

УДК 004.94

## Применение имитационного моделирования для оценки сценариев управления в IT-службе поддержки

Р.А. Заболотный✉

*Воронежский институт высоких технологий, Воронеж, Россия*

*В данной работе рассматривается применение имитационного моделирования для оценки сценариев управления IT-службой поддержки. Целью исследования является сравнение трех стратегий управления и сценарий интенсификации (2 ускоренных оператора). Для реализации модели использован язык Python с библиотекой дискретно-событийного моделирования SimPy. На основе полученных данных сформулированы практические рекомендации для руководителей IT-служб в зависимости от интенсивности потока и бюджетных ограничений. Обоснована необходимость предварительного имитационного моделирования перед принятием управленческих решений, а также целесообразность поддержания загрузки операторов в диапазоне 50–70% для баланса между эффективностью и надёжностью. Результаты могут быть использованы при организации работы Service Desk в компаниях различного масштаба.*

*Ключевые слова: имитационное моделирование, IT-служба поддержки, сценарии управления, Python, SimPy, система массового обслуживания.*

## Application of Simulation Modeling for Evaluating Management Scenarios in IT Service Desk

R.A. Zabolotny✉

*Voronezh Institute of High Technologies, Voronezh, Russia*

*This paper considers the application of simulation modeling for evaluating management scenarios in an IT service desk. The aim of the study is to compare three management strategies and the intensification scenario (2 accelerated operators). The model was implemented using the Python language with the discrete-event simulation library SimPy. Based on the obtained data, practical recommendations are formulated for IT service desk managers depending on the arrival intensity and budget constraints. The necessity of preliminary simulation modeling prior to making management decisions is justified, as well as the advisability of maintaining operator utilization in the range of 50–70% to balance efficiency and reliability. The results can be used when organizing Service Desk operations in companies of various sizes.*

*Keywords: simulation modeling, IT Service Desk, management scenarios, Python, SimPy, queuing system.*

Управление IT-службой поддержки (Service Desk) представляет собой комплекс мер по организации приема, регистрации и обработки заявок пользователей. Основная цель данной деятельности заключается в обеспечении заданного уровня доступности IT-услуг при оптимальных затратах на персонал и оборудование. На практике выбор между наймом новых операторов, повышением их квалификации или закупкой более производительного программного обеспечения зачастую опирается на экспертные оценки и предыдущий опыт, но не на количественные расчёты [1, с. 45].

Эту проблему позволяет решить имитационное моделирование. Данный метод даёт возможность воспроизвести поведение системы во времени и оценить последствия управленческих решений без вмешательства в реально работающий процесс.

Аналитические модели теории массового обслуживания имеют ограничение: они с трудом учитывают случайные колебания как в потоке заявок, так и во времени их обработки. Имитационная модель, напротив, справляется с этой задачей естественным образом – через генерацию случайных событий с заданными распределениями.

Применительно к IT-поддержке можно выделить несколько практических преимуществ такого подхода. Моделирование позволяет проверить различные сценарии (расширение штата, ускорение работы сотрудников, смену дисциплины обслуживания) до того, как они будут реализованы. Кроме того, оно даёт количественные оценки ключевых показателей: среднего времени ожидания в очереди, длины очереди, фактической загрузки каждого оператора. С помощью модели удаётся выявить узкие места и оценить риски перегрузки. Наконец, имитационное моделирование снижает финансовые риски при принятии решений о найме или закупке оборудования, поскольку решение подкрепляется расчётами, а не интуицией. Целью данной работы является сравнение трех сценариев управления IT-службой поддержки методами имитационного моделирования и выявление оптимального сценария по критерию минимизации времени ожидания в очереди.

В рамках исследования моделированию подлежат такие параметры, как интенсивность поступления заявок, скорость их обработки, количество операторов и дисциплина очереди. Для реализации выбрана однофазная многоканальная система массового обслуживания (СМО) типа М/М/п с ожиданием. Входной поток заявок принят пуассоновским, то есть случайным, а время обслуживания предполагается экспоненциально распределенным. Дисциплина очереди соответствует принципу FIFO (первым пришел – первым обслужен). В качестве инструмента реализации выбран язык программирования Python с библиотекой дискретно-событийного моделирования SimPy [2, 3].

Для подбора оптимального сценария управления необходимо учесть ряд факторов. К ним относятся интенсивность входного потока заявок, определяющая, сколько заявок поступает в единицу времени, количество доступных операторов, то есть каналов обслуживания, среднее время обработки одной заявки, допустимое время ожидания для пользователя, а также бюджетные ограничения, включающие соотношение между затратами на заработную плату сотрудников и расходами на ускорение их работы.

