Министерство Образования Российской Федерации

ЮЖНО-РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ЭКОНОМИКИ И СЕРВИСА (ЮРГУЭС)

Саакян Г.Р.

ЛЕКЦИИ

ТЕОРИЯ МАССОВОГО ОБСЛУЖИВАНИЯ

для студентов экономических специальностей очной, заочной и дистанционной форм обучения

Шахты 2006

ТЕОРИЯ МАССОВОГО ОБСЛУЖИВАНИЯ

Введение

Сложный характер рыночной экономики и современный уровень предъявляемых к ней требований стимулируют использование более серьезных методов анализа ее теоретических и практических проблем. В последние десятилетия значительный вес в экономических исследованиях приобрели математические методы. Математическое моделирование все более и более становится одним из основных и наиболее плодотворных методов изучения экономических процессов и объектов. Математический анализ экономических задач органично превращается в часть экономики. Положительная оценка этого подтверждается и тем, что начиная с 1969 г. Нобелевские премии в области экономики присуждаются, как правило, за экономико-математические исследования.

Одним из важных разделов экономико-математического моделирования является *теория массового обслуживания*, представляющая собой теоретические основы эффективного конструирования и эксплуатации систем массового обслуживания. Системы массового обслуживания (СМО) встречаются во многих областях экономики (производство, техника, военная область, быт и др.) и предназначены для многократного использования при выполнении однотипных задач.

В борьбу за клиента в современной экономике вкладываются огромные средства. По оценкам западных экономистов, завоевание фирмой нового клиента обходится ей в 6 раз дороже, чем удержание существующих покупателей. А если клиент ушел неудовлетворенным, то на его возвращение приходится потратить в 25 раз больше средств. Во многих случаях неудовлетворенность клиента вызвана неудачной организацией его обслуживания (слишком долгое ожидание в очереди, отказ в обслуживании и т.д.). Использование теории массового обслуживания позволяет фирме избежать подобных неприятностей.

Основоположником теории массового обслуживания считается датский ученый А. К. Эрланг. Являясь сотрудником Копенгагенской телефонной компании, он опубликовал в 1909 году работу «Теория вероятностей и телефонные переговоры», в которой решил ряд задач по теории систем массового обслуживания с отказами.

Значительный вклад в создание и разработку общей теории массового обслуживания внес выдающийся советский математик Александр Яковлевич Хинчин (1984 – 1959), который предложил сам термин *теория массового обслуживания*. В зарубежной литературе чаще используется название *теория очередей*.

Предмет, цель и задачи теории массового обслуживания

Во многих областях производства, бытового обслуживания, экономики и финансов важную роль играют системы[1] специального вида, реализующие многократное выполнение однотипных задач. Подобные системы называют *системами массового обслуживания* (СМО). В качестве примеров СМО в финансово-экономической сфере можно привести системы, представляющие собой банки, страховые организации, налоговые инспекции, аудиторские службы. В сфере производства и обслуживания примерами СМО могут служить: различные системы связи (в том числе телефонные станции), погрузочно-разгрузочные комплексы (порты, товарные станции), автозаправочные станции, магазины, парикмахерские, билетные кассы, пункты обмена валюты, ремонтные мастерские, больницы и т.д. Такие системы как компьютерные сети, системы сбора, хранения и обработки информации, транспортные системы, автоматизированные производственные участки и, в военной области, системы противовоздушной или противоракетной обороны также могут рассматриваться как своеобразные СМО.

[1] Теоретически в общем случае *система* определяется как целостное множество взаимосвязанных элементов, которое нельзя расчленить на независимые подмножества.

Каждая СМО включает в свою структуру некоторое число обслуживающих устройств (единиц, приборов, линий), которые называют *каналами обслуживания*. Роль каналов могут играть лица, выполняющие те или иные операции (кассиры, операторы, продавцы, парикмахеры и т.д.), линии связи, автомашины, краны, ремонтные бригады, железнодорожные пути, бензоколонки и т.д.

Каждая СМО предназначена для обслуживания (выполнения) некоторого *потока*[2] *заявок* (или *требований*), поступающих на вход системы большей частью не регулярно, а в случайные моменты времени. Обслуживание заявок, в общем случае, также длится не постоянное, заранее известное, а случайное время. После обслуживания заявки канал освобождается и готов к приему следующей заявки. Случайный характер потока и времени их обслуживания приводит к неравномерной загруженности СМО: в некоторые промежутки времени на входе СМО могут скапливаться необслуженные заявки (они либо становятся в очередь, либо покидают СМО необслуженными), в другие же периоды при свободных каналах на входе СМО заявок не будет, что приводит к недогрузке СМО, т.е. к простаиванию каналов.

Схема СМО изображена на рисунке 1.

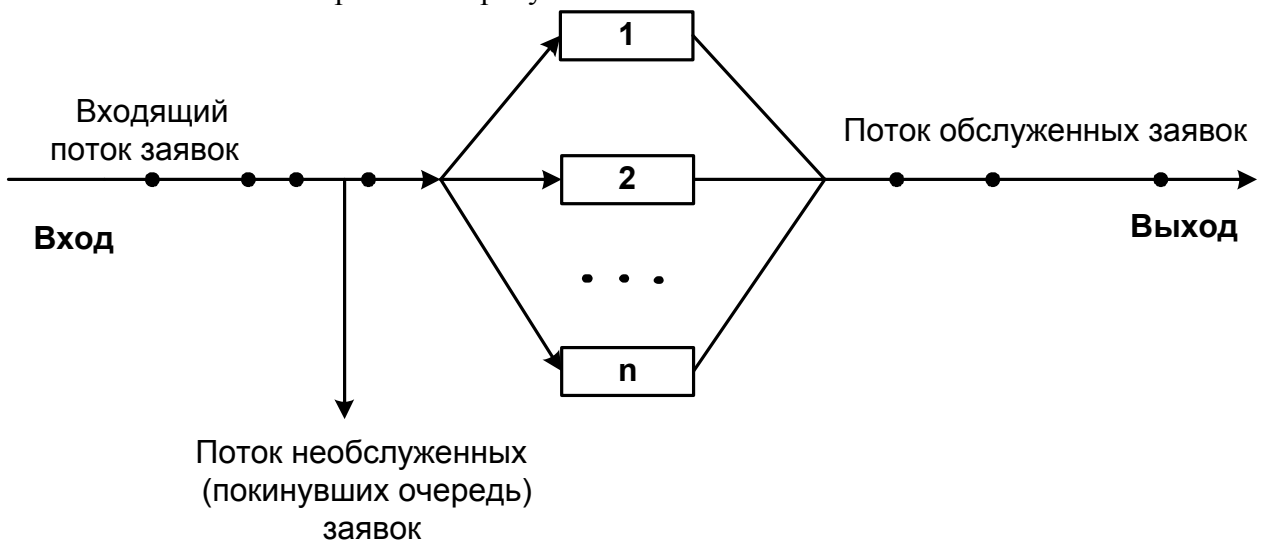

Рисунок 1.

Таким образом, во всякой СМО можно выделить следующие основные элементы:

1) входящий поток заявок;
2) очередь;
3) каналы обслуживания;
4) выходящий поток обслуженных заявок.

Каждая СМО в зависимости от своих параметров: характера потока заявок, числа каналов обслуживания и их производительности, а также от правил организации работы, обладает определенной *эффективностью функционирования* (пропускной способностью), позволяющей ей более или менее успешно справляться с потоком заявок.

Предметом изучения теории массового обслуживания являются СМО.

Цель теории массового обслуживания – выработка рекомендаций по рациональному построению СМО, рациональной организации их работы и регулированию потока заявок для обеспечения высокой эффективности функционирования СМО.

Для достижения этой цели ставятся *задачи теории массового обслуживания*, состоящие в установлении зависимостей эффективности функционирования СМО от ее организации (параметров): характера потока заявок, числа каналов и их производительности и правил работы СМО.

---

[2] *Потоком событий* (в данном случае, заявок) называют последовательность событий, наступающих одно за другим в какие-то заранее неизвестные, случайные моменты времени.

В качестве характеристик эффективности функционирования СМО можно выбрать три основные группы (обычно средних) показателей:

1. Показатели эффективности использования СМО:

1.1. Абсолютная пропускная способность СМО – среднее число заявок, которое сможет обслужить СМО в единицу времени.

1.2. Относительная пропускная способность СМО – отношение среднего числа заявок, обслуживаемых СМО в единицу времени, к среднему числу поступивших за это же время заявок.

1.3. Средняя продолжительность периода занятости СМО.

1.4. Коэффициент использования СМО – средняя доля времени, в течение которого СМО занята обслуживанием заявок, и т.п.

2. Показатели качества обслуживания заявок:

2.1. Среднее время ожидания заявки в очереди.

2.2. Среднее время пребывания заявки в СМО.

2.3. Вероятность отказа заявке в обслуживании без ожидания.

2.4. Вероятность того, что вновь поступившая заявка немедленно будет принята к обслуживанию.

2.5. Закон распределения времени ожидания заявки в очереди.

2.6. Закон распределения времени пребывания заявки в СМО.

2.7. Среднее число заявок, находящихся в очереди.

2.8. Среднее число заявок, находящихся в СМО, и т.п.

3. Показатели эффективности функционирования пары «СМО – клиент», где под «клиентом» понимают всю совокупность заявок или некий их источник. К числу таких показателей относится, например, средний доход, приносимый СМО в единицу времени, и т.п.

Случайный характер потока заявок и длительности их обслуживания порождает в СМО *случайный процесс*.

Определение. *Случайным процессом* (или *случайной функцией*) называется соответствие, при котором каждому значению аргумента (в данном случае – моменту из промежутка времени проводимого опыта) ставится в соответствие случайная величина (в данном случае – состояние СМО).

Поэтому для решения задач теории массового обслуживания необходимо изучить случайный процесс, протекающий в СМО, т.е. необходимо построить и проанализировать его математическую модель. Математический анализ работы СМО существенно упрощается, если этот случайный процесс удовлетворяет определенным условиям, которые будут рассмотрены ниже.

Классификация систем массового обслуживания

Системы массового обслуживания делятся на типы (или классы) по ряду признаков.

По числу каналов СМО подразделяют на *одноканальные* (когда имеется один канал обслуживания) и *многоканальные*, точнее n-канальные (когда количество каналов $n \geq 2$). Здесь и далее будем полагать, что каждый канал одновременно может обслуживать только одну заявку и, если не оговорено специально, каждая находящаяся под обслуживанием заявка обслуживается только одним каналом. Многоканальные СМО могут состоять из однородных каналов, либо из разнородных, отличающихся длительностью обслуживания одной заявки. Практически время обслуживания каналом одной заявки $T_{об}$ является непрерывной случайной величиной. Однако при условии абсолютной однородности поступающих заявок и каналов время обслуживания может быть и величиной постоянной ($T_{об} = const$).

По дисциплине обслуживания СМО подразделяют на три класса:

1. СМО *с отказами*, в которых заявка, поступившая на вход СМО в момент, когда все каналы заняты, получает «отказ» и покидает СМО («пропадает»). Чтобы эта заявка все же была обслужена, она должна снова поступить на вход СМО и рассматриваться при этом как заявка, поступившая впервые. Примером СМО с отказами может служить работа АТС: если набранный телефонный номер (заявка, поступившая на вход) занят, то заявка получает отказ, и, чтобы дозвониться по этому номеру, следует его набрать еще раз (заявка поступает на вход как новая).

