочтительным выбором для большинства поставщиков услуг становится микросервисная архитектура [1–5]. В отличие от монолитной архитектуры, являющейся структурой со связанными в единое целое компонентами, микросервисная архитектура представляет собой систему, состоящую из отдельных структурных элементов — микросервисов, которые могут иметь свои собственные базы данных. При этом предполагается, что клиент, направляя свой запрос в систему, может инициировать процедуру одновременного обращения к нескольким микросервисам с собственной базой данных, в каждую из которых требуется внести необходимые изменения. В этом случае говорят о распределенных транзакциях [6]. Например, процесс перевода денежных средств может задействовать два микросервиса: один списывает средства со счета отправителя, а второй зачисляет их на нужный расчетный счет получателя [1, 7, 8].

Использование микросервисной архитектуры имеет свои преимущества и недостатки [4, 5]. В частности, разбиение сложной архитектуры на более простые и независимые элементы облегчает добавление новых микросервисов и обеспечивает масштабируемость, однако усложняет координацию транзакций; распределение по различным узлам, в том числе и параллельным, снижает общую нагрузку на систему и повышает производительность, уменьшая время отклика системы, но при этом порождает сложности с синхронизацией и согласованностью данных [4, 5, 9].

Несмотря на неизбежно сопутствующие трудности, характерные для распределенных систем, опыт внедрения описанных технологий для организации рабочих процессов управления транзакциями (финансовыми транзакциями) на примере платформы PayPal является довольно успешным [10, 11].

Таким образом, на первый план выходит необходимость адекватно прогнозировать показатели производительности систем транзакционных услуг при меняющейся рабочей нагрузке, что в свою очередь позволит оценить ее надежность, а также повысить запас прочности и заложить потенциал адаптации к меняющимся условиям внешней среды и требованиям пользователей.

Традиционно для оценки характеристик качества функционирования различных телекоммуникационных систем используются инструменты теории массового обслуживания. Так, в работе [12] предлагается использовать сети Джексона для математического моделирования и первичного анализа систем транзакционных услуг, содержащих параллельные узлы, а для более сложных вариантов распределений (неэкспоненциальных) предлагается использовать имитационное моделирование. В статье [13] для исследования характеристик рабочих процессов транзакционных услуг рассматриваются математические модели с узлами более сложной архитектуры (G/G/1), однако без учета параллелизма.

В данной статье предлагается математическая модель для оценки среднего времени отклика систем последовательных транзакционных услуг с микросервисной архитектурой, содержащей параллельные узлы в виде сети массового обслуживания с линейной топологией, некоторые узлы которой представляют собой так называемые fork-join-структуры. При этом рассматривается случай, когда время обслуживания характеризуется распределением Парето, а входящий в систему поток является пуассоновским. Несмотря на то, что в работе исследуется частный случай подобной системы, предложенный метод для оценки характеристик модели позволяет распространить его и на другие варианты вероятностных распределений для интервалов времени между поступлениями очередных запросов и длительностей интервалов времени их обслуживания.

Для определения характеристик узлов непараллельной структуры сетей линейной топологии используется несколько вариантов аппроксимаций, представляющих собой известные классические результаты. В целом же предполагается обращение к методу декомпозиции — оценке показателей производительности каждого узла системы по отдельности с дальнейшим использованием полученных результатов для оценки характеристик сети в итоге.

Для оценки же характеристик параллельных узлов (fork-join) применяется комплексный подход, включающий методы интеллектуального анализа данных, который позволяет получить хорошее качество приближения для аналитического выражения.

Fork-join-структура представляет собой узел, при поступлении на который запрос разделяется на подзапросы, каждый из которых направляется на обслуживание в отдельный подузел, причем время обслуживания всего запроса является максимумом из всех времен пребывания подзапросов в своих подузлах. За счет такой организации обслуживания запроса повышается производительность по сравнению с последовательным выполнением операций, при этом выигрыш во времени получается значительным. Поэтому подобные структуры довольно популярны, даже несмотря на сложности, связанные с техниче-

ской реализацией корректного процесса разбиения запроса на более мелкие задачи.

Математическая модель системы транзакционных услуг с параллельными узлами

Итак, рассмотрим распределенное транзакционное приложение. Поскольку предполагается, что возможно одновременное обращение к нескольким микросервисам, то оно будет содержать параллельный узел, который будет моделироваться с помощью fork-join-системы массового обслуживания, содержащей K подузлов. Переход рабочего процесса к следующему узлу будет означать завершение всех необходимых операций на каждом из микросервисов системы данного узла.

Допустим, что входящий в систему поток является пуассоновским, причем средняя длительность интервала между соседними поступлениями требований равна $1/\lambda$, а длительность интервала обслуживания на приборе как каждого параллельного подузла, так и каждого следующего узла системы имеет распределение Парето со следующей функцией распределения:

$$B_{Pa}(t) = 1 - \left(\frac{\alpha - 1}{\alpha}\frac{1}{t}\right)^{\alpha}, \; t \geq \frac{\alpha - 1}{\alpha} \qquad (1)$$

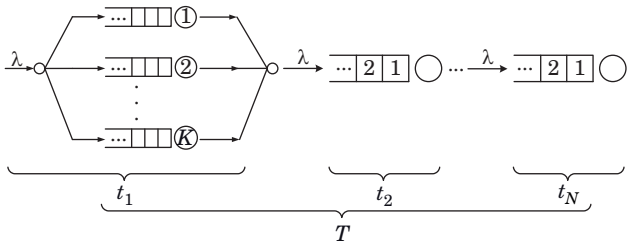
со средним значением $b_{Pa} = 1$, вторым моментом $b_{Pa}^{(2)} = \dfrac{(\alpha - 1)^2}{\alpha(\alpha - 2)}$ и параметром $\alpha > 3$.

