

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра теории функций и стохастического анализа

**АНАЛИЗ И ОПТИМИЗАЦИЯ СИСТЕМЫ МАССОВОГО
ОБСЛУЖИВАНИЯ С ОЖИДАНИЕМ
АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ**

Студента 4 курса 412 группы
направления 01.03.02 — Прикладная математика и информатика

механико-математического факультета

Гусева Семёна Сергеевича

Научный руководитель

доцент, к. ф.-м. н., доцент

С. С. Волосивец

Заведующий кафедрой

д. ф.-м. н., доцент

С. П. Сидоров

Саратов 2025

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	3
1 Основное содержание работы	4
ЗАКЛЮЧЕНИЕ	15

ВВЕДЕНИЕ

Теория массового обслуживания (ТМО) является важным разделом прикладной математики, который находит применение в различных областях, таких как телекоммуникации, транспорт, логистика, управление производственными процессами и другие. Основная цель ТМО — анализ и оптимизация систем, в которых возникают очереди, что позволяет повысить их эффективность и снизить затраты.

Цель бакалаврской работы: Изучить методы теории массового обслуживания, применить их для анализа реальной системы и сформулировать практические рекомендации по её оптимизации.

Задачи бакалаврской работы:

- рассмотреть основные параметры СМО: длина очереди, время ожидания, интервалы занятости;
- исследовать системы массового обслуживания (СМО) с ожиданием;
- рассмотреть прикладную задачу анализа СМО для склада запасных частей;
- рассчитать ключевые характеристики: среднее время ожидания, вероятность простоя, средняя длина очереди;
- провести оптимизацию числа серверов на основе стоимостной функции.

Данная бакалаврская работа состоит из двух разделов:

В первом разделе представлен анализ систем массового обслуживания, таких как $M/M/s$, системы с бесконечным числом обслуживающих устройств. Особое внимание уделено изучению параметров, характеризующих работу этих систем, включая длину очереди, длительность ожидания и интервалы занятости.

Во втором разделе сформулирована и решена задача вычислительного эксперимента, с использованием методов теории вероятностей и математической статистики. Цель вычислительного эксперимента показать важность применения методов теории массового обслуживания для анализа и оптимизации реальных систем. Представлена программная реализация на языке Python.

1 Основное содержание работы

Образование очереди возможно, когда клиенты, нуждающиеся в обслуживании, прибывают к определенному обслуживаемому устройству. В качестве примеров очередей можно назвать абонентские вызовы, поступающие на телефонный коммутатор; пациентов, ожидающих приема у врача; станки, ожидающие обслуживания рабочим; автомобили, ожидающие у дорожного пересечения. Система массового обслуживания полностью описывается заданием: 1) входящего потока, 2) дисциплины очереди и 3) порядка обслуживания. Эти понятия можно определить следующим образом:

Входящий поток описывает порядок прибытия клиентов и поступления их в систему. Число клиентов может быть ограниченным или неограниченным, и они могут прибывать поодиночке или группами. Если на обслуживание поступает каждый прибывающий клиент, то будем иметь систему с ожиданием. Однако возможно, что система не примет клиента, когда она полностью занята; в таком случае имеем систему с потерями. Клиент может покинуть систему также и в том случае, если в момент прибытия он застает слишком длинную (или, возможно, слишком короткую) очередь; такое поведение называется неприсоединением к очереди. Наконец, клиент может оказаться вынужденным покинуть систему, если в системе не оказывается места для ожидания.

Дисциплина очереди определяет порядок образования очереди и характер поведения клиентов в процессе ожидания. Простейшей дисциплиной является «первым прибыл - первым обслужен», согласно которой клиенты обслуживаются в порядке их прибытия. При обратном порядке обслуживания имеем дисциплину «прибыл последним - обслужен первым». Возможны и некоторые другие дисциплины очереди. Кроме того, возможны многочисленные различия в поведении клиентов; так, например, прождав в течение некоторого времени, «нетерпеливый» клиент может уйти из системы (т. е. он может покинуть систему до начала обслуживания).