2. СМО *с ожиданием* (*неограниченным ожиданием* или *очередью*). В таких системах заявка, поступившая в момент занятости всех каналов, становится в очередь и ожидает освобождения канала, который примет ее к обслуживанию. Каждая заявка, поступившая на вход, в конце концов будет обслужена. Такие СМО часто встречаются в торговле, в сфере бытового и медицинского обслуживания, на предприятиях (например, обслуживание станков бригадой наладчиков).

3. СМО *смешанного типа* (*с ограниченным ожиданием*). Это такие системы, в которых на пребывание заявки в очереди накладываются некоторые ограничения.

Эти ограничения могут накладываться на *длину очереди*, т.е. максимально возможное число заявок, которые одновременно могут находиться в очереди. В качестве примера такой системы можно привести мастерскую по ремонту автомобилей, имеющую ограниченную по размерам стоянку для неисправных машин, ожидающих ремонта.

Ограничения ожидания могут касаться *времени пребывания заявки в очереди*, по истечению которого она выходит из очереди и покидает систему, либо касаться *общего времени пребывания заявки в* СМО (т.е. суммарного времени пребывания заявки в очереди и под обслуживанием).

В СМО с ожиданием и в СМО смешанного типа применяются различные схемы обслуживания заявок из очереди. Обслуживание может быть *упорядоченным*, когда заявки из очереди обслуживаются в порядке их поступления в систему, и *неупорядоченным*, при котором заявки из очереди обслуживаются в случайном порядке. Иногда применяется *обслуживание с приоритетом*, когда некоторые заявки из очереди считаются приоритетными и поэтому обслуживаются в первую очередь.

По ограничению потока заявок СМО делятся на *замкнутые* и *открытые*.

Если поток заявок ограничен и заявки, покинувшие систему, могут в нее возвращаться, то СМО является *замкнутой*, в противном случае – *открытой*. Классическим примером замкнутой СМО служит работа бригады наладчиков в цеху. Станки являются источниками заявок на обслуживание, и их количество ограничено, наладчики – каналы обслуживания. После проведения ремонтных работ вышедший из строя станок снова становится источником заявок на обслуживание. В открытой СМО характеристики потока заявок не зависят от того, в каком состоянии сама СМО (сколько каналов занято). В замкнутой СМО – зависят. Так, в рассмотренном выше примере интенсивность потока «заявок» со стороны станков (т.е. количество заявок в единицу времени) зависит от того, сколько их неисправно и ждет наладки.

По количеству этапов обслуживания СМО делятся на *однофазные* и *многофазные* системы. Если каналы СМО однородны, т.е. выполняют одну и ту же операцию обслуживания, то такие СМО называются *однофазными*. Если каналы обслуживания расположены последовательно и они неоднородны, так как выполняют различные операции обслуживания (т.е. обслуживание состоит из нескольких последовательных этапов или фаз), то СМО называется *многофазной*. Примером работы многофазной СМО является обслуживание автомобилей на станции технического обслуживания (мойка, диагностирование и т.д.). Далее будем рассматривать только однофазные СМО.

Случайные процессы с дискретными состояниями

Случайный процесс, протекающий в СМО, состоит в том, что система в случайные моменты времени переходит из одного состояния в другое: меняется число занятых каналов, число заявок, стоящих в очереди, и т.п. Это означает, что СМО представляет собой физическую систему дискретного типа с конечным (или счетным) множеством состояний[3], а переход системы из одного состояния в другое происходит скачком, в момент, когда осуществляется какое-то событие (приход новой заявки, освобождение канала, уход заявки из очереди и т.п.).

Рассмотрим физическую систему X с не более, чем счетным множеством состояний

$$x_1, x_2, ..., x_n, ...$$

В любой момент времени t система X может быть в одном из этих состояний. Обозначим $p_k(t)$ $(k = 1,2,...,n,...)$ вероятность того, что в момент t система будет находиться в состоянии x_k. Очевидно, для любого t

$$\sum_k p_k(t) = 1.$$

Случайные процессы с дискретными состояниями (не более, чем счетным множеством состояний) бывают двух типов: с дискретным или непрерывным временем. Первые отличаются тем, что переходы из состояния в состояние могут происходить только в строго определенные, разделенные конечными интервалами моменты времени $t_1, t_2, ...$ Случайные процессы с непрерывным временем отличаются тем, что переход системы из состояния в состояние возможен в любой момент времени t.

В качестве примера дискретной системы X, в которой протекает случайный процесс с непрерывным временем, рассмотрим группу из n самолетов, совершающих налет на территорию противника, обороняемую системой ПВО. Ни момент обнаружения группы, ни момент начала работы пусковых установок системы ПВО заранее не известны. Различные состояния системы соответствуют различному числу пораженных самолетов в составе группы:

x_0 – не уничтожено ни одного самолета,

x_1 – уничтожен ровно один самолет,

……

x_n – уничтожены все n самолетов.

Схема возможных состояний системы и возможных переходов из состояния в состояние показана на рисунке 2 (такая схема называется *графом состояний*).

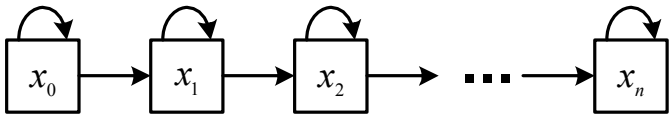

Рисунок 2.

Стрелками показаны возможные переходы системы из состояния в состояние. Закругленная стрелка, направленная из состояния x_k в него же, означает, что система может не только перейти в соседнее состояние x_{k+1}, но и остаться в прежнем. Для данной системы характерны *необратимые переходы* (уничтоженные самолеты не восстанавливаются); в связи с этим из состояния x_n никакие переходы в другие состояния уже невозможны.

Отметим, что граф состояний на рис. 2 показывает только переходы из состояния в соседнее состояние и не показывает «перескоки» через состояние: эти перескоки отброшены

[3] В математике *счетным* называется бесконечное множество, элементы которого можно перенумеровать, т.е. записать в виде последовательности $a_1, a_2, ..., a_n, ...$ Если множество конечное или счетное, то его называют *не более, чем счетным*.

как практически невозможные. Действительно, для того чтобы система «перескочила» через состояние, нужно, чтобы строго одновременно были поражены два или более самолета, а вероятность такого события равна нулю.

Случайные процессы, протекающие в СМО, как правило, представляют собой процессы с непрерывным временем. Это связано со случайностью потока заявок. В противоположность системе с необратимыми переходами, рассмотренной в предыдущем примере, для СМО характерны *обратимые переходы*: занятый канал может освободиться.

В качестве примера рассмотрим одноканальную СМО (например, одну телефонную линию), в которой заявка, заставшая канал занятым, не становится в очередь, а покидает систему (получает «отказ»). Это – дискретная система с непрерывным временем и двумя возможными состояниями:

x_0 – канал свободен,

x_1 – канал занят.

Переходы из состояния в состояние обратимы. Граф состояний показан на рисунке 3.

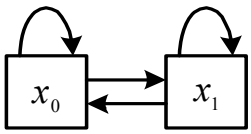

Рисунок 3.

Для того чтобы описать случайный процесс, протекающий в дискретной системе с непрерывным временем, прежде всего нужно проанализировать причины, вызывающие переход системы из состояния в состояние. Для СМО основным фактором, обусловливающим протекающие в ней процессы, является поток заявок. Поэтому математическое описание любой СМО начинается с потока заявок.

Потоки событий

Под *потоком событий* понимается последовательность однородных событий, следующих одно за другим в какие-то случайные моменты времени (например, поток вызовов на телефонной станции, поток покупателей, поток заказных писем, поступающих в почтовое отделение и т.п.).

Поток характеризуется *интенсивностью* λ – частотой появления событий или средним числом событий, поступающих в СМО в единицу времени.

Поток событий называется *регулярным*, если события следуют одно за другим через определенные равные промежутки времени. Например, поток изделий на конвейере сборочного цеха (с постоянной скоростью движения) является регулярным. Такой поток сравнительно редко встречается в реальных системах, но представляет интерес как предельный случай. Типичным для системы массового обслуживания является случайный поток заявок.

В этом пункте мы рассмотрим потоки событий, обладающие некоторыми особенно простыми свойствами. Для этого введем ряд определений.

1. Поток событий называется *стационарным*, если его вероятностные характеристики не зависят от времени. В частности, интенсивность стационарного потока есть величина постоянная: $\lambda(t) = \lambda$. Это отнюдь не значит, что фактическое число событий, появляющееся в единицу времени, постоянно, – нет, поток неизбежно (если только он не регулярный) имеет какие-то случайные сгущения и разрежения. Важно, что для стационарного потока эти сгущения и разрежения не носят закономерного характера: на один участок длины 1 может попасть больше, на другой – меньше событий, но среднее число событий, приходящееся на единицу времени, постоянно и от времени не зависит.

2. Поток событий называется *потоком без последействия*, если для любых двух непересекающихся участков времени τ_1 и τ_2 число событий, попадающих на один из них, не зависит от числа событий попавших на другой. По сути, это означает, что события, образующие поток, появляются в те или иные моменты времени независимо друг от друга, вызванные каждое своими собственными причинами. Например, поток пассажиров, входящих в метро, практически не имеет последействия. А вот поток покупателей, отходящих с покупками от прилавка, уже имеет последействие (хотя бы потому, что интервал времени между отдельными покупателями не может быть меньше, чем минимальное время обслуживания каждого из них).
3. Поток событий называется *ординарным*, если вероятность попадания на малый (элементарный) участок времени Δt двух и более событий пренебрежимо мала по сравнению с вероятностью попадания одного события. Другими словами, поток событий ординарен, если события появляются в нем поодиночке, а не группами. Например, поток поездов, подходящих к станции, ординарен, а поток вагонов – неординарен.

Поток событий называется простейшим (или стационарным пуассоновским), если он одновременно стационарен, ординарен и не имеет последействия. Название «простейший» объясняется тем, что СМО с простейшими потоками имеет наиболее простое математическое описание. Между прочим, самый простой, на первый взгляд, регулярный поток не является «простейшим», так как обладает последействием: моменты появления событий в таком потоке связаны жесткой, функциональной зависимостью. Без специальных усилий по поддержанию его регулярности такой поток обычно не создается.

Простейший поток в качестве предельного возникает в теории случайных процессов столь же естественно, как в теории вероятностей нормальное распределение получается в качестве предельного для суммы случайных величин: *при наложении (суперпозиции) достаточного большого числа n независимых, стационарных и ординарных потоков (сравнимых между собой по интенсивностям λ_i $(i = 1,2,...,n)$) получается поток, близкий к простейшему с интенсивностью λ, равной сумме интенсивностей входящих потоков, т.е.*

$$\lambda = \sum_{i=1}^{n} \lambda_i .$$

Название «пуассоновский» связано с тем, что при соблюдении 1 – 3 число событий, попадающих на любой фиксированный интервал времени, будет распределено по закону Пуассона. Покажем это с помощью элементарных рассуждений.