Выбор распределения Парето, т. е. степенного распределения, обуславливается распространенностью его применения для моделирования процессов, протекающих в различного рода сетях передачи данных. Что касается пуассоновского входящего потока, то, несмотря на некоторые его ограничения, он все еще используется для построения моделей и первичной оценки параметров инфокоммуникационных систем [14–18].

После прохождения параллельного узла рабочий процесс последовательно переходит от одного узла системы к другому до N-го узла включительно, пока не завершит все необходимые операции по обслуживанию запроса пользователя системы.

Схема описанной модели функционирования приложения (рис. 1) представляет собой сеть массового обслуживания линейной архитектуры, но с параллельным узлом.

Одним из основных методов исследования сетей массового обслуживания является метод декомпозиции, который предполагает анализ отдельных фрагментов сети изолированно.



■ ***Рис. 1.*** Схема модели транзакционного приложения с микросервисной архитектурой и параллельным узлом

■ ***Fig. 1.*** Transactional application model diagram with microservice architecture and parallel node

Причем под фрагментом иногда может подразумеваться не только один узел, но и некоторая их совокупность.

Такой подход позволяет получить точные аналитические решения лишь для ограниченного класса сетей, к которым относятся открытые экспоненциальные сети Джексона и некоторые их расширения. В остальных же случаях, которых существенное большинство, данный способ предполагает приближенный анализ. При этом, разумеется, нельзя не отметить, что в отдельных ситуациях точный аналитический подход все-таки возможен, но из-за высокой размерности пространства состояний исследуемых систем, как правило, является нерациональным.

Оценка среднего времени отклика системы

Одним из наиболее важных показателей производительности системы является ее среднее время отклика. Корректная оценка этой характеристики важна для провайдеров (поставщиков услуг) в связи с необходимостью соблюдения соглашения о качестве оказываемых ими услуг (Quality of Service, QoS). Кроме того, на основе полученных оценок выстраивается стратегия выделения необходимого количества ресурсов под выполнение соответствующих задач, поскольку поддержание работоспособности системы является затратной статьей, что сказывается на общей стоимости предоставляемых провайдером услуг и его конкурентоспособности [19].

Среднее время отклика всей сети определяется суммой средних времен прохождения запроса через каждый отдельный узел

$$T = t_1 + t_2 + \ldots + t_N. \qquad (2)$$

Согласно методу декомпозиции остается определить величину t_i для каждого имеющегося в системе узла, $i = 1, \ldots, N$. Поэтому на первый

план выходят методы, предполагающие одномерную диффузионную аппроксимацию узлов типа G/G/1. При этом стоит отметить, что иногда они допускают довольно серьезные относительные погрешности приближения в зависимости от величины загрузки узлов и выбранных конкретных типов распределений для входящего потока и времен обслуживания.

Кроме того, информация о распределении входящего потока и, соответственно, его первых и вторых моментах доступна только для самого первого узла, который в данном случае представляет собой систему типа fork-join, т. е. параллельный узел.

Для остальных же узлов, учитывая неограниченные емкости накопителей, а также линейную архитектуру сети, можем допустить, что среднее время между соседними поступлениями требований будет таким же, как и на первом узле, а именно $1/\lambda$.

Изучение характеристик выходных потоков узлов рассматриваемой сети и, в частности, самого первого узла, имеющего параллельную структуру, представляет собой отдельную непростую задачу, поэтому допущения, касающиеся оценки моментов выходящих потоков, будут накладывать определенные ограничения, влияющие на точность получаемого решения.

Что касается вторых моментов, а точнее, коэффициентов вариации CV_i для входящих в i-й узел потоков ($i = 2, ..., N$), необходимых для оценки среднего времени пребывания в каждом из узлов, то здесь можно воспользоваться некоторыми известными приближениями. Согласно [20–22]:

$$CV_i = CV_{Pa_{i-1}}; \tag{3}$$

$$CV_i = \rho_{i-1}\left(1-\rho_{i-1}\right) + \rho_{i-1}^2 CV_{Pa_{i-1}}^2 + \left(1-\rho_{i-1}\right)CV_{i-1}^2; \tag{4}$$

$$CV_i = CV_{i-1}^2 + 2\rho_{i-1}CV_{Pa_{i-1}}^2 - \rho_{i-1}\left(CV_{i-1}^2 + CV_{Pa_{i-1}}^2\right)\times \times f\left(\rho_{i-1},\, CV_{i-1},\, CV_{Pa_{i-1}}\right); \tag{5}$$

$$CV_i = \rho_{i-1}^2 CV_{Pa_{i-1}}^2 + \left(1-\rho_{i-1}^2\right)CV_{i-1}^2, \tag{6}$$

где ρ_{i-1} — загрузка $(i-1)$-го узла, которая в рамках рассматриваемой модели идентична для всех узлов: $\rho_i = \rho = \lambda$, $i = 1, ..., N$; $CV_{Pa_{i-1}}$ — коэффициент вариации времени обслуживания:

$$CV_{Pa_i} = CV_{Pa} = \frac{1}{\sqrt{\alpha(\alpha-2)}},\ i=1,\ ...,\ N,$$

причем допустим, что это справедливо и для первого узла; функцию $f(\rho, CV_i, CV_{Pa})$ определим далее по тексту.

