# **Тема 2. Задачи в условиях неопределенности**

## Системы массового обслуживания: понятия, примеры, модели

Основные понятия теории массового обслуживания

***Теория массового обслуживания*** (ТМО) — научная дисциплина, занимающаяся математическим моделированием постоянно повторяющихся однотипных задач в одних и тех же условиях их появления, анализом способов их решения в целях повышения эффективности этих решений.

***Система массового обслуживания*** (СМО) — система, предназначенная для многократно повторяющегося (многоразового) использования при решении однотипных задач.

***Процесс обслуживания*** — решение такой системой возникающей типовой задачи.

***Задача обслуживания*** — повторяющаяся типовая задача, возникающая в СМО.

***Заявка*** (требование) — объект обслуживания в СМО, для которого решается типовая задача обслуживания.

***Поток заявок*** — последовательность объектов, подлежащих обслуживанию в СМО.

***Канал обслуживания*** — обслуживающая единица, т. е. элемент системы, предназначенный для решения задачи обслуживания.

***Очередь*** — последовательность заявок, ожидающих обслуживания.

***Дисциплина обслуживания*** — правила, определяющие порядок выбора заявок из числа поступивших и порядок их распределения между каналами обслуживания.

***Показатели эффективности СМО*** — показатели, описывающие способность СМО справляться с потоком заявок.

**Структура, классификация и предмет теории СМО**

*Структура СМО*

В структуре СМО выделяют следующие элементы:

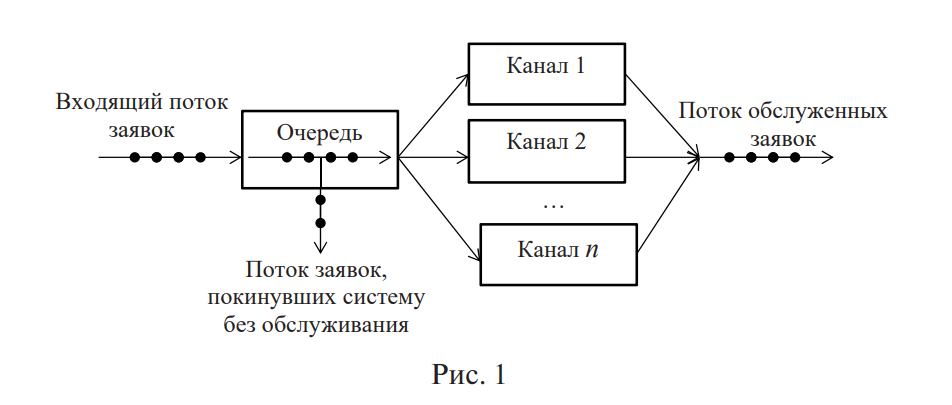
* входящий поток заявок;
* очередь (если ее наличие предусмотрено системой);
* каналы обслуживания;
* выходящий поток заявок.

Выходящий поток может делиться на поток обслуженных заявок и поток заявок, оставшихся без обслуживания.

Последний, в свою очередь, может делиться на:

* поток заявок, которым было отказано в обслуживании;
* поток заявок, покинувших систему из-за превышения лимита времени нахождения в системе.

Структурная схема СМО представлена на рис. 1.



*Классификация СМО*

***По типу обслуживания выделяют два основных класса СМО:***

* СМО с отказами — СМО, в которых заявка, поступившая в тот момент, когда все каналы СМО заняты, получает отказ в обслуживании, покидает СМО и в дальнейшем процессе обслуживания не участвует. Эти СМО называют также системами без очереди;
* СМО с ожиданием — СМО, в которых заявка, поступившая в момент, когда все каналы системы заняты, не уходит, а становится в очередь на обслуживание. Эти СМО называют также системами с очередью. СМО с ожиданием подразделяются на:
* СМО с неограниченным ожиданием (с неограниченной очередью);
* СМО с ограниченным ожиданием двух типов: с ограничением на длину очереди и с ограничением на время пребывания заявки в очереди (или на общее время пребывания заявки в системе).