Рассмотрим на оси времени Ot простейший поток событий как неограниченную последовательность случайных точек (рис. 4).

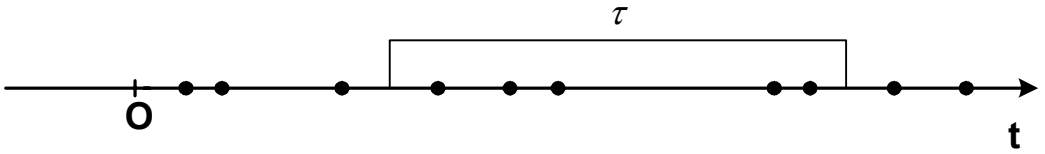

Рисунок 4.

Пусть случайная величина X выражает число событий (точек), попадающих на произвольный промежуток времени τ. Покажем, что *случайная величина X распределена по закону Пуассона.*

□ Разобьем мысленно временной промежуток τ на n равных элементарных отрезков $\Delta t = \tau / n$. Математическое ожидание числа событий, попадающих на элементарный отрезок Δt, очевидно, равно $\lambda \cdot \Delta t$, где λ – интенсивность потока (т.к. на единицу длины попадает в среднем λ точек). Согласно свойству ординарности потока можно пренебречь вероятностью попадания на элементарный (т.е. малый) отрезок двух и более событий. Поэтому мате-

матическое ожидание $\lambda \cdot \Delta t$ числа точек, попадающих на участок Δt, будет приближенно (с точностью до бесконечно малых высшего порядка при $\Delta t \to 0$) равно вероятности попадания на него одной точки (или, что в наших условиях равнозначно, хотя бы одной).

Будем считать элементарный отрезок Δt «занятым», если в нем появилось событие потока, и «свободным», если не появилось. Вероятность того, что отрезок $\Delta t = \tau/n$ окажется «занятым», равна $\lambda \Delta t = \lambda\tau/n$; вероятность того, что он окажется «пустым», равна $1 - \lambda\tau/n$ (чем меньше Δt, тем точнее равенства).

Число занятых элементарных отрезков, т.е. число X событий на всем временном промежутке τ, можно рассматривать как случайную величину, имеющую биномиальный закон распределения (с параметрами n и $p = \lambda\tau/n$), а, следовательно, по формуле Бернулли

$$P(X = m) = C_n^m \left(\frac{\lambda\tau}{n}\right)^m \left(1 - \frac{\lambda\tau}{n}\right)^{n-m}.$$

(Необходимое для возникновения биномиального закона условие независимости испытаний, в данном случае – независимость n элементарных отрезков относительно события «отрезок занят», обеспечивается свойством отсутствия последействия потока).

Известно, что при неограниченном увеличении числа элементарных отрезков Δt, т.е. при $n \to \infty$, $p = \dfrac{\lambda\tau}{n} \to 0$ и постоянном значении произведения $np = n\dfrac{\lambda\tau}{n} = \lambda\tau$ биномиальное распределение стремится к распределению Пуассона с параметром $a = \lambda\tau$:

$$P_m(\tau) = \frac{(\lambda\tau)^m}{m!} e^{-\lambda\tau}. \qquad (1)$$

От этого свойства закона Пуассона – выражать биномиальное распределение при большом числе опытов и малой вероятности события – происходит его название, часто применяемое в учебниках статистики: *закон редких явлений*.

В частности, вероятность того, что за время τ не произойдет ни одного события ($m = 0$), равна

$$P_0(\tau) = e^{-\lambda\tau}. \blacksquare \qquad (2)$$

Пример. На автоматическую телефонную станцию поступает простейший поток вызовов с интенсивностью $\lambda = 1{,}2$ вызовов в минуту. Найти вероятность того, что за две минуты: а) не придет ни одного вызова; б) придет ровно один вызов; в) придет хотя бы один вызов.

Решение. а) Случайная величина X – число вызовов за две минуты – распределена по закону Пуассона с параметром $\lambda\tau = 1{,}2 \cdot 2 = 2{,}4$. Вероятность того, что вызовов не будет ($m = 0$), по формуле (2):

$$P_0(2) \approx e^{-2{,}4} \approx 0{,}091.$$

б) Вероятность одного вызова ($m = 1$) по формуле (1):

$$P_1(2) \approx 2{,}4 \cdot 0{,}091 \approx 0{,}218.$$

в) Вероятность хотя бы одного вызова:

$$P(X \geq 1) = 1 - P(X = 0) = 1 - P_0(2) \approx 1 - 0{,}091 = 0{,}909.$$

Найдем распределение интервала времени T между двумя произвольными соседними событиями простейшего потока.

В соответствии с формулой (2) вероятность того, что на участке времени длиной t не появится ни одного из последующих событий, равна

$$P(T \geq t) = e^{-\lambda t},$$

а вероятность противоположного события, т.е. функция распределения случайной величины T, есть

$$F(t) = P(T < t) = 1 - e^{-\lambda t}. \quad (3)$$

Функция распределения (3) определяет показательный (экспоненциальный) закон распределения. Таким образом, *интервал времени между двумя произвольными соседними событиями простейшего потока имеет показательное распределение*, для которого математическое ожидание равно среднему квадратичному отклонению случайной величины:

$$a = \sigma = \frac{1}{\lambda},$$

и обратно по величине интенсивности потока λ.

Для простейшего потока с интенсивностью λ вероятность попадания на *элементарный (малый)* отрезок времени Δt хотя бы одного события потока равна согласно (3):

$$P_{\Delta t} = P(T < \Delta t) = 1 - e^{-\lambda \Delta t} \approx \lambda \Delta t. \quad (4)$$

Эта приближенная формула, получаемая заменой функции $e^{-\lambda \Delta t}$ лишь двумя первыми членами ее разложения в ряд по степеням Δt, тем точнее, чем меньше Δt.

Понятие марковского случайного процесса

Математический анализ работы СМО существенно упрощается, если процесс этой работы – марковский. Случайный процесс, протекающий в системе S, называется *марковским* (или процессом без последействия), если он обладает следующим свойством: для каждого момента времени t_0 вероятность любого состояния системы в будущем (при $t > t_0$) зависит только от ее состояния в настоящем (при $t = t_0$) и не зависит от того, когда и каким образом система перешла в это состояние, т.е. не зависит от ее поведения в прошлом (при $t < t_0$).

Ранее мы уже упоминали об аналогичном свойстве некоторых потоков событий (отсутствии последействия). Не надо понимать марковское свойство случайного процесса как полную независимость «будущего» от «прошлого»; нет, в общем случае «будущее» зависит от «настоящего», т.е. вероятности $p_i(t)$ при $t > t_0$ зависят от того, в каком состоянии s_i находится система в настоящем (при $t = t_0$); само же это «настоящее» зависит от «прошлого», от того, как вела себя система S при $t < t_0$. Это можно сформулировать следующим образом: для марковского случайного процесса «будущее» зависит от «прошлого» только через «настоящее» (рис. 5). При фиксированном «настоящем» условные вероятности всех состояний системы в «будущем» не зависят от предыстории процесса, т.е. от того, когда и как система S к моменту t_0 пришла в состояние s_i.

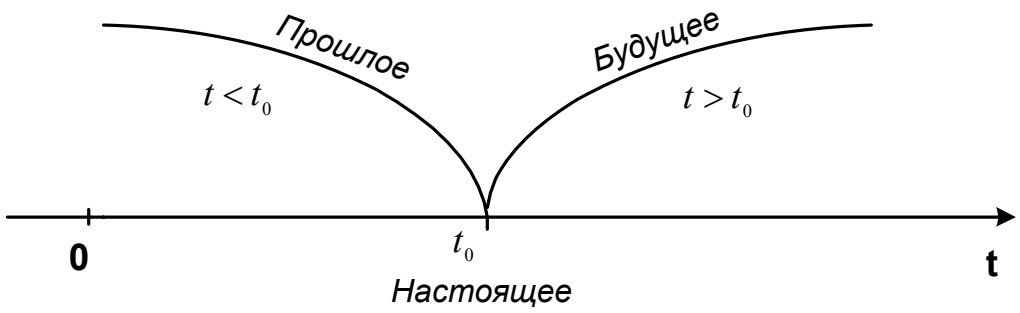

Рисунок 5.

Пример 1 (немарковского случайного процесса). Возьмем ранее рассмотренную систему, представляющую собой группу из n самолетов, совершающих налет на территорию противника, обороняемую системой ПВО. Состояние системы в «будущем» зависит от того, когда и каким образом система пришла в «настоящее» состояние. В данном случае нельзя не учитывать предысторию процесса, а именно, как быстро часть самолетов данной группы была уничтожена системой ПВО.

Пример 2 (немарковского случайного процесса). Рассмотрим процесс игры в шахматы; система S – группа шахматных фигур. Состояние системы характеризуется числом фигур (обеих сторон) и позицией на шахматной доске в момент времени t_0. Будущее состояние системы (в момент $t > t_0$) зависит не только от состояния в «настоящем», но и от того, когда и, главное, каким образом система пришла в это состояние. А именно, если один из противников имеет материальное и/или позиционное преимущество, то важно знать, случайно или закономерно получено это преимущество, как развивалась партия (т.е. изменялись состояния системы) и т.д., поскольку от ответов на эти вопросы зависит информация о квалификации шахматистов, а следовательно, возможность предсказать изменение состояний системы.

На практике марковские процессы в чистом виде обычно не встречаются, но нередко приходится иметь дело с процессами, которые можно приближенно считать марковскими, т.е. такие, для которых влиянием «предыстории» можно пренебречь. Кроме того, существуют приемы, позволяющие сводить немарковские случайные процессы к марковским. Например, можно вводить в состав параметров, характеризующих настоящее состояние системы, те параметры из прошлого, от которых зависит будущее (в этом случае говорят о «марковизации» случайного процесса). Правда, такая процедура нередко приводит к сильному усложнению математического аппарата. Существуют и другие приемы сведения немарковских случайных процессов к марковским.

Уравнения Колмогорова. Предельные вероятности состояний

Если все потоки событий, переводящие систему S из состояния в состояние, – простейшие, то процесс, протекающий в системе, будет марковским[4]. Это и естественно, так как простейший поток не обладает последействием: в нем «будущее» не зависит от «прошлого».

Рассмотрим математическое описание марковского случайного процесса с дискретными состояниями и непрерывным временем на следующем примере.

Пример. Техническое устройство S состоит из двух узлов, каждый из которых в случайный момент времени может выйти из строя (отказать), после чего мгновенно начинается ремонт узла, продолжающийся заранее неизвестное случайное время.

Возможные состояния системы можно перечислить: S_0 – оба узла исправны; S_1 – первый узел ремонтируется, второй исправен; S_2 – второй узел ремонтируется, первый исправен; S_3 – оба узла ремонтируются.

Будем полагать, что все переходы системы из состояния S_i в S_j происходят под воздействием простейших потоков событий с интенсивностями λ_{ij} $(i, j = 0,1,2,3)$; так, переход системы из состояния S_0 в S_1 будет происходить под воздействием потока отказов первого узла, а обратный переход из состояния S_1 в S_0 – под воздействием потока «окончаний ремонтов» первого узла и т.п.