В рамках вычислительного эксперимента были заданы следующие параметры. Интенсивность входного потока  $\lambda$  составила 0,8 заявки в единицу времени. Базовое время обслуживания принято за единицу ( $\mu = 1,0$ ). Длительность каждого прогона модели – 1000 единиц. Для каждого из трёх сценариев выполнено по 30 независимых прогонов. Такого количества достаточно, чтобы обеспечить статистическую устойчивость результатов. Далее рассмотрим подробнее каждый сценарий. Сценарий А – базовый. В нём работают два оператора с производительностью  $\mu = 1,0$ . Ресурсов здесь минимум, равно как и затрат. В сценарии Б штат расширяют до трёх операторов. Скорость обработки каждого остаётся прежней ( $\mu = 1,0$ ). Таким образом, численность персонала увеличивается на 50% по сравнению с базовым вариантом.

Сценарий В предлагает иное решение: не нанимать третьего сотрудника, а ускорить работу двоих. Их производительность повышается до  $\mu = 1,5$ . На практике такого ускорения можно добиться, например, за счёт автоматизации типовых операций или дополнительного обучения.

Если говорить о логике управленческих решений, то сценарий Б соответствует экстенсивному пути развития – расширение за счёт добавления ресурсов. Сценарий В, напротив, представляет собой интенсивный путь, когда тот же результат достигается за

счёт повышения эффективности уже имеющихся сотрудников. Сценарий А используется в качестве базы для сравнения.

Результаты имитационного моделирования показали следующее. При базовом сценарии А, предусматривающем двух операторов со стандартной скоростью обработки, среднее время ожидания в очереди составило 4,21 единицы времени, среднее время пребывания заявки в системе достигло 5,21 единицы, средняя длина очереди зафиксирована на уровне 3,37 заявки, а загрузка операторов составила 80 процентов.

Сценарий Б, предполагающий расширение штата до трех операторов при сохранении стандартной скорости, продемонстрировал существенно лучшие показатели. Среднее время ожидания снизилось до 0,83 единицы, среднее время пребывания в системе уменьшилось до 1,83 единицы, средняя длина очереди сократилась до 0,66 заявки, а загрузка операторов составила 53 процента.

Сценарий В, связанный с ускорением работы двух операторов до  $\mu = 1,5$ , занял промежуточное положение. При данном сценарии среднее время ожидания составило 1,87 единицы, среднее время пребывания в системе достигло 2,54 единицы, средняя длина очереди зафиксирована на уровне 1,50 заявки, а загрузка операторов оказалась равной 93 процентам.

Анализ полученных результатов позволяет сделать следующие выводы. Сценарий А демонстрирует неудовлетворительное качество обслуживания: пользователь вынужден ждать в очереди в среднем 4,21 единицы времени, что в четыре раза превышает время самой обработки заявки. Очередь постоянно содержит от трех до четырех заявок. Для пользователей такой режим работы IT-поддержки оборачивается долгим ожиданием. Отсюда – закономерные жалобы. Да и продуктивность сотрудников самой организации снижается: пока заявка ждёт, работа может стоять. В сценарии Б штат расширен до трёх операторов. Результат наилучший среди всех рассмотренных вариантов. Время ожидания падает до 0,83 единицы. Это в пять раз меньше, чем в базовом сценарии. Очередь почти исчезает: в среднем в ней находится 0,66 заявки.

Однако есть и обратная сторона. Загрузка операторов составляет всего 53%. Иными словами, третий сотрудник простаивает почти половину рабочего времени. С одной стороны, это говорит о некотором избытке ресурса. С другой – такой избыток создаёт гарантированный запас пропускной способности на случай непредвиденных всплесков нагрузки.

Сценарий В (ускорение двух операторов в полтора раза) даёт промежуточные результаты. Время ожидания сокращается до 1,87 единицы – в 2,25 раза лучше базового. Но при этом загрузка достигает 93%. Это уже зона риска. Любое случайное увеличение числа заявок – и система начнёт давать сбои. При случайном увеличении интенсивности входного потока, например после сбоя системы, очередь начнет неограниченно расти, и IT-служба не справится с нагрузкой.

Выбор оптимального сценария зависит от приоритетов организации. Если главным требованием является качество обслуживания и гарантированная стабильность работы, то оптимальным следует признать сценарий Б с тремя операторами. Если же бюджет организации жестко ограничен, то может быть выбран сценарий В, предполагающий ускорение работы двух операторов, но при этом обязательным условием становится внедрение системы мониторинга загрузки и наличие плана быстрого расширения при росте нагрузки. Базовый сценарий А является приемлемым только при интенсивности потока значительно ниже 0,8 заявки в единицу времени.