Выражения (3)–(6) используются для оценки среднего времени пребывания в i-м узле [23]

$$t_i \approx \frac{\rho}{2(1-\rho)}\left(CV_i^2 + CV_{Pa}^2\right)f\left(\rho,\, CV_i,\, CV_{Pa}\right) + 1,\quad i = 2,\ ...,\ N, \tag{7}$$

где

$$f(\rho,\, CV_i,\, CV_{Pa}) = \begin{cases} \exp\left\{-\dfrac{2(1-\rho)}{3\rho}\dfrac{\left(1-CV_i^2\right)^2}{CV_i^2 + CV_{Pa}^2}\right\}, & \text{если } CV_i \le 1; \\ \exp\left\{-(1-\rho)\dfrac{CV_i^2 - 1}{CV_i^2 + 4CV_{Pa}^2}\right\}, & \text{если } CV_i > 1. \end{cases} \tag{8}$$

Для времени пребывания в самом первом узле, который представляет собой параллельную структуру, состоящую из K подузлов, воспользуемся приближением, представленным в работах [24, 25]:

$$t_1 \approx 1 + \frac{(\alpha-1)^2}{2\alpha(\alpha-2)}\frac{\rho}{1-\rho} + \left(K^{\frac{1}{\alpha}} - 1\right)\times \times \left(1{,}25918 + 0{,}36996\alpha - 1{,}97400\rho - 0{,}28495\alpha\rho + 1{,}40841\rho^2 - 0{,}01122\alpha^2\right)\times \times \sqrt{\frac{1}{\alpha(\alpha-2)} + \frac{(\alpha-1)^3}{3\alpha^2(\alpha-3)}\frac{\rho}{1-\rho} + \frac{(\alpha-1)^4}{4\alpha^2(\alpha-2)^2}\frac{\rho^2}{(1-\rho)^2}}. \tag{9}$$

Данное выражение показывает хорошее качество приближения для значений параметров модели $\alpha \in [4; 10]$, $\rho \in [0{,}1; 0{,}9]$ и числа подузлов параллельной структуры $K = 2, ..., 20$, при этом средняя относительная погрешность приближения составляет около 1,6 %, а максимальная не превышает 4 %.

Стоит отметить, что анализ представленной сети и оценка ее характеристик в случае показательного распределения времени обслуживания со средним значением $b = 1$ и, соответственно, загрузкой $\rho = \lambda < 1$ на каждом из приборов не будет представлять серьезной сложности, если допустить, что поток запросов, поступающий на второй узел, будет также пуассоновским. Тогда времена пребывания в каждом i-м узле, начиная со второго и заканчивая последним, будут иметь показательное распределение с параметром $(1 - \lambda)$, а общее суммарное время прохождения запроса через $(N - 1)$ узел сети составит $\dfrac{N-1}{1-\lambda}$.

Единственная трудность — в определении времени пребывания на первом параллельном узле сети, для которого в случае $K > 2$ нет точных решений [14, 26–31]. Однако в этой ситуации предлагается воспользоваться приближением, полученным в работе [27], которое уточняет наиболее известную аппроксимацию для fork-join-системы с пуассоновским входящим потоком и экспоненциальным распределением времени обслуживания из [26]. Таким образом, выражение для оценки среднего времени пребывания в сети T, учитывая, что $\rho = \lambda$, будет иметь вид

$$T \approx \frac{N-1}{1-\lambda} + \frac{\lambda}{1-\lambda}\left(\frac{H_K}{H_2} - 1\right) \times$$
$$\times \left(0{,}08720 - 0{,}07024\left(\frac{H_K}{H_2} - 1\right) + 0{,}0964\lambda\right) +$$
$$+ \left[\frac{H_K}{H_2} + \frac{4}{11}\left(1 - \frac{H_K}{H_2}\right)\lambda\right]\frac{12-\lambda}{8}\frac{1}{1-\lambda}, \qquad (10)$$

где $H_K = \sum_{i=1}^{K} 1/i$ — частичная сумма гармонического ряда; K — число подузлов в первом параллельном узле.

Численный эксперимент

В данном разделе проверим работоспособность предложенной аппроксимации среднего времени отклика для модели распределенных транзакционных приложений и ее качество, учитывая сделанные предположения о параметрах первого параллельного узла, используемых при аппроксимации коэффициентов вариации входящего потока для последующих узлов сети.