Механизм обслуживания описывает характер распределения клиентов по обслуживаемым устройствам. В общем случае имеется $s \geq 1$ обслуживающих устройств. Мы предполагаем, что число обслуживающих устройств в системе всегда одно и то же, однако возможны случаи, когда в определенные

моменты времени в системе отсутствуют одно или несколько устройств. Если $s = \infty$, то все клиенты будут приняты на обслуживание непосредственно после прибытия, и очередь не образуется. Если имеется конечное число обслуживающих устройств, то они назначаются клиентам в определенном порядке. В частности, дисциплина очереди «первым прибыл - первым обслужен» означает, что то из s обслуживающих устройств, которое освобождается первым, обслуживает клиента, находящегося в начале очереди. В системах с групповым обслуживанием клиенты обслуживаются партиями, а не поодиночке. Кроме того, в некоторых системах обслуживание клиентов может производиться на основании приоритетов определенного порядка.

Допустим, что имеющиеся обслуживающие устройства спроектированы таким образом, что с любой предполагаемой перегрузкой, например с такой, как наплыв транспорта в часы пик, они в состоянии справиться. Рассмотрим перегрузку, вызываемую такими непрогнозируемыми факторами, как нерегулярность входящего потока или механизма обслуживания, и предположим, что эта нерегулярность носит стохастический характер. Следовательно, задачей является изучение статистических флуктуаций в процессах массового обслуживания, направленное на правильное понимание системы, и предполагаем, что на основании данного исследования можно будет предпринять лучшее планирование работы обслуживающих устройств.

В настоящей работе рассмотрим главным образом системы с ожиданием, где совокупность клиентов бесконечна, клиенты прибывают поодиночке, не отказываются поступать в систему, не покидают систему до окончания обслуживания и не имеют приоритетов. Рассматриваемые системы определяются следующими допущениями:

1. Входящий поток. Пусть последовательные клиенты прибывают в моменты времени $t_0, t_1, t_2, \dots$. Обозначим длительность промежутков между последовательными прибытиями через $u_r = t_r - t_{r-1}$ ($r = 1, 2, \dots$). Полагаем, что $u_1, u_2, u_3, \dots$ независимые в совокупности случайные величины, имеющие одну и ту же функцию распределения $A(u)$ ($0 \leq u < \infty$). Кратко это условие можно сформулировать следующим образом: последовательность случайных величин $\{u_r, r = 1, 2, \dots\}$ представляет собой процесс восстановления.

Это допущение означает, что модель прибытия клиентов является ста-

ционарной во времени. Однако это допущение не всегда справедливо. Например, телефонные вызовы поступают на коммутатор более часто в часы занятости, а врачи имеют большой поток пациентов в определенное время года. Однако в большинстве случаев, по-видимому, целесообразно предположить, что входящий поток не меняется с течением времени.

2. Дисциплина очереди - «первым прибыл - первым обслужен».

3. Порядок обслуживания. Число обслуживаемых устройств равно s , где $1 \leq s < \infty$. Время, в течение которого какой-либо определенный клиент находится на обслуживании, называется длительностью обслуживания. Допустим, что длительности обслуживания $u_1, u_2, u_3, \dots$ последовательных клиентов - независимые в совокупности случайные величины, не зависящие также от совокупности случайных величин $u_1, u_2, u_3, \dots$, и что они имеют одну и ту же функцию распределения $B(v) (0 \leq v < \infty)$.

Эти допущения относительно порядка обслуживания означают, что с точки зрения потребности в обслуживании все клиенты одинаковы. Если же, например, на некоторых почтовых отделениях у разных окошек оказываются различные виды услуг, то допускаем, что каждый клиент направляется к соответствующему окошку, и имеется по существу несколько независимых систем массового обслуживания. Кроме того, эти допущения означают, что обслуживающее устройство не прекращает и не меняет темпа обслуживания в какой-либо момент времени и что на его работу ни в коей мере не влияет длина очереди (например, оно не ускоряет обслуживания, если наблюдается образование большой очереди).

Пусть $E(u_i) = a$ обозначает среднюю длительность промежутков времени между последовательными моментами прибытия клиентов, а $E(v_i) = b$ - среднюю длительность обслуживания. Примем также, что $0 < a < \infty$ и $0 < b < \infty$. Для системы с одним обслуживающим устройством (однолинейной системы) отношение средней длительности обслуживания к средней длительности промежутка времени между последовательными моментами прибытия клиентов

$$\rho = \frac{b}{a}, \quad (0 < \rho < \infty),$$

называется загрузкой и выражается в эрлангах (международная единица на-

грузки телефонной станции); эта величина является безразмерной. Применяя к последовательности $\{u_r\}$ известный результат теории восстановления, находим, что при $t \rightarrow \infty$ математическое ожидание числа клиентов, прибывающих за единицу времени, составляет

$$n \approx \frac{1}{a},$$

поэтому можно сказать, что загрузка ρ равна математическому ожиданию числа клиентов, прибывающих за среднее время обслуживания; это оправдывает определение показателя загрузки системы. Величина ρ является важным показателем при анализе систем массового обслуживания. Для s -канальной системы обслуживания ($s < \infty$) загрузка системы определяется как

$$\rho = \frac{b}{as}.$$

В качестве частных случаев входящего потока рассмотрим следующие:

1. Детерминированный, или регулярный, входящий поток D . В данном случае клиенты прибывают через равные промежутки времени длительностью a , поэтому длительность промежутков времени между последовательными моментами прибытия клиентов имеет функцию распределения

$$A(u) = \begin{cases} 0, & \text{если } u < a, \\ 1, & \text{если } u \geq a, \end{cases} \quad (0 < a < \infty). \quad (1)$$

2. Марковский или пуассоновский входящий поток M . В данном случае прибытие клиентов следует пуассоновскому процессу с интенсивностью λ ($0 < \lambda < \infty$). Таким образом, число клиентов, прибывающих в интервале $(0, t)$, распределено по закону Пуассона

$$e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \quad (n = 0, 1, \dots).$$

На основании известного свойства этого распределения промежутки времени между последовательными моментами прибытия клиентов имеют показательное распределение

$$dA(u) = \lambda e^{-\lambda u} du, \quad (0 < u < \infty). \quad (2)$$

Поскольку это распределение обладает марковским свойством

$$P \{u < u_i < u + du \mid u_i \geq u_0\} = \frac{dA(u)}{1 - A(u_0)} = \lambda e^{-\lambda(u-u_0)} du. \quad (3)$$

Такой поток называется марковским; иначе его называют также чисто случайным.

3. Эрланговский входящий поток E_k . В данном случае распределение длительности промежутков времени между последовательными моментами прибытия клиентов имеет вид

$$dA(u) = \frac{(\lambda k)^k}{(k-1)!} e^{-\lambda k u} u^{k-1} du, \quad (0 < u < \infty), \quad (4)$$

где k - положительное целое число.

При $k = 1$ это распределение переходит в показательное, задаваемое формулой (2). При $k \rightarrow \infty$ характеристическая функция распределения (4) будет иметь вид

$$\int_0^\infty e^{i\theta u} dA(u) = \left(1 - \frac{i\theta}{k\lambda}\right)^{-k} \rightarrow e^{i\theta/\lambda},$$

где $e^{i\theta/\lambda}$ - характеристическая функция распределения (1) при $a = \lambda^{-1}$. Это показывает, что при $k \rightarrow \infty$ эрланговский поток стремится к регулярному. В общем случае, когда $1 < k < \infty$, распределение (4) является k -кратной сверткой распределения (2), и поэтому можно считать, что оно изображает процесс прибытия клиентов, который состоит из k последовательных фаз. Время, необходимое для каждой фазы, имеет показательное распределение (2), независимое от других фаз, и поступление следующего клиента происходит только после завершения этих k фаз. Когда поступает новый клиент, процесс начинается с первой фазы.

Вычислительный эксперимент

Постановка задачи

На складе запасных частей производятся замена вышедших из строя механических узлов ставки новыми. Заявки на замену поступают в соответствии с пуассоновским распределением вероятностей с математическим ожиданием 17.5 заявок в час. Штат обслуживающего персонала на складе запасных частей включает двух человек. Каждый работник склада способен удовлетворить в среднем 10 заявок в час; длительность выполнения заявок распределена по экспоненциальному закону. Вычислите:

- a) среднее время от момента подачи заявки до момента получения требуемой запасной части;
- b) долю времени, в течение которого обслуживающий персонал на складе бездействует;
- c) долю времени, в течение которого бездействует один из работников склада;
- d) вероятность того, что заявку на замену узла не удастся удовлетворить сразу;
- e) среднее число заявок, ожидающих выполнения.

Пусть затраты, связанные с добавлением к штату обслуживающего персонала на складе запасных частей одного человека, оцениваются в 6 ед. в час, а потери из-за простоя ставки в период замены тех или иных вышедших из строя узлов составляют 30 ед. в час. Сколько человек должен включать штат обслуживающего персонала на складе, чтобы минимизировать суммарные экономические потери?

Рассмотрим систему массового обслуживания $M/M/s$, где:

- Интенсивность входящего потока заявок: $\lambda = 17.5$ заявок в час.
- Интенсивность обслуживания одним работником: $\mu = 10$ заявок в час.
- Число обслуживающих каналов: $s = 2$.