***По характеру поступающего потока заявок СМО делят на:***

* немарковские системы;
* марковские системы.

***По числу каналов СМО делят на:***

* одноканальные;
* многоканальные.

***По количеству этапов обслуживания их делят на:***

* однофазные — заявка проходит только одну стадию (фазу) обработки, все каналы выполняют одну и ту же операцию;
* многофазные — заявка проходит последовательно несколько стадий обработки в системе, каждый канал проводит заявку через все необходимые ей фазы.

***По схеме обслуживания заявок из очереди (по дисциплине очереди) СМО*** ***делят на:***

* СМО с упорядоченной очередью (обслуживание в соответствии с определенным порядком);
* СМО с неупорядоченным (случайным) выбором заявок из очереди;
* СМО с обслуживанием заявок в соответствии с приоритетом.

Очередь может быть упорядочена разными способами: в порядке поступления заявок (заявка, поступившая первой, обрабатывается первой и т.д.), в обратном порядке (заявка, поступившая последней, обрабатывается первой и т. д.) и другими способами.

При обслуживании заявок с приоритетом в первую очередь обслуживаются наиболее важные заявки.

Абсолютный приоритет означает, что более важная заявка «вытесняет» из-под обслуживания менее важную.

Относительный приоритет означает, что более важная заявка получает лучшее место в очереди, т. е. выстраивается ранжированная по важности очередь заявок.

Приоритет может быть статичным (не меняться с течением времени) и динамичным (повышаться с увеличением времени ожидания в очереди с разным коэффициентом для разных заявок).

***По ограничению потока заявок выделяют:***

* открытые СМО, в них заявки поступают в неограниченном количестве из источников вне самой системы;
* замкнутые СМО (количество источников ограничено), в которых источники заявок включены в саму систему; заявка из такого источника, покинувшая систему, через какое-то время может снова в нее возвратиться.

***По времени обслуживания заявок рассматривают:***

* СМО со случайным временем обслуживания каждой заявки;
* СМО с заданным временем обслуживания заявок, которое может быть одинаковым для всех заявок и разным для разных заявок.

***По способу формирования множества заявок выделяют:***

* СМО с фиксированной очередью. В них очередь формируется до начала работы системы и в процессе ее работы новые заявки в очередь не поступают;
* СМО с потоком заявок. Заявки поступают в такую систему в продолжение всего времени работы системы. Поток заявок в этом случае может быть:
* регулярным, когда заявки поступают в систему в заранее фиксированные моменты времени или через определенные промежутки времени;
* случайным, когда заявки поступают в систему в случайные, заранее не определенные моменты времени.

*Предмет теории массового обслуживания*

Предметом теории массового обслуживания является построение математических моделей, связывающих заданные условия работы СМО (число каналов, их производительность, характер потока заявок, организация работы и т.п.) с показателями эффективности функционирования СМО.

К числу показателей эффективности функционирования СМО от‑

носят три основных группы показателей.

1. Основные показатели эффективности использования СМО:

* среднее число заявок, обслуживаемых в единицу времени;
* средняя доля заявок, обслуживаемых СМО в единицу времени, от среднего числа заявок, поступающих в СМО за то же время;
* средняя продолжительность периода занятости СМО;
* средняя доля времени, в течение которого СМО занята обслуживанием заявок;
* средняя доля времени, в течение которого СМО простаивает, т. е. не занята обслуживанием заявок.

2. Основные показатели качества обслуживания:

* среднее время ожидания обслуживания;
* среднее время пребывания заявки в СМО;
* вероятность отказа в обслуживании без ожидания;
* вероятность того, что прибывшая заявка немедленно будет принята к обслуживанию;
* среднее число заявок, находящихся в СМО;
* среднее число заявок в очереди;
* вероятность того, что число заявок в очереди превысит определенное значение.

