

0185.05915
К89

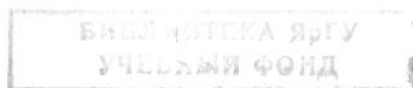
Министерство образования и науки Российской Федерации
Федеральное агентство по образованию
Ярославский государственный университет им. П.Г. Демидова

В.А. Кузнецова, Е.В. Никулина

Введение в теорию массового обслуживания

Текст лекций

Ярославль 2005



УДК 579.2
ББК В 183.53я73
К 89

Рекомендовано
Редакционно-издательским советом университета
в качестве учебного издания. План 2005 года

Рецензенты:
кафедра математического анализа ЯГПУ им. К.Д. Ушинского;
кандидат физ.-мат. наук, доцент кафедры алгебры
ЯГПУ им. К.Д. Ушинского Т.Л. Трошина

К89 Кузнецова В.А., Никулина Е.В. Введение в теорию массового обслужи-
вания: Текст лекций / В.А. Кузнецова, Е.В. Никулина; Яросл. гос. ун-т. – Ярославль:
ЯрГУ, 2005. – 60 с.
ISBN 5-8397-0406-7

Целью данного пособия является первоначальное ознакомление с основными понятиями и идеями теории массового обслуживания, с областями применения рассматриваемых теоретических предложений. Основное внимание уделяется системам массового обслуживания, работающим в стационарном режиме и имеющим входящий пуассоновский поток. Теоретические положения иллюстрированы примерами, приведены упражнения для самостоятельной работы.

Предназначено для студентов, обучающихся по специальности Математика (дисциплина „Теория массового обслуживания“, блок ОПД), очной формы обучения.

©Ярославский государственный
университет, 2005

УДК 579.2
ББК В 183.53я73

©В.А. Кузнецова, Е.В. Никулина, 2005

ISBN 5-8397-0406-7

Оглавление

1. Общая характеристика задач теории массового обслуживания	4
2. Временная диаграмма СМО. Формула Литтла	8
3. Процессы гибели и размножения	11
4. Процесс чистого размножения	14
5. Применение процесса гибели и размножения к различным системам массового обслуживания	16
6. Входящий поток заявок и его свойства	21
7. Основные характеристики пуассоновского потока	24
8. Распределение Эрланга	28
9. Системы, описываемые процессами гибели и размножения, в стационарном режиме	31
10. Классическая система массового обслуживания: $M/M/1$	36
11. Система $M/M/\infty$	40
12. Система $M/M/n$	42
13. Система $M/M/1/V$	47
14. Система $M/M/n$ с n обслуживающими приборами и с потерями	49
15. Системы $M/M/1/\infty/m$, $M/M/\infty/\infty/m$, $M/M/n/V/m$	52
15.1. Система $M/M/1/\infty/m$	52
15.2. Система $M/M/\infty/\infty/m$	53
15.3. Система $M/M/n/V/m$	53
16. Метод этапов. Эрланговское распределение	55
17. Статистическое моделирование СМО	56

1. Общая характеристика задач теории массового обслуживания

В последние десятилетия в математике возникли многочисленные новые направления исследований (теория игр, теория графов, теория сплайнов, теория массового обслуживания, и т.д.). Своим возникновением они обязаны и развитию самой математики, и потребностям практики.

Часто в обычной обстановке приходится считаться не только с возможностью появления случайных влияний, которые налагаются на некоторые закономерности, но возникает такая ситуация, что именно случайные воздействия являются определяющими для всего дальнейшего процесса. Задачи теории массового обслуживания (далее – ТМО) относятся именно к этому типу.

Первые задачи ТМО были рассмотрены сотрудником Копенгагенской телефонной компании, ученым Эрлангом (1878-1929) в период между 1908 и 1922 годами. Стояла задача упорядочить работу телефонной станции и заранее рассчитать качество обслуживания потребителей в зависимости от числа используемых устройств.

Рассмотрим схематически телефонную сеть того времени. Имеется телефонный узел (*обслуживающий прибор*), на котором телефонистки время от времени соединяют отдельные номера телефонов друг с другом. Системы массового обслуживания (далее СМО) могут быть двух видов: с ожиданием и без ожидания (т.е. с потерями). В первом случае вызов (*требование, заявка*), пришедший на станцию в момент, когда занята нужная линия, остается ждать момента соединения. Во втором случае он „покидает систему“ и не требует забот СМО.

Далее обратим внимание на следующие два факта.

Во-первых, моменты вызовов нельзя заранее фиксировать, и поступают на станцию они случайно. Может случиться, что за какой-то короткий промежуток времени их поступит очень много, и наоборот, будет длительный срок бездействия системы. Когда это произойдет - предсказать невозможно.

Во-вторых, если разговор начался (т.е. началось обслуживание), то нельзя предсказать, когда он кончится. Иначе говоря, длительность занятости линии каким-либо разговором представляет собой случайную величину.

Таким образом, в итоге получается двойная случайность. В этих условиях надо оценить, сколько технических систем и обслуживающего персонала необходимо взять для заданного качества обслуживания.

С задачами такого типа приходится сталкиваться достаточно часто.

1. Рассмотрим работу скорой медицинской помощи. Имеется несколько врачебных бригад (приборов), обслуживающих вызовы. Если все бригады заняты, то пациент случайное время ждет врача. Сами вызовы тоже появляются в случайные моменты времени, т.е. здесь полная аналогия с работой телефонной сети. Вопрос: сколько нужно врачебных бригад, чтобы станция скорой помощи своевременно обслуживала пациентов и чтобы в то же время врачи не очень долго были свободны?
2. При работе промышленного предприятия надо учитывать, что механизмы и технические устройства время от времени ломаются и требуют времени для ремонта. Длительность работы механизма, как и длительность ремонта, представляют собой случайные величины. Вопрос: сколько надо иметь ремонтников и как организовать их работу?

3. Рассмотрим морские перевозки грузов. За каждые десять лет объем морских перевозок удваивается, поэтому организация оптимального использования судов и портовых сооружений является одной из важнейших задач, а именно, задача состоит в максимизации объема перевозок при минимальных затратах (т.е. надо сократить простои судов перед погрузкой и разгрузкой). В то же время, например, для сухогрузных судов дальнего плавания (в отличие от пассажирских) нет точного расписания прибытия судна в порт назначения, так как возможны различные случайные факторы: встречный ветер, ураган, задержка с погрузкой и выгрузкой в промежуточных портах, или, наоборот, попутный ветер. В результате длительность рейса и длительность погрузо-разгрузочных работ зависят от случайных обстоятельств, т.е. опять в условиях двойной случайности нужно рассчитать количество необходимых причалов и погрузочных сооружений для перевозок заданного объема грузов.

4. Рассмотрим работу прибора для определения интенсивности ядерного излучения – счетчик Гейгера. Частица, попавшая в счетчик, запирает его для подсчета новых частиц. Время записывания зависит от энергии этой частицы и ее характера и является случайной величиной. Сам поток частиц тоже нерегулярен, и промежутки между двумя последовательно попадающими частицами – тоже случайная величина. Таким образом, часть частиц остается незарегистрированной, следовательно, надо в показания счетчика вносить поправку, зависящую от случайных величин. Здесь мы имеем систему с потерями, и возникает вопрос о числе „потерянных“ заявок.

К системам массового обслуживания приводит и множество других задач: простой судов перед шлюзами, покупателей в магазинах, ожидание транспорта, и т.д. Очереди – бич современной жизни, поэтому во многих странах сейчас большое внимание обращено на изучение закономерностей образования очередей и их рассасывания.

В упомянутых выше двух типах СМО: „с потерями“ и „с ожиданием“ различаются и характеристики. В первой важна вероятность „отказа“, а во второй – среднее время ожидания в очереди. Причем, среднее время ожидания является важной, но не единственной характеристикой. Здесь важно знать величину возможного разброса длительностей ожидания, среднюю длину очереди, распределение длины очереди. Системы с ожиданием тоже различаются между собой, так как может быть система с ожиданием и в то же время с потерями (например, входим в магазин и отказываемся от обслуживания из-за слишком большой очереди). Причем, часто интересует не время ожидания, а время пребывания в системе – время ожидания плюс время обслуживания. Такие вопросы возникают, например, при проектировании разного рода информационных систем, так как информация обладает свойством старения.

Итак, можно выделить три типа СМО с ожиданием и с ограничениями:

- 1) Заявки в очереди остаются до тех пор, пока очередь не достигнет некоторого фиксированного объема (ограничение на длину очереди). Например, прием заказов в ателье ограничен – не более десяти.
- 2) Заявки ожидают в очереди до тех пор, пока время ожидания не превысит заданной величины, после чего „покидают систему“ (ограничение на время ожидания). Например, ожидание междугороднего разговора, если абонент не отвечает, не превышает двух часов.
- 3) Время пребывания заявки в системе (время ожидания начала обслуживания плюс время обслуживания) не должно превосходить некоторого фиксированно-

го значения. При этом заявка может покинуть систему, так и не дождавшись начала обслуживания, а может покинуть систему, начав обслуживаться, но не дождавшись конца обслуживания.

Во всех этих случаях интерес представляет не только средняя длительность ожидания, но и средняя величина потерь заявок за данный промежуток времени. В последнем случае интерес представляет вероятность того, что заявка, уже начавшая обслуживаться, покинет систему необслуженной (с незаконченным обслуживанием).

На практике СМО с ожиданием различаются дисциплиной очереди, под которой подразумевается правило выбора требований из очереди. Рассмотрим сначала случай, когда все поступающие требования одинаковы. В этом случае нет смысла делить требования на классы и предоставлять преимущества требованиям одного класса перед требованиями другого. Единственным признаком в таком случае, по которому определяется дисциплина очереди, служит порядок поступления требований.

Типичными являются следующие дисциплины:

1. Первым пришел – первым обслужен, т.е. требования обслуживаются в порядке их поступления в СМО. В англоязычной литературе эта дисциплина обозначается FIFO (first in – first out). Примером может служить обслуживание в супермаркете.
2. Последним пришел – первым обслужен; английское обозначение LIFO (last in – first out). Данная дисциплина очереди имеет место, например, при проверке контролером изделий, которые накапливаются перед ним стопкой, так что последнее изделие проверяется первым.
3. Случайный и равновероятный выбор на обслуживание среди всех имеющихся в очереди требований, например вызов студентов к доске.

Теперь предположим, что требования обладают отличительными признаками, которые могут быть приписаны каждому из них. Это могут быть:

- а) определенная стоимость единицы времени ожидания данного требования или, по крайней мере, „показатель срочности“, который характеризует требуемую срочность обработки;
- б) особый закон распределения длительности обслуживания, отличный от закона, относящегося к случайно выбранному требованию;
- в) оба вышеуказанных признака одновременно.

Обычно поступают следующим образом. Все требования разбиваются на классы. Требования i -го класса между собой неразличимы и характеризуются стоимостью единицы времени ожидания в очереди и законом распределения длительности обслуживания. Требованиям i -го класса предоставляется приоритет в обслуживании перед требованиями j -го класса при $i < j$. При этом будем говорить, что требования i -го класса имеют более высокий приоритет по сравнению с требованиями j -го класса (более низкого приоритета). Используются два типа приоритета:

- 1) *относительный*. В момент окончания обслуживания предыдущей заявки следующим на обслуживание из очереди выбирается требование класса с наименьшим номером. Требования высокого приоритета не вытесняют с прибора требований низкого приоритета во время обслуживания последних. Требования одно-

го класса обслуживаются в порядке их поступления в систему. Данной дисциплиной обладает, например, очередь к врачу, состоящая из пациентов, не обладающих какими-либо льготами, и пациентов-льготников;

- 2) *абсолютный*. Если в момент поступления в систему требования i -го класса имеется хотя бы один прибор, занятый обслуживанием требования класса с большим номером, то вновь поступившее требование вытесняет обслуживаемое, т.е. система начинает обслуживать требование i -го класса, а вытесненное требование возвращается в очередь и повторно поступает на обслуживание лишь после завершения обслуживания всех требований более высоких приоритетов. В принципе могут существовать такие СМО, в которых требование, снятое с обслуживания, покидает систему.

Реальные системы, встречающиеся на практике, как правило, очень сложны и включают ряд этапов обслуживания, т.е. представляют собою многофазную СМО. На каждой фазе может произойти отказ от дальнейшего обслуживания, или же заявки некоторых типов могут обслуживаться с разными преимуществами по отношению к заявкам других типов. Может случиться, что приборы СМО по тем или иным причинам прекращают работу (для ремонта отдельных звеньев, подналадки). В некоторых системах заявки, получившие отказ или уже обслуженные, вновь через какое-то время могут возвратиться в систему. Именно с таким положением встречаемся в телефонии, при больничном обслуживании.

Представим в виде схемы типичную СМО на примере ремонтного предприятия. Заявки, поступающие в систему, проходят такую последовательность операций: осмотр с целью определения характера ремонта, направление в те или иные ремонтные бригады с указанием последовательности действий, проверка качества ремонта, сдача заказчику. В процессе выполнения той или иной операции иногда выясняется, что износ так велик, что восстановление невозможно, или изделие приходится возвращать на предыдущие этапы обработки для доделок. Условимся, что имеются лишь две операции. Тогда схема функционирования выглядит следующим образом:

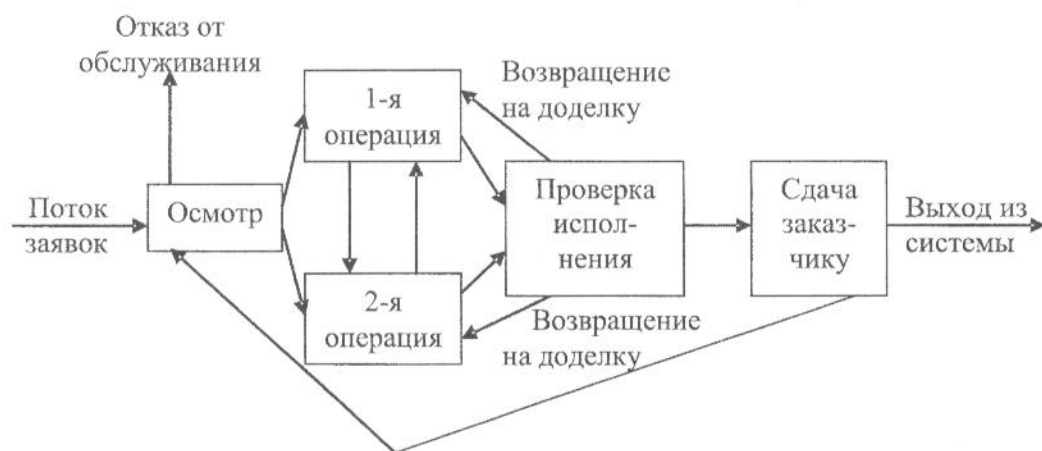


Рис. 1

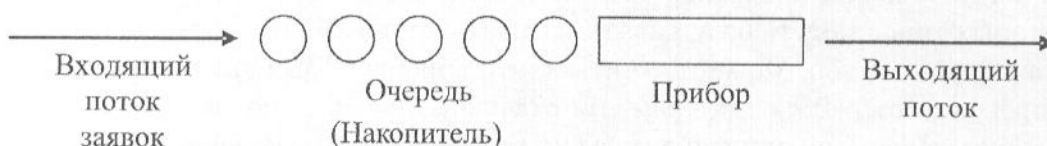
Упражнения

- 1. Привести примеры СМО с ожиданием, с потерями, с различными дисциплинами очереди.

- 2. На конкретных примерах объяснить смысл терминов: „обслуживающий прибор“, „заявка“, „очередь“.
- 3. В чем состоит различие СМО с относительным приоритетом и абсолютным? Привести примеры.

