

# ОБ ОЦЕНКЕ ХАРАКТЕРИСТИК МОДЕЛИ ТРАНЗАКЦИОННЫХ ПРИЛОЖЕНИЙ

Горбунова А.В.

Институт проблем управления им. В.А. Трапезникова РАН, Москва, Россия  
avgorbunova@list.ru

*Аннотация.* В статье рассматривается математическая модель распределенных транзакционных приложений с микросервисной архитектурой в виде сети массового обслуживания с последовательными узлами, один из которых имеет параллельную структуру типа fork-join. На основе метода декомпозиции для анализа сетей массового обслуживания предлагается подход к оценке среднего времени отклика рассматриваемой системы.

*Ключевые слова:* транзакционные приложения, микросервисная архитектура, распределенные системы, параллельные операции, сеть массового обслуживания, fork-join структура.

## Введение

Транзакционные приложения представляют собой системы, обслуживающие или оказывающие транзакционные услуги, под которыми в большинстве случаев подразумевается обработка финансовых или коммерческих операций. Примерами транзакционных приложений могут являться системы онлайн-банкинга, системы электронной коммерции, системы 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. 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. Т. е. 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]. В отличие от монолитной архитектуры, которая представляет собой единую структуру, компоненты которой связаны в единое целое, 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 [2]. 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, 3, 4].

Использование микросервисной архитектуры имеет свои преимущества и недостатки. 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, однако при этом порождает сложности с синхронизацией и согласованностью данных [5, 6].

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 [7, 8].

Таким образом, на первый план выходит необходимость адекватного прогнозирования показателей 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 [9] 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 [10] 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 оценки характеристик узлов непараллельной структуры сетей линейной топологии используются несколько вариантов аппроксимаций, представляющих собой известные классические результаты. В целом же предполагается использование метода декомпозиции – оценка 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 с использованием методов 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 технической реализацией корректного процесса разбиения запроса на более мелкие задачи.

## 1. 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

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

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

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

Схема описанной модели функционирования приложения имеет вид, показанный на рисунке 1, т. е. представляет собой сеть массового обслуживания линейной архитектуры, но с параллельным узлом.

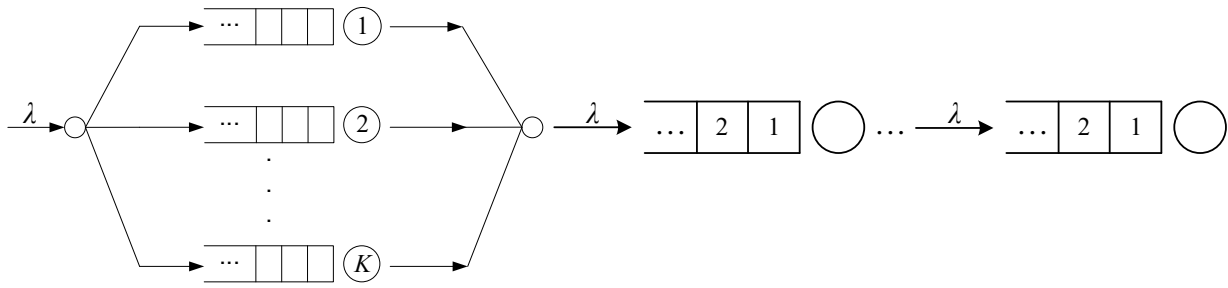


Рис. 1. Схема математической модели транзакционного приложения с микросервисной архитектурой и параллельным узлом

Одним из основных методов исследования сетей массового обслуживания является метод декомпозиции, который предполагает анализ отдельных фрагментов сети изолированно. 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 так называемые открытые экспоненциальные сети Джексона и некоторые их расширения. В остальных же случаях, которых существенно 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.

## 2. 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 (англ. 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 статьёй, то сказывается на общей стоимости предоставляемых провайдером услуг и его конкурентоспособности [11].

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

$$T = t_1 + t_2 + \dots + t_N. \quad (2)$$

Согласно методу декомпозиции остается определить величину  $t_i$  для каждого имеющегося в системе узла,  $i = 1, \dots, N$ . 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  $1/\lambda$ .

Что касается вторых моментов, а точнее коэффициентов вариации  $CV_i$  для входящих в  $i$ -й узел потоков ( $i = 2, \dots, N$ ), необходимых для оценки среднего времени пребывания в каждом из узлов, то здесь можно воспользоваться некоторыми известными приближениями. В частности [12-14],

$$CV_i = CV_{Pa_{i-1}}, \quad (3)$$

$$CV_i = \rho_{i-1}(1 - \rho_{i-1}) + \rho_{i-1}^2 CV_{Pa_{i-1}}^2 + (1 - \rho_{i-1}) CV_{i-1}^2, \quad (4)$$

$$CV_i = CV_{i-1}^2 + 2\rho_{i-1} CV_{Pa_{i-1}}^2 - \rho_{i-1} (CV_{i-1}^2 + CV_{Pa_{i-1}}^2) f(\rho_{i-1}, CV_{i-1}, CV_{Pa_{i-1}}), \quad (5)$$

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

где  $\rho_{i-1}$  – это нагрузка ( $i-1$ )-го узла, которая в рамках рассматриваемой модели идентична для всех узлов и равна  $\rho_i = \rho = \lambda$ ,  $i = 1, \dots, N$ , величина  $CV_{Pa_{i-1}}$  – коэффициент вариации времени обслуживания

и определяется выражением  $CV_{Pa_i} = CV_{Pa} = 1/\sqrt{\alpha(\alpha - 2)}$ , причем допустим, что это справедливо и для первого узла. Что касается функции  $f(\rho, CV_i, CV_{Pa})$ , то ее определим далее по тексту.