Требуется:

1. Вычислить среднее время от момента подачи заявки до момента получения требуемой запасной части.
2. Определить долю времени, в течение которого обслуживающий персонал бездействует.

3. Найти вероятность того, что заявка не будет удовлетворена сразу.
4. Рассчитать среднее число заявок в очереди.
5. Оптимизировать число работников s , минимизируя суммарные экономические потери, где:
 - Затраты на добавление одного работника: $c_2 = 6$ у.е./час.
 - Потери из-за простоя станка: $c_1 = 30$ у.е./час.

Решение задачи

Теоретические основы для решения:

Для системы $M/M/s$ с интенсивностью входящего потока λ и интенсивностью обслуживания μ справедливы следующие формулы:

1. **Коэффициент загрузки системы:**

$$\rho = \frac{\lambda}{s\mu}.$$

Условие стационарности: $\rho < 1$.

2. **Вероятность нулевого состояния** (все каналы свободны):

$$P_0 = \left[\sum_{k=0}^{s-1} \frac{(\lambda/\mu)^k}{k!} + \frac{(\lambda/\mu)^s}{s!(1-\rho)} \right]^{-1}.$$

3. **Вероятность того, что заявка не будет удовлетворена сразу** (все каналы заняты):

$$P_{\text{ож}} = \frac{(\lambda/\mu)^s}{s!(1-\rho)} P_0.$$

4. **Среднее число заявок в очереди:**

$$L_q = \frac{(\lambda/\mu)^s \rho}{s!(1-\rho)^2} P_0.$$

5. **Среднее время пребывания заявки в системе:**

$$W = \frac{1}{\mu} + \frac{L_q}{\lambda}.$$

6. Доля времени простоя системы:

$$P_{\text{простой}} = P_0.$$

7. Доля времени, когда занят ровно один канал:

$$P_1 = \frac{(\lambda/\mu)^1}{1!} P_0.$$

Расчет характеристик для $s = 2$

— Коэффициент загрузки:

$$\rho = \frac{17.5}{2 \cdot 10} = 0.875.$$

— Вероятность нулевого состояния:

$$P_0 = \left[1 + \frac{17.5}{10} + \frac{(17.5/10)^2}{2!(1 - 0.875)} \right]^{-1} = \frac{1}{14.75} \approx 0.0678.$$

— Вероятность ожидания:

$$P_{\text{ож}} = \frac{(1.75)^2}{2 \cdot 0.125} \cdot 0.0678 \approx 0.83.$$

— Среднее число заявок в очереди:

$$L_q = \frac{(1.75)^2 \cdot 0.875}{2 \cdot (0.125)^2} \cdot 0.0678 \approx 5.8.$$

— Среднее время пребывания в системе:

$$W = \frac{1}{10} + \frac{5.8}{17.5} \approx 0.1 + 0.33 = 0.43 \text{ часа} \approx 26 \text{ минут}.$$

— Доля времени простоя:

$$P_{\text{простой}} = P_0 \approx 0.068\%.$$

- Доля времени, когда занят ровно один канал:

$$P_1 = 1.75 \cdot 0.0678 \approx 0.1186.$$

Оптимизация с учётом стоимостной функции

Для оптимизации числа работников склада будем использовать следующую стоимостную функцию:

$$F(s) = c_1 \cdot W \cdot \lambda + c_2 \cdot (s - \bar{s}_{\text{зан}}),$$

где:

- $c_1 = 30$ у.е./час·заявка — потери из-за простоя станка в час на одну заявку;
- $c_2 = 6$ у.е./час — затраты на содержание одного работника в час;

Здесь:

- Первое слагаемое $c_1 \cdot W \cdot \lambda$ отражает общие потери из-за простоя станков, пропорциональные среднему времени обслуживания и количеству заявок.
- Второе слагаемое $c_2 \cdot (s - \bar{s}_{\text{зан}})$ учитывает затраты на содержание свободных работников.

Вычислим $F(s)$ для $s = 2, 3, 4$:

В таблицах 1, 2, 3 представлены результаты анализа системы массового обслуживания М/М/с для разного количества серверов.