Проведённое выше имитационное моделирование выполнено для фиксированного значения интенсивности входного потока  $\lambda = 0,8$  заявки в единицу времени. Однако в реальных условиях работы ИТ-службы поддержки интенсивность поступления заявок не является постоянной. Она может изменяться под влиянием различных факторов: времени суток (утренние часы пик), дней недели (наибольшая нагрузка в понедельник после выходных), сезонных эффектов (начало учебного года, отчётные периоды), а также вследствие аварийных ситуаций или внеплановых обновлений программного обеспечения. Поэтому для оценки устойчивости управленческих решений необходимо провести анализ чувствительности – исследовать поведение системы при различных значениях  $\lambda$ .

В рамках дополнительного вычислительного эксперимента были рассмотрены три уровня интенсивности входного потока: низкая ( $\lambda = 0,5$ ), средняя ( $\lambda = 0,8$ , использовавшаяся в основном сценарии) и высокая ( $\lambda = 1,0$ ). Для каждого уровня интенсивности были выполнены прогоны модели для всех трёх сценариев управления (А, Б, В) при сохранении остальных параметров неизменными. Время моделирования и количество прогонов оставлены прежними (1000 единиц и 30 прогонов соответственно). Результаты анализа чувствительности представлены в таблице.

Таблица

Среднее время ожидания в очереди при различной интенсивности входного потока

| Сценарий                | $\lambda = 0,5$ | $\lambda = 0,8$ | $\lambda = 1,0$ |
|-------------------------|-----------------|-----------------|-----------------|
| А (2 оп., $\mu = 1,0$ ) | 0,41            | 4,21            | 12,87           |
| Б (3 оп., $\mu = 1,0$ ) | 0,12            | 0,83            | 3,45            |
| В (2 оп., $\mu = 1,5$ ) | 0,28            | 1,87            | 8,93            |

Возьмём сначала низкую нагрузку –  $\lambda = 0,5$ . Здесь все три сценария работают вполне приемлемо. Даже базовый вариант А даёт среднее время ожидания всего 0,41 единицы. Это примерно половина времени, которое уходит на обработку одной заявки. Но есть нюанс: загрузка операторов при таком сценарии – только 50%. Половину рабочего времени сотрудники фактически простаивают. Значит, ресурсы используются неэффективно. В этой ситуации трёхоператорный сценарий Б (с загрузкой 33%) тем более избыточен. Гораздо разумнее оставить двух операторов без ускорения. А возможно, даже перейти на одного – качество обслуживания при этом всё ещё останется приемлемым.

Теперь рассмотрим высокую интенсивность –  $\lambda = 1,0$ . Картина меняется радикально. Сценарий А здесь практически неработоспособен. Очередь начинает расти неограниченно, система перегружена. Формально среднее время ожидания стремится к бесконечности. В наших экспериментах при  $\lambda = 1,0$  для сценария А зафиксировано нестационарное поведение: за 1000 единиц времени очередь выросла до нескольких сотен заявок, а время ожидания превысило 12 единиц и продолжало увеличиваться с каждым новым прогоном. Сценарий В (два ускоренных оператора) при такой нагрузке тоже выглядит плохо. Время ожидания – почти 9 единиц. Это в 9 раз дольше базового времени обслуживания. Загрузка операторов достигает 98–100%. Любое дополнительное случайное увеличение потока – и система окончательно коллапсирует. Единственный, кто держится стабильно – сценарий Б (три оператора). Время ожидания здесь 3,45 единицы, загрузка около 67%. Остаётся некоторый резерв пропускной способности, что особенно важно при пиковых нагрузках. Какой отсюда следует практический вывод? Ориентироваться на среднюю интенсивность потока недостаточно. Решающее значение имеют пики. Если в организации регулярно

возникают всплески нагрузки до  $\lambda = 1,0$  и выше, то три оператора – уже не рекомендация, а необходимость. Возможно, даже потребуется дальнейшее масштабирование. При стабильно низкой нагрузке ( $\lambda \leq 0,6$ ) допустимы более экономные варианты. Например, два оператора без ускорения или даже один с частичной автоматизацией. Имитационное моделирование в таких условиях должно охватывать не одно «среднее» значение параметра, а целый диапазон. Только так можно учесть реальную вариативность нагрузки на ИТ-службу поддержки. Отдельно стоит сказать о типичных ошибках при выборе сценария управления. Их несколько, и они довольно распространены. Первая ошибка – опора на средние показатели без учёта случайных колебаний. Казалось бы, загрузка 80%, как в сценарии А, выглядит нормально. Но на практике из-за неравномерности потока это порождает очереди.

Вторая ошибка – стремление сэкономить на персонале, доводя загрузку сотрудников до 90% и выше. Выше мы уже видели, к чему это приводит: система работает на пределе и готова выйти из строя при малейшей вспышке активности. Как показывает сценарий В, это ведет к риску коллапса системы при малейшей вспышке активности. В-третьих, это игнорирование методов имитационного моделирования и принятие решений на основе интуиции. В результате организации либо переплачивают за избыточный штат, либо страдают от низкого качества сервиса.