2. Временная диаграмма СМО. Формула Литтла

Рассмотрим СМО с такой дисциплиной обслуживания, когда заявки обслуживаются в порядке поступления (FIFO). При этом поток требований, поступающих в систему, называется *входящим потоком*, покидающих – *выходящим потоком*. Будем считать, что в СМО один прибор и имеется ординарный поток заявок, т.е. в одно и то же время не может поступить более одной заявки. В виде схемы вышесказанное можно изобразить следующим образом:



Построим временную диаграмму системы, удовлетворяющей указанным условиям. Введем следующие обозначения:

- c_n — n -я заявка, поступающая в СМО,
- w_n — время ожидания n -й заявки в очереди,
- x_n — время обслуживания n -й заявки,
- τ_n — момент поступления n -й заявки в СМО,
- s_n — время пребывания n -й заявки в СМО,
- t_n — промежуток времени между заявками c_{n-1} и c_n , т.е. $t_n = \tau_n - \tau_{n-1}$.

Горизонтальными линиями на диаграмме обозначим очередь и прибор, время при этом движется вправо:

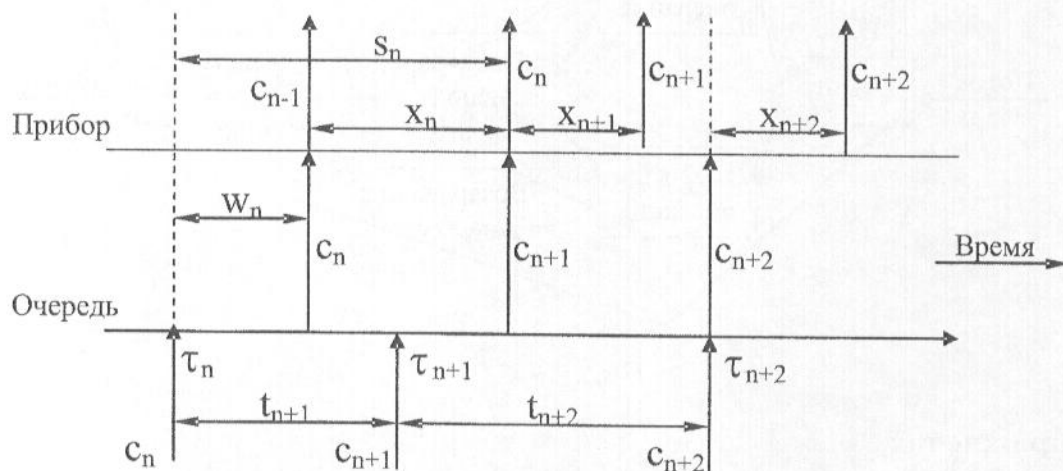


Рис. 2

Средний промежуток времени t^* между двумя соседними заявками (т.е. $t_n \rightarrow t^*$) иногда обозначают через $\frac{1}{\lambda}$, т.е.

$$t^* = \frac{1}{\lambda},$$

λ называют *интенсивностью* поступления заявок в систему, или интенсивностью входного потока.

Далее найдем соотношение между интенсивностью входного потока, средним числом заявок в системе и средним временем пребывания заявки в СМО.

Пусть $\alpha(t)$ – число заявок, поступающих в промежутке $(0, t)$; $\delta(t)$ – число заявок, покинувших систему в промежутке $(0, t)$. Изобразим в координатной плоскости (k, t) (k – число заявок, t – момент времени) диаграмму, наглядно показывающую, в каком состоянии находится СМО в каждый момент времени. Заметим, что на оси числа заявок берутся только целые числа.

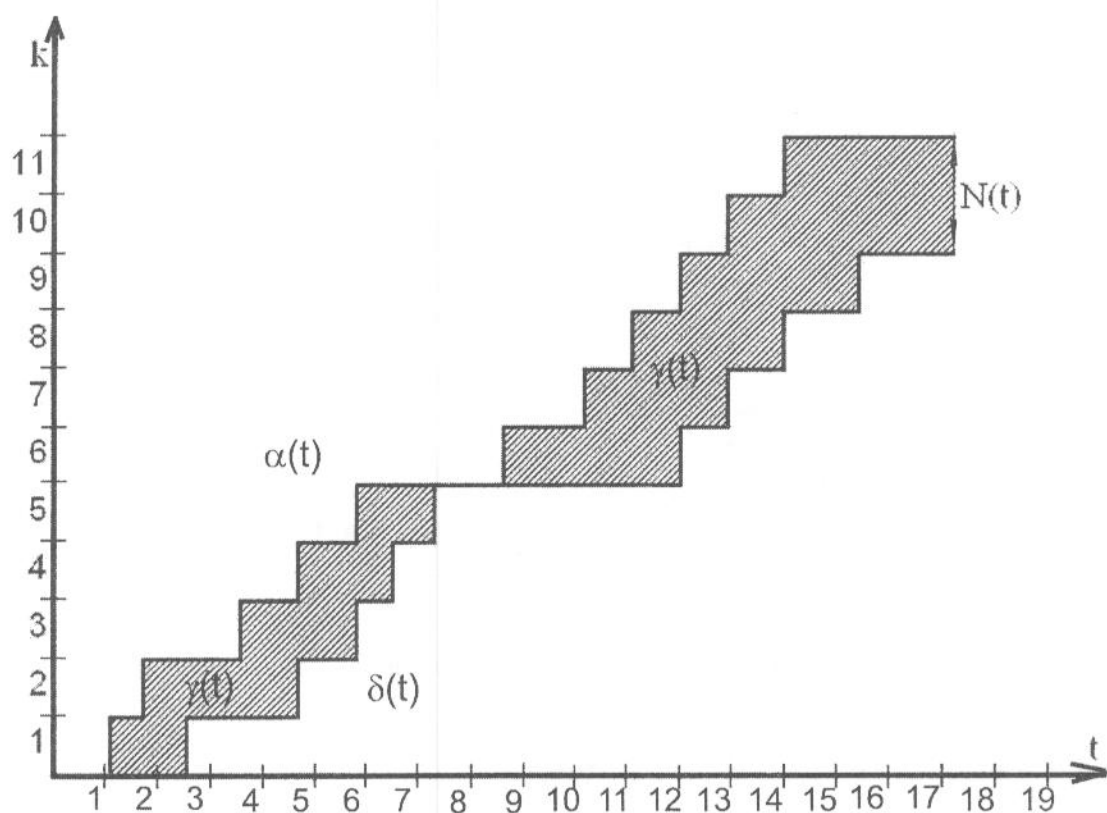


Рис. 3

Заметим, что заштрихованная площадь равна суммарному количеству времени, проведенному всеми заявками в системе за период $(0, t)$, так как, если рассмотреть отдельно любой из заштрихованных прямоугольников, то его площадь численно равна длине стороны по оси времени (поскольку высота равна единице). Обозначим всю заштрихованную площадь через $\gamma(t)$.

Тогда:

$\frac{\gamma(t)}{\alpha(t)}$ – среднее время, проведенное одной заявкой в системе, усредненное по всем

заявкам, находившимся в системе в промежутке $(0, t)$, обозначим:

$$T_t = \frac{\gamma(t)}{\alpha(t)},$$

λ_t – интенсивность поступления заявок в единицу времени обозначим:

$$\lambda_t = \frac{\alpha(t)}{t},$$

N_t^* – среднее число заявок в системе в промежутке $(0, t)$:

$$N_t^* = \frac{\gamma(t)}{t}.$$

Но тогда :

$$N_t^* = \lambda_t T_t. \quad (2.1)$$

Подчеркнем, что T_t , λ_t , N_t^* существенно зависят от t , т.е. от длины промежутка, на котором происходит усреднение. Например, для нашего рисунка $\lambda_3 = \frac{\alpha(3)}{3} = \frac{2}{3}$, а $\lambda_4 = \frac{3}{4}$.

Предположим теперь, что система такова, что существуют пределы:

$$\lim_{t \rightarrow \infty} \lambda_t = \lambda \quad \text{и} \quad \lim_{t \rightarrow \infty} T_t = T,$$

тогда из (2.1) следует, что существует предел N_t^* , который обозначим через N^* – среднее число заявок в системе.

Итак, имеем:

$$\boxed{N^* = \lambda T} \quad - \text{ формула Литтла.} \quad (2.2)$$

Среднее число требований в системе равно произведению интенсивности входного потока на среднее время пребывания заявки в системе.

Этот результат не зависит от распределения входящего потока, от распределения времени обслуживания, от дисциплины обслуживания.

Заметим, что если очередь понимать как отдельную СМО, то:

$$N_q^* = \lambda w,$$

где w – среднее время ожидания, N_q^* – среднее число заявок в очереди"; аналогично, если работу прибора понимать как отдельную СМО без рассмотрения накопителя, то:

$$N_p^* = \lambda x^*,$$

где x^* – среднее время обслуживания, а N_p^* – среднее число заявок на приборе. Если учитывать очередь и прибор, то $T = x^* + w$ и

$$N^* = \lambda(x^* + w).$$

Далее рассмотрим одну из важных характеристик СМО – коэффициент использования системы, обозначаемый через ρ .

Для СМО с одним прибором коэффициент использования определим как среднее число заявок в приборе, т.е. $\rho = \frac{\lambda}{\mu} = \lambda x^*$, позднее мы увидим, что для продуктивности работы СМО с неограниченной очередью необходимо, чтобы $0 \leq \rho < 1$. В случае, когда $0 \leq \rho < 1$, его можно рассматривать также как долю времени занятости прибора. Например, если в среднем в единицу времени поступает 3 заявки, а обслуживается в единицу времени в среднем 6 заявок, то прибор в среднем занят

$\frac{3}{6} = \frac{1}{2}$ долю времени. Если p_0 – вероятность того, что прибор свободен (т.е. в системе 0 заявок), то $1 - p_0$ – вероятность занятости прибора, и в случае, когда $0 \leq \rho < 1$, она равна ρ , т.е. $\lambda x^* = 1 - p_0 = \rho$.

Если в системе n приборов, то $\rho = \frac{\lambda}{n\mu} = \frac{\lambda x^*}{n}$, т.е. ρ – отношение числа заявок, в среднем поступающих в систему в единицу времени, к среднему числу заявок, обслуживаемых всеми приборами в единицу времени. Также ρ можно рассматривать как среднее значение доли занятых приборов. В случае системы с неограниченной очередью для продуктивности ее работы также необходимо, чтобы $0 \leq \rho < 1$.

Упражнения

- 1. Что называют интенсивностью входного потока?
- 2. Почему для среднего времени, проведенного одной заявкой в системе, верна формула $T_t = \frac{\gamma(t)}{\alpha(t)}$, а не формула следующего вида: $T_t = \frac{\gamma(t)}{\delta(t)}$?
- 3. На рисунке найдите T_4 , λ_4 , N_4^* , T_8 , λ_8 , N_8^* .
- 4. Выпишите формулу Литтла и ее частные случаи, объясните смысл входящих в нее обозначений.
- 5. Что такое коэффициент использования для системы с одним прибором, с n приборами?

3. Процессы гибели и размножения

Рассмотрим такой процесс, когда в каждый случайный момент может быть одна из трех возможностей: либо происходит поступление в точности одной заявки (рождение и увеличение популяции на одну особь), либо в точности одна заявка покидает систему, будучи обслуженной в течение промежутка времени, длительность которого была случайной величиной (гибель и уменьшение популяции на 1), либо не происходит изменения численности популяции ни в ту, ни в другую сторону.

Назовем *состоянием СМО в данный момент* общее число требований, находящихся в системе в этот момент. Тогда система может иметь следующие состояния:

- E_0 – в системе нет заявок,
- E_1 – 1 заявка,
- E_2 – 2 заявки, и т.д.

$E_k(t)$ – событие, состоящее в том, что в системе в момент t есть k заявок ($k=0,1,2,\dots$).

Если в некоторый момент t_0 прибыла новая заявка, то положение состояния системы изменилось на единицу в сторону возрастания; если закончилось обслуживание заявки, то произошло изменение на единицу в сторону уменьшения популяции, иначе – состояние системы осталось тем же. Сказанное можно обозначить следующим образом:

$$E_k(t-0) \rightarrow E_{k+1}(t+0),$$

$$E_k(t-0) \rightarrow E_{k-1}(t+0),$$

$$E_k(t-0) \rightarrow E_k(t+0).$$

Для процессов гибели и размножения характерны лишь переходы в соседние состояния. При этом пусть вероятности переходов из состояния $E_k(t)$ в $E_{k+1}(t+h)$, в $E_{k-1}(t+h)$ и в $E_k(t+h)$ за малый промежуток времени h удовлетворяют соотношениям:

$$\begin{aligned} P\{E_k(t) \rightarrow E_{k+1}(t+h)\} &= \lambda_k h + o(h), \\ P\{E_k(t) \rightarrow E_{k-1}(t+h)\} &= \mu_k h + o(h), \text{ тогда} \\ P\{E_k(t) \rightarrow E_k(t+h)\} &= 1 - (\lambda_k + \mu_k)h + o(h). \end{aligned} \quad (3.1)$$

Величины λ_k и μ_k не являются вероятностями переходов, они лишь „управляют“ такими вероятностями. Эти величины неотрицательны, может оказаться, что некоторые из них равны нулю (далее мы рассмотрим такие случаи). В частности, величина μ_0 всегда равна нулю, поскольку система не может перейти в отрицательное состояние. Они зависят от того состояния системы, в котором она находилась и в какое переходит, но не зависят от продолжительности нахождения системы в этом состоянии и не зависят также от момента времени t , в который рассматривается функционирование системы. Наличие величины $o(h)$ в соотношениях означает, что вероятность перехода не в соседнее состояние за малый промежуток времени есть величина бесконечно малая по сравнению с длиной этого промежутка. Таким образом, непосредственный переход не в соседнее состояние практически невозможен.

Все вышесказанное представляет собой жесткие условия, тем не менее, оказывается, процесс гибели и размножения играет колоссальную роль в теории массового обслуживания. Подавляющее большинство задач теории массового обслуживания в той или иной степени опирается на этот процесс.

Пусть $p_k(t)$ – вероятность того, что в момент t объем популяции равен k , т.е. система находится в состоянии $E_k(t)$. Оказывается, вероятности $p_k(t)$ подчинены системе дифференциальных уравнений. Найдем уравнение, которому удовлетворяет $p_0(t)$. Рассмотрим $p_0(t+h)$.

В момент $t+h$ система может приобрести состояние E_0 в одном из двух случаев: в момент t она имела состояние E_0 , т.е. $E_0(t) \rightarrow E_0(t+h)$, или в момент t она была в состоянии E_1 , т.е. произошел переход: $E_1(t) \rightarrow E_0(t+h)$.

Тогда:

$$p_0(t+h) = p_0(t)p\{E_0(t) \rightarrow E_0(t+h)\} + p_1(t)p\{E_1(t) \rightarrow E_0(t+h)\} + o(h).$$

В соответствии с (3.1) получаем:

$$\begin{aligned} p_0(t+h) &= p_0(t)[1 - \lambda_0 h + o(h)] + p_1(t)[\mu_1 h + o(h)] + o(h), \\ \frac{p_0(t+h) - p_0(t)}{h} &= -\lambda_0 p_0(t) + \mu_1 p_1(t) + o(h). \end{aligned}$$

Предположим, что все $p_k(t)$ дифференцируемы, тогда существует предел отношения $\frac{p_0(t+h) - p_0(t)}{h}$ при $h \rightarrow 0$, и он равен $p_0'(t)$, т.е.

$$p_0'(t) = -\lambda_0 p_0(t) + \mu_1 p_1(t).$$

Аналогично, найдем $p_k(t+h)$ при $k \geq 1$:

$$p_k(t+h) = p_{k-1}(t)p\{E_{k-1}(t) \rightarrow E_k(t+h)\} + p_k(t)p\{E_k(t) \rightarrow E_k(t+h)\} +$$

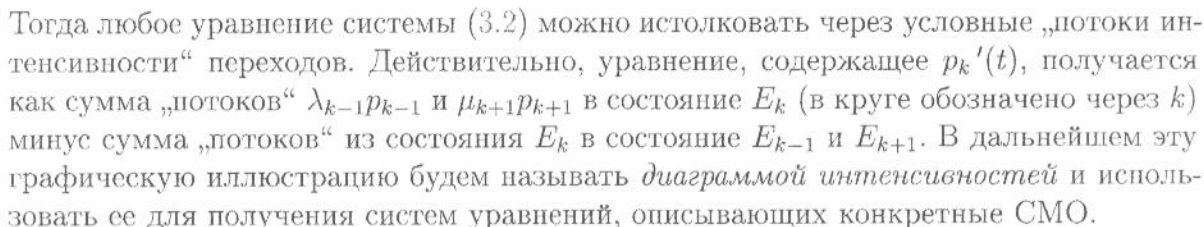
тогда:

Получили требуемую систему дифференциальных уравнений:

Последнее алгебраическое уравнение означает, что в каждый момент времени t система находится в одном из возможных состояний.