Граф состояний системы с проставленными у стрелок интенсивностями называют *размеченным* (рис. 6).

[4] Простейший характер потоков – достаточное, но не необходимое условие для того, чтобы процесс был марковским.

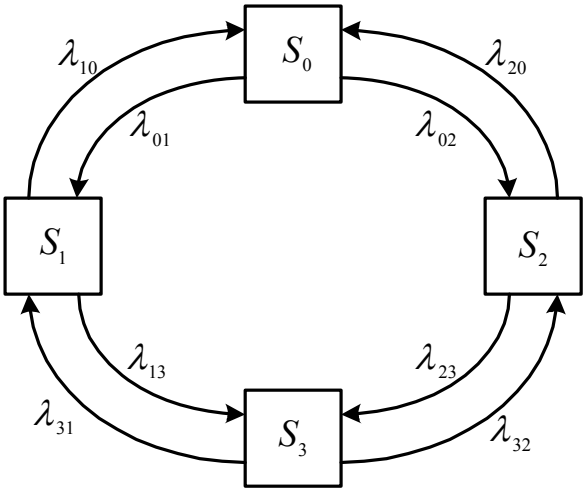

Рисунок 6.

На графе отсутствуют стрелки из S_0 в S_3 и из S_1 в S_2. Это объясняется тем, что выходы узлов из строя предполагаются независимыми друг от друга и, например, вероятностью одновременного выхода из строя двух узлов (переход из S_0 в S_3) или одновременного окончания ремонтов двух узлов (переход из S_3 в S_0) можно пренебречь.

Напомним, что *вероятностью i-го состояния* называется вероятность $p_i(t)$ того, что в момент t система будет находиться в состоянии S_i. При этом для $\forall t$

$$\sum_{i=0}^{3} p_i(t) = 1.$$

Рассмотрим систему в момент t и, задав малый промежуток Δt, найдем вероятность $p_0(t+\Delta t)$ того, что система в момент $t+\Delta t$ будет находиться в состоянии S_0. Это достигается разными способами: либо 1) система в момент t с вероятностью $p_0(t)$ находилась в состоянии S_0, а за время Δt не вышла из него; либо 2) система в момент t с вероятностями $p_1(t)$ (или $p_2(t)$) находилась в состоянии S_1 или S_2 и за время Δt перешла в состояние S_0.

1) Найдем вероятность первого варианта. Вывести систему из состояния S_0 (см. рис.6) можно суммарным простейшим потоком (при наложении двух простейших потоков, как уже отмечалось, получается опять простейший поток) с интенсивностью $\lambda_{01}+\lambda_{02}$, т.е. в соответствии с (4), с вероятностью, приближенно равной $(\lambda_{01}+\lambda_{02})\Delta t$. А вероятность того, что система не выйдет из состояния S_0, равна $1-(\lambda_{01}+\lambda_{02})\Delta t$. Вероятность того, что система будет находиться в состоянии S_0 и не выйдет из него за время Δt (т.е. вероятность первого варианта), равна по теореме умножения вероятностей:

$$p_0(t)\left(1-(\lambda_{01}+\lambda_{02})\Delta t\right).$$

2) Найдем вероятность второго варианта. Под действием потока интенсивностью λ_{10} (или λ_{20}) (см. рис. 6) система перейдет в состояние S_0 с вероятностью, приближенно равной $\lambda_{10}\Delta t$ (или $\lambda_{20}\Delta t$). Вероятность того, что система будет находиться в состоянии S_0, по этому способу равна $p_1(t)\lambda_{10}\Delta t$ (или $p_2(t)\lambda_{20}\Delta t$).

Применяя теорему сложения вероятностей (для попарно несовместных событий), получим:

$$p_0(t+\Delta t) = p_1(t)\lambda_{10}\Delta t + p_2(t)\lambda_{20}\Delta t + p_0(t)\big(1-(\lambda_{01}+\lambda_{02})\Delta t\big),$$

откуда

$$\frac{p_0(t+\Delta t)-p_0(t)}{\Delta t} = p_1(t)\lambda_{10} + p_2(t)\lambda_{20} - (\lambda_{01}+\lambda_{02})p_0(t).$$

Переходя к пределу при $\Delta t \to 0$ (приближенные равенства, связанные с применением формулы (4), перейдут в точные), получим в левой части уравнения производную $p_0'(t)$ (обозначим ее для простоты p_0'):

$$p_0' = \lambda_{10}p_1 + \lambda_{20}p_2 - (\lambda_{01}+\lambda_{02})p_0.$$

Получено дифференциальное уравнение первого порядка. Рассуждая аналогично для других состояний системы S, можно получить *систему дифференциальных уравнений Колмогорова для вероятностей состояний*:

$$\begin{cases} p_0' = \lambda_{10}p_1 + \lambda_{20}p_2 - (\lambda_{01}+\lambda_{02})p_0, \\ p_1' = \lambda_{01}p_0 + \lambda_{31}p_3 - (\lambda_{10}+\lambda_{13})p_1, \\ p_2' = \lambda_{02}p_0 + \lambda_{32}p_3 - (\lambda_{20}+\lambda_{23})p_2, \\ p_3' = \lambda_{13}p_1 + \lambda_{23}p_2 - (\lambda_{31}+\lambda_{32})p_3. \end{cases} \qquad (5)$$

Сформулируем правило составления уравнений Колмогорова. *В левой части каждого из них стоит производная вероятности i-го состояния. В правой части – сумма произведений вероятностей всех состояний, из которых идут стрелки в данное состояние, умноженных на интенсивности соответствующих потоков событий, минус суммарная интенсивность всех потоков, выводящих систему из данного состояния, умноженная на вероятность данного (i-го) состояния.*

В системе (5) независимых уравнений на единицу меньше общего числа уравнений. Поэтому для решения системы необходимо добавить уравнение

$$\sum_{i=0}^{3} p_i(t) = 1.$$

Особенность решения дифференциальных уравнений вообще состоит в том, что требуется задавать так называемые начальные условия, в данном случае – вероятности состояний системы в начальный момент $t = 0$. Так, например, систему уравнений (5) естественно решать при условии, что в начальный момент оба узла исправны и система находилась в состоянии S_0, т.е. при начальных условиях $p_0(0) = 1$, $p_1(0) = p_2(0) = p_3(0) = 0$.

Как решать подобные уравнения? Вообще говоря, системы линейных дифференциальных уравнений с постоянными коэффициентами можно решать аналитически, но это удобно, когда число уравнений не превосходит двух (иногда – трех). Если уравнений больше, обычно их решают численно – вручную или на ЭВМ.

Уравнения Колмогорова дают возможность найти все вероятности состояний как *функции времени*. Особый интерес представляют вероятности системы $p_i(t)$ в *предельном стационарном режиме*, т.е. при $t \to \infty$, которые называются *предельными* (*финальными*) вероятностями состояний.

В теории случайных процессов доказывается, что *если число состояний системы конечно и из каждого из них можно (за конечное число шагов) перейти в любое другое состояние, то предельные вероятности существуют.*

Предельная вероятность состояния S_i имеет четкий смысл: она показывает *среднее относительное время пребывания системы в этом состоянии*. Например, если предельная

вероятность состояния S_0, т.е. $p_0 = 0{,}5$, то это означает, что в среднем половину времени система находится в состоянии S_0.

Как же вычислить предельные вероятности? Очень просто. Так как предельные вероятности постоянны, то, заменяя в уравнениях Колмогорова их производные нулевыми значениями, получим систему линейных алгебраических уравнений, описывающих стационарный режим. Для системы S с графом состояний, изображенном на рис. 6, такая система уравнений имеет вид:

$$\begin{cases} (\lambda_{01} + \lambda_{02}) p_0 = \lambda_{10} p_1 + \lambda_{20} p_2, \\ (\lambda_{10} + \lambda_{13}) p_1 = \lambda_{01} p_0 + \lambda_{31} p_3, \\ (\lambda_{20} + \lambda_{23}) p_2 = \lambda_{02} p_0 + \lambda_{32} p_3, \\ (\lambda_{31} + \lambda_{32}) p_3 = \lambda_{13} p_1 + \lambda_{23} p_2. \end{cases} \qquad (6)$$

Систему можно составить непосредственно по размеченному графу состояний, если руководствоваться правилом, согласно которому *слева в уравнениях стоит предельная вероятность данного состояния p_i, умноженная на суммарную интенсивность всех потоков, ведущих из данного состояния, а справа – сумма произведений интенсивностей всех потоков, входящих в i-е состояние, на вероятности тех состояний, из которых эти потоки исходят.*

Эту систему из четырех уравнений с четырьмя неизвестными p_0, p_1, p_2, p_3, казалось бы, вполне можно решить. Но вот беда: уравнения (6) однородны (не имеют свободного члена) и, значит, определяют неизвестные только с точностью до произвольного множителя. К счастью, мы можем воспользоваться нормировочным условием

$$\sum_{i=0}^{3} p_i = 1$$

и с его помощью решить систему. При этом одно (любое) из уравнений можно отбросить (оно вытекает как следствие из остальных).

Пример 1. Найти предельные вероятности для системы S (см. рисунок 6 и соответствующий пример) при $\lambda_{01} = 1,\ \lambda_{02} = 2,\ \lambda_{10} = 2,\ \lambda_{13} = 2,\ \lambda_{20} = 3,\ \lambda_{23} = 1,\ \lambda_{31} = 3,\ \lambda_{32} = 2.$

Решение. Система алгебраических уравнений, описывающих стационарный режим для данной СМО, имеет вид (6) или

$$\begin{cases} 3p_0 = 2p_1 + 3p_2, \\ 4p_1 = p_0 + 3p_3, \\ 4p_2 = 2p_0 + 2p_3, \\ p_0 + p_1 + p_2 + p_3 = 1. \end{cases} \qquad (7)$$

(Здесь вместо одного «лишнего» уравнения системы (6) записали нормировочное условие).

Решив систему (7), получим $p_0 = 0{,}4$, $p_1 = 0{,}2$, $p_2 = 0{,}27$, $p_3 = 0{,}13$, т.е. в предельном стационарном режиме система S в среднем 40% времени будет находиться в состоянии S_0 (оба узла исправны), 20% – в состоянии S_1 (первый узел ремонтируется, второй работает), 27% – в состоянии S_2 (второй узел ремонтируется, первый работает) и 13% времени – в состоянии S_3 (оба узла ремонтируются).

Пример 2. Найти средний чистый доход от эксплуатации в стационарном режиме системы S в условиях предыдущего примера. Если известно, что в единицу времени исправная работа первого и второго узлов приносит доход соответственно в 10 и 6 ден. ед., а их ремонт требует затрат соответственно в 4 и 2 ден. ед. Оценить экономическую эффективность имеющейся возможности уменьшения вдвое среднего времени ремонта каждого из двух уз-

лов, если при этом придется вдвое увеличить затраты на ремонт каждого узла (в единицу времени).