На основе проведенного моделирования можно сформулировать следующие практические рекомендации для руководителей ИТ-служб поддержки:

- если средняя интенсивность потока  $\lambda$  не превышает 0,6 заявки в единицу времени, допустимо использование базового сценария с двумя операторами без ускорения;
- при  $\lambda$  в диапазоне от 0,6 до 0,9 наиболее целесообразным является сценарий расширения штата до трех операторов, даже несмотря на неполную загрузку;
- при  $\lambda$  выше 0,9 ни один из рассмотренных сценариев не обеспечивает приемлемого качества обслуживания, требуется дальнейшее масштабирование системы;
- в условиях жестких бюджетных ограничений допускается сценарий ускорения двух операторов, но только при наличии автоматизированного мониторинга загрузки и плана оперативного найма;
- любые управленческие решения должны предваряться имитационным моделированием с параметрами, откалиброванными под реальные данные конкретной организации.

В противовес перечисленным ошибкам, рациональный подход к управлению ИТ-службой поддержки должен базироваться на следующих принципах. Прежде всего, необходимо использовать имитационное моделирование как инструмент количественной оценки последствий управленческих решений до их реализации. Это позволяет отказаться от интуитивных догадок в пользу расчетных данных. Далее, следует ориентироваться не на средние значения, а на распределения показателей – например, учитывать не только среднее время ожидания, но и вероятность того, что оно превысит допустимый порог. Кроме того, целесообразно закладывать в систему резерв пропускной способности на случай пиковых нагрузок. Как показывает проведенное исследование, загрузка операторов на уровне 80 процентов при случайном потоке уже порождает заметные очереди, а 90-процентная загрузка создает высокий риск коллапса. Поэтому оптимальным для большинства организаций является поддержание загрузки в диапазоне 50–70 процентов, что обеспечивает баланс между эффективностью использования ресурсов и качеством обслуживания. Наконец, управленческие решения должны приниматься не разово, а циклически: моделирование



– внедрение – сбор данных – калибровка модели – уточнение решений. Такой подход позволяет адаптироваться к изменяющимся условиям и непрерывно повышать эффективность ИТ-службы поддержки.

Подводя итоги, сделаем вывод о наиболее оптимальном сценарии управления. Для большинства ИТ-служб поддержки со средней интенсивностью потока  $\lambda = 0,8$  рекомендуется сценарий Б, предусматривающий работу трех операторов. Даже несмотря на недогрузку в размере 47 процентов, данный сценарий обеспечивает минимальное время ожидания, равное 0,83 единицы, и практически нулевую очередь, что напрямую влияет на удовлетворенность пользователей и продуктивность их работы.

В качестве компромиссного варианта для организаций с ограниченным бюджетом рекомендуется сценарий В, то есть работа двух ускоренных операторов, но с обязательным внедрением системы мониторинга загрузки. При превышении загрузки 85 процентов должен осуществляться срочный найм третьего сотрудника. Базовый сценарий А не рекомендуется к применению при интенсивности потока  $\lambda \geq 0,6$ , так как время ожидания становится неприемлемо высоким.

На практике организации могут использовать комбинацию подходов, а именно базовый штат из двух операторов плюс возможность быстрого привлечения третьего сотрудника в часы пик за счет аутсорсинга или внутреннего резерва. Имитационное моделирование в данном контексте позволяет точно рассчитать пороги срабатывания такого гибридного сценария и обосновать управленческое решение количественными данными.

## СПИСОК ИСТОЧНИКОВ

1. Лычкина Н.Н. Имитационное моделирование экономических процессов: учеб. пособие / Н.Н. Лычкина. – М.: ИНФРА-М, 2020. – 254 с.
2. Бади́ка Е.М. Использование языка программирования Python в имитационном моделировании / Е.М. Бади́ка, Д.А. Зырянов, С.Г. Бабчи́нецкий // Chronos. – 2022. – Т. 7, № 6 (68). – С. 18–23.
3. Фунг В. Моделирование компьютерной системы с использованием библиотеки SimPy на Python / В. Фунг, В.А. Богатырев // Сборник тезисов докладов конгресса молодых ученых. – СПб.: Университет ИТМО, 2024. – URL: <https://kmu.itmo.ru/digests/article/12036> (дата обращения: 02.06.2026).
4. Кокорин С.В. Сравнение аналитического и численного методов расчёта моментов распределения времени пребывания заявок в сетях массового обслуживания / С.В. Кокорин, Ю.И. Рыжиков // Труды СПИИРАН. – 2012. – № 4 (23). – С. 369–380.

## ИНФОРМАЦИЯ ОБ АВТОРЕ

**Заболотний Роман Александрович**, студент, Воронежский институт высоких технологий, Воронеж, Россия.  
e-mail: [xbg13@yandex.ru](mailto:xbg13@yandex.ru)