Конкретный вид системы зависит от функций λ_k и μ_k . Для решения системы надо задать начальные условия $p_k(0)$ при $k = 0, 1, 2, \dots$.

Графически изобразим различные состояния системы с помощью кругов, в которых обозначено состояние, стрелками обозначим переходы из состояния E_k в E_{k-1} и в E_{k+1} :



Таким образом, процесс гибели и размножения, рассматриваемый с сильными вышеприведенными ограничениями, может быть использован как для описания жизнедеятельности некоторой биологической системы (популяции), так и для функционирования систем массового обслуживания.

- 1. Охарактеризуйте процесс гибели и размножения.
- 2. Запишите систему дифференциальных уравнений (3.2) с помощью диаграммы интенсивности переходов.
- 3. Составьте уравнения СМО, имеющей следующую диаграмму интенсивностей:



4. Процесс чистого размножения

Обратимся теперь к такой системе, в которой не происходит обслуживание заявок (особи не умирают). Это означает, что $\mu_k = 0$ при любом k , т.е., по существу, рассматривается лишь входной поток заявок в систему (особи рождаются, осуществляется процесс чистого размножения). Введем еще одно ограничение: пусть при любом k $\lambda_k = \lambda$. Тогда система (3.2) имеет вид:

$$\begin{cases} p_0'(t) = -\lambda p_0(t), \\ \dots\dots\dots \\ p_k'(t) = \lambda p_{k-1}(t) - \lambda p_k(t), \quad k \geq 1, \\ \dots\dots\dots \\ \sum_{k=0}^{\infty} p_k(t) = 1. \end{cases}$$

Пусть процесс начинается в тот момент, когда система пуста, т.е. $k = 0$, и

$$p_k(0) = \begin{cases} 1, & \text{при } k = 0, \\ 0, & \text{при } k \neq 0. \end{cases}$$

Из первого уравнения системы найдем $p_0(t)$:

$$p_0(t) = Ce^{-\lambda t},$$

при $t = 0$ $p_0(0) = 1 = Ce^0 \Rightarrow C = 1$, т.е.

$$p_0(t) = e^{-\lambda t}.$$

Далее:

$$p_1'(t) = \lambda e^{-\lambda t} - \lambda p_1(t),$$

$$p_1'(t) + \lambda p_1(t) - \lambda e^{-\lambda t} = 0,$$

$$p_1(t) = e^{-\lambda t}(C + \lambda \int e^{-\lambda t} e^{\lambda t} dt) = e^{-\lambda t}(C + \lambda t),$$

при $t = 0$ $p_1(0) = 0 \implies 0 = C \implies$

$$p_1(t) = \lambda t e^{-\lambda t}.$$

Используем метод математической индукции. Предположим, что $p_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$, найдем $p_{k+1}(t)$.

$$p_{k+1}'(t) = \lambda p_k(t) - \lambda p_{k+1}(t) = \frac{\lambda^{k+1} t^k}{k!} e^{-\lambda t} - \lambda p_{k+1}(t),$$

$$p_{k+1}'(t) + \lambda p_{k+1}(t) - \frac{\lambda^{k+1} t^k}{k!} e^{-\lambda t} = 0,$$

$$p_{k+1}(t) = e^{-\lambda t} (C + \frac{\lambda^{k+1}}{k!} \int t^k e^{-\lambda t} e^{\lambda t} dt) = e^{-\lambda t} (C + \frac{\lambda^{k+1}}{(k+1)!} t^{k+1}),$$

$$C = 0 \implies p_{k+1}(t) = \frac{(\lambda t)^{k+1}}{(k+1)!} e^{-\lambda t}.$$

Таким образом, для всех k :

$$p_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad (4.1)$$

Читателю эта формула должна быть известна из курса теории вероятностей как пуассоновский закон распределения дискретной случайной величины. В данном случае в роли этой случайной величины выступает количество заявок, поступающих в систему за промежуток длины t .

Рассмотрим далее общий случай процесса чистого размножения, когда при любом k λ_k зависит от состояния системы. Тогда система (3.2) имеет вид:

$$\begin{cases} p_0'(t) = -\lambda_0 p_0(t), \\ \dots\dots\dots \\ p_k'(t) = \lambda_{k-1} p_{k-1}(t) - \lambda_k p_k(t), \quad k \geq 1, \\ \dots\dots\dots \\ \sum_{k=0}^{\infty} p_k(t) = 1. \end{cases}$$

По-прежнему предположим, что процесс начинается с нулевой популяции, т.е.

$$p_k(0) = \begin{cases} 1, & \text{при } k = 0, \\ 0, & \text{при } k \neq 0. \end{cases}$$

Тогда имеем:

$$p_0(t) = e^{-\lambda_0 t}.$$

Далее:

$$\begin{aligned} p_1'(t) &= \lambda_0 p_0(t) - \lambda_1 p_1(t), \\ p_1'(t) + \lambda_1 p_1(t) - \lambda_0 p_0(t) &= 0, \text{ где } p_0(t) = e^{-\lambda_0 t}, \end{aligned}$$

тогда

$$\begin{aligned} p_1(t) &= e^{-\lambda_1 t} \cdot \left[C + \lambda_0 \int e^{(-\lambda_0 + \lambda_1)t} dt \right], \\ p_1(t) &= e^{-\lambda_1 t} \cdot \left[C + \frac{\lambda_0}{\lambda_1 - \lambda_0} e^{(\lambda_1 - \lambda_0)t} \right]. \end{aligned}$$

Пусть $t = 0$, тогда:

$$\begin{aligned} 0 &= C + \frac{\lambda_0}{\lambda_1 - \lambda_0}, \\ C &= -\frac{\lambda_0}{\lambda_1 - \lambda_0}, \\ p_1(t) &= \frac{\lambda_0}{\lambda_1 - \lambda_0} \left[e^{-\lambda_0 t} - e^{-\lambda_1 t} \right]. \end{aligned}$$

Аналогично,

$$\begin{aligned} p_2'(t) &= \lambda_1 p_1(t) - \lambda_2 p_2(t), \\ p_2'(t) + \lambda_2 p_2(t) - \lambda_1 p_1(t) &= 0, \\ p_2(t) &= e^{-\lambda_2 t} \cdot \left[C + \frac{\lambda_0 \lambda_1}{\lambda_1 - \lambda_0} \cdot \left(\frac{1}{\lambda_2 - \lambda_0} e^{(\lambda_2 - \lambda_0)t} - \frac{1}{\lambda_2 - \lambda_1} e^{(\lambda_2 - \lambda_1)t} \right) \right]. \end{aligned}$$

Если $t = 0$, то:

$$0 = C + \frac{\lambda_0 \lambda_1}{\lambda_1 - \lambda_0} \left(\frac{1}{\lambda_2 - \lambda_0} - \frac{1}{\lambda_2 - \lambda_1} \right), \text{ тогда}$$

$$p_2(t) = \frac{\lambda_0 \lambda_1}{\lambda_1 - \lambda_0} \cdot \left[\frac{e^{-\lambda_0 t} - e^{-\lambda_2 t}}{\lambda_2 - \lambda_0} - \frac{e^{-\lambda_1 t} - e^{-\lambda_2 t}}{\lambda_2 - \lambda_1} \right].$$

Аналогично вычисляются $p_k(t)$ при $k \geq 3$. Сделанные выкладки приведены для того, чтобы показать, что вычисления даже при чистом размножении в общем случае технически громоздки.

Упражнения

- 1. Охарактеризуйте процесс чистого размножения в случае, когда для любого k $\lambda_k = \lambda$, и в случае, когда λ_k зависят от состояния системы.
- 2. Докажите формулу (4.1).

5. Применение процесса гибели и размножения к различным системам массового обслуживания

Рассмотрим примеры вывода системы дифференциальных уравнений, характеризующих динамику изменения состояний процесса, для некоторых частных типов СМО.

Пример 1. Система с потерями.

Пусть в системе имеются n одинаковых приборов. Поступающая заявка, застающая хотя бы один свободный прибор, сразу идет на обслуживание. Выбор приборов произволен. Если все приборы заняты, заявка покидает систему необслуженной, т.е. очереди в системе нет.

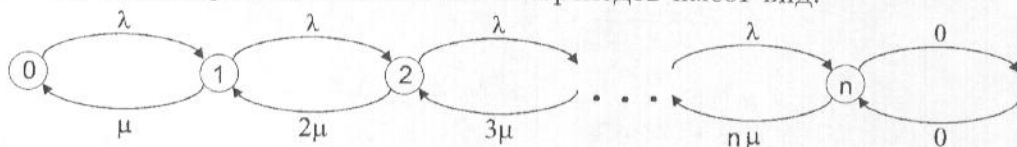
Положим $\lambda_0 = \lambda_1 = \dots = \lambda_{n-1} = \lambda$. В этом случае будем говорить, что *поток однороден во времени*, т.е. имеет постоянную интенсивность. Очевидно, $\lambda_n = 0$, т.к. $(n+1)$ -я заявка уходит из системы необслуженной.

Будем считать, что все приборы имеют в среднем одинаковую производительность, равную μ , т.е., если прибор занят, то вероятность того, что он освободится за время h , в терминах процесса „гибели и размножения“, равна вероятности убывания популяции на единицу, т.е. $\mu h + o(h)$. Если заняты k приборов, то вероятность того, что освободится хотя бы один из них за время h , равна $k\mu h + o(h)$.

Отсюда имеем:

$$\begin{aligned} \mu_k &= k\mu \quad \text{для } k = 0, 1, \dots, n, \\ \mu_k &= 0 \quad \text{для } k > n, \text{ т.к. очереди нет.} \end{aligned}$$

В этом случае диаграмма интенсивности переходов имеет вид:



Далее легко записать систему дифференциальных уравнений:

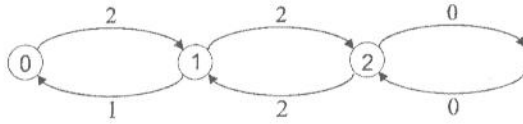
$$\begin{cases} p_0'(t) = -\lambda p_0(t) + \mu p_1(t), \\ \dots \\ p_k'(t) = \lambda p_{k-1}(t) - (\lambda + k\mu)p_k(t) + (k+1)\mu p_{k+1}(t), \quad \text{при } 1 \leq k \leq n-1, \\ \dots \\ p_n'(t) = \lambda p_{n-1}(t) - n\mu p_n(t), \\ \sum_{k=0}^n p_k(t) = 1. \end{cases}$$

Заметим, что для такой СМО система дифференциальных уравнений конечна.

Рассмотрим теперь примеры систем с конкретными числовыми данными.

- а) Пусть в системе с потерями $n = 2, \mu = 1, \lambda = 2, p_0(0) = 1, p_1(0) = 0, p_2(0) = 0$.

Диаграмма интенсивностей переходов будет иметь вид:



Получаем следующую систему дифференциальных уравнений:

$$\begin{cases} p_0'(t) = -2p_0(t) + p_1(t), \\ p_1'(t) = 2p_0(t) - 3p_1(t) + 2p_2(t), \\ p_2'(t) = 2p_1(t) - 2p_2(t), \\ p_0(t) + p_1(t) + p_2(t) = 1. \end{cases}$$

Система линейно зависима, можно не рассматривать третье уравнение. Из алгебраического уравнения получаем: $p_2(t) = 1 - p_0(t) - p_1(t)$ и подставляем его во второе уравнение. Получим:

$$\begin{cases} p_0'(t) = -2p_0(t) + p_1(t), \\ p_1'(t) = -5p_1(t) + 2, \\ p_2(t) = 1 - p_0(t) - p_1(t). \end{cases}$$

Из второго уравнения, учитывая начальные условия, найдем $p_1(t)$, затем $p_0(t)$ и $p_2(t)$ из первого и третьего уравнений соответственно. Получим:

$$\begin{aligned} p_0(t) &= \frac{1}{5} + \frac{2}{3}e^{-2t} + \frac{2}{15}e^{-5t}, \\ p_1(t) &= \frac{2}{5} - \frac{2}{5}e^{-5t}, \\ p_2(t) &= \frac{2}{5} - \frac{2}{3}e^{-2t} + \frac{4}{15}e^{-5t}. \end{aligned}$$

- б) В системе с потерями $n = 2, \mu = 1, \lambda = 3, p_0(0) = 1, p_1(0) = 0, p_2(0) = 0$.

Соответствующая система дифференциальных уравнений примет вид:

$$\begin{cases} p_0'(t) = -3p_0(t) + p_1(t), \\ p_1'(t) = 3p_0(t) - 4p_1(t) + 2p_2(t), \\ p_2'(t) = 3p_1(t) - 2p_2(t), \\ p_0(t) + p_1(t) + p_2(t) = 1. \end{cases}$$

Система уравнений линейно зависима, выберем для ее решения первое, второе и четвертое уравнения. Тогда:

$$\begin{cases} p_0'(t) = -3p_0(t) + p_1(t), \\ p_1'(t) = p_0(t) - 6p_1(t) + 2, \\ p_2(t) = 1 - p_0(t) - p_1(t). \end{cases}$$

Перейдем к решению системы следующих дифференциальных уравнений:

$$\begin{cases} p_0'(t) = -3p_0(t) + p_1(t), \\ p_1'(t) = p_0(t) - 6p_1(t) + 2. \end{cases}$$

Характеристическое уравнение имеет вид:

$$\begin{vmatrix} -3-\lambda & 1 \\ 1 & -6-\lambda \end{vmatrix} = 0.$$

Отсюда:

$$\lambda_{1,2} = -\frac{9}{2} \pm \frac{\sqrt{13}}{2}.$$

Тогда в качестве собственных можно взять следующие векторы:

$$\overline{H}_1 = \left\{ -\frac{3}{2} - \frac{\sqrt{13}}{2}; -1 \right\},$$

$$\overline{H}_2 = \left\{ -\frac{3}{2} + \frac{\sqrt{13}}{2}; -1 \right\}.$$

Общее решение однородной системы дифференциальных уравнений будет иметь вид:

$$p_0(t) = C_1 \left(-\frac{3}{2} - \frac{\sqrt{13}}{2} \right) e^{\left(-\frac{9}{2} + \frac{\sqrt{13}}{2} \right) t} + C_2 \left(-\frac{3}{2} + \frac{\sqrt{13}}{2} \right) e^{\left(-\frac{9}{2} - \frac{\sqrt{13}}{2} \right) t},$$

$$p_1(t) = C_1(-1)e^{\left(-\frac{9}{2} + \frac{\sqrt{13}}{2} \right) t} + C_2(-1)e^{\left(-\frac{9}{2} - \frac{\sqrt{13}}{2} \right) t}.$$

Частное решение неоднородной системы будем искать в следующем виде:

$$\begin{aligned} p_0 &= a, \\ p_1 &= b, \end{aligned} \quad \text{где } a, b \in R.$$