3. Основные показатели экономической эффективности функционирования СМО:

* средний доход, приносимый СМО в единицу времени;
* отношение среднего дохода к затратам на содержание и функционирование СМО в единицу времени (т. е. средний доход, приходящийся на единицу затрат);
* потеря дохода в результате отказа части заявок в обслуживании.

СМО с фиксированной очередью (задачи упорядочения)

***Задача минимизации штрафа за задержку обслуживания.***

Будем рассматривать один пункт обслуживания с поступившим на него определенным количеством заявок (n заявок), т.е. одноканальную СМО с образовавшейся очередью. Каждая заявка имеет свое определенное время обслуживания. Задержка обслуживания каждой заявки влечет за собой определенный для каждой заявки штраф. Необходимо так выстроить последовательность обслуживания имеющихся заявок, чтобы минимизировать суммарный штраф.

Примеры подобных систем: грузовой причал морского порта с пришедшими к нему для разгрузки судами, которые имеют разный объем с различными по сроку хранения и реализации грузами. В силу этого у них отличаются время обслуживания и штраф за задержку обслуживания на единицу времени (час, сутки и т.п.). То же самое для грузовой платформы ж/д станции с подошедшими для разгрузки поездами или вагонами, имеющими разнокачественный груз: продовольственные товары, промышленные изделия, горюче-смазочные материалы и т.п. Такая же ситуация может возникнуть при обработке документации различного объема, важности и срочности.

Итак, имеется один пункт обслуживания и n ожидающих обслуживания заявок. Каждая заявка имеет свое время обслуживания и штраф за его задержку. Обозначим время обслуживания i‑й заявки как , а штраф за задержку обслуживания этой заявки на единицу времени (минута, час, сутки и т.п.) — ; № — номер заявки. Сведем все данные в табличную форму (см. ниже).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № | 1 | 2 | 3 | … | 4 |
|  |  |  |  | … |  |
|  |  |  |  | … |  |

*Показатель эффективности рассматриваемой СМО* — суммарный штраф за задержку обслуживания. Разные последовательности обслуживания заявок дают в общем случае разный суммарный штраф, т. е. разную эффективность СМО. Решение задачи сводится к определению такого порядка обслуживания (установлению такой последовательности обслуживания заявок), который обеспечивает минимальный суммарный штраф за задержку обслуживания всех этих заявок.

Такие задачи называются задачами упорядочения.

Число возможных различных последовательностей из n заявок равно числу всевозможных перестановок из n элементов (n!). Это число может быть весьма велико (например, 10! = 3628800), но оно конечно. Следовательно, среди этих последовательностей найдется по крайней мере одна такая, которая даст решение поставленной задачи, т. е. обеспечит минимальный суммарный штраф обслуживания. Такую последовательность будем называть оптимальной. Произвольную последовательность заявок обозначим символом α. Символом обозначим оптимальную последовательность. Может оказаться, что существует несколько оптимальных последовательностей. Это будет означать, что данная задача имеет не единственное решение, т. е. будут существовать несколько последовательностей обслуживания заявок, обеспечивающих одинаковый минимальный суммарный штраф за задержку обслуживания. Суммарный штраф от некоторой последовательности α обслуживания заявок обозначим S(α). Тогда суммарный штраф от оптимальной последовательности будет равен S().

Поскольку оптимальная последовательность обслуживания заявок существует у каждой задачи рассматриваемого типа, будем считать, что она нами уже найдена (хотя бы одна для каждой задачи) и что заявки пронумерованы в соответствии с этой последовательностью,

т.е. = (1, 2, 3, …, n). Суммарный штраф S () = S (1, 2, 3, …, n) = min S(α). Рассмотрим найденную последовательность = (1, 2, 3, …, k, k+1, …, n). Очевидно, что любые два элемента конечной последовательности мож‑ но поменять местами, последовательно меняя местами два соседних элемента. Следовательно, любая последовательность заявок может быть получена из оптимальной последовательности путем конечного числа изменения мест двух соседних элементов. Это значит, что вывод, полученный при рассмотрении изменения штрафа после перестановки двух произвольных соседних заявок в оптимальной последовательности, можно распространить на произвольную последовательность заявок, которая отличается от рассмотренной оптимальной.