Таблица 1 – Параметры системы при $s = 2$ серверах

Параметр	Значение
Среднее время ожидания W	0.43 ч
Среднее число занятых серверов $\bar{s}_{\text{зан}}$	1.75
Число свободных серверов $s - \bar{s}_{\text{зан}}$	0.25
Итоговая стоимость системы $F(2)$	227.25 у.е./ч

При двух серверах система перегружена, что приводит к высокой сто-

имости из-за длительного ожидания клиентов (основная составляющая стоимости - 225.75 у.е./ч). Код программы для расчета параметров представлен в приложении Б. Описание программы представлено в приложении А.

Таблица 2 – Параметры системы при $s = 3$ серверах

Параметр	Значение
Коэффициент загрузки ρ	0.583
Вероятность простоя P_0	0.116
Средняя длина очереди L_q	0.34
Среднее время ожидания W	0.119 ч
Среднее число занятых серверов $\bar{s}_{зан}$	1.75
Число свободных серверов $s - \bar{s}_{зан}$	1.25
Итоговая стоимость системы $F(3)$	69.98 у.е./ч

Добавление третьего сервера значительно снижает стоимость системы (в 3.25 раза). Стоимость ожидания уменьшается до 62.48 у.е./ч, но появляется стоимость простоя 7.5 у.е./ч.

Таблица 3 – Параметры системы при $s = 4$ серверах

Параметр	Значение
Коэффициент загрузки ρ	0.4375
Вероятность простоя P_0	0.132
Средняя длина очереди L_q	0.04
Среднее время ожидания W	0.102 ч
Среднее число занятых серверов $\bar{s}_{зан}$	1.75
Число свободных серверов $s - \bar{s}_{зан}$	2.25
Итоговая стоимость системы $F(4)$	67.05 у.е./ч

Четвертый сервер дает незначительное улучшение (стоимость снижа-

ется всего на 2.93 у.е./ч). Дальнейшее увеличение числа серверов становится неэффективным.

В таблице 4 представлен сравнительный анализ систем массового обслуживания $M/M/s$ при разном количестве серверов.

Таблица 4 – Сравнительный анализ систем $M/M/s$ при разном количестве серверов

Параметр	Количество серверов (s)		
	2	3	4
Коэффициент загрузки ρ	0.875	0.583	0.438
Вероятность простоя P_0 (%)	6.78	11.60	13.20
Среднее время ожидания W (мин)	25.9	7.2	6.1
Среднее число занятых серверов $\bar{s}_{зан}$	1.75	1.75	1.75
Среднее число свободных серверов	0.25	1.25	2.25
Средняя длина очереди L_q	5.80	0.34	0.04
Итоговая стоимость $F(s)$ (у.е./час)	227.25	69.98	67.05
Экономия стоимости относительно $s=2$ (%)	–	69.2	70.5
Сокращение времени ожидания в очереди (%)	–	72.2	76.4

- При $s=2$ система перегружена ($\rho = 0.875$), что приводит к длинным очередям и высокой стоимости
- $s=3$ дает оптимальный баланс - стоимость снижается на 69.2% при значительном улучшении времени ожидания
- $s=4$ обеспечивает минимальную стоимость, но экономический выигрыш незначителен (+2.93 у.е. к $s=3$)

На основе теоретических формул для системы $M/M/s$ были рассчитаны ключевые характеристики и проведена оптимизация числа работников. Оптимальное число работников — $s = 3$, так как при $s = 4$ уменьшение потерь $c_1 \cdot W \cdot \lambda$ не компенсирует рост затрат на содержание дополнительного работника.

ЗАКЛЮЧЕНИЕ

Основные результаты теоретической части:

- Исследованы системы $M/M/s$, где определены предельные распределения длины очереди и длительности ожидания. Показано, что при загрузке системы $\rho < 1$ система достигает стационарного режима, а при $\rho \geq 1$ очередь неограниченно возрастает.
- Рассмотрены системы с бесконечным числом обслуживающих устройств, для которых получено пуассоновское распределение количества занятых устройств в стационарном режиме.

Основные результаты практической части:

- Проведён расчёт характеристик системы $M/M/s$.
- Оптимизировано число обслуживающих каналов.
- Реализован алгоритм расчёта на Python, включая визуализацию зависимостей потерь от числа каналов.

Работа демонстрирует эффективность методов ТМО для оптимизации систем с очередями. Результаты показывают, что увеличение числа каналов с 2 до 3 значительно улучшает показатели системы. Дальнейшие исследования могут включать анализ нестационарных потоков или систем с приоритетами. Результаты работы могут быть полезны для принятия управленческих решений в аналогичных системах.