Для этого рассмотрим модель сети транзакционного приложения, число последовательных узлов N в которой меняется от трех до 10, а количество параллельных подузлов K первого узла равно пяти. Уровень загрузки каждого узла $\rho = \lambda \in [0{,}05; 0{,}90]$ с шагом 0,05. Для оценки величины t_i, $i = 1, \ldots, N$, будем использовать выражение (7), где величина коэффициента вариации для входящего в узел потока будет вычисляться по одной из формул (3)–(6), а для оценки среднего времени пребывания в первом узле, соответственно, формулу (9).

Будем сравнивать значения, рассчитанные по аналитическим формулам, и результаты имитационного моделирования математической модели транзакционного приложения, описанной выше. Для имитационного моделирования используется программная среда Python. Для повышения качества симуляции число запросов, пропускаемых через систему, в рамках одного запуска модели для определения одного значения математического ожидания случайной величины времени отклика будет составлять 5 млн для низких значений уровня загруженности сети, т. е. меньших 0,50, и 10 млн для значений загрузки от 0,50 и выше.

Для оценки качества аппроксимации рассмотрим среднюю относительную погрешность приближения, а также ее максимальное значение:

$$\text{MAPE} = \frac{1}{L}\sum_{j=1}^{L}\left|\frac{T_j^* - T_j}{T_j}\right| \cdot 100\%;$$

$$\text{MaxAPE} = \max_{1 \le j \le L}\left|\frac{T_j^* - T_j}{T_j}\right| \cdot 100\%,$$

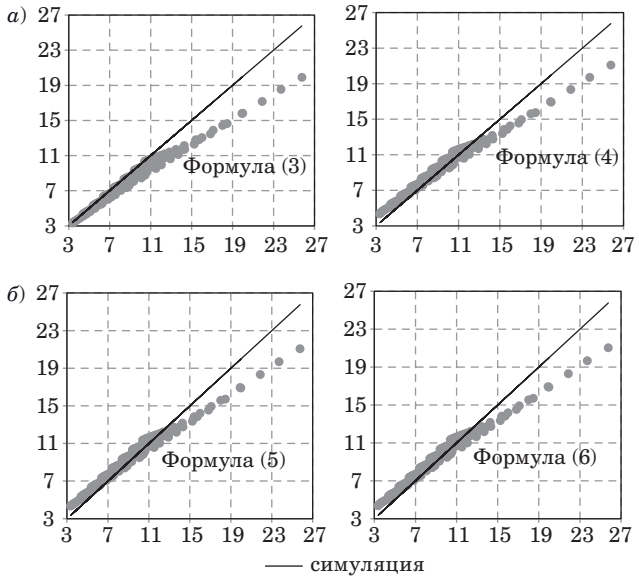
где L — общее количество наборов входных параметров, для которых делаются расчеты, в данном случае $L = 144$; T_j^* — оценка среднего значения времени отклика, рассчитанная по аналитическим формулам (2)–(9); T_j — значение среднего времени отклика модели системы, рассчитанное на статистических данных имитационного моделирования для j-го набора значений входных параметров (значения уровня загрузки системы ρ и общего числа узлов системы).

Значения заявленных типов погрешностей приближения для всех уровней загрузки системы $\rho \in [0{,}05; 0{,}90]$ представлены в табл. 1. Несмотря на то, что средняя погрешность аппроксимации остается в рамках инженерной, т. е. не превышает или совсем незначительно превышает 10 %, ее максимальное значение слишком велико. Поэтому проведем более детальный анализ использования выражений (3)–(6) для расчетов при оценке показателей узлов со второго по 10-й, так как средняя погрешность приближения для расчетов времени пребывания запроса в первом узле не превышает 2 % [23]. На рис. 2, *а* и *б* наглядно сравниваются результаты имитационно-

■ ***Таблица 1.*** Погрешности аппроксимации для среднего времени отклика системы, рассчитанные с помощью формул (3)–(6) для коэффициентов вариации и для значений загрузки системы $\rho \in [0{,}05; 0{,}90]$

■ ***Table 1.*** Approximation errors for the average system response time, calculated using formulas (3)–(6) for the variation coefficients and for the system load values $\rho \in [0.05; 0.90]$

Расчетная формула	MAPE, %	MaxAPE, %
(3)	8,068	22,674
(4)	10,094	30,173
(5)	9,859	29,341
(6)	10,047	30,165



■ ***Рис. 2.*** Сравнение результатов имитационного моделирования величины среднего времени отклика с результатами расчетов по аналитическим формулам (3), (4) (*а*) и (5), (6) (*б*)

■ ***Fig. 2.*** Comparison of the results of simulation modeling of the average response time with the results of calculations using analytical formulas using formulas (3), (4) (*а*) and (5), (6) (*б*)

го моделирования и расчетов по аналитическим формулам.

Формула (3) показывает наилучшее приближение для низких уровней загрузки системы до 0,50 включительно, что отражено в табл. 2 и на соответствующем графике. В отношении формул (4)–(6) ситуация ожидаемо противоположная и свойственная в целом для оценки характеристик подобных узлов (типа G/G/1), т. е. показывает лучшее качество приближения для более высоких уровней нагрузки на узлы, причем на более низком уровне она завышается, а на более высоком уровне занижается. Результаты вычислений схожи, хотя лучший из трех демонстрирует формула (4).