Поменяем местами k‑ю и k+1‑ю заявки в оптимальной последовательности = (1, 2, 3, …, k, k+1, …, n). В результате получим последовательность

α = (1, 2, 3, …, k–1, k+1, k, k+2, …, n), которая в общем случае уже не является оптимальной. Возможно, что она останется оптимальной, но это будет частным случаем конкретной задачи. В любом случае суммарный штраф от новой последовательности будет не меньше, чем штраф от прежней оптимальной последовательности обслуживания заявок.

Таким образом,

S() = S(1, 2, 3, …, k, k+1, …, n) ≤ S(1, 2, 3, …, k–1, k+1, k, k+2,…, n) = S(α). Возможное увеличение штрафа равно ΔS (α) = S (α) — S (α0) ≥ 0.

Найдем величину ΔS (α). Заявка с номером k, переместившись на k+1‑е место, будет теперь ждать пока идет обслуживание k+1‑й заявки, занявшей ее место. Это время равно , штраф за задержку обслуживания k‑й заявки на единицу времени равен . Общий штраф в результате этого перемещения увеличится на величину . Одновременно с этим общий штраф уменьшится за счет перемещения k+1‑й заявки на k‑е место на величину , т. к. время ожидания обслуживания k+1‑й заявки уменьшится на , поскольку ей не нужно будет ждать обслуживания k‑й заявки, а штраф за задержку обслуживания на единицу времени у k+1‑й заявки равен .

Итак, величина общего штрафа изменится на ΔS (α) = – ≥ 0. Номер k был выбран произвольно в пределах от 1 до n – 1. Это значит, что неравенство = – ≥ 0 определяет оптимальную последовательность, если выполняется при любом k = 1, 2, 3, …, n – 1. В этом случае перестановка любых двух соседних заявок приводит к увеличению суммы общего штрафа, если для таких заявок неравенство выполнялось как строгое, и последовательность обслуживания перестает быть оптимальной. Если же для каких-то соседних заявок это неравенство переходит в равенство, то перестановка таких заявок не меняет оптимальности полученной последовательности и общая сумма штрафа остается прежней, т.е. минимальной. В этом случае получаем второе решение нашей задачи. Представим неравенство = – ≥ 0 в виде ≥ . Для любой заявки время ее обслуживания > 0, по‑ этому > 0 и > 0. Разделим обе части неравенства на > 0. Получим характеризующее оптимальную последовательность заявок неравенство

≥ , которое должно выполняться для любого k = 1, 2, 3, …, n – 1.

Отношение будем называть *относительным штрафом* i‑й заявки. Таким образом, чтобы последовательность обслуживания заявок была оптимальной, требуется выполнение последовательности неравенств

≥ ≥ ≥ ... ≥ .

Отсюда следует правило построения оптимальной последовательности обслуживания заявок, обеспечивающей минимальный суммарный штраф за задержку обслуживания: *первыми должны обслуживаться заявки, имеющие больший относительный штраф*.

***Пример***.

К причалу морского порта подошли для разгрузки 10 судов. Время разгрузки (в часах) и штраф за задержку обслуживания (в каких-то денежных единицах за час ожидания, например в тыс. руб.) отражены ниже.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| № | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|  | 6 | 8 | 7 | 5 | 3 | 2 | 4 | 1 | 3 | 5 |
|  | 4 | 2 | 3 | 6 | 1 | 4 | 5 | 3 | 6 | 3 |

Нужно найти порядок обслуживания, минимизирующий суммарный штраф ожидания разгрузки, определить величину этого штрафа и сравнить ее с суммарным штрафом от обслуживания в порядке исходной нумерации заявок.

***Решение.***

Сначала найдем суммарный штраф при обслуживании заявок впорядке их исходной нумерации. Время ожидания обслуживания k‑й по очереди заявки равно (k) = + + … + . Штраф за ожидание обслуживания k первых по очереди заявок можно найти следующим образом:

S(1, 2, …, k) = · 0 + + ( + ) + … + ( + + … + ).