Решение. Из предыдущего примера следует, что в среднем первый узел исправно работает долю времени, равную $p_0 + p_2 = 0{,}4 + 0{,}27 = 0{,}67$, а второй узел – $p_0 + p_1 = 0{,}4 + 0{,}2 = 0{,}6$. В то же время первый узел находится в ремонте в среднем долю времени, равную $p_1 + p_3 = 0{,}2 + 0{,}13 = 0{,}33$, а второй узел – $p_2 + p_3 = 0{,}27 + 0{,}13 = 0{,}4$. Поэтому средний чистый доход в единицу времени от эксплуатации системы, т.е. разность между доходами и затратами, равен

$$Д = 0{,}67 \cdot 10 + 0{,}6 \cdot 6 - 0{,}33 \cdot 4 - 0{,}4 \cdot 2 = 8{,}18 \text{ ден. ед.}$$

Уменьшение вдвое среднего времени ремонта каждого из узлов будет означать увеличение вдвое интенсивностей потока «окончаний ремонтов» каждого узла. Это следует из равенства $a = \frac{1}{\lambda}$ для показательного распределения (потоков) с параметром λ, о котором упоминалось ранее. Напомним, что a – это математическое ожидание случайной величины T – промежутка времени между произвольными двумя соседними событиями в простейшем потоке. Таким образом, теперь интенсивности потоков событий будут равны: $\lambda_{10} = 4$, $\lambda_{20} = 6$, $\lambda_{31} = 6$, $\lambda_{32} = 4$ (остальные остались прежними). При этом система линейных алгебраических уравнений (6), описывающая стационарный режим системы S, вместе с нормировочным условием примет вид:

$$\begin{cases} 3p_0 = 4p_1 + 6p_2, \\ 6p_1 = p_0 + 6p_3, \\ 7p_2 = 2p_0 + 4p_3, \\ p_0 + p_1 + p_2 + p_3 = 1. \end{cases}$$

Решив систему, получим $p_0 = 0{,}6$, $p_1 = 0{,}15$, $p_2 = 0{,}2$, $p_3 = 0{,}05$.

Учитывая, что $p_0 + p_2 = 0{,}6 + 0{,}2 = 0{,}8$, $p_0 + p_1 = 0{,}6 + 0{,}15 = 0{,}75$, $p_1 + p_3 = 0{,}15 + 0{,}05 = 0{,}2$, $p_2 + p_3 = 0{,}2 + 0{,}05 = 0{,}25$, а затраты на ремонт первого и второго узлов составляют теперь соответственно 8 и 4 ден. ед., вычислим средний чистый доход в единицу времени:

$$Д_1 = 0{,}8 \cdot 10 + 0{,}75 \cdot 6 - 0{,}2 \cdot 8 - 0{,}25 \cdot 4 = 9{,}9 \text{ ден. ед.}$$

Так как $Д_1$ больше $Д$ примерно на 21% ($\frac{9{,}9 - 8{,}18}{8{,}18} \cdot 100\% \approx 21\%$), то экономическая целесообразность ускорения ремонтов узлов очевидна.

Процессы гибели и размножения

В теории массового обслуживания широко распространен специальный класс случайных процессов – так называемые *процессы гибели и размножения*. Название это связано с рядом биологических задач, где этот процесс служит математической моделью изменения численности биологических популяций.

Граф состояний процесса гибели и размножения имеет вид, показанный на рисунке 7.

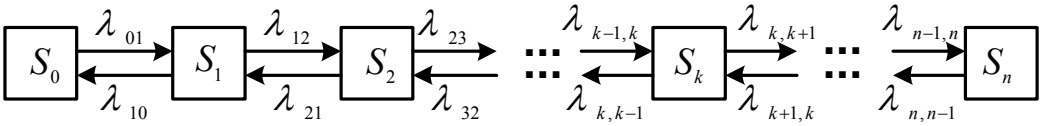

Рисунок 7.

Рассмотрим упорядоченное множество состояний системы $S_0, S_1, ..., S_n$. Переходы могут осуществляться из любого состояния только в состояния с соседними номерами, т.е. из состояния S_k возможны переходы либо в состояние S_{k-1}, либо в состояние S_{k+1}[5].

Предположим, что все потоки событий, переводящие систему по стрелкам графа, простейшие с соответствующими интенсивностями $\lambda_{k,k+1}$ или $\lambda_{k+1,k}$.

По графу, представленному на рис. 7, составим и решим алгебраические уравнения для предельных вероятностей состояний (их существование вытекает из возможности перехода из каждого состояния в каждое другое и конечности числа состояний).

В соответствии с ранее сформулированным правилом составления таких уравнений получим:

для состояния S_0

$$\lambda_{01} p_0 = \lambda_{10} p_1, \tag{8}$$

для состояния S_1 —

$$(\lambda_{12} + \lambda_{10}) p_1 = \lambda_{01} p_0 + \lambda_{21} p_2,$$

которое с учетом (8) приводится к виду

$$\lambda_{12} p_1 = \lambda_{21} p_2.$$

Аналогично, записывая уравнения для предельных вероятностей других состояний, можно получить следующую систему уравнений:

$$\begin{cases} \lambda_{01} p_0 = \lambda_{10} p_1, \\ \lambda_{12} p_1 = \lambda_{21} p_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \lambda_{k-1,k} p_{k-1} = \lambda_{k,k-1} p_k, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \lambda_{n-1,n} p_{n-1} = \lambda_{n,n-1} p_n, \end{cases} \tag{9}$$

к которой добавляется нормировочное условие

$$p_0 + p_1 + \ldots + p_n = 1. \tag{10}$$

Решим эту систему уравнений. Из первого уравнения (9) выразим p_1 через p_0:

$$p_1 = \frac{\lambda_{01}}{\lambda_{10}} p_0. \tag{11}$$

Из второго, с учетом (11), получим:

$$p_2 = \frac{\lambda_{12}}{\lambda_{21}} p_1 = \frac{\lambda_{12}\lambda_{01}}{\lambda_{21}\lambda_{10}} p_0; \tag{12}$$

из третьего, с учетом (12),

$$p_3 = \frac{\lambda_{23}\lambda_{12}\lambda_{01}}{\lambda_{32}\lambda_{21}\lambda_{10}} p_0,$$

и вообще, для любого k (от 1 до n):

---

[5] При анализе численности популяций считают, что состояние S_k соответствует численности популяции, равной k, и переход системы из состояния S_k в состояние S_{k+1} происходит при рождении одного члена популяции, а переход в состояние S_{k-1} – при гибели одного члена популяции.

$$p_k = \frac{\lambda_{k-1,k}\ldots\lambda_{12}\lambda_{01}}{\lambda_{k,k-1}\ldots\lambda_{21}\lambda_{10}}\,p_0\,. \qquad (13)$$

Обратим внимание на формулы для вероятностей $p_1, p_2, ..., p_n$: числители представляют собой произведения всех интенсивностей, стоящих у стрелок, ведущих слева направо (от начала и до данного состояния S_k); знаменатели – произведения всех интенсивностей, стоящих у стрелок, ведущих справа налево (из состояния S_k и до начала).

Таким образом, все вероятности состояний $p_0, p_1, p_2, ..., p_n$ выражены через одну из них (p_0). Подставим эти выражения в нормировочное условие (10). Получим, вынося за скобки p_0:

$$p_0\left(1+\frac{\lambda_{01}}{\lambda_{10}}+\frac{\lambda_{12}\lambda_{01}}{\lambda_{21}\lambda_{10}}+\ldots+\frac{\lambda_{n-1,n}\ldots\lambda_{12}\lambda_{01}}{\lambda_{n,n-1}\ldots\lambda_{21}\lambda_{10}}\right)=1,$$

откуда можно получить выражение для p_0:

$$p_0=\left(1+\frac{\lambda_{01}}{\lambda_{10}}+\frac{\lambda_{12}\lambda_{01}}{\lambda_{21}\lambda_{10}}+\ldots+\frac{\lambda_{n-1,n}\ldots\lambda_{12}\lambda_{01}}{\lambda_{n,n-1}\ldots\lambda_{21}\lambda_{10}}\right)^{-1}. \qquad (14)$$

Заметим, что слагаемые в правой части (14) представляют собой не что иное, как последовательные коэффициенты при p_0 в формулах для вероятностей $p_1, p_2, ..., p_n$.

СМО с отказами

В качестве показателей эффективности СМО с отказами будем рассматривать:

A[6] – *абсолютную пропускную способность* СМО, т.е. среднее число заявок, обслуживаемых в единицу времени;

Q[7] – *относительную пропускную способность*, т.е. среднюю долю пришедших заявок, обслуживаемых системой (или вероятность того, что пришедшая заявка будет обслужена);

$P_{отк}$ – *вероятность отказа* – вероятность того, что заявка покинет СМО необслуженной;

$\overline{k}$ – *среднее число занятых каналов* (для многоканальной системы).

1.Одноканальная система с отказами. Рассмотрим следующую задачу. Имеется один канал, на который поступает поток заявок с интенсивностью λ. Поток обслуживаний имеет интенсивность μ. Найти предельные вероятности состояний системы и показатели ее эффективности.

Здесь и в дальнейшем будем предполагать, что все потоки событий, переводящие СМО из состояния в состояние, – простейшие. К ним относится и поток обслуживаний – поток заявок, обслуживаемых одним непрерывно занятым каналом. Поскольку среднее время между двумя произвольными соседними событиями простейшего потока обратно по величине интенсивности потока, а для потока обслуживаний это время есть время обслуживания (одной заявки), то среднее время обслуживания $\overline{T}_{об}=1/\mu$.

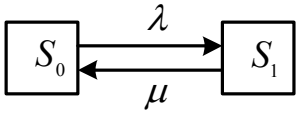

Рисунок 8.

[6] A – первая буква английского *absolute* – абсолютный.

[7] Q – первая буква английского *quota* – доля, часть, квота.

Система S (СМО) имеет два состояния: S_0 – канал свободен, S_1 – канал занят. Размеченный граф состояний представлен на рисунке 8.

В предельном стационарном режиме система алгебраических уравнений (6) для вероятностей состояний имеет вид (см. правило составления таких уравнений):

$$\begin{cases} \lambda\, p_0 = \mu\, p_1, \\ \mu\, p_1 = \lambda\, p_0, \end{cases}$$

т.е. система вырождается в одно уравнение. Учитывая нормировочное условие $p_0 + p_1 = 1$, найдем из полученной предельные вероятности состояний:

$$p_0 = \frac{\mu}{\lambda + \mu}, \quad p_1 = \frac{\lambda}{\lambda + \mu}.$$

Предельные вероятности состояний p_0 и p_1 можно выразить через средние времена простоя канала $\overline{T}_{пр}$ и обслуживания одной заявки $\overline{T}_{об}$. Для этого в формулы для вероятностей следует подставить $\mu = 1/\overline{T}_{об}$ и $\lambda = 1/\overline{T}_{пр}$. В результате получим

$$p_0 = \frac{\overline{T}_{пр}}{\overline{T}_{об} + \overline{T}_{пр}}, \quad p_1 = \frac{\overline{T}_{об}}{\overline{T}_{об} + \overline{T}_{пр}}.$$

Предельные вероятности выражают среднее относительное время пребывания системы в состоянии S_0 (когда канал свободен) и S_1 (когда канал занят), т.е. определяют соответственно относительную пропускную способность Q системы и вероятность отказа $P_{отк}$:

$$Q = \frac{\mu}{\lambda + \mu},$$

$$P_{отк} = \frac{\lambda}{\lambda + \mu}.$$

Пояснение. Почему $Q = p_0$? В самом деле, p_0 есть вероятность того, что заявка будет принята к обслуживанию (система находится в состоянии S_0, т.е. канал свободен). Всего в единицу времени приходит в среднем λ заявок и из них обслуживается λp_0 заявок. Тогда доля обслуживаемых заявок по отношению ко всему потоку заявок определяется величиной

$$Q = \frac{\lambda\, p_0}{\lambda} = p_0.$$

Абсолютную пропускную способность (или, иначе, среднее число заявок, поступающих в СМО в единицу времени) найдем, умножив относительную пропускную способность Q на интенсивность потока заявок:

$$A = \lambda Q = \frac{\lambda\mu}{\lambda + \mu}.$$

Пример. В фирму поступает простейший поток заявок на телефонные переговоры с интенсивностью $\lambda = 90$ вызовов в час, а средняя продолжительность разговора по телефону $\overline{T}_{об} = 2$ мин. Определить показатели эффективности работы СМО (телефонной связи) при наличии одного телефонного номера.