Подставим $p_0 = a, p_1 = b$ в систему $\begin{cases} p_0'(t) = -3p_0(t) + p_1(t), \\ p_1'(t) = p_0(t) - 6p_1(t) + 2. \end{cases}$

Получим:

$$p_0 = a = \frac{2}{17},$$

$$p_1 = b = \frac{6}{17}.$$

Тогда общее решение неоднородной системы будет иметь вид:

$$p_0(t) = C_1 \left(-\frac{3}{2} - \frac{\sqrt{13}}{2} \right) e^{\left(-\frac{9}{2} + \frac{\sqrt{13}}{2} \right) t} + C_2 \left(-\frac{3}{2} + \frac{\sqrt{13}}{2} \right) e^{\left(-\frac{9}{2} - \frac{\sqrt{13}}{2} \right) t} + \frac{2}{17},$$

$$p_1(t) = C_1(-1)e^{\left(-\frac{9}{2} + \frac{\sqrt{13}}{2} \right) t} + C_2(-1)e^{\left(-\frac{9}{2} - \frac{\sqrt{13}}{2} \right) t} + \frac{6}{17}.$$

Используя начальные условия, найдем C_1 и C_2 :

$$C_1 = \frac{-24}{17\sqrt{13}} + \frac{3}{17},$$

$$C_2 = \frac{24}{17\sqrt{13}} + \frac{3}{17}.$$

Итак,

$$p_0(t) = \frac{15\sqrt{13} + 33}{34\sqrt{13}} e^{(-\frac{9}{2} + \frac{\sqrt{13}}{2})t} + \frac{15\sqrt{13} - 33}{34\sqrt{13}} e^{(-\frac{9}{2} - \frac{\sqrt{13}}{2})t} + \frac{2}{17},$$

$$p_1(t) = \frac{24 - 3\sqrt{13}}{17\sqrt{13}} e^{(-\frac{9}{2} + \frac{\sqrt{13}}{2})t} - \frac{24 + 3\sqrt{13}}{17\sqrt{13}} e^{(-\frac{9}{2} - \frac{\sqrt{13}}{2})t} + \frac{6}{17},$$

$$p_2(t) = -\frac{9\sqrt{13} + 81}{34\sqrt{13}} e^{(-\frac{9}{2} + \frac{\sqrt{13}}{2})t} - \frac{9\sqrt{13} - 81}{34\sqrt{13}} e^{(-\frac{9}{2} - \frac{\sqrt{13}}{2})t} + \frac{9}{17}.$$

Обратим внимание на то, что, получив значения $p_0(t), p_1(t), p_2(t)$, мы можем ответить практически на все вопросы, касающиеся данной системы. Например, значение $p_0(t)$ — это вероятность того, что система пуста, величина $p_2(t)$ означает вероятность того, что оба прибора заняты, т.е. поступившая заявка получит отказ. Вероятность занятости хотя бы одного прибора задается суммой $p_1(t) + p_2(t)$ или разностью $1 - p_0(t)$, среднее число заявок в системе или, что то же самое, среднее число занятых приборов найдем, взяв: $0p_0(t) + 1p_1(t) + 2p_2(t)$.

Пример 2. Система с ожиданием.

Пусть имеются n приборов с одинаковой средней производительностью. Поступающая заявка, застающая занятыми все приборы, ждет *неограниченно* до тех пор, пока не обслужится.

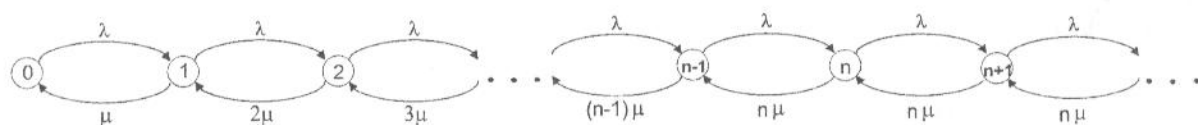
Предполагаем:

$$\lambda_k = \lambda \quad \text{при любом } k.$$

На основании рассуждений, приведенных в первом примере, для μ_k получаем следующие выражения:

$$\begin{aligned} \mu_k &= k\mu \quad \text{при } 0 \leq k \leq n, \\ \mu_k &= n\mu \quad \text{при } k > n. \end{aligned}$$

Тогда диаграмма будет выглядеть следующим образом:



Уравнения, управляющие изменением вероятностей состояний системы обслуживания во времени, запишутся так:

$$\left\{ \begin{aligned} p_0'(t) &= -\lambda p_0(t) + \mu p_1(t), \\ &\dots\dots\dots \\ p_k'(t) &= \lambda p_{k-1}(t) - (\lambda + k\mu)p_k(t) + (k+1)\mu p_{k+1}(t), \quad 1 \leq k \leq n-1, \\ &\dots\dots\dots \\ p_k'(t) &= \lambda p_{k-1}(t) - (\lambda + n\mu)p_k(t) + n\mu p_{k+1}(t), \quad k \geq n, \\ &\dots\dots\dots \\ \sum_{k=0}^{\infty} p_k(t) &= 1. \end{aligned} \right.$$

Пример 3. Замкнутая система.

Пусть r рабочих имеют одинаковую квалификацию и, в среднем, одинаковую производительность, равную μ (т.е. каждый рабочий за единицу времени ремонтирует в среднем μ станков). На заводе имеются n станков, которые время от времени выходят из строя и ремонтируются этими рабочими. Одновременно каждый рабочий ремонтирует не более одного станка. Если сломаны k станков, будем говорить, что система находится в состоянии E_k . Тогда система может иметь состояния E_0, E_1, \dots, E_n . Предполагается, что рабочих меньше, чем станков, т.е. $r < n$. Если работает только один из n станков и вероятность его поломки определить как $\lambda h + o(h)$, то в случае состояния E_k , т.е. работы $(n-k)$ станков, вероятность поломки какого-нибудь одного из них будет иметь вид:

$$p\{E_k(t) \rightarrow E_{k+1}(t+h)\} = (n-k)\lambda h + o(h), \quad k \leq n.$$

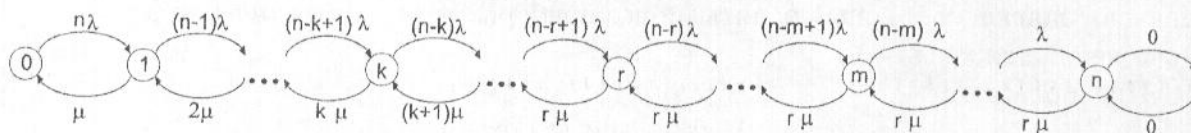
Если сломаны k станков, т.е. заняты k рабочих, то вероятность того, что освободится один из них:

$$p\{E_k(t) \rightarrow E_{k-1}(t+h)\} = k\mu h + o(h), \quad k \leq r.$$

Для $k = k+1, \dots, n$ имеем:

$$p\{E_k(t) \rightarrow E_{k-1}(t+h)\} = r\mu h + o(h).$$

Диаграмма интенсивностей переходов выглядит следующим образом:



а система уравнений имеет вид:

$$\begin{cases} p_0'(t) = -n\lambda p_0(t) + \mu p_1(t), \\ \dots \\ p_k'(t) = (n-k+1)\lambda p_{k-1}(t) - [\lambda(n-k) + k\mu]p_k(t) + (k+1)\mu p_{k+1}(t), & 1 \leq k \leq r-1, \\ \dots \\ p_k'(t) = (n-k+1)\lambda p_{k-1}(t) - [\lambda(n-k) + r\mu]p_k(t) + r\mu p_{k+1}(t), & r \leq k \leq n-1, \\ \dots \\ p_n'(t) = -r\mu p_n(t) + \lambda p_{n-1}(t), \\ \sum_{k=0}^n p_k(t) = 1. \end{cases}$$

Упражнения

1. Работают две СМО. В каждой системе имеются два прибора одинаковой средней производительности. Система M_1 – с потерями, с отсутствием очереди, в ней средняя интенсивность входного потока равна одной заявке в час, а средняя производительность каждого прибора равна трем заявкам в час. Система M_2 – с ожиданием, с неограниченной очередью, со средней интенсивностью входного потока, равной трем заявкам в час, и средней производительностью у каждого прибора, равной двум заявкам в час. Для каждой системы изобразите диаграмму интенсивностей и выпишите систему дифференциальных уравнений.

- 2. Рабочий обслуживает группу из шести автоматов. В среднем автомат останавливается один раз в час. Обслуживание одного автомата занимает у рабочего в среднем двенадцать минут. В каких состояниях может находиться данная система? К какому из рассмотренных в пункте 5 типов СМО она относится? Изобразите для данной системы диаграмму интенсивностей и составьте систему дифференциальных уравнений.

6. Входящий поток заявок и его свойства

Природа требований, поступающих в СМО, может быть самой разнообразной, однако чаще всего рассматривают *однородные* требования, различающиеся лишь моментами поступления в систему, поэтому входящий (входной) поток данной СМО можно рассматривать как множество моментов поступления в систему требований. Например, входящий поток станции скорой помощи - множество моментов телефонных вызовов врача.

Будем говорить, что происходит событие в момент τ , если в этот момент поступает требование. Тогда множество моментов, когда происходят события, называется потоком однородных событий. Заметим, что в действительности может случиться так, что поток событий является разнородным, например, когда заявка случайно оказывается двойной, тройной и т.д. (т.е. состоит из двух, трех и т.д. одновременно поступающих заявок, например заказы при покупке билетов в кинотеатре), или заявки различаются по каким-то другим признакам, определяющим разную дисциплину обслуживания в СМО.

Поток называется *регулярным*, если события следуют друг за другом через определенные фиксированные промежутки времени. Такой поток может рассматриваться как предельный случай и редко встречается в ТМО. В ТМО, как правило, приходится иметь дело со случайным потоком событий.

Среди различных свойств, которыми может обладать входящий поток СМО, особенно важны свойства стационарности, ординарности и свойство отсутствия последствия.

1. Поток называется *стационарным*, если вероятность попадания того или иного количества заявок в промежуток времени длины t зависит лишь от количества этих требований и длины промежутка t и не зависит от места расположения этого промежутка на временной оси, т.е. вероятность появления k заявок в промежутке $(\tau, \tau + t)$ является функцией только от k и t .

Свойство стационарности выражает неизменность вероятностного режима потока во времени. Кроме этого, стационарный поток характеризуется постоянным средним числом заявок, поступающих в единицу времени. На практике встречаются потоки заявок, которые на ограниченном участке времени являются стационарными. Например, поток вызовов скорой медицинской помощи между 10 и 12 часами. Этот же поток в течение суток не может считаться стационарным в обычное время (не в период эпидемий, когда он стационарен в течение суток), так как, например, после 2³⁰ до утра среднее число вызовов меньше, чем днем (по данным Ярославской станции скорой помощи). Поток вызовов междугородней телефонной станции в период между 10 и 13 часами может считаться стационарным и не может быть стационарным в течение суток (т.к. ночью вызовов меньше, чем днем). В действительности, реальные физические потоки

стационарны на ограниченном участке времени и распространение этого участка до бесконечности - лишь удобный прием, применяемый в целях упрощения анализа.

2. Поток называется *поток без последействия*, если для любых двух непересекающихся отрезков времени число заявок, поступающих в один из них, не зависит от числа заявок в другом отрезке, т.е. вероятность поступления k заявок за промежуток $(\tau, \tau + t)$ не зависит от того, сколько заявок и когда поступали до этого промежутка. Другими словами, условная вероятность поступления k требований за промежуток $(\tau, \tau + t)$ при любом предположении о поступлениях заявок до этого промежутка совпадает с безусловной вероятностью поступления k требований за указанный промежуток.

Условие отсутствия последействия – очень сильное условие, означающее, что заявки поступают в систему независимо друг от друга. Примером потока, обладающего последействием, может служить регулярный поток. В качестве потока без последействия можно рассматривать поток пассажиров, являющихся на трамвайную остановку, расположенную в центре города, а не рядом с предприятиями. Тогда причины, обуславливающие приход отдельного пассажира в тот, а не в другой момент времени, не связаны с аналогичными причинами для других пассажиров.

Заметим, что выходной из системы поток обычно обладает последействием, даже если входной его не имел. Рассмотрим СМО с одним обслуживающим прибором (*одноканальную*), для которой время обслуживания одной заявки определено и равно $t_{\text{обсл.}}$. Тогда минимальный интервал времени между заявками, покидающими СМО, равен $t_{\text{обсл.}}$. Наличие данного интервала уже приводит к последействию. Если известно, что заявка в какой-то момент t_1 покинула систему, то следующая обслуженная заявка появится не ранее чем в момент $t_1 + t_{\text{обсл.}}$, т.е. будет зависимость между числом заявок на непересекающихся отрезках времени. Последействие, присущее выходному потоку, необходимо учитывать в многофазных СМО, когда выходной поток для одной фазы системы является входным для другой.

3. Поток называется *ординарным*, если вероятность попадания на малый промежуток времени h двух или большего числа заявок бесконечно мала по сравнению с h , т.е

$$\lim_{h \rightarrow 0} \frac{p_{>1}(h)}{h} = 0 \quad \text{или} \quad p_{>1}(h) = o(h).$$

Примером ординарного потока может служить процесс рождения, изученный ранее (вспомним, что этот процесс характеризуется тем, что вероятность рождения более чем одной особи за малый промежуток времени есть величина бесконечно малая по сравнению с длиной этого промежутка). Отсюда следует, что вероятность появления за промежуток времени h двух или более заявок бесконечно мала по сравнению с вероятностью появления одной заявки. Действительно, в процессе рождения вероятность появления одной заявки за время h равна $\lambda h + o(h)$, тогда:

$$\lim_{h \rightarrow 0} \frac{p_{>1}(h)}{\lambda h + o(h)} = \lim_{h \rightarrow 0} \frac{p_{>1}(h)}{\lambda h} = \frac{1}{\lambda} \lim_{h \rightarrow 0} \frac{p_{>1}(h)}{h} = 0.$$

Ординарность потока означает невозможность одновременного поступления нескольких заявок. Например, поток клиентов в парикмахерскую - ординарный поток, а поток в ЗАГС для регистрации брака - неординарный. Однако, если в неординарном потоке заявки однородные (т.е. все - пары, как в последнем примере, или все заявки - тройки, и т.п.), то этот поток можно свести к ординарному, рассмотрев поток пар, троек и т.п.

Если $K(t)$ - число требований, поступающих в систему за промежуток $(0, t)$, то график изменения этой случайной величины есть график неубывающей целой функции от t (см. рис. 4).

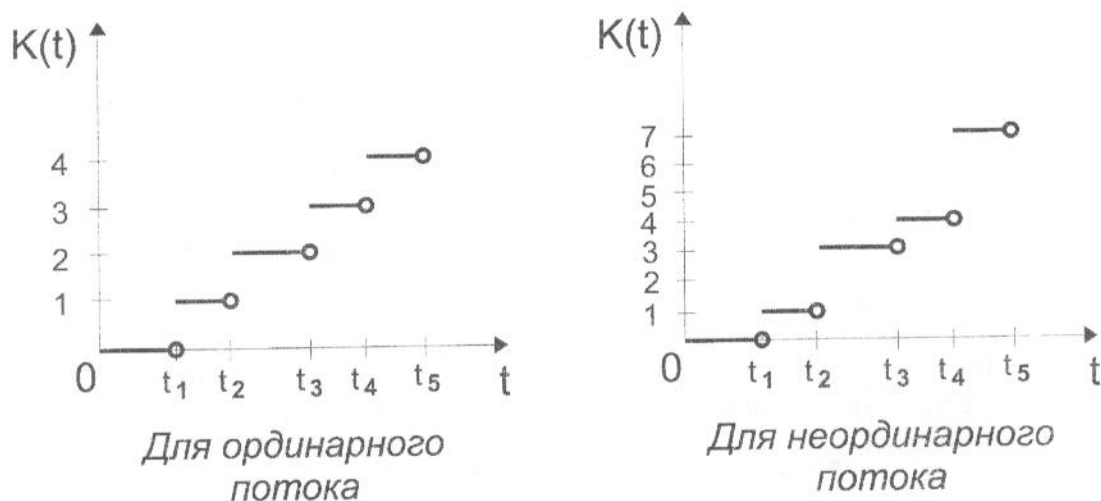


Рис. 4

Если поток обладает всеми тремя перечисленными свойствами, то он называется *простейшим*, или *пуассоновским потоком*.