Первая заявка имеет нулевое ожидание, вторая ждет, пока будет обслужена первая заявка, третья заявка ожидает обслуживания первых двух заявок и т.д. Общий штраф за задержку обслуживания всех заявок в последовательности исходной нумерации α = (1, 2, …, n) равен

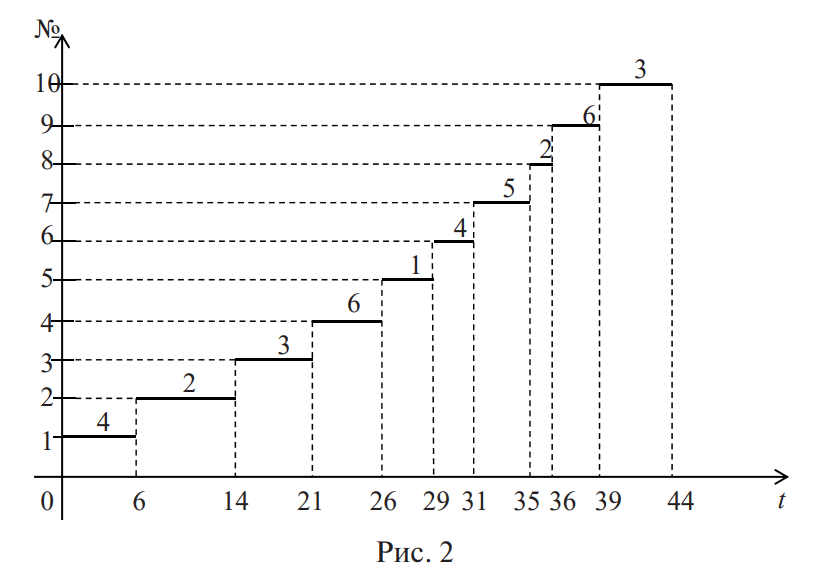
S(α) = S(1, 2, …, n) = ·0 + + ( + ) + … + ( + + … + ).

Для упрощения подсчета и его наглядности можно воспользоваться диаграммой Ганта. В диаграммах Ганта процесс, занимающий определенный интервал времени, изображается на графике отрезком со‑ ответствующей длины в выбранном масштабе. По оси x откладываем текущее время, по оси y — номера заявок в исходном варианте. Время обслуживания каждой заявки изображаем отрезком прямой, параллельным оси абсцисс, на месте соответствующей заявки. Последовательность выполнения заявок отражается на графике последовательностью соответствующих отрезков.

Ниже представлена диаграмма Ганта для последовательности обслуживания в исходном порядке (рис. 2).

Над каждым отрезком, изображающим процесс обслуживания заявки, поставим значение штрафа , соответствующего этой заявке. На оси абсцисс отразим время протекания процесса обслуживания. Фиксируем время окончания обслуживания очередной заявки.

Это время является также временем ожидания обслуживания следую‑ щей заявки. Первая заявка заканчивает обслуживание через 6 ч после начала процесса обслуживания, вторая — через 14 ч (6 + 8), третья — через 21 ч (14 + 7) и т.д. Соответственно, у первой заявки ожидание отсутствует, вторая заявка ждет обслуживания 6 ч, третья — 14 ч, чет‑ вертая — 21 ч, пятая — 26 ч и т.д.



Подсчитаем суммарный штраф обслуживания:

S(α) ­= 4·0 + 2·6 + 3·14 + 6·21 + 1·26 + 4·29 + 5·31 + 2·35 + 6·36 + 3·39 = 880

Найдем оптимальную последовательность обслуживания заявок. Для этого рассчитаем относительный штраф / каждой заявки и расположим заявки в порядке уменьшения их относительного штрафа (см.ниже).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| № | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|  | 6 | 8 | 7 | 5 | 3 | 2 | 4 | 1 | 3 | 5 |
|  | 4 | 2 | 3 | 6 | 1 | 4 | 5 | 2 | 6 | 3 |
|  | 2/3 | 1/4 | 3/7 | 6/5 | 1/3 | 2 | 5/4 | 2 | 2 | 3/5 |
|  | 6 | 10 | 8 | 5 | 9 | 1 | 4 | 2 | 3 | 7 |

- оптимальное расположение (последовательность) обработки заявок.