Решение. Интенсивность потока обслуживаний $\mu = 1/\overline{T}_{об} = 1/2 = 0{,}5\ (1/мин) = 30\ (1/ч)$. Относительная пропускная способность СМО $Q = 30/(90+30) = 0{,}25$, т.е. в среднем только 25% поступающих заявок осуществят переговоры по телефону. Соответст-

венно вероятность отказа составит $P_{отк} = 1 - 0{,}25 = 0{,}75$. Абсолютная пропускная способность СМО $A = 90 \cdot 0{,}25 = 22{,}5$, т.е. в среднем в час будут обслужены 22,5 заявки на переговоры. Очевидно, что при наличии только одного телефонного номера СМО будет плохо справляться с потоком заявок.

2.Многоканальная система с отказами (задача Эрланга). Здесь мы рассмотрим одну из первых по времени, «классических» задач теории массового обслуживания; эта задача возникла из практических нужд телефонии и была решена в 1909 г. датским инженером-математиком А.К. Эрлангом. Задача ставится так: имеется n каналов (линий связи), на которые поступает поток заявок с интенсивностью λ. Поток обслуживаний каждого канала имеет интенсивность μ. Найти предельные вероятности состояний системы и показатели ее эффективности.

Система S (СМО) имеет следующие состояния (нумеруем их по числу заявок, находящихся в системе): S_0, S_1,…, S_n, где S_k – состояние системы, когда в ней находится k заявок, т.е. занято k каналов.

Граф состояний СМО соответствует процессу гибели и размножения (рис. 9):

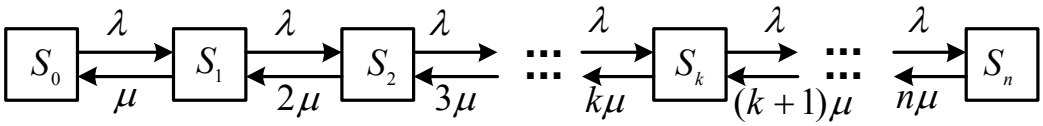

Рисунок 9.

Поток заявок последовательно переводит систему из любого левого состояния в соседнее правое с одной и той же интенсивностью λ. Интенсивность же потока обслуживаний, переводящих систему из любого правого состояния в соседнее левое, постоянно меняется в зависимости от состояния. Действительно, если СМО находится в состоянии S_2 (два канала заняты), то она может перейти в состояние S_1 (один канал занят), когда закончит обслуживание либо первый, либо второй канал, т.е. суммарная интенсивность их потоков обслуживаний будет 2μ. Аналогично суммарный поток обслуживаний, переводящий СМО из состояния S_3 (три канала заняты) в S_2, будет иметь интенсивность 3μ, т.е. может освободиться любой из трех каналов, и т.д.

В формуле (14) для схемы гибели и размножения получим для предельной вероятности состояния

$$p_0 = \left(1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2!\mu^2} + \ldots + \frac{\lambda^k}{k!\mu^k} + \ldots + \frac{\lambda^n}{n!\mu^n}\right)^{-1}, \qquad (15)$$

где члены разложения $\frac{\lambda}{\mu}$, $\frac{\lambda^2}{2!\mu^2}$,…, $\frac{\lambda^n}{n!\mu^n}$ – коэффициенты при p_0 в выражениях для предельных вероятностей $p_1, p_2, \ldots, p_n$.

Заметим, что в формулу (15) интенсивности λ и μ входят не по отдельности, а только в виде отношения λ/μ. Обозначим

$$\lambda/\mu = \rho$$

и будем называть величину ρ *приведенной интенсивностью потока заявок* или *интенсивностью нагрузки канала*. Она выражает среднее число заявок, приходящих за среднее время обслуживания одной заявки. Пользуясь этим обозначением, перепишем формулу (15) в виде:

$$p_0 = \left(1 + \rho + \frac{\rho^2}{2!} + ... + \frac{\rho^n}{n!}\right)^{-1}. \qquad (16)$$

При этом

$$p_1 = \rho\, p_0,\ p_2 = \frac{\rho^2}{2!} p_0, \dots,\ p_n = \frac{\rho^n}{n!} p_0. \qquad (17)$$

Формулы (16) и (17) для предельных вероятностей получили названия *формул Эрланга* в честь основателя теории массового обслуживания.

Вероятность отказа СМО есть предельная вероятность того, что все n каналов системы будут заняты, т.е.

$$P_{отк} = p_n = \frac{\rho^n}{n!} p_0.$$

Отсюда находим относительную пропускную способность – вероятность того, что заявка будет обслужена:

$$Q = 1 - P_{отк\,n} = 1 - \frac{\rho^n}{n!} p_0.$$

Абсолютную пропускную способность получим, умножая интенсивность потока заявок λ на Q:

$$A = \lambda Q = \lambda\left(1 - \frac{\rho^n}{n!} p_0\right). \qquad (18)$$

Осталось только найти среднее число занятых каналов $\bar{k}$. Эту величину можно было бы найти «впрямую», как математическое ожидание дискретной случайной величины с возможными значениями $0, 1, ..., n$ и вероятностями этих значений p_0, p_1, ..., p_n:

$$\bar{k} = 0 \cdot p_0 + 1 \cdot p_1 + 2 \cdot p_2 + ... + n \cdot p_n = \sum_{k=0}^{n} k p_k.$$

Подставляя сюда выражения (17) для p_k и выполняя соответствующие преобразования, мы, в конце концов, получили бы формулу для $\bar{k}$. Однако среднее число занятых каналов можно найти проще, если учесть, что абсолютная пропускная способность A системы есть не что иное, как интенсивность потока *обслуженных* системой заявок (в единицу времени). Так как каждый занятый канал обслуживает в среднем μ заявок (в единицу времени), то среднее число занятых каналов

$$\bar{k} = \frac{A}{\mu}$$

или, учитывая (18):

$$\bar{k} = \rho\left(1 - \frac{\rho^n}{n!} p_0\right).$$

Пример. В условиях предыдущего примера определить оптимальное число телефонных номеров в фирме, если условием оптимальности считать удовлетворение из каждых 100 заявок на переговоры в среднем не менее 90 заявок.

Решение. Интенсивность нагрузки канала $\rho = 90/30 = 3$, т.е. за время среднего (по продолжительности) телефонного разговора $\overline{T}_{об} = 2\ мин$ поступает в среднем 3 заявки на переговоры.

Будем постепенно увеличивать число каналов (телефонных номеров) $n = 2, 3, 4,...$ и определим для получаемой n-канальной СМО характеристики обслуживания. Значения характеристик СМО сведем в таблицу.

Показатели эффективности	Обозначение	Число каналов (телефонных номеров)					
		1	2	3	4	5	6
Относительная пропускная способность	***Q***	0,25	0,47	0,65	0,79	0,90	0,95
Абсолютная пропускная способность	***A***	22,5	42,3	58,8	71,5	80,1	85,3

По условию оптимальности $Q \geq 0{,}9$, следовательно, в фирме необходимо установить 5 телефонных номеров (в этом случае $Q = 0{,}90$). При этом в час будут обслуживаться в среднем 80 заявок ($A = 80{,}1$), а среднее число занятых телефонных номеров (каналов) $\bar{k} = A/\mu = 80{,}1/30 \approx 2{,}67$.

Тут уже проглядывает некоторый намек на *оптимизацию*. В самом деле, содержание каждого канала в единицу времени обходится в какую-то сумму. Вместе с тем, каждая обслуженная заявка приносит какой-то доход (если речь идет о СМО, для которых этот доход можно оценить). Умножая этот доход на среднее число заявок A, обслуживаемых в единицу времени, мы получим средний доход от СМО в единицу времени. Естественно, при увеличении числа каналов этот доход растет, но растут и расходы, связанные с содержанием каналов. Что перевесит – увеличение доходов или расходов? Это зависит от условий операции, т.е. от «платы за обслуживание заявки» и от стоимости содержания канала. Зная эти величины, можно найти оптимальное число каналов, наиболее экономически эффективное.

СМО с ожиданием (с очередью)

1. Одноканальная СМО с ожиданием и ограничением на длину очереди. На практике довольно часто встречаются одноканальные СМО с очередью (врач, обслуживающий пациентов; кассир, выдающий зарплату; телефон-автомат на улице и т.д.). В теории массового обслуживания одноканальные СМО с очередью также занимают особое место: именно к таким СМО относится большинство полученных до сих пор аналитических формул для немарковских систем.

Рассмотрим одноканальную СМО, на вход которой поступает простейший поток заявок с интенсивностью λ. Предположим, что поток обслуживаний также простейший с интенсивностью μ. Это означает, что непрерывно занятый канал обслуживает в среднем μ заявок в единицу времени. Заявка, поступившая в СМО в момент, когда канал занят, в отличие от СМО с отказами, не покидает систему, а становится в очередь и ожидает обслуживания.

Далее предполагаем, что в данной системе имеется ограничение на длину очереди, под которой понимается максимальное число мест в очереди, а именно, предполагаем, что в очереди могут находиться максимум $m \geq 1$ заявок. Поэтому заявка, пришедшая на вход СМО, в момент, когда в очереди уже стоят m заявок, получает отказ и покидает систему необслуженной.

Таким образом, рассматриваемая СМО относится к системам *смешанного типа с ограничением на длину очереди*.

Пронумеруем состояния СМО по числу заявок, находящихся в системе, т.е. под обслуживанием и в очереди:

S_0 – канал свободен (следовательно, очереди нет);

S_1 – канал занят и очереди нет, т.е. в СМО находится (под обслуживанием) одна заявка;

S_2 – канал занят и в очереди стоит одна заявка;

……..

S_{m+1} – канал занят и в очереди m заявок.

Граф состояний данной СМО представлен на рис. 10 и совпадает с графом , описывающим процесс гибели и размножения, с тем отличием, что при наличии только одного канала обслуживания все интенсивности потоков обслуживаний равны μ.