Зададим вероятность появления одной заявки за малый промежуток времени $(t, t + h)$ длины h следующей формулой:

$$\tilde{p}_1(t) = \lambda h + o(h). \quad (6.1)$$

Заметим, что $\tilde{p}_1(t)$ означает не вероятность наличия одной заявки в системе в момент времени t , а именно вероятность появления одной заявки, при этом до этого заявок могло быть сколько угодно и прибавилась еще одна, т.е. система из состояния $E_k(t)$ перешла в состояние $E_k(t + h)$ при любом k ($E_k(t) \rightarrow E_{k+1}(t + h)$).

Из формулы (6.1) и условий, налагаемых на процесс гибели и размножения, следует выполнение всех трех свойств простейшего потока для процесса чистого размножения, рассматриваемого как поток. Поскольку процесс чистого размножения определяется распределением Пуассона (см. п. 4, формула 4.1), то простейший поток часто называют пуассоновским.

Пуассоновский поток среди потоков играет особую роль, с ним сталкиваются значительно чаще, чем это можно предположить.

Оказывается, что при взаимном наложении большого числа независимых стационарных ординарных потоков с любым последствием получается поток, сколь угодно близкий к простейшему. Требуется лишь, чтобы складываемые потоки оказывали на сумму равномерно малое влияние (имели малую интенсивность, т.е. малый вклад каждого потока в общую сумму) (теорема А.Я. Хинчина).

Суммирование понимается в том смысле, что все моменты поступления заявок в каждом потоке сносятся на одну ось (как суммарный поток можно рассматривать, например, поток автомашин на дороге с достаточно мощным движением или поток требований, образующих общий спрос на конкурентном рынке). Пусть потоки $\Pi_1, \Pi_2, \dots, \Pi_n$ сравнимы по своему влиянию на суммарный поток (но не сравнимы по интенсивности со всем суммарным потоком). Тогда поток $\sum_{k=1}^n \Pi_k$ стационарен и ординарен, а последствие при росте n должно слабеть. „Удельный вес“ точек каждого отдельного потока с ростом n будет убывать. Каждая из точек, попадающих в два непересекающихся временных отрезка, случайным образом может оказаться принадлежащей тому или иному потоку, остальные точки тоже случайным образом принадлежат разным потокам и появляются на этих двух отрезках независимо друг от друга.

Практически, оказывается, достаточно сложить 4-5 потоков, чтобы получить поток, с которым можно работать как с пуассоновским.

Упражнения

- 1. Приведите примеры конкретных СМО, имеющих следующие входящие потоки: регулярный, стационарный, ординарный, без последствия.
- 2. Какими свойствами обладает поток посетителей ресторана?
- 3. Пусть ось времени разбита на примыкающие друг к другу равные маленькие отрезки длиной δ , стремящейся к нулю. На каждом отрезке производится испытание, которое оканчивается удачей с вероятностью p и неудачей с вероятностью $q = 1 - p$. Можно ли тогда последовательность отрезков, соответствующих удаче, рассматривать как последовательность требований из пуассоновского потока?
- 4. В аэропорт прибывает пуассоновский поток самолетов, в среднем 2 самолета за 5 минут. Найти вероятность того, что за 15 минут придут 3 самолета.
- 5. По шоссе мимо наблюдателя движется в одном направлении пуассоновский поток машин. Известно, что вероятность отсутствия машин в течение 5 минут равна 0,5. Требуется найти вероятность того, что за 10 минут мимо наблюдателя пройдет не более 2 машин.

7. Основные характеристики пуассоновского потока

Процесс Пуассона будем рассматривать как модель входящего в СМО потока. Тогда $P_k(t)$ означает вероятность появления k требований за промежуток времени $(0, t)$.

Зависимость $P_k(t)$ от k и от λt может быть представлена следующим образом:

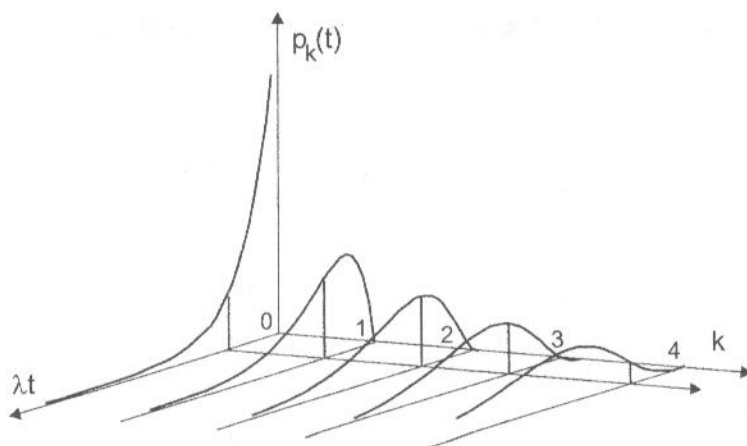


Рис. 5

Эта же зависимость от t при $\lambda = 1$ имеет вид:

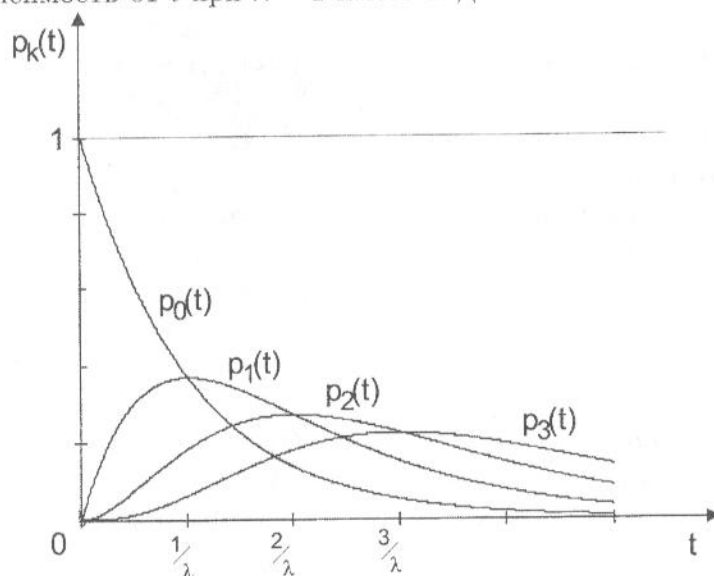


Рис. 6

В любой точке t (λt) выполняется равенство:

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

Это означает, что в любой точке t (λt) сумма всего бесконечного семейства графиков равна единице, т.е. суммарный график (см. рис. 6) представляет собой горизонтальную прямую, проходящую на высоте, равной единице.

Пусть K – дискретная случайная величина – число заявок, появившихся за промежуток $(0, t)$. Найдем ее математическое ожидание $M(K)$:

$$\begin{aligned} M(K) &= \sum_{k=0}^{\infty} k p_k(t) = \sum_{k=0}^{\infty} k \frac{(\lambda t)^k}{k!} e^{-\lambda t} = e^{-\lambda t} \lambda t \sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} = e^{-\lambda t} \lambda t \sum_{m=0}^{\infty} \frac{(\lambda t)^m}{m!} = \\ &= \lambda t e^{-\lambda t} e^{\lambda t} = \lambda t. \end{aligned}$$

Итак,

$$M(K) = \lambda t. \quad (7.1)$$

Вычислим дисперсию $D(K)$ данной случайной величины по формуле: $D(K) = M(K^2) - M^2(K)$.

Найдем $M(K^2)$ – начальный момент второго порядка:

$$\begin{aligned} M(K^2) &= \sum_{k=0}^{\infty} k^2 p_k(t) = \sum_{k=0}^{\infty} k^2 \frac{(\lambda t)^k}{k!} e^{-\lambda t} = \lambda t \sum_{k=1}^{\infty} k \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t} = \\ &= \lambda t \sum_{k=1}^{\infty} (k+1-1) \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t} = \lambda t e^{-\lambda t} \left(\sum_{k=1}^{\infty} (k-1) \frac{(\lambda t)^{k-1}}{(k-1)!} + \sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!} \right) = \\ &= \lambda t e^{-\lambda t} \left(\lambda t \sum_{m=0}^{\infty} \frac{(\lambda t)^m}{m!} + e^{\lambda t} \right) = \lambda t e^{-\lambda t} (\lambda t e^{\lambda t} + e^{\lambda t}) = \lambda t (\lambda t + 1). \end{aligned}$$

Тогда:

$$D(K) = \lambda t (\lambda t + 1) - (\lambda t)^2 = \lambda t.$$

$$D(K) = \lambda t. \quad (7.2)$$

Таким образом, и среднее значение, и дисперсия для пуассоновского распределения одинаковы и равны λt .

Рассмотрим теперь распределение длины промежутков между соседними заявками пуассоновского потока.

Пусть τ – случайная величина, равная интервалу времени между соседними заявками, поступающими в СМО. Ее функцию распределения и плотность обозначим соответственно через $A(t)$ и $a(t)$. Тогда, по определению плотности, $a(t)h + o(h)$ есть вероятность того, что следующая заявка поступит не ранее чем через t секунд после последнего пришедшего требования и не позднее чем через $t + h$ секунд, где h – достаточно мало.

$A(t)$ есть вероятность того, что $\tau \leq t$, т.е. $A(t) = p(\tau \leq t)$ – вероятность того, что время между соседними заявками не больше t . Тогда $A(t) = 1 - p(\tau > t)$. Но $p(\tau > t)$ означает вероятность того, что за промежуток $(0, t)$ не пришло ни одной заявки, т.е. $p(\tau > t) = p_0(t) = e^{-\lambda t}$ (из формулы 4.1). Отсюда

$$A(t) = 1 - e^{-\lambda t}, \text{ где } t \geq 0.$$

Тогда

$$a(t) = A'(t) = \lambda e^{-\lambda t}, \text{ где } t \geq 0.$$

Это так называемое *показательное распределение*.

Итак, пуассоновский поток заявок характеризуется *показательным распределением промежутков времени между моментами поступления этих заявок*.

Обратим внимание на то, что показательное распределение характеризует *непрерывную* случайную величину, в отличие от пуассоновского, характеризующего *дискретную* случайную величину.

Рассмотрим математическое ожидание случайной величины τ . Поскольку это непрерывная случайная величина, то

$$\begin{aligned} M(\tau) &= \int_0^{\infty} t a(t) dt = \lambda \int_0^{\infty} e^{-\lambda t} t dt = \lambda \left[-\frac{t}{\lambda} e^{-\lambda t} \right]_0^{\infty} + \frac{1}{\lambda} \int_0^{\infty} e^{-\lambda t} dt = \lambda \left[-\frac{t}{\lambda} e^{-\lambda t} \right]_0^{\infty} + \\ &+ \frac{1}{\lambda} \left(-\frac{1}{\lambda} e^{-\lambda t} \right) \Big|_0^{\infty} = \frac{1}{\lambda}. \end{aligned}$$

$$\boxed{M(\tau) = \frac{1}{\lambda}.} \quad (7.3)$$

Итак, среднее время между соседними заявками равно $\frac{1}{\lambda}$, где λ – интенсивность входного потока. Заметим, что этот результат был очевиден с самого начала, поскольку λ – среднее число заявок, поступающих в единицу времени.

Найдем дисперсию случайной величины τ по формуле: $D(\tau) = M(\tau^2) - M^2(\tau)$. $M(\tau^2)$ – начальный момент второго порядка, который для непрерывной случайной величины имеет вид: $M(\tau^2) = \int_{-\infty}^{\infty} t^2 f(t) dt$, где $f(t)$ – плотность.

Тогда в нашем случае:

$$M(\tau^2) = \lambda \int_0^{\infty} t^2 e^{-\lambda t} dt = \lambda \left[-\frac{t^2}{\lambda} e^{-\lambda t} \Big|_0^{\infty} + \frac{2}{\lambda} \int_0^{\infty} e^{-\lambda t} t dt \right] =$$

$$= \lambda \left[-\frac{t^2}{\lambda} e^{-\lambda t} \Big|_0^{\infty} + \frac{2}{\lambda} \left(-\frac{t}{\lambda} e^{-\lambda t} - \frac{1}{\lambda^2} e^{-\lambda t} \right) \Big|_0^{\infty} \right] = \frac{2}{\lambda^2}.$$

$$D(\tau) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

$$\boxed{D(\tau) = \frac{1}{\lambda^2}}. \quad (7.4)$$

Замечательной чертой показательного распределения является отсутствие последствия, т.е. закон распределения вероятностей для интервала времени между соседними заявками совпадает с законом распределения вероятностей для интервала времени между произвольным фиксированным моментом t_0 и ближайшим моментом поступления заявки. Другими словами, если промежуток времени, распределенный по показательному закону, уже длился некоторое время t_0 , то это никак не влияет на распределение оставшейся части промежутка, закон распределения будет таким же, как и для всего промежутка T .

Действительно, пусть T – случайный промежуток, функцией распределения которого является $A(t) = 1 - e^{-\lambda t}$. Пусть T уже продолжается некоторое время t_0 , т.е. произошло событие $T > t_0$. Найдем при этом предположении закон распределения оставшейся части $T - t_0$. Обозначим его через $A^{t_0}(t)$, т.е. $A^{t_0}(t) = p(T - t_0 < t \mid T > t_0)$. Докажем, что $A^{t_0}(t)$ не зависит от t_0 .

Рассмотрим вероятность произведения событий: $T - t_0 < t$ и $T > t_0$, т.е.

$$p((T - t_0 < t) \cdot (T > t_0)) = p(T > t_0) \cdot p(T - t_0 < t \mid T > t_0) = p(T > t_0) \cdot A^{t_0}(t).$$

Отсюда:

$$A^{t_0}(t) = \frac{p((T - t_0 < t) \cdot (T > t_0))}{p(T > t_0)}.$$

Но произведение событий $T - t_0 < t$ и $T > t_0$ означает $t_0 < T < t + t_0$, а тогда его вероятность равна: $A(t + t_0) - A(t_0)$, причем $p(T > t_0) = 1 - p(T \leq t_0) = 1 - A(t_0)$, тогда:

$$A^{t_0}(t) = \frac{A(t + t_0) - A(t_0)}{1 - A(t_0)} = \frac{1 - e^{-\lambda(t+t_0)} - 1 + e^{-\lambda t_0}}{1 - 1 + e^{-\lambda t_0}} = 1 - e^{-\lambda t}.$$

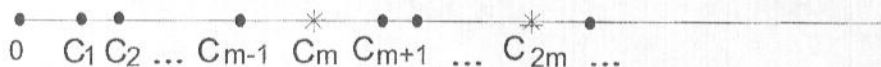
Упражнения

- 1. Покажите, что точки максимума функций $p_k(t)$ с ростом k ($k \geq 1$) образуют арифметическую прогрессию.
- 2. Докажите формулы (7.1) и (7.2).

- 3. Пусть имеется пуассоновский поток с параметром $\lambda = 3$ (минуты). Найдите вероятность того, что длина интервала между соседними требованиями составляет от двух до трех минут.

8. Распределение Эрланга

Рассмотрим новую случайную величину – промежуток времени между моментами поступления в пуассоновском потоке двух требований, не следующих непосредственно одно за другим, а с *пропуском* $m - 1$ требований. Иначе говоря, „просеиваем в пуассоновском потоке“ $m - 1$ требование и учитываем лишь требование c_m , затем вновь „просеиваем“ $m - 1$ требование, берем требование c_{2m} и т.д.:



Таким образом, возникает новый поток, состоящий из требований c_m, c_{2m}, \dots . Остальные „просеянные“ требования как бы забыты.