Последовательность относительных штрафов данной задачи имеет вид:

2 = 2 = 2 > 5/4 > 6/5 > 2/3 > 3/5 > 3/7 > 1/3 > 1/4.

Для заявок № 6, № 8, № 9 относительный штраф одинаковый (равен 2), поэтому эти три заявки можно менять между собой местами. Таким образом, существует 6 (3!) равнозначных решений поставленной задачи, т. е. 6 оптимальных последовательностей обслуживания, обеспечивающих одинаковый минимальный размер штрафа:

= (6, 8, 9, 7, 4, 1, 10, 3, 5, 2);

= (8, 9, 6, 7, 4, 1, 10, 3, 5, 2);

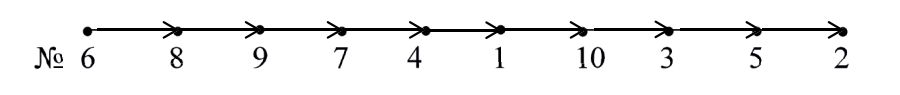
= (9, 6, 8, 7, 4, 1, 10, 3, 5, 2);

= (8, 6, 9, 7, 4, 1, 10, 3, 5, 2);

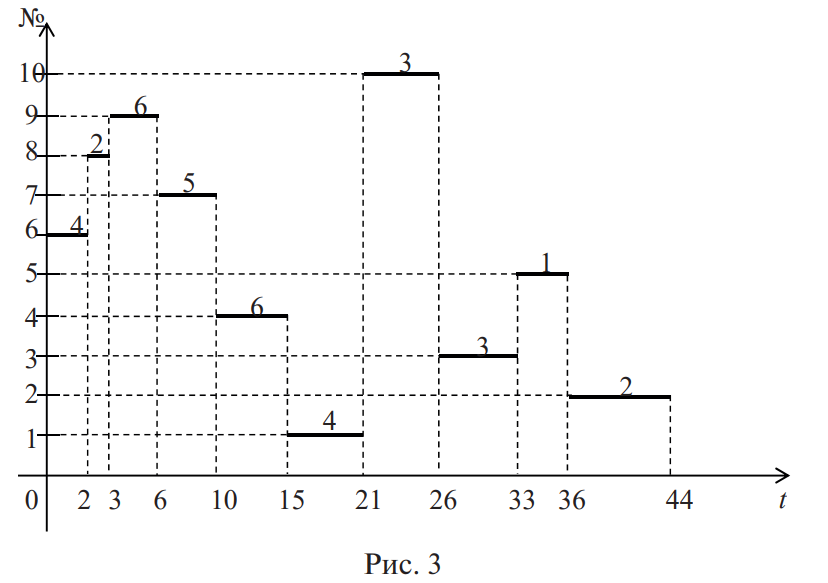
= (6, 9, 8, 7, 4, 1, 10, 3, 5, 2);

= (9, 8, 6, 7, 4, 1, 10, 3, 5, 2).

Одна из оптимальных последовательностей обработки заявок (с сохранением их исходной нумерации) такова: = (6, 8, 9, 7, 4, 1, 10, 3, 5, 2). Изобразим графически последовательность обработки заявок:



Ниже представлена диаграмма Ганта для оптимальной последовательности (рис. 3):



Штраф для оптимальной последовательности обслуживания заявок равен

= S(α) – S() = S(α) - =880 – 418 = 462.

Штраф для оптимальной последовательности обслуживания заявок меньше на 462 денежных единицы (тыс. руб.) штрафа для исходной последовательности.