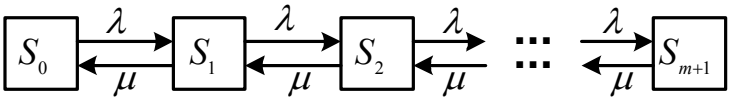

Рисунок 10.

Для описания предельного режима работы СМО можно воспользоваться изложенными ранее правилами и формулами. Запишем сразу выражения, определяющие предельные вероятности состояний:

$$\begin{cases} p_k = \rho^k \cdot p_0, \ k = 1,2,...,m+1, \\ p_0 = \left(1 + \rho + \rho^2 + ... + \rho^{m+1}\right)^{-1}, \end{cases}$$

где $\rho = \lambda / \mu$ – интенсивность нагрузки канала.

Если $\lambda = \mu$, то получаем $p_0 = p_1 = ... = p_{m+1} = 1/(m+2)$.

Пусть теперь $\lambda \neq \mu$ ($\rho \neq 1$). Выражение для p_0 можно в данном случае записать проще, пользуясь тем, что в знаменателе стоит сумма $m+2$ членов геометрической прогрессии со знаменателем ρ:

$$p_0 = \frac{1-\rho}{1-\rho^{m+2}} \ .$$

Заметим, что при $m = 0$ мы переходим к уже рассмотренной одноканальной СМО с отказами. В этом случае $p_0 = (1-\rho)/(1-\rho^2) = \mu/(\lambda+\mu)$ (как и было получено ранее).

Определим основные характеристики одноканальной СМО с ожиданием: относительную и абсолютную пропускную способность, вероятность отказа, а также среднюю длину очереди и среднее время ожидания заявки в очереди.

Поступившая на вход СМО заявка получает отказ тогда и только тогда, когда канал занят и в очереди ожидают m заявок, т.е. когда система находится в состоянии S_{m+1}. Поэтому вероятность отказа определяется вероятностью появления состояния S_{m+1}:

$$P_{отк} = p_{m+1} = \begin{cases} \dfrac{\rho^{m+1}(1-\rho)}{1-\rho^{m+2}}, & если \quad \rho \neq 1; \\ \dfrac{1}{m+2}, & если \quad \rho = 1. \end{cases}$$

Относительная пропускная способность, или доля обслуживаемых заявок, поступающих в единицу времени, определяется выражением:

$$Q = 1 - P_{отк} = \begin{cases} \dfrac{1-\rho^{m+1}}{1-\rho^{m+2}}, & если \quad \rho \neq 1; \\ \dfrac{m+1}{m+2}, & если \quad \rho = 1. \end{cases}$$

Заметим, что относительная пропускная способность Q совпадает со средней долей принятых (т.е. не получивших отказ) в систему заявок среди всех поступивших, поскольку заявка попавшая в очередь непременно будет обслужена.

Абсолютная пропускная способность системы

$$A = \lambda Q.$$

Среднее число заявок $L_{оч}$, стоящих в очереди на обслуживание определяется как математическое ожидание дискретной случайной величины k – числа заявок, стоящих в очереди:

$$L_{оч} = M(k).$$

Случайная величина k принимает значения 0, 1, 2, … , m, вероятности которых определяются вероятностями состояний системы p_k. Таким образом, закон распределения дискретной случайной величины k имеет следующий вид:

k	0	1	2	…	m
p	$p_0 + p_1$	p_2	p_3	…	p_{m+1}

Поэтому по определению математического ожидания дискретной случайной величины (с учетом формул для вероятностей состояний) получаем:

$$M(k) = 0 \cdot (p_0 + p_1) + 1 \cdot p_2 + 2 \cdot p_3 + ... + m \cdot p_{m+1} =$$

$$= \sum_{j=1}^{m} j p_{j+1} = \sum_{j=1}^{m} j \rho^{j+1} p_0 = \rho^2 p_0 \sum_{j=1}^{m} j \rho^{j-1}. \qquad (19)$$

Предположим, что $\rho \neq 1$. Очевидно имеем:

$$\sum_{j=1}^{m} j \rho^{j-1} = \sum_{j=1}^{m} \frac{d}{d\rho} \rho^j = \frac{d}{d\rho} \sum_{j=1}^{m} \rho^j.$$

Но сумма $\sum_{j=1}^{m} \rho^j$ представляет собой сумму первых m членов геометрической прогрессии $\rho, \rho^2, \rho^3, ..., \rho^m$: $\sum_{j=1}^{m} \rho^j = \frac{\rho(1-\rho^m)}{1-\rho} = \frac{\rho - \rho^{m+1}}{1-\rho}$, $\rho \neq 1$. Тогда

$$\sum_{j=1}^{m} j \rho^{j-1} = \frac{d}{d\rho} \sum_{j=1}^{m} \rho^j = \frac{d}{d\rho} \left(\frac{\rho - \rho^{m+1}}{1-\rho} \right) = \frac{1 - \rho^m (m+1-m\rho)}{(1-\rho)^2} \quad (\rho \neq 1). \qquad (20)$$

Подставив выражение (20) в (19), найдем:

$$M(k) = \rho^2 p_0 \frac{1 - \rho^m (m+1-m\rho)}{(1-\rho)^2},$$

или, используя равенство $p_0 = \frac{1-\rho}{1-\rho^{m+2}}$ (полученное при $\rho \neq 1$), имеем

$$M(k) = \frac{\rho^2 \left(1 - \rho^m (m+1-m\rho)\right)}{(1-\rho)(1-\rho^{m+2})}.$$

Если же $\rho = 1$, то из равенства (19)

$$M(k) = p_0 \sum_{j=1}^{m} j,$$

а учитывая, что в этом случае $p_0 = 1/(m+2)$ и $\sum_{j=1}^{m} j = \frac{m(m+1)}{2}$ (сумма m членов арифметической прогрессии), окончательно получаем

$$M(k) = \frac{m(m+1)}{2(m+2)} \qquad (\rho = 1).$$

Итак, среднее число заявок в очереди

$$L_{оч} = \begin{cases} \dfrac{\rho^2\left(1-\rho^m(m+1-m\rho)\right)}{(1-\rho)(1-\rho^{m+2})}, & \text{если} \quad \rho \neq 1; \\ \dfrac{m(m+1)}{2(m+2)}, & \text{если} \quad \rho = 1. \end{cases} \tag{21}$$

Важной характеристикой СМО с ожиданием является среднее время ожидания заявки в очереди $\overline{T}_{оч}$. Пусть $T_{оч}$ – непрерывная случайная величина, представляющая собой время ожидания заявки в очереди. Среднее время ожидания заявки в очереди вычислим как математическое ожидание этой случайной величины:

$$\overline{T}_{оч} = M(T_{оч}).$$

Для вычисления математического ожидания воспользуемся формулой полного математического ожидания: если об условиях опыта можно сделать n (попарно) несовместных гипотез $H_1, H_2, ..., H_n$, то полное математическое ожидание случайной величины X может быть вычислено по формуле

$$M(X) = \sum_{k=1}^{n} P(H_k) M(X \mid H_k),$$

где $M(X \mid H_k)$ – условное математическое ожидание величины X при гипотезе H_k [Вентцель, Овчаров «Прикладные задачи теории вероятностей». – М.: Радио и связь, 1983, с.77].

Рассмотрим $m+2$ несовместных гипотез H_k, $k = 0,1,..., m+1$, состоящих в том, что СМО находится соответственно в состояниях S_k, $k = 0,1,..., m+1$. Вероятности этих гипотез $p(H_k) = p_k$, $k = 0,1,..., m+1$.

Если заявка поступает в СМО при гипотезе H_0, т.е. когда СМО находится в состоянии S_0, в котором канал свободен, то заявке не придется стоять в очереди и, следовательно, условное математическое ожидание $M(T_{оч} \mid H_0)$ случайной величины $T_{оч}$ при гипотезе H_0, совпадающее со средним временем ожидания заявки в очереди при гипотезе H_0, равно нулю.

Для заявки, поступившей в СМО при гипотезе H_1, т.е. когда СМО находится в состоянии S_1, в котором канал занят, но очереди нет, условное математическое ожидание $M(T_{оч} \mid H_1)$ случайной величины $T_{оч}$ при гипотезе H_1, совпадающее со средним временем ожидания заявки в очереди при гипотезе H_1, будет равно среднему времени обслуживания одной заявки $\overline{T}_{об} = 1/\mu$.

Условное математическое ожидание $M(T_{оч} \mid H_2)$ случайной величины $T_{оч}$ при гипотезе H_2, т.е. при условии, что заявка поступила в СМО, находящуюся в состоянии S_2, в котором канал занят и в очереди уже ждет одна заявка, равно $2/\mu$ (удвоенному среднему вре-

мени обслуживания, поскольку нужно обслужить две заявки: ту, которая находится в канале обслуживания, и ту, которая ждет в очереди). И так далее.

Если заявка поступит в систему при гипотезе H_m, т.е. когда канал занят и в очереди ждут $m-1$ заявок, то $M(T_{оч} \mid H_m) = m/\mu$.

Наконец, заявка, пришедшая в СМО при гипотезе H_{m+1}, т.е. когда канал занят, m заявок стоят в очереди, и свободных мест в очереди больше нет, получает отказ и покидает систему. Поэтому в этом случае $M(T_{оч} \mid H_{m+1}) = 0$.

Следовательно, по формуле полного математического ожидания, среднее время ожидания заявки в очереди

$$\overline{T}_{оч} = M(T_{оч}) = \sum_{k=0}^{m+1} p(H_k) \cdot M(T_{оч} \mid H_k) = \sum_{k=1}^{m} p_k \cdot \frac{k}{\mu} = \frac{1}{\mu} \sum_{k=1}^{m} k p_k .$$

Подставляя сюда выражения для вероятностей p_k ($k = 1,2,...,m$), получаем:

$$\overline{T}_{оч} = \frac{1}{\mu} \sum_{k=1}^{m} k \rho^k p_0 = \frac{\rho p_0}{\mu} \sum_{k=1}^{m} k \rho^{k-1} . \qquad (22)$$

Если интенсивность нагрузки канала $\rho \neq 1$, то из равенства (22) с учетом формул (20), (21), а также выражения для p_0 находим:

$$\overline{T}_{оч} = \frac{\rho}{\mu} \cdot \frac{1-\rho}{1-\rho^{m+2}} \cdot \frac{1-\rho^m (m+1-m\rho)}{(1-\rho)^2} =$$

$$= \frac{\rho^2 (1-\rho^m (m+1-m\rho)}{\mu\rho(1-\rho^{m+2})(1-\rho)} = \frac{L_{оч}}{\lambda} .$$

Если же $\rho = 1$, то, подставляя в равенство (22) выражение $p_0 = 1/(m+2)$, значение суммы $\sum_{k=1}^{m} k = m(m+1)/2$, используя формулу (21) при $\rho = 1$ и учитывая, что в данном случае $\mu = \lambda$, будем иметь

$$\overline{T}_{оч} = \frac{m(m+1)}{2\lambda(m+2)} = \frac{L_{оч}}{\lambda} .$$

Итак, для любого ρ получаем формулу для среднего времени пребывания заявки в очереди, которая называется формулой Литтла:

$$\overline{T}_{оч} = \frac{L_{оч}}{\lambda},$$

т.е. *среднее время ожидания заявки в очереди* $\overline{T}_{оч}$ *равно среднему числу заявок в очереди* $L_{оч}$, *деленному на интенсивность* λ *входящего потока заявок.*

Пример. На автозаправочной станции (АЗС) имеется одна колонка. Площадка при станции, на которой машины ожидают заправку, может вместить не более трех машин одновременно, и если она занята, то очередная машина, прибывшая к станции, в очередь не становится, а проезжает на соседнюю АЗС. В среднем машины прибывают на станцию каждые 2 мин. Процесс заправки одной машины продолжается в среднем 2,5 мин. Определить основные характеристики системы.