■ ***Таблица 2.*** Погрешности аппроксимации для среднего времени отклика системы, рассчитанные по формулам (3)–(6) для коэффициентов вариации и различных значений загрузки системы

■ ***Table 2.*** Approximation errors for the average system response time calculated using formulas (3)–(6) for the variation coefficients and different system load values

Расчетная формула	MAPE, %	MaxAPE, %
(3), $\rho \in [0{,}05; 0{,}50]$	3,242	7,945
(4), $\rho \in [0{,}55; 0{,}90]$	7,005	18,106
(5), $\rho \in [0{,}55; 0{,}90]$	7,025	18,263
(6), $\rho \in [0{,}55; 0{,}90]$	7,077	18,370

Тем не менее, резюмируя, можно отметить пригодность представленных в работе аналитических формул для первичной оценки среднего времени отклика модели распределенных транзакционных приложений, особенно учитывая, что они не требуют больших вычислительных затрат и дают быстрый результат.

Заключение

Рассмотрена математическая модель распределенных транзакционных приложений с микросервисной архитектурой и параллельными узлами в виде сети массового обслуживания с линейной архитектурой и параллельным узлом. Предполагается, что время обслуживания на узлах имеет распределение Парето, а входящий поток является пуассоновским.

На основе метода декомпозиции проведена оценка среднего времени отклика каждого узла в отдельности, что позволяет получить приближение для среднего времени отклика всей сети. Оценка для непараллельных узлов выполнена с помощью формулы для оценки среднего времени отклика для систем типа G/G/1, в которой фигурируют коэффициенты вариации для входящего потока и времен обслуживания. В отличие от коэффициентов вариации для времени обслуживания в узлах сети, которые известны в силу постановки задачи, сложность представляет оценка коэффициентов вариации для времен между соседними поступлениями запросов между узлами сети, поэтому для их оценки использовано несколько типов приближений. Несмотря на то, что формула для оценки среднего времени отклика внутренних узлов сети может иногда давать не вполне удовлетворительные результаты в условиях слабой загрузки системы и зависит от типа распределения, проведенные вычислительные эксперименты позволяют говорить о приемлемом качестве приближения для случая распределения Парето.

Для оценки среднего времени отклика fork-join-узла применен комплексный подход, включающий имитационное моделирование, визуальный анализ данных и оптимизацию (метод подробно описан в [23]), а средняя погрешность приближения не превышает 2 %. Так, средняя погрешность аппроксимации для среднего времени отклика модели системы транзакционных микросервисных приложений со смешанной последовательной/параллельной структурой в случае более высоких уровней загрузки сети (выше 55 %) не превышает 10 %, а в случае более низкой загрузки (ниже 50 %) — соответственно 5 %.

Таким образом, предложенный подход позволяет оценить среднее время отклика для модели системы транзакционных микросервисных при-

ложений, точность его будет выше за счет качественной оценки времени отклика fork-join-узла.

Кроме того, при таком подходе возможно провести оценку показателей системы и с другими типами распределений, по крайней мере первичную, что позволит поставщикам услуг получить необходимые прогнозы и использовать их при проектировании подобных систем.

Рассмотренная в статье модель транзакционного приложения имеет линейную топологию после узла типа fork-join, при этом реальные взаимодействия микросервисов могут представлять собой сложные графы. Соответственно, изучение сетей более сложной архитектуры, т. е. отличной от линейной, может являться одним из возможных направлений для будущих исследований, так же как и оценка моментов изучаемых случайных величин более высокого порядка, например дисперсии.