Заметим, что случайная величина T_i принимает значения промежутков времени между моментами поступления заявок c_{i-1} и c_i , между c_{m+i-1} и c_{m+i} , между c_{2m+i-1} и c_{2m+i} , и т.д. Для T_1 момент поступления $c_{i-1} = c_0$ есть начало отчета времени.

Найдем закон распределения суммы:

$$T = T_1 + T_2 + \dots + T_m = \sum_{k=1}^m T_k.$$

Случайные величины T_k независимы и имеют одну и ту же плотность $a(t) = \lambda e^{-\lambda t}$. Преобразование Лапласа плотности $\lambda e^{-\lambda t}$ имеет вид: $\lambda \int_0^\infty e^{-\lambda t} e^{-st} dt$. С другой стороны, этот интеграл можно рассматривать как математическое ожидание случайной величины e^{-sT_k} :

$$M(e^{-sT_k}) = \lambda \int_0^\infty e^{-(\lambda+s)t} dt = -\frac{\lambda}{\lambda+s} e^{-(\lambda+s)t} \Big|_0^\infty = \frac{\lambda}{\lambda+s}, \quad \text{где } k = 1, \dots, m.$$

$$M(e^{-sT}) = M(e^{-s \sum_{k=1}^m T_k}) = M(e^{-sT_1} \cdot e^{-sT_2} \dots e^{-sT_m}) = \left(\frac{\lambda}{\lambda+s} \right)^m.$$

Последнее равенство выполняется, так как математическое ожидание от произведения случайных независимых величин равно произведению математических ожиданий этих величин.

Преобразование Лапласа плотности $a_m(t)$ равно $\int_0^\infty a_m(t) e^{-st} dt$, этому же интегралу равно математическое ожидание случайной величины e^{-sT} , поскольку функции плотности случайных величин e^{-sT} и T совпадают, т.е.

$$M(e^{-sT}) = \left(\frac{\lambda}{\lambda+s} \right)^m = \int_0^\infty a_m(t) e^{-st} dt.$$

Тогда $M(e^{-sT})$ есть образ, а соответствующий оригинал этого образа по таблицам преобразования Лапласа имеет вид:

$$a_m(t) = \frac{\lambda(\lambda t)^{m-1}}{(m-1)!} e^{-\lambda t}, \quad t \geq 0. \quad (8.1)$$

$a_m(t)$ есть плотность вероятностей для интервала времени между двумя непосредственно следующими друг за другом моментами поступления заявок c_m, c_{2m}, \dots , т.е. рассматривается новый поток. Его называют *поток Эрланга*. Если в простейшем потоке „просеиваются“ каждые $m - 1$ точки, т.е. сохраняется каждая m -я точка, то поток называется *поток Эрланга $(m - 1)$ -го порядка*. При $m = 1$ для плотности получаем $\lambda e^{-\lambda t}$, т.е. обычное показательное распределение. Итак, поток Эрланга – это пуассоновский поток с „просеиванием“. Про формулу (8.1) говорят, что она задает распределение Эрланга.

При $m > 1$ распределение Эрланга достигает наибольшего значения в точке $t > 0$:

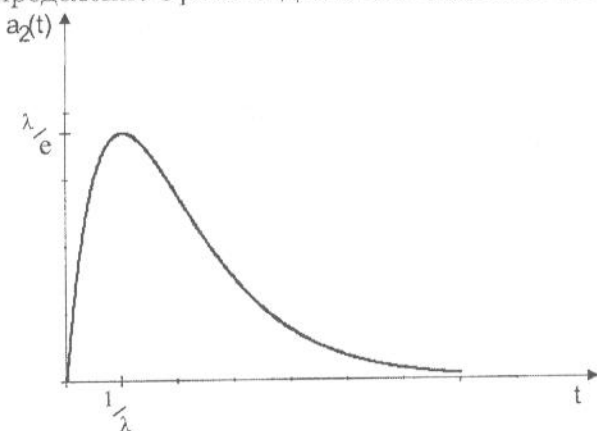


Рис. 7

В случае, когда $m = 1$, наибольшее значение достигается при $t = 0$:

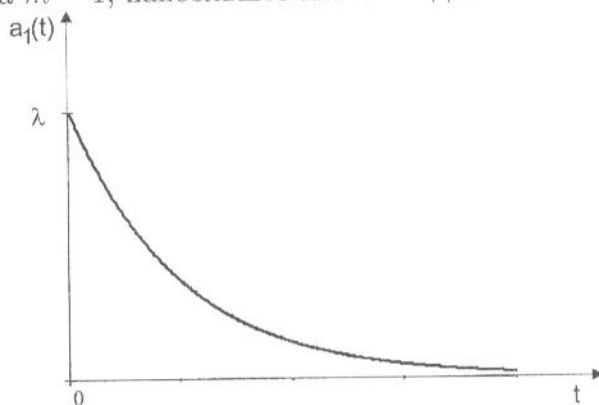


Рис. 8

Пример.

В СМО, содержащую 2 прибора, поступают заявки в пуассоновском потоке, причем, заявки с нечетными номерами обслуживаются первым прибором, а с четными – вторым. Тогда плотности вероятностей длин интервалов времени между моментами поступления заявок на каждый из этих приборов определяются из (8.1) при $m = 2$ (пропускается или „просеивается“ одна заявка).

В эрланговском потоке мы уже говорим о распределении плотности вероятности для интервала времени между соседними заявками (все „просеянные“ забыты). Так как $M(T_k) = \frac{1}{\lambda}$ и $D(T_k) = \frac{1}{\lambda^2}$ (см. показательное распределение), то:

$$M\left(\sum_{k=1}^m T_k\right) = \frac{m}{\lambda}, \quad D\left(\sum_{k=1}^m T_k\right) = \frac{m}{\lambda^2}.$$

При $m \rightarrow \infty$ математическое ожидание и дисперсия интервала времени между момен-

Упражнения

- 1. Какой поток называют потоком Эрланга $(m - 1)$ -го порядка?
- 2. По шоссе в одном направлении движется пуассоновский поток машин с интенсивностью две машины в минуту. Пост ДПС на развилке направляет каждую третью машину в объезд по боковой дороге. Найдите среднюю длину интервала между машинами на боковой дороге.
- 3. Охарактеризуйте процесс чистой гибели и выпишите соответствующую систему дифференциальных уравнений.

9. Системы, описываемые процессами гибели и размножения, в стационарном режиме

Систему массового обслуживания можно описать с помощью процесса гибели и размножения. Например, имеется приемная врача, состоящая из очереди в данный кабинет и обслуживания врачом в кабинете. Момент прихода пациента в очередь можно рассматривать как поступление заявки в систему или как рождение нового члена популяции. Популяция в данном случае - совокупность всех людей, ожидающих приема, а также и тех, кто уже находится в кабинете у врача. Уход пациента после осмотра врачом можно интерпретировать как уход заявки из системы или гибель одного члена популяции.

Рассмотрим случай, когда СМО работает в стационарном режиме, т.е. *когда с ростом t вероятности $p_k(t)$ становятся постоянными* или, что то же самое, когда существуют пределы

$$\lim_{t \rightarrow \infty} p_k(t) = p_k,$$

где p_k - вероятность того, что в *произвольный момент* достаточно отдаленного будущего в системе будет находиться k заявок. То, что p_k не зависит от времени, *не означает*, что в предельном случае процесс не может переходить из одного состояния в другое. Число членов популяции изменяется во времени, но *вероятность* пребывания системы через *достаточно большое* время в состоянии E_k равна величине p_k .

Поскольку p_k являются константами, то $p_k'(t)$ равны нулю, и, значит, система дифференциальных уравнений (3.2) превращается в алгебраическую, которая имеет вид:

$$\begin{cases} 0 = -\lambda_0 p_0 + \mu_1 p_1, & \text{при } k = 0, \\ 0 = \lambda_{k-1} p_{k-1} - (\lambda_k + \mu_k) p_k + \mu_{k+1} p_{k+1}, & \text{при } k \geq 1. \end{cases} \quad (9.1)$$

Эти уравнения для стационарного режима, так же как и ранее, могут быть получены непосредственно из диаграммы переходов.

Из данной системы легко установить, что в стационарном режиме поток должен удовлетворять условию *сохранения* в том смысле, что входящий поток должен быть „равен“ выходящему, т. е.

$$\lambda_k p_k = \mu_{k+1} p_{k+1}, \text{ где } k \geq 0. \quad (9.2)$$

Обратно, из (9.2) следует (9.1). Действительно, запишем (9.2) в однородном виде:

$$0 = -\lambda_k p_k + \mu_{k+1} p_{k+1},$$

добавим $\lambda_{k-1}p_{k-1}$ и вычтем $\mu_k p_k$, каждое уравнение системы при этом не изменится, так как $\lambda_{k-1}p_{k-1} = \mu_k p_k$, тогда получим:

$$0 = \lambda_{k-1}p_{k-1} - \lambda_k p_k - \mu_k p_k + \mu_{k+1}p_{k+1}.$$

Тогда:

$$0 = \lambda_{k-1}p_{k-1} - (\lambda_k + \mu_k)p_k + \mu_{k+1}p_{k+1}. \quad (9.3)$$

(9.3) совпадает с (9.1), т.е. система (9.2) равносильна системе (9.1). Следовательно, решив систему разностных уравнений, мы получим решение системы (9.1).

Найдем решения p_k ($k \geq 1$), выразив их через p_0 .

Из (9.2) для p_1 имеем:

$$p_1 = \frac{\lambda_0}{\mu_1} p_0.$$

Воспользуемся методом математической индукции. Предположим, что в случае номера k p_k имеет вид:

$$p_k = \frac{\lambda_0 \dots \lambda_{k-1}}{\mu_1 \dots \mu_k} p_0.$$

Из (9.2):

$$p_{k+1} = \frac{\lambda_k}{\mu_{k+1}} p_k.$$

Тогда:

$$p_{k+1} = \frac{\lambda_0 \dots \lambda_{k-1} \lambda_k}{\mu_1 \dots \mu_k \mu_{k+1}} p_0.$$

Таким образом:

$$p_k = \frac{\lambda_0 \dots \lambda_{k-1}}{\mu_1 \dots \mu_k} p_0, \quad \text{или}$$

$$p_k = \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} p_0, \quad k = 1, \dots \quad (9.4)$$

Это решение системы (9.1) в установившемся стационарном режиме. Из (9.4) следует, что $p_0 > 0$, так как в противном случае вероятности нахождения системы в любом конечном состоянии будут равны нулю, т.е. система всегда будет иметь бесконечное множество заявок и нельзя будет описать ее характеристики. Требование $p_0 > 0$ означает, что система должна обязательно время от времени опустошаться, только тогда и может быть ненулевое значение вероятности p_0 .

Все вероятности выражаются через константу p_0 . Найдем ее:

$$\sum_{k=0}^{\infty} p_k = 1, \quad \text{тогда}$$

$$p_0 + \frac{\lambda_0}{\mu_1} p_0 + \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} p_0 + \dots + \frac{\lambda_0 \lambda_1 \dots \lambda_{k-1}}{\mu_1 \mu_2 \dots \mu_k} p_0 + \dots = 1,$$

$$p_0 \left(1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right) = 1.$$

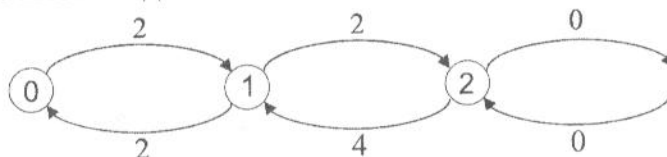
Отсюда:

$$p_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}} \quad (9.5)$$

Отличие p_0 от нуля означает необходимость сходимости ряда $\sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}$. Следовательно, для стационарного режима, начиная с некоего номера k : $\frac{\lambda_k}{\mu_{k+1}} < 1$.

Пример.

Рассмотрим СМО с потерями, с отсутствием очереди, для которой $n = 2$, $\lambda = 2$, $\mu = 2$, работающую в стационарном режиме. Для нее диаграмма интенсивностей переходов будет иметь вид:



Тогда получим:

$$\begin{cases} p_1 = \frac{2}{2} p_0, \\ p_2 = \frac{2 \cdot 2}{2 \cdot 4} p_0, \\ p_0 + p_1 + p_2 = 1, \end{cases}$$

отсюда: $p_0 = \frac{2}{5}$, $p_1 = \frac{2}{5}$, $p_2 = \frac{1}{5}$.

Найдем среднее число заявок в системе по формуле:

$$N = 0 \cdot p_0 + 1 \cdot p_1 + 2 \cdot p_2.$$

Тогда:

$$N = 0 \cdot \frac{2}{5} + \frac{2}{5} + 2 \cdot \frac{1}{5} = 0,8.$$

В пункте 5 рассмотрен пример аналогичной задачи для СМО, работающей не в стационарном режиме и имеющей параметры: $n = 2$, $\mu = 1$, $\lambda = 3$. При $t \rightarrow \infty$ соответствующие вероятности будут равны: $p_0 = \frac{2}{17}$, $p_1 = \frac{6}{17}$, $p_2 = \frac{9}{17}$. Тогда $N =$

$$= 1 \cdot \frac{6}{17} + 2 \cdot \frac{9}{17} = 1 \frac{7}{17} - \text{математическое ожидание числа заявок в СМО, работающей}$$

в стационарном режиме или, что то же самое, среднее число занятых приборов. В случае СМО, работающей в стационарном режиме, и имеющей параметры: $n = 2$, $\lambda =$

$$= 2$$

$= 2$, $\mu = 1$ (см. пункт 5), соответствующие вероятности равны: $p_0 = \frac{1}{5}$, $p_1 = \frac{2}{5}$, $p_2 = \frac{2}{5}$, а среднее число заявок в системе $= \frac{6}{5}$.

Заметим, что, изобразив графики функций $p_0(t)$, $p_1(t)$, $p_2(t)$, $N(t)$, можно определить, в какой момент времени работы системы устанавливается практически стационарный режим. В случае СМО с потерями, имеющей параметры: $n = 2$, $\lambda = 3$, $\mu = 1$, можно считать, что это случится в момент $t = 1.8$ (см. рис. 9).

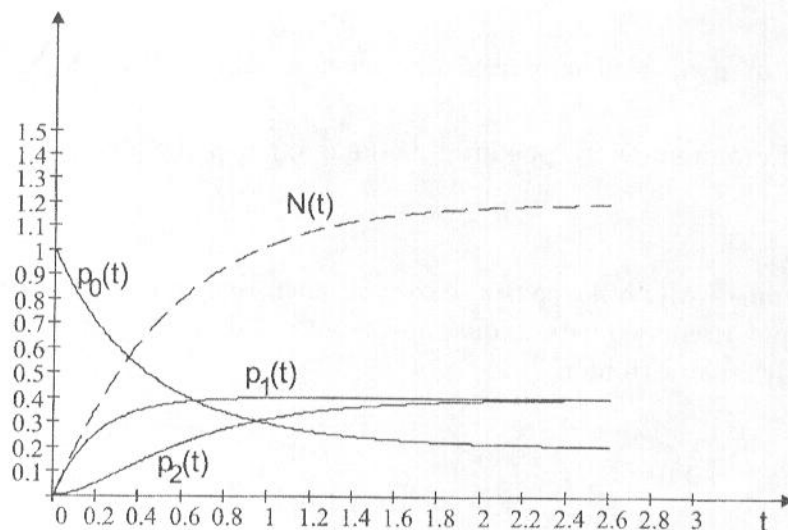


Рис. 9

Полученные для рассмотренных СМО данные можно систематизировать в следующей таблице:

n	λ	μ	p_0	p_1	p_2	N	ρ
2	2	2	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{4}{5}$	$\frac{1}{2}$
2	2	1	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{6}{5}$	1
2	3	1	$\frac{2}{17}$	$\frac{6}{17}$	$\frac{9}{17}$	$1\frac{7}{17}$	$\frac{3}{2}$

В последнем столбце стоят значения коэффициента использования для каждой СМО. Таким образом, проследив зависимость N от ρ , можно сделать следующее естественное предположение: чем больше нагрузка системы (ρ), тем больше в среднем число занятых приборов в системе.