Решение. Математической моделью данной АЗС является одноканальная СМО с ожиданием и ограничением на длину очереди ($m = 3$). Предполагается, что поток машин, подъезжающих к АЗС для заправки, и поток обслуживаний – простейшие.

Поскольку машины прибывают в среднем через каждые 2 мин, то интенсивность входящего потока равна $\lambda = 1/2 = 0,5$ (машин в минуту). Среднее время обслуживания одной

машины $\overline{T}_{об} = 2{,}5$ мин, следовательно, интенсивность потока обслуживаний $\mu = 1/2{,}5 = 0{,}4$ (машины в минуту).

Определяем интенсивность нагрузки канала: $\rho = \lambda/\mu = 0{,}5/0{,}4 = 1{,}25$.

Вычисляем вероятность отказа $P_{отк} = \dfrac{\rho^4(1-\rho)}{1-\rho^5} \approx 0{,}297$, откуда относительная пропускная способность $Q = 1 - P_{отк} \approx 1 - 0{,}297 = 0{,}703$ и абсолютная пропускная способность $A = \lambda Q \approx 0{,}5 \cdot 0{,}703 \approx 0{,}352$.

Среднее число машин, ожидающих в очереди на заправку

$$L_{оч} = \frac{\rho^2\left(1-\rho^3(4-3\rho)\right)}{(1-\rho)(1-\rho^5)} \approx 1{,}559.$$

Среднее время ожидания машины в очереди находим по формуле Литтла

$$\overline{T}_{оч} = L_{оч}/\lambda \approx 1{,}559/0{,}5 = 3{,}118.$$

Таким образом, из анализа работы СМО следует, что из каждых 100 подъезжающих машин 30 получают отказ ($P_{отк} \approx 29{,}7\%$), т.е. обслуживаются 2/3 заявок. Поэтому необходимо либо сократить время обслуживания одной машины (увеличить интенсивность потока обслуживаний), либо увеличить число колонок, либо увеличить площадку для ожидания. Оптимальное решение принимается с учетом затрат, связанных соответственно с увеличением штата обслуживающего персонала (увеличение производительности канала), с расширением площадки для ожидания или приобретением дополнительной колонки, и потерь, связанных с потерей заявок на обслуживание.

2. Одноканальная СМО с (неограниченным) ожиданием. Проанализируем работу одноканальной СМО с ожиданием без ограничений на длину очереди и на время ожидания в очереди. По-прежнему будем предполагать, что входящий поток и поток обслуживаний являются простейшими и имеют интенсивности λ и μ соответственно.

Такая система представляет собой предельный случай системы, рассмотренной в предыдущем пункте, при $m \to \infty$. Таким образом, длина очереди станет бесконечной и в соответствии с этим бесконечным станет число состояний СМО. Размеченный граф состояний представлен на рис. 11.

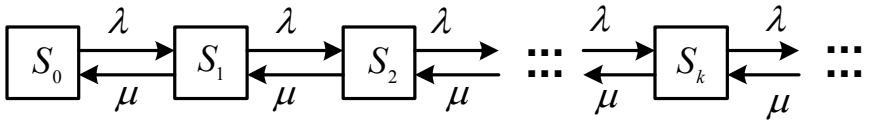

Рисунок 11.

Если отказаться от ограничения на длину очереди, то случаи $\rho < 1$ и $\rho \geq 1$ начинают существенно различаться.

Если $\lambda > \mu$ ($\rho > 1$), т.е. среднее число заявок, поступивших в систему за единицу времени, больше среднего числа обслуживаемых заявок за то же время при непрерывно работающем канале, то очевидно, что очередь неограниченно растет. В этом случае предельный режим не устанавливается и предельных вероятностей состояний не существует (точнее, они равны нулю).

В случае $\lambda = \mu$ ($\rho = 1$) только при условии, что входящий поток заявок и поток обслуживаний регулярные (т.е. заявки поступают в СМО через равные интервалы времени, и время обслуживания одной заявки является постоянным, равным интервалу времени между поступлениями заявок), очереди вообще не будет и канал будет обслуживать заявки непрерывно. Но как только входящий поток или поток обслуживаний перестает быть регулярным и приобретает элементы случайности, очередь начинает расти до бесконечности.

Поэтому далее при рассмотрении указанных систем будем предполагать, что $\lambda < \mu$, т.е. $\rho < 1$. При этом условии с течением времени устанавливается предельный режим, и предельные вероятности состояний существуют.

Устремляя m к бесконечности в формулах для вероятностей состояний (полученных для СМО с ограниченной длиной очереди при $\rho < 1$), находим выражения для предельных вероятностей состояний рассматриваемой СМО:

$$p_k = \lim_{m\to\infty} \rho^k \; p_0 = \rho^k \lim_{m\to\infty} \frac{1-\rho}{1-\rho^{m+2}} = \rho^k (1-\rho); \quad k = 0,1,2,.. \tag{23}$$

Предельные вероятности (23) удовлетворяют нормировочному условию

$$p_0 + p_1 + p_2 + ... = 1.$$

В самом деле,

$$\sum_{k=0}^{\infty} p_k = \sum_{k=0}^{\infty} \rho^k (1-\rho) = (1-\rho) \sum_{k=0}^{\infty} \rho^k .$$

Но ряд $\sum_{k=0}^{\infty} \rho^k$ представляет собой сумму бесконечно убывающей геометрической прогрессии с первым членом $b_1 = \rho^0 = 1$ и знаменателем $\rho < 1$. Поэтому

$$\sum_{k=0}^{\infty} \rho^k = \frac{1}{1-\rho} \text{ и, следовательно, } \sum_{k=0}^{\infty} p_k = (1-\rho) \cdot \frac{1}{1-\rho} = 1.$$

При отсутствии ограничений на очередь каждая заявка, поступившая в СМО, рано или поздно будет обслужена. Поэтому вероятность отказа равна нулю: $P_{отк} = 0$.

Следовательно, вероятность того, что поступившая заявка будет принята в систему, так же как и относительная пропускная способность Q, равна единице:

$$Q = 1 - P_{отк} = 1.$$

Тогда для абсолютной пропускной способности A (и интенсивности выходящего потока) будем иметь: $A = \lambda Q = \lambda$, т.е. интенсивности входящего и выходящего потоков совпадают.

Среднее число заявок в очереди $L_{оч}$ получим из формулы (21) при $\rho < 1$ переходом к пределу при $m \to \infty$:

$$L_{оч} = \lim_{m\to\infty} \frac{\rho^2 \left(1 - \rho^m (m+1-m\rho)\right)}{(1-\rho)(1-\rho^{m+2})} = \lim_{m\to\infty} \frac{\rho^2 \left(1 - m\rho^m (1+1/m-\rho)\right)}{1-\rho}.$$

Известно, что бесконечно малая ρ^m ($\rho < 1$, $m \to \infty$) является бесконечно малой более высокого порядка, чем бесконечно малая m^{-1} ($\rho^m = o(m^{-1})$), т.е. $m\rho^m \to 0$ при $m \to \infty$. Следовательно, $L_{оч} = \dfrac{\rho^2}{1-\rho}$.

Среднее время ожидания заявки в очереди по формуле Литтла равно

$$\overline{T}_{оч} = \frac{L_{оч}}{\lambda} = \frac{\rho^2}{\lambda(1-\rho)} = \frac{\rho^2}{\mu\rho(1-\rho)} = \frac{\rho}{\mu(1-\rho)}.$$

Наконец, среднее время пребывания заявки в СМО $\overline{T}_{СМО}$ складывается из среднего времени заявки в очереди $\overline{T}_{оч}$ и среднего времени обслуживания заявки $\overline{T}_{об}$:

$$\overline{T}_{СМО} = \overline{T}_{оч} + \overline{T}_{об} = \frac{\rho}{\mu(1-\rho)} + \frac{1}{\mu} = \frac{1}{\mu(1-\rho)} = \frac{\rho}{\lambda(1-\rho)}.$$

Пример. В парикмахерской работает только один мужской мастер. Среднее время стрижки одного клиента составляет 20 мин. Клиенты в среднем приходят каждые 25 мин. Средняя стоимость стрижки составляет 60 руб. Как в первую смену с 9 до 15, так и во вторую – с 15 до 21, работают по одному мастеру. Провести анализ работы системы обслуживания. Определить ежедневный «чистый» доход каждого мастера, если он получает только 30% от выручки (остальное уходит на оплату аренды помещения, налоги, амортизацию оборудования и проч.).

Решение. Интенсивность входящего потока $\lambda = 2{,}4$ клиента/ч, интенсивность потока обслуживаний $\mu = 1/\overline{T}_{об} = 1/20\ мин = \dfrac{1}{(1/3)ч} = 3$ клиента/ч. Находим:

интенсивность нагрузки (канала) мастера $\rho = \lambda/\mu = 0{,}8$;

долю времени (вероятность) простоя мастера $p_0 = 1 - \rho = 1 - 0{,}8 = 0{,}2$;

вероятность того, что мастер занят работой $p_{зан} = 1 - p_0 = 1 - 0{,}2 = 0{,}8$;

среднее число клиентов в очереди $L_{оч} = \dfrac{\rho^2}{1-\rho} = \dfrac{0{,}8^2}{1-0{,}2} = 3{,}2$ клиента;

среднее время ожидания в очереди $\overline{T}_{оч} = \dfrac{L_{оч}}{\lambda} = \dfrac{3{,}2}{2{,}4} = 1{,}34$ мин;

среднее время пребывания клиентов в парикмахерской

$$\overline{T}_{СМО} = \overline{T}_{оч} + \overline{T}_{об} = 1{,}34 + 20 = 21{,}34 \text{ мин.}$$

Система работает вполне удовлетворительно. Поскольку $\rho < 1$, то режим работы системы устойчивый, 20% рабочего времени мастер не занят, а остальные 80% времени занят работой, длина очереди 3,2 клиента небольшая, а среднее время пребывания клиента в парикмахерской всего 21,34 мин.

Каждый мастер занимается обслуживанием клиентов в среднем ежедневно в течение

$$0{,}8 \cdot (15 - 9) = 4{,}8\ ч = 288\ мин.$$

За это время он обслужит $288/20 = 14{,}4$ клиента, поэтому ежедневная выручка в среднем составит $14{,}4 \cdot 60 = 864$ руб. Ежедневный «чистый» доход каждого мастера в среднем составляет $864 \cdot 0{,}3 = 259{,}2$ руб.