Литература

1. **Бондаренко А. С., Зайцев К. С.** Использование систем управления контейнерами для построения распределенных облачных информационных систем с микросервисной архитектурой. *Международный журнал гуманитарных и естественных наук*, 2022, № 64, с. 62–65. doi:10.24412/2500-1000-2022-1-1-62-65
2. **Miao K., Li J., Hong W., Chen M.** A microservice-based big data analysis platform for online educational applications. *Scientific Programming*, 2020, vol. 2020, pp. 1–13. doi:10.1155/2020/6929750
3. **Hao J., Zhao J., Li Y.** Research on decompostion method of relational database oriented to microservice refactoring. *2023 24st Asia-Pacific Network Operations and Management Symposium (APNOMS)*, 2023, pp. 282–285.
4. **Berardi D., Giallorenzo S., Mauro J., Melis A., Montesi F., Prandini M.** Microservice security: A systematic literature review. *PeerJ Computer Science*, 2022, vol. 8, p. :e779. doi:10.7717/peerj-cs.779
5. **Velepucha V., Flores P.** A survey on microservices architecture: Principles, patterns and migration challenges. *IEEE Access*, 2023, vol. 11, pp. 88339–88358. doi:10.1109/ACCESS.2023.3305687
6. **Harrison G., Marshall A., Custer C.** *Architecting Distributed Transactional Applications*. O'Reilly Media, Incorporated, 2023. 41 p.
7. **Гольчевский Ю. В., Ермоленко А. В.** Актуальность использования микросервисов при разработке информационных систем. *Вестник Сыктывкарского университета. Серия 1. Математика. Механика. Информатика*, 2020, № 2, с. 25–36. EDN: MYITJK
8. **Артамонов Ю. С., Востокин С. В.** Разработка распределенных приложений сбора и анализа данных на базе микросервисной архитектуры. *Известия Самарского научного центра РАН*, 2016, т. 18, № 4-4, с. 688–693. EDN: YGSQTV
9. **Фомин Д. С., Бальзамов А. В.** Проблематика обработки транзакций при использовании микросервисной архитектуры. *Известия высших учебных заведений. Поволжский регион. Технические науки*, 2021, № 2, с. 15–23. doi:10.21685/2072-3059-2021-2-2
10. **Никонов А. А., Стельмашонок Е. В.** Анализ внедрения современных цифровых технологий в финансовой сфере. *Научно-технические ведомости СПбГПУ. Экономические науки*, 2018, № 4, с. 111–119. doi:10.18721/JE.11408, EDN: UYUPJQ
11. **Chatterjee P.** Cloud-native architecture for high-performance payment system. *TIJER-International Research Journals (TIJER)*, 2023, vol. 10, no. 4, pp. 345–358. doi:10.2139/ssrn.5101076
12. **Редругина Н. М.** Метод вычисления временных характеристик обслуживания в сервисных платформах инфокоммуникационных транзакционных услуг с параллельной обработкой запросов. *Труды учебных заведений связи*, 2023, № 3, с. 82–90. doi:10.31854/1813-324X-2023-9-3-82-90
13. **Редругина Н. М., Зарубин А. А.** Модели и методы расчета временных характеристик слабосвязанных транзакционных услуг. *Наукоемкие технологии в космических исследованиях Земли*, 2024, № 2, с. 4–12. doi:10.36724/2409-5419-2024-16-2-4-12
14. **Nguyen M., Alesawi S., Li N., Che H., Jiang H.** A black-box fork-join latency prediction model for data-intensive applications. *IEEE Transactions on Parallel and Distributed Systems*, 2020, vol. 31, no. 9, pp. 1983–2000. doi:10.1109/TPDS.2020.2982137
15. **Gorbunova A. V., Vishnevsky V. M., Larionov A. A.** Evaluation of the end-to-end delay of a multiphase queuing system using artificial neural networks. *Lecture Notes in Computer Science*, 2020, vol. 12563, pp. 631–642. doi:10.1007/978-3-030-66471-8_48
16. **Кутузов О. И., Татарникова Т. М.** К оцениванию и сопоставлению очередей классических и фрактальных систем массового обслуживания. *Информационно-управляющие системы*, 2016, № 2, с. 48–55. doi:10.15217/issn1684-8853.2016.2.48
17. **Задорожный В. Н., Захаренкова Т. Р.** Методы планирования имитационных экспериментов при моделировании фрактальных очередей. *Омский научный вестник*, 2016, № 3, с. 87–92. EDN: VWXULR
18. **Рыжиков Ю. И.** Теория очередей и распределение Парето. *Труды Военно-космической академии им. А. Ф. Можайского*, 2015, № 648, с. 28–43. EDN: UZMKMF
19. **Горбунова А. В., Вишневский В. М.** Оценка времени отклика среды для вычислений с интенсивным использованием данных. *Информационно-*

управляющие системы, 2022, № 4, с. 12–19. doi:10.31799/1684-8853-2022-4-12-19

20. **Reiser M., Kobayashi H.** Accuracy of the diffusion approximation for some queuing systems. *IBM Journal of Research and Development*, 1974, vol. 18, no. 2, pp. 110–124.
21. **Gelenbe E., Pujolle G.** The behaviour of a single queue in a general queueing network. *Acta Informatica*, 1976, vol. 7, no. 2, pp. 123–136.
22. **Kuhn P.** Analysis of complex queuing networks by decomposition. *Proceedings of the 8th International Teletraffic Congress*, 1976, pp. 1–8.
23. **Kraemer W., Langenbach-Belz M.** Approximate formulae for the delay in the queueing system GI|G|1. *Proceedings of the 8th International Teletraffic Congress*, 1976, vol. 235, pp. 1–8.
24. **Gorbunova A. V., Lebedev A. V.** Nonlinear approximation of characteristics of a fork–join queueing system with Pareto service as a model of parallel structure of data processing. *Mathematics and Computers in Simulation*, 2023, vol. 214, pp. 409–428. doi:10.1016/j.matcom.2023.07.029
25. **Gorbunova A. V., Lebedev A. V.** On the features of service rate control in fork-join queueing system. *Automation and Remote Control*, 2024, vol. 85, no. 12, pp. 1184–1198. doi:10.31857/S0005117924120043
26. **Nelson R., Tantawi A. N.** Approximate analysis of fork/join synchronization in parallel queues. *IEEE Transactions on Computers*, 1988, vol. 37, pp. 739–743. doi:10.1109/12.2213
27. **Gorbunova A. V., Lebedev A. V.** On estimating the characteristics of a fork-join queueing system with Poisson input and exponential service times. *Advances in Systems Science and Applications*, 2023, vol. 23, no. 2, pp. 99–114. doi:10.25728/assa.2023.23.2.1351
28. **Thomasian A.** Analysis of fork/join and related queueing systems. *ACM Computing Surveys (CSUR)*, 2014, vol. 47, pp. 17:1–17:71. doi:10.1145/2628913
29. **Varki E., Merchant A., Chen H.** *The M/M/1 fork-join queue with variable subtasks*. https://www.cs.unh.edu/~varki/publication/2002-nov-open.pdf (дата обращения: 05.05.2024).
30. **Qiu Z., Perez J. F., Harrison P. G.** Beyond the mean in fork-join queues: Efficient approximation for response-time tails. *Performance Evaluation*, 2015, vol. 91, pp. 99–116. doi:10.1016/j.peva.2015.06.007
31. **Wang W., Harchol-Balter M., Jiang H., Scheller-Wolf A., Srikant R.** Delay asymptotics and bounds for multitask parallel jobs. *Queueing Systems*, 2019, vol. 91, pp. 207–239. doi:10.1007/s11134-018-09597-5