Кроме случайных величин, характеризующих входной поток, в СМО исследуется еще одна непрерывная случайная величина – время обслуживания одной заявки ($T_{\text{обсл}}$). Обозначим ее функцию распределения $A(t)$:

$$A(t) = p(T_{\text{обслуж}} \leq t)$$

Для практики особый интерес представляет случай, когда эта случайная величина распределена по показательному закону, т.е.

$$a(t) = \mu e^{-\mu t} \quad (t > 0), \quad \text{где}$$

μ – величина, обратная среднему времени обслуживания одной заявки:

$$\mu = \frac{1}{M(T_{\text{обслуж}})}.$$

На практике существуют условия, в которых время обслуживания действительно распределяется по закону, близкому к показательному. Это прежде всего относится к задачам, где обслуживание состоит из ряда „попыток“, каждая из которых приводит к необходимому результату с какой-то вероятностью p .

Например, идет обстрел какой-то цели и обслуживание заканчивается в момент ее поражения. Обстрел ведется независимыми выстрелами со средней скорострельностью λ выстрелов в единицу времени. Каждый выстрел поражает цель с вероятностью, равной p . Предположим, что выстрелы происходят в случайные моменты

времени и образуют простейший поток с плотностью λ . Тогда поток выстрелов разделится на успешные и безуспешные. Поток успешных выстрелов тоже будет простейшим с плотностью $\Lambda = \lambda p$ (каждый выстрел независимо от других может стать поражающим с вероятностью p). Вероятность того, что цель будет поражена до момента t , равна:

$$A(t) = p(T_{\text{обслуж}} < t) = 1 - p(T_{\text{обслуж}} \geq t) = 1 - e^{-\lambda p t}.$$

Тогда $a(t) = \lambda p e^{-\lambda p t}$, т.е. имеем показательное распределение с параметром λp .

К „показательному“ типу обслуживания можно отнести и обслуживание по устранению неисправностей технических устройств, когда поиски неисправной детали осуществляются рядом проверок; задачи, когда „обслуживание“ заключается в обнаружении какого-либо объекта радиолокатором, если объект с определенной вероятностью может быть обнаружен при каждом цикле обзора. Например, считается, что лучший метод отыскания неисправностей в телевизоре - метод проб. Практика показывает, что если при отыскании неисправностей в сложных системах пытаются осмыслить логически причину, проследить всю электрическую цепь прохождения сигналов, то этот метод требует значительно большего времени, чем метод проб, конечно, с обдумыванием наиболее удачных проб (например, метод направленного перебора). Последний - наиболее общепринятый у настройщиков электронной аппаратуры. Если пробы независимы, то поток попыток - простейший, и поток успешных попыток - тоже простейший.

Показательным законом хорошо описываются и те случаи, когда плотность распределения времени обслуживания убывает при возрастании t . Это происходит тогда, когда основная масса заявок обслуживается очень быстро и значительные задержки в обслуживании наблюдаются редко. Например, в киоске подавляющее большинство покупателей обслуживается быстро, покупая газету, и лишь иногда происходит задержка, когда начинают покупать журнал (смотрят его перед покупкой). Аналогично, в сбербанке (долго - если аккредитив или завещание) или на почте (покупка марок и конвертов и отправление заказного письма или перевода).

Оказывается, что пропускная способность системы сравнительно мало зависит от закона распределения времени обслуживания, а зависит, главным образом, от среднего значения времени обслуживания. Поэтому в теории массового обслуживания часто пользуются допущением, что время обслуживания распределено по показательному закону, так как это позволяет упростить математический аппарат.

Упражнения

- 1. Охарактеризуйте СМО, работающую в стационарном режиме.
- 2. Докажите формулы (9.4) и (9.5).
- 3. Система с двумя приборами, имеющими среднюю одинаковую производительность – две заявки в минуту, и среднюю интенсивность входного потока, равную трем заявкам в минуту, работает в стационарном режиме. Изобразить диаграмму интенсивности переходов и найти соответствующие вероятности состояний, если в очереди может стоять не более одной заявки. Найти вероятность того, что:
 - а) заявка получит отказ,
 - б) заявке придется стоять в очереди.

10. Классическая система массового обслуживания: M/M/1

Сначала обратимся к аббревиатуре рассматриваемой системы.

В обозначении СМО часто используется аббревиатура следующего вида:

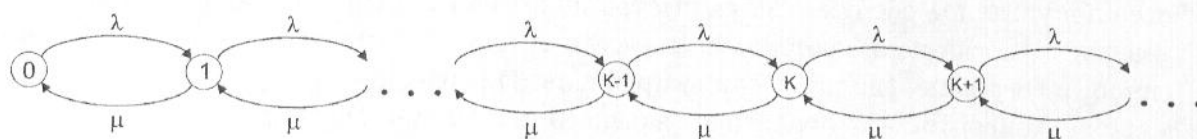
Буква / Буква / Число / Число.

Первая буква обозначает функцию распределения заявок входного потока, в частности, для показательного распределения используется буква M , вторая буква – закон распределения времени обслуживания одной заявки, первое число обозначает число приборов в системе, второе – объем накопителя. в некоторых случаях используются еще дополнительные символы, например, число входных потоков, и т.д. Следовательно, классическая система – это система с показательным входным потоком, показательным временем обслуживания заявок и одним прибором. Заметим, что, при всей ее простоте, она может без больших погрешностей применяться для расчета основных характеристик многих систем массового обслуживания.

Пусть в процессе размножения и гибели все λ_k и все μ_k равны, соответственно, λ и μ . Этот процесс размножения и гибели представляет систему массового обслуживания с одним обслуживающим прибором, пуассоновским входным потоком и показательным законом распределения времени обслуживания, т.е.:

$$\begin{cases} \lambda_k = \lambda, \\ \mu_k = \mu, \end{cases} \quad \begin{cases} A(t) = 1 - e^{-\lambda t}, \\ B(x) = 1 - e^{-\mu x}. \end{cases} \quad (10.1)$$

Диаграмма интенсивности переходов выглядит следующим образом:



Очевидно, средняя длина промежутка времени между соседними требованиями равна $\frac{1}{\lambda}$, среднее время обслуживания равно $\frac{1}{\mu}$, так как обе случайные величины распределены по показательному закону.

Рассмотрим функционирование системы в стационарном режиме. Тогда формулы (9.4) и (9.5) с учетом (10.1) примут вид:

$$p_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^k}, \quad (10.2)$$

$$p_k = p_0 \left(\frac{\lambda}{\mu}\right)^k, \quad k \geq 0. \quad (10.3)$$

Для того чтобы система $M/M/1$ не переполнялась, вероятность p_0 должна быть больше нуля, т.е. $\frac{\lambda}{\mu} < 1$, т.е. интенсивность поступления должна быть меньше интенсивности обслуживания: $\lambda < \mu$. В этом случае ряд $\sum_{k=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^k$ сходится и его сум-

ма равна $\frac{\frac{\lambda}{\mu}}{1 - \frac{\lambda}{\mu}}$ как сумма геометрической убывающей прогрессии со знаменателем $\frac{\lambda}{\mu} < 1$ и первым членом $-\frac{\lambda}{\mu}$. Тогда:

$$p_0 = \frac{1}{1 + \frac{\frac{\lambda}{\mu}}{1 - \frac{\lambda}{\mu}}} = 1 - \frac{\lambda}{\mu} = 1 - \rho,$$

где ρ – коэффициент использования системы, по определению равный произведению интенсивности входного потока на среднее время обслуживания заявки, т.е. $\rho = \lambda \bar{x}$, в нашем случае $\bar{x} = \frac{1}{\mu}$.

Тогда:

$$p_k = (1 - \rho)\rho^k, \quad k \geq 0, \quad 0 \leq \rho < 1. \quad (10.4)$$

Формула (10.4) указывает на то, что вероятность p_k зависит не от λ и μ , а от их отношения. На рис. 10 приведены графики p_k для трех значений ρ : $\rho = \frac{1}{6}$, $\rho = \frac{1}{2}$ и $\rho = \frac{5}{6}$.

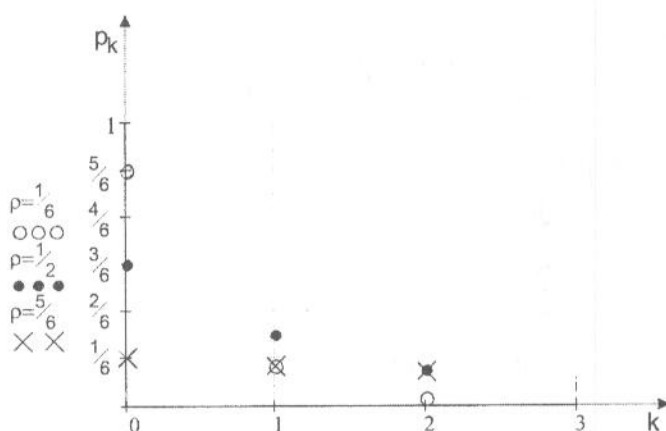


Рис. 10

Из рисунка видно, что, чем ближе загрузка системы (ρ) к нулю, тем вероятности нахождения в СМО большего количества заявок быстрее убывают, а при ρ , близких к единице, „точки вероятностей“, напротив, располагаются „более полого“.

Найдем среднее число требований в системе $M/M/1$. Это значит, найдем математическое ожидание случайной величины K – числа требований, находящихся в системе (значения этой величины обозначим через k):

$$\begin{aligned} N &= \sum_{k=0}^{\infty} k p_k = \sum_{k=0}^{\infty} k (1 - \rho) \rho^k = (1 - \rho) \rho \sum_{k=1}^{\infty} k \rho^{k-1} = (1 - \rho) \rho \sum_{k=0}^{\infty} \frac{d \rho^k}{d \rho} = (1 - \rho) \rho \frac{\sum_{k=0}^{\infty} \rho^k}{d \rho} = \\ &= (1 - \rho) \rho \frac{d}{d \rho} \frac{1}{1 - \rho} = (1 - \rho) \rho \frac{1}{(1 - \rho)^2} = \frac{\rho}{1 - \rho}, \end{aligned}$$

Итак:

$$N = \frac{\rho}{1 - \rho}. \quad (10.5)$$

Изобразим график математического ожидания числа заявок как функции от ρ :

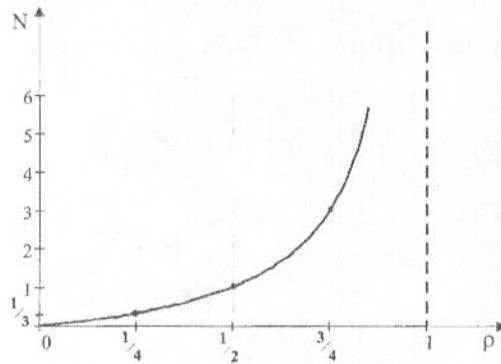


Рис. 11

Из графика видно, что при увеличении коэффициента использования растет среднее число заявок в системе, и, чем ближе ρ к единице, тем быстрее растет N .

Рассмотрим следующую важную характеристику СМО – дисперсию числа заявок в системе. По определению $D(K) = M[(K - M(K))^2]$, значит, в нашем случае:

$$\begin{aligned} D(K) &= \sum_{k=0}^{\infty} (k-N)^2 p_k = \sum_{k=0}^{\infty} k^2 p_k - 2 \sum_{k=0}^{\infty} k N p_k + \sum_{k=0}^{\infty} N^2 p_k = (1-\rho) \sum_{k=0}^{\infty} k^2 \rho^k - 2\rho \sum_{k=0}^{\infty} k \rho^k + \\ &+ \frac{\rho^2}{1-\rho} \sum_{k=0}^{\infty} \rho^k = (1-\rho) \rho \sum_{k=1}^{\infty} k^2 \rho^{k-1} - 2\rho \sum_{k=0}^{\infty} k \rho^k + \frac{\rho^2}{1-\rho} \sum_{k=0}^{\infty} \rho^k = (1-\rho) \rho \sum_{k=1}^{\infty} \frac{d}{d\rho} k^2 \int \rho^{k-1} d\rho - \\ &- 2\rho \frac{\rho}{(1-\rho)^2} + \frac{\rho^2}{1-\rho} \frac{1}{1-\rho} = (1-\rho) \rho \frac{d}{d\rho} \sum_{k=1}^{\infty} k \rho^k - \frac{2\rho^2}{(1-\rho)^2} + \frac{\rho^2}{(1-\rho)^2} = (1-\rho) \rho \frac{d}{d\rho} \frac{\rho}{(1-\rho)^2} - \\ &- \frac{\rho^2}{(1-\rho)^2} = (1-\rho) \rho \frac{1+\rho}{(1-\rho)^3} - \frac{\rho^2}{(1-\rho)^2} = \frac{\rho + \rho^2}{(1-\rho)^2} - \frac{\rho^2}{(1-\rho)^2} = \frac{\rho}{(1-\rho)^2}. \end{aligned}$$

Итак:

$$D(K) = \frac{\rho}{(1-\rho)^2}. \quad (10.6)$$

Изобразим график дисперсии числа заявок в системе как функцию от ρ :

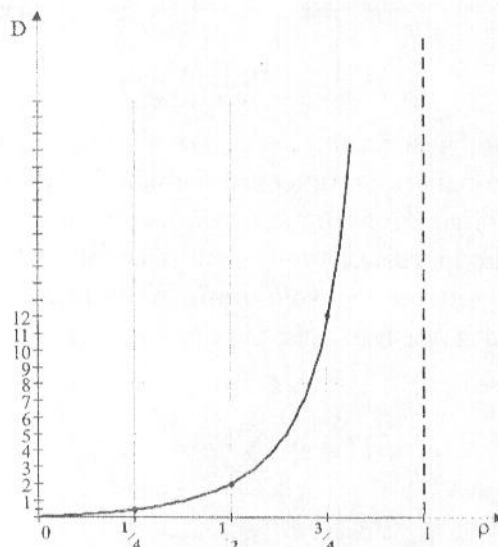


Рис. 12

Таким образом, чем ближе ρ к нулю, тем меньше отличается значение случайной

величины K — количество заявок в системе — от ее среднего значения (очевидно, сравнительно небольшого по величине), и наоборот, когда ρ близко к единице, рассеяние относительно среднего значения случайной величины K быстро возрастает.

Далее рассмотрим математическое ожидание времени пребывания заявки в системе $M/M/1$. По формуле Литтла: $N = T\lambda$, значит:

$$T = \frac{\rho}{(1 - \rho)\lambda} = \frac{\frac{1}{\mu}}{1 - \rho}.$$

Построим график среднего времени пребывания заявки в системе как функцию от ρ :

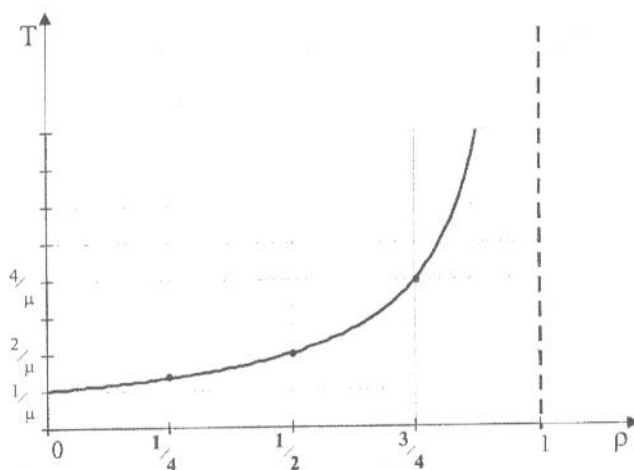


Рис. 13

Из графика видно, что при ρ , близком к нулю, заявка практически не стоит в очереди, а обслуживается сразу, и среднее время ее пребывания в системе практически совпадает со средним временем ее обслуживания — $\frac{1}{\mu}$. Это хорошо согласуется с практикой. Действительно, если заявок поступает мало, а обслуживаются они очень быстро, то практически каждая заявка сразу идет на прибор и находится в системе столько времени, сколько она обслуживается.