Выражения (3) – (6) используются для оценки среднего времени пребывания в  $i$ -м узле,  $i = 2, \dots, N$ , т. е. [15]

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

где

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

Что касается времени пребывания в самом первом узле, который представляет собой параллельную структуру, состоящую из  $K$  узлов, то здесь воспользуемся приближением, представленным в работах [16, 17], а именно

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

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

### 3. Заключение

В статье рассматривается 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 проводится оценка среднего времени отклика каждого узла в отдельности, что позволяет получить приближение для среднего времени отклика всей сети. Оценка для непараллельных узлов проводится с помощью формулы для оценки среднего времени отклика для систем типа  $G|G|1$ , в которой фигурируют коэффициенты вариации для входящего и обслуживающего потоков. 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 используется несколько типов приближений. Несмотря на то, что формула для оценки среднего времени внутренних узлов сети может иногда давать не вполне удовлетворительные результаты в условиях слабой загрузки системы и зависит от типа распределения, исследования и численные эксперименты, проведенные в работе [18], позволяют говорить о приемлемом качестве приближения для случая распределения Парето.

Что касается среднего времени отклика fork-join узла, то для его оценки используется комплексный подход, включающий имитационное моделирование, визуальный анализ данных и оптимизацию (подход подробно описан в [16]), а средняя погрешность приближения не превышает 2%.

Таким образом, предложенный подход позволяет оценить среднее время отклика для модели системы транзакционных вычислений, точность его будет выше за счет качественной оценки времени отклика в 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.

## 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.
2. *Harrison G., Marshall A., Custer C.* Architecting Distributed Transactional Applications. – O'Reilly Media, Incorporated, 2023. – 41 p.
3. *Гольчевский Ю.В., Ермоленко А.В.* Актуальность использования микросервисов при разработке информационных систем // Вестник Сыктывкарского университета. Серия 1. Математика. Механика. Информатика. – 2020. – № 2(35). – С. 25–36.
4. *Артамонов Ю.С., Востокин С.В.* Разработка распределенных приложений сбора и анализа данных на базе микросервисной архитектуры // Известия Самарского научного центра РАН. – 2016. – № 4. – С. 688–693.
5. *Фомин Д.С., Бальзамов А.В.* Проблематика обработки транзакций при использовании микросервисной архитектуры // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2021. – № 2(58). – С. 15–23.
6. *Шкрябин Г.Д.* Проблемы обеспечения целостности данных в микросервисной архитектуре на примере распределенных систем. <https://www.it-world.ru/cionews/sd5qmwelj5wgggg8c80kkgk44gw4w8w.html> (дата обращения 25.05.2025).
7. *Никонов А.А., Стельмашонок Е.В.* Анализ внедрения современных цифровых технологий в финансовой сфере // π-Economy. – 2018. – № 4. – С. 688–693.
8. *Pushpalika Chatterjee.* Cloud-Native Architecture for High-Performance Payment System // TIJER-International Research Journals (TIJER). – 2023. – Vol. 10, № 4. – P. 345–358.
9. *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.
10. *Редругина Н.М., Зарубин А.А.* Модели и методы расчета временных характеристик слабосвязанных транзакционных услуг // Научные технологии в космических исследованиях Земли. – 2024. – № 2. – С. 4–12.
11. *Горбунова А.В., Вишневский В.М.* Оценка времени отклика среды для вычислений с интенсивным использованием данных // Информационно-управляющие системы. – 2022. – № 4(119). – С. 12–19.
12. *Reiser M., Kobayashi H.* Accuracy of the diffusion approximation for some queuing systems // IBM Journal of Research and Development. – 1974. – Vol. 18, № 2. – P. 110–124.
13. *Gelenbe E., Pujolle G.* The behaviour of a single queue in a general queueing network // Acta Informatica. – 1976. – Vol. 7, № 2. – P. 123–136.
14. *Kuhn P.* Analysis of complex queuing networks by decomposition // Proceedings of the 8th International Teletraffic Congress. – 1976. – P. 1–8.
15. *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. – P. 1–8.
16. *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. – P. 409–428.
17. *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, № 12. – P. 1184–1198.
18. *Gorbunova A.V., Vishnevsky V.M., Larionov A.A.* Evaluation of the End-to-End Delay of a Multiphase Queueing System Using Artificial Neural Networks // Lecture Notes in Computer Science. – 2020. – Vol. 12563. – P. 631–642.