UDC 004.032
doi:10.31799/1684-8853-2025-6-42-50
EDN: EGLAUQ

Evaluation of the characteristics of a distributed transactional application model with microservice architecture and fork-join structures

A. V. Gorbunova[a], PhD, Phys.-Math., Senior Researcher, orcid.org/0000-0002-9183-0426, avgorbunova@list.ru
[a]V. A. Trapeznikov Institute of Control Sciences of RAS, 65, Profsoyuznaya St., 117997, Moscow, Russian Federation

Introduction: A microservice architecture, which allows applications to be built as a set of independent microservices that work together to fulfill a common client request, has recently become the basis or even the standard for deploying complex systems that affect multiple physical structures and devices. In addition, the introduction of parallel service scenarios into such systems, especially in the case of high-load systems, makes it possible to increase their efficiency and performance. **Purpose:** To develop a mathematical model of distributed transactional applications with a microservice architecture and parallel nodes and to evaluate such a performance indicator as the average response time. **Results:** We present a mathematical model of distributed transactional applications with a microservice architecture in the form of a queueing network with sequential nodes, one of which has a parallel structure with several subnodes, the number of those is more than two. Based on the decomposition method for analyzing queueing networks, an approach to estimating the average response time of the system under consideration is proposed. We use known results for assessing individual nodes of a G/G/1 type network, as well as nodes with division and with parallel service. The results of the computational experiments allow drawing conclusions about the admissibility of the proposed approach, and obtaining recommendations regarding the applicability of various formulas for different load levels, in particular, for those whose average approximation error does not exceed 10%. **Practical relevance:** The proposed model and the method of its research can be used for the initial assessment and prediction of the average response time of transactional applications with parallel nodes at different load levels and, as a result, can contribute to maintaining the required quality of service for users of transactional applications.

Keywords — transactional applications, microservice architecture, distributed systems, parallel operations, queueing network, average response time, fork-join, G/G/1.

For citation: Gorbunova A. V. Evaluation of the characteristics of a distributed transactional application model with microservice architecture and fork-join structures. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2025, no. 6, pp. 42–50 (In Russian). doi:10.31799/1684-8853-2025-6-42-50, EDN: EGLAUQ

References

1. Bondarenko A. S., Zaytsev K. S. Using container management systems to build distributed cloud information systems with microservice architecture. *International Journal of Humanities and Natural Sciences,* 2022, vol. 1-1(64). pp. 62–65. (In Russian). doi:10.24412/2500-1000-2022-1-1-62-65
2. Miao K., Li J., Hong W., Chen M. A microservice-based big data analysis platform for online educational applications. *Scientific Programming*, 2020, vol. 2020, pp. 1–13. doi:10.1155/2020/6929750
3. Hao J., Zhao J., Li Y. Research on decompostion method of relational database oriented to microservice refactoring. *2023 24st Asia-Pacific Network Operations and Management Symposium (APNOMS)*, 2023, pp. 282–285.
4. Berardi D., Giallorenzo S., Mauro J., Melis A., Montesi F., Prandini M. Microservice security: A systematic literature review. *PeerJ Computer Science*, 2022, vol. 8, p. :e779. doi:10.7717/peerj-cs.779
5. Velepucha V., Flores P. A survey on microservices architecture: Principles, patterns and migration challenges. *IEEE Access*, 2023, vol. 11, pp. 88339–88358. doi:10.1109/ACCESS.2023.3305687
6. Harrison G., Marshall A., Custer C. *Architecting Distributed Transactional Applications*. O'Reilly Media, Incorporated, 2023. 41 p.
7. Golchevskiy Yu. V., Yermolenko A. V. The relevance of using microservices in the development of information systems. *Vestnik Syktyvkarskogo Universiteta. Seriya 1: Matematika. Mekhanika. Informatika*, 2020, vol. 2(35), pp. 25–36 (In Russian). EDN: MYITJK
8. Artamonov Yu. S., Vostokin S. V. Development of distributed applications for data collection and analysis on the basis of a microservice architecture. *Izvestiya Samarskogo Nauchnogo Centra RAN*, 2016, vol. 18, no. 4-4, pp. 688–693 (In Russian). EDN: YGSQTV
9. Fomin D. S., Bal'zamov A. V. The problem of transaction processing using microservice architecture. *University Proceedings. Volga Region. Engineering Sciences*, 2021, no. 2, pp. 15–23 (In Russian). doi:10.21685/2072-3059-2021-2-2
10. Nikonov A. A., Stelmashonok E. V. Analysis of modern digital technologies' implementation in the financial sphere. *St. Petersburg State Polytechnical University Journal. Economics*, 2018, vol. 11, no. 4, pp. 111–119 (In Russian). doi:10.18721/JE.11408, EDN: UYUPJQ
11. Chatterjee P. Cloud-native architecture for high-performance payment system. *TIJER-International Research Journals (TIJER)*, 2023, vol. 10, no. 4, pp. 345–358. doi:10.2139/ssrn.5101076
12. Redrugina N. M. Method for time characteristics calculating in the service platforms of info-communication transactional services with parallel requests processing. *Proceedings of Telecommunication Universities*, 2023, vol. 9, no. 3, pp. 82–90 (In Russian). doi:10.31854/1813-324X-2023-9-3-82-90
13. Redrugina N. M., Zarubin A. A. Models and methods for calculating the temporal characteristics of loosely coupled transactional services. *High Technologies in Earth Space Research*, 2024, vol. 16, no. 2, pp. 4–12 (In Russian). doi:10.36724/2409-5419-2024-16-2-4-12
14. Nguyen M., Alesawi S., Li N., Che H., Jiang H. A black-box fork-join latency prediction model for data-intensive applications. *IEEE Transactions on Parallel and Distributed Systems*, 2020, vol. 31, no. 9, pp. 1983–2000. doi:10.1109/TPDS.2020.2982137
15. Gorbunova A. V., Vishnevsky V. M., Larionov A. A. Evaluation of the end-to-end delay of a multiphase queuing system using artificial neural networks. *Lecture Notes in Computer Science*, 2020, vol. 12563, pp. 631–642. doi:10.1007/978-3-030-66471-8_48
16. Kutuzov O. I., Tatarnikova T. M. Evaluation and comparison of queues in classical and fractal queuing systems. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2016, no. 2, pp. 48–55 (In Russian). doi:10.15217/issn1684-8853.2016.2.48
17. Zadorozhnyi V. N., Zakharenkova T. R. Methods for planning simulation experiments in modeling fractal queues. *Omsk Scientific Bulletin*, 2016, no. 3(147), pp. 87–92 (In Russian). EDN: VWXULR
18. Ryzhikov Yu. I. Queuing theory and Pareto distribution. *Trudy Voenno-kosmicheskoj akademii im. A. F. Mozhajskogo*, 2015, no. 648, pp. 28–43 (In Russian). EDN: UZMKMF
19. Gorbunova A. V., Vishnevsky V. M. Estimating the response time of a data-intensive computing environment. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2022, no. 4, pp. 12–19 (In Russian). doi:10.31799/1684-8853-2022-4-12-19
20. Reiser M., Kobayashi H. Accuracy of the diffusion approximation for some queuing systems. *IBM Journal of Research and Development*, 1974, vol. 18, no. 2, pp. 110–124.
21. Gelenbe E., Pujolle G. The behaviour of a single queue in a general queueing network. *Acta Informatica*, 1976, vol. 7, no. 2, pp. 123–136.
22. Kuhn P. Analysis of complex queuing networks by decomposition. *Proceedings of the 8th International Teletraffic Congress*, 1976, pp. 1–8.
23. Kraemer W., Langenbach-Belz M. Approximate formulae for the delay in the queueing system GI|G|1. *Proceedings of the 8th International Teletraffic Congress*, 1976, vol. 235, pp. 1–8.
24. Gorbunova A. V., Lebedev A. V. Nonlinear approximation of characteristics of a fork-join queueing system with Pareto service as a model of parallel structure of data processing. *Mathematics and Computers in Simulation*, 2023, vol. 214, pp. 409–428. doi:10.1016/j.matcom.2023.07.029
25. Gorbunova A. V., Lebedev A. V. On the features of service rate control in fork-join queueing system. *Automation and Remote Control*, 2024, vol. 85, no. 12, pp. 1184–1198. doi:10.31857/S0005117924120043
26. Nelson R., Tantawi A. N. Approximate analysis of fork/join synchronization in parallel queues. *IEEE Transactions on Computers*, 1988, vol. 37, pp. 739–743. doi:10.1109/12.2213
27. Gorbunova A. V., Lebedev A. V. On estimating the characteristics of a fork-join queueing system with Poisson input and exponential service times. *Advances in Systems Science and Applications*, 2023, vol. 23, no. 2, pp. 99–114. doi:10.25728/assa.2023.23.2.1351
28. Thomasian A. Analysis of fork/join and related queueing systems. *ACM Computing Surveys (CSUR)*, 2014, vol. 47, pp. 17:1–17:71. doi:10.1145/2628913
29. Varki E., Merchant A., Chen H. *The M/M/1 fork-join queue with variable subtasks*. Available at: https://www.cs.unh.edu/~varki/publication/2002-nov-open.pdf (accessed 5 May 2024).
30. Qiu Z., Perez J. F., Harrison P. G. Beyond the mean in fork-join queues: Efficient approximation for response-time tails. *Performance Evaluation*, 2015, vol. 91, pp. 99–116. doi:10.1016/j.peva.2015.06.007
31. Wang W., Harchol-Balter M., Jiang H., Scheller-Wolf A., Srikant R. Delay asymptotics and bounds for multitask parallel jobs. *Queueing Systems*, 2019, vol. 91, pp. 207–239. doi:10.1007/s11134-018-09597-5