Графики показывают, что при ρ , близких к единице, среднее число заявок в системе и среднее время пребывания в системе неограниченно возрастают.

Из полученной формулы для T легко увидеть, что $T = \frac{1}{\mu - \lambda}$. Отсюда можно получить формулу для вычисления среднего времени ожидания заявки в системе:

$$W = T - \frac{1}{\mu} = \frac{1}{\mu - \lambda} - \frac{1}{\mu}.$$

Найдем среднюю длину очереди, посмотрев на очередь как на отдельную СМО и используя для ее вычисления формулу Литтла:

$$\lambda W = \lambda \left(\frac{1}{\mu - \lambda} - \frac{1}{\mu} \right) = \lambda \frac{\mu - \mu + \lambda}{\mu(\mu - \lambda)} = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\rho\lambda}{\mu - \lambda} = \frac{\rho\lambda}{\mu(1 - \frac{\lambda}{\mu})} = \frac{\rho^2}{1 - \rho} = \rho N.$$

Очевидно, что при ρ , стремящимся к единице, очередь неограниченно растет.

Пример.

Имеется система $M/M/1$, причем $\lambda = 2$, $\mu = 3$.

- а) Найти вероятность того, что в очереди будет больше двух заявок:

$$\sum_{k=4}^{\infty} p_k = (1 - \rho) \frac{\rho^4}{1 - \rho} = \rho^4 = \frac{16}{81}.$$

- б) Найти вероятность того, что в очереди ровно m заявок:

$$p_{m+1} = (1 - \rho) \rho^{m+1} = \frac{1}{3} \cdot \left(\frac{2}{3}\right)^{m+1}.$$

- в) Найти вероятность того, что заявке не придется ждать:

$$p_0 = 1 - \rho = \frac{1}{3}.$$

Упражнения

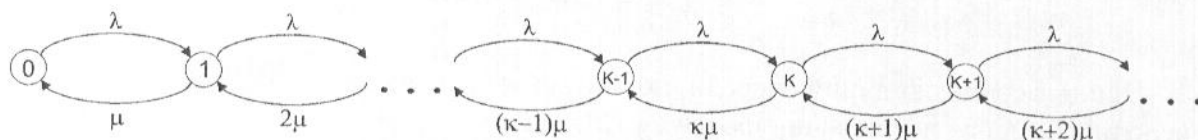
- 1. Охарактеризуйте систему массового обслуживания $M/M/1$, изобразите диаграмму интенсивности переходов и выпишите формулы для вероятностей p_0 и p_k , когда $k \geq 1$.
- 2. Докажите формулы (10.5) и (10.6).
- 3. Имеется система $M/M/1$, причем $\lambda = 1$, $\mu = 3$. Найдите среднее время пребывания заявки в системе, среднюю длину очереди. Как изменятся найденные значения, если интенсивность входного потока положить равной двум?

11. Система $M/M/\infty$

Систему $M/M/\infty$ можно рассматривать как систему, в которой имеется бесконечное число приборов и для каждой поступившей заявки сразу же найдется прибор, ее обслуживающий; с другой стороны, ее можно истолковывать как такую, в которой прибор немедленно обслуживает заявку, т.е. интенсивность обслуживания линейно растет с ростом числа поступающих заявок. Можно положить:

$$\begin{aligned} \lambda_k &= \lambda, & k &= 0, 1, 2, \dots \\ \mu_k &= k\mu, & k &= 1, 2, \dots \end{aligned}$$

Диаграмма интенсивностей переходов имеет вид:



Находясь в условиях стационарного режима, в случае нашей системы для выражения p_k получаем следующую формулу:

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} = p_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}.$$

Тогда:

$$p_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \frac{1}{k!} \left(\frac{\lambda}{\mu}\right)^k} = \frac{1}{e^{\frac{\lambda}{\mu}}} = e^{-\frac{\lambda}{\mu}}, \quad (11.1)$$

$$p_k = \frac{1}{k!} \left(\frac{\lambda}{\mu}\right)^k e^{-\frac{\lambda}{\mu}}. \quad (11.2)$$

Следовательно, число заявок в системе $M/M/\infty$ описывается распределением Пуассона.

Найдем математическое ожидание и дисперсию числа заявок в системе:

$$N = \frac{\lambda}{\mu}, \quad D(K) = \frac{\lambda}{\mu},$$

так как в случае распределения Пуассона, заданного формулой $p_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$, указанные характеристики равны λt . Тогда среднее время пребывания заявки в системе равно:

$$T = \frac{N}{\lambda} = \frac{1}{\mu}.$$

Это выражение для среднего времени пребывания заявки в системе естественно, поскольку заявка не ждет начала обслуживания, поэтому и среднее время ее пребывания равно среднему времени обслуживания одной заявки.

Пример.

Пусть на некоторую станцию медицинской скорой помощи поступают за час в среднем два вызова, причем поток вызовов – пуассоновский. Будем считать, что после получения вызова станция немедленно высылает врача больному, а время, затрачиваемое на оказание помощи и дорогу в оба конца, имеет показательное распределение с математическим ожиданием – 1,5 часа. Состояние СМО определяется количеством врачей, находящихся на обслуживании (включая дорогу). Найти вероятность того, что хотя бы один врач находится на обслуживании, дать рекомендации относительно количества врачей.

По условию задачи, работу станции скорой помощи можно интерпретировать как систему $M/M/\infty$, причем $\lambda = 2$, $\frac{1}{\mu} = 1,5$. p_k – вероятность того, что на обслужи-

вании в установившемся режиме находятся k врачей, причем $p_k = \frac{1}{k!} \left(\frac{\lambda}{\mu}\right)^k e^{-\frac{\lambda}{\mu}}$, где $k \geq 0$. Так как $\frac{\lambda}{\mu} = 2 \cdot 1,5 = 3$, то из формулы (11.2):

$$p_k = \frac{1}{k!} 3^k e^{-3}.$$

Отсюда:

P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
0,05	0,149	0,224	0,224	0,168	0,101	0,05	0,022	0,008

Понятно тогда, что наиболее вероятными являются состояния, когда заняты одно- временно 2 или 3 врача.

Вероятность того, что хотя бы один врач находится на обслуживании, равна

$$1 - p_0 = 1 - 0,05 = 0,95.$$

Если при той же средней продолжительности обслуживания в единицу времени в среднем поступает 1 вызов, то получаются следующие значения p_k :

p_0	p_1	p_2	p_3	p_4	p_5
0,223	0,335	0,251	0,126	0,047	0,014

При данных условиях наиболее вероятным является состояние, когда занят только один врач. Вероятность того, что занят хотя бы один врач, равна 0,777.

Упражнения

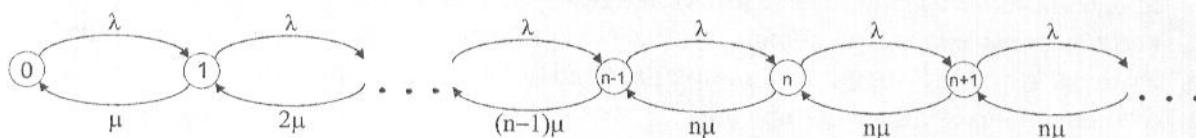
- 1. Докажите формулу (11.2).
- 2. Приведите пример СМО $M/M/\infty$.

12. Система $M/M/n$

Рассмотрим систему, функционирующую в стационарном режиме, имеющую пуассоновский входной поток со средней интенсивностью λ , неограниченный объем накопителя (длину очереди) и n приборов с одинаковой средней производительностью, равной μ . Тогда:

$$\begin{aligned} \lambda_k &= \lambda, & k &\geq 0, \\ \mu_k &= k\mu, & 1 \leq k < n, \\ \mu_k &= n\mu, & k &\geq n. \end{aligned}$$

Диаграмма интенсивности переходов имеет вид:



Найдем распределение числа заявок в системе. Поскольку режим работы системы – стационарный, то $p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}$.

Тогда:

$$\begin{aligned} p_k &= p_0 \frac{1}{k!} \left(\frac{\lambda}{\mu} \right)^k, & k < n, \\ p_k &= p_0 \prod_{i=0}^{n-1} \frac{\lambda}{(i+1)\mu} \prod_{i=n}^{k-1} \frac{\lambda}{n\mu} = p_0 \frac{1}{n!} \left(\frac{\lambda}{\mu} \right)^n \frac{1}{n^{k-1-(n-1)}} \left(\frac{\lambda}{\mu} \right)^{k-1-(n-1)} = \\ &= p_0 \frac{1}{n! n^{k-n}} \left(\frac{\lambda}{\mu} \right)^k, & k \geq n. \end{aligned}$$

Обозначим: $\frac{\lambda}{n\mu} = \rho$. Это есть отношение среднего числа заявок, поступающих в единицу времени, к числу заявок, которые может обслужить система за единицу времени при условии непрерывной работы всех n приборов, т.е. ρ – коэффициент использования системы.

Тогда p_k имеют следующий вид:

$$p_k = p_0 \frac{n^k}{k!} \rho^k \quad \text{для } k < n,$$

$$p_k = p_0 \frac{n^n}{n!} \rho^k \quad \text{для } k \geq n.$$

Далее найдем p_0 , учитывая, что для стационарного режима справедлива формула:

$$p_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}.$$

Тогда:

$$p_0 = \left[1 + \sum_{k=1}^{n-1} \frac{(n\rho)^k}{k!} + \sum_{k=n}^{\infty} \frac{n^n \rho^k}{n!} \right]^{-1} = \left[1 + \sum_{k=1}^{n-1} \frac{(n\rho)^k}{k!} + \frac{n^n}{n!} (\rho^n + \rho^{n+1} + \rho^{n+2} + \dots) \right]^{-1} =$$

$$= \left[\sum_{k=0}^{n-1} \frac{(n\rho)^k}{k!} + \frac{n^n}{n!} \frac{\rho^n}{1-\rho} \right]^{-1} = \left[\sum_{k=0}^{n-1} \frac{(n\rho)^k}{k!} + \frac{(n\rho)^n}{n!(1-\rho)} \right]^{-1}.$$

Заметим, что если $\rho \geq 1$, то в данных вычислениях бесконечно убывающей геометрической прогрессии не возникает, т.е. ряд $\sum_{k=n}^{\infty} \rho^k$ расходится, и значит вероятность того, что в системе нет ни одного требования, равна нулю, но тогда будет равна нулю и вероятность того, что в системе имеется фиксированное конечное (но любое) число заявок. В таком случае длина очереди становится бесконечной – система не справляется с обслуживанием. Следовательно, надо принять, что $\rho < 1$, интенсивность λ входного потока достаточно мала по сравнению с величиной $n\mu$ – общей производительностью всей системы, когда работают все n приборов.

Рассмотрим вероятность того, что поступившая в систему заявка окажется в очереди. Это есть вероятность того, что в системе имеется не менее n заявок, или вероятность того, что все приборы заняты:

$$\pi = \sum_{k=n}^{\infty} p_k = \sum_{k=n}^{\infty} p_0 \frac{n^n \rho^k}{n!} = \sum_{k=n}^{\infty} \frac{n^n \rho^k}{n! \left[\sum_{k=0}^{n-1} \frac{(n\rho)^k}{k!} + \frac{(n\rho)^n}{n!(1-\rho)} \right]} = \frac{\frac{(n\rho)^n}{n!} \frac{1}{1-\rho}}{\sum_{k=0}^{n-1} \frac{(n\rho)^k}{k!} + \frac{(n\rho)^n}{n!(1-\rho)}}.$$

Эту формулу используют в телефонии, она определяет вероятность того, что поступающий на n линий вызов не застанет ни одной свободной линии.

Вычислим различные средние характеристики рассматриваемой СМО, которые для единообразия будем обозначать через M_i .

Пусть R – длина очереди, r – значение этой случайной величины, тогда через M_1 обозначим $M(R)$. Если r – число заявок в очереди, то:

$$r = \begin{cases} 0, & \text{если число заявок в системе меньше или равно } n, \\ k - n, & \text{если число заявок в системе больше } n. \end{cases}$$

Тогда:

$$\begin{aligned} M_1 &= \sum_{r=1}^{\infty} r p_{n+r} = \sum_{r=1}^{\infty} r p_0 \frac{n^n \rho^n \rho^r}{n!} = p_n \sum_{r=1}^{\infty} r \rho^r = p_n \rho \sum_{r=1}^{\infty} r \rho^{r-1} = p_n \rho \frac{d}{d\rho} \sum_{r=1}^{\infty} \rho^r = \\ &= p_n \rho \frac{d}{d\rho} \left(\frac{\rho}{1-\rho} \right) = p_n \frac{\rho}{(1-\rho)^2}. \end{aligned}$$

Отсюда видно, что когда $\rho \rightarrow 1$, то длина очереди растет до бесконечности, если загрузка системы ρ близка к единице, то малейшее увеличение загрузки приводит к очень быстрому возрастанию очереди.

Пример.

Если $n = 1$, то $p_1 = (1 - \rho)\rho$, так как $p_1 = p_0\rho$, а $p_0 = 1 - \rho$. Тогда:

$$M_1 = \frac{\rho^2}{1 - \rho}.$$

Если $n = 2$, то $p_0 = \left[1 + 2\rho + \frac{(2\rho)^2}{2(1-\rho)} \right]^{-1} = \frac{1-\rho}{1+\rho}$, $p_2 = \frac{1-\rho}{1+\rho} 2\rho^2$. Тогда:

$$M_1 = \frac{1-\rho}{1+\rho} 2\rho^2 \frac{\rho}{(1-\rho)^2} = \frac{2\rho^3}{1-\rho^2}.$$

Приведем таблицу изменения средней длины очереди в зависимости от загрузки системы:

$n \backslash \rho$	0,5	0,6	0,7	0,8	0,9	0,95
1	0,5	0,9	1,63	3,2	8,1	18,05
2	0,33	0,675	1,345	2,844	7,674	17,587

Средняя длина очереди очень быстро возрастает, когда загрузка близка к единице. Увеличение числа приборов уменьшает среднюю длину очереди при малых значениях коэффициента использования, с ростом загрузки этот относительный выигрыш быстро убывает.

Обозначим: M_2 – среднее число заявок, находящихся в системе (в очереди и на обслуживании), тогда:

$$\begin{aligned} M_2 &= \sum_{k=1}^{\infty} k p_k = p_0 \sum_{k=1}^{n-1} k \frac{(n\rho)^k}{k!} + p_0 \sum_{k=n}^{\infty} k \frac{n^n \rho^k}{n!} = p_0 \sum_{k=1}^{n-1} \frac{(n\rho)^k}{(k-1)!} + p_n \sum_{k=n}^{\infty} k \rho^{k-n} = \\ &= p_0 \sum_{k=1}^{n-1} \frac{(n\rho)^k}{(k-1)!} + p_n \sum_{i=0}^{\infty} (n+i) \rho^i = p_0 \sum_{k=1}^{n-1} \frac{(n\rho)^k}{(k-1)!} + p_n \sum_{i=0}^{\infty} n \rho^i + p_n \sum_{i=0}^{\infty} i \rho^i = \\ &= p_0 \sum_{k=1}^{n-1} \frac{(n\rho)^k}{(k-1)!} + \frac{p_n n}{1-\rho} + \frac{p_n \rho}{(1-\rho)^2}. \end{aligned}$$

Найдем M_3 – среднее число свободных от обслуживания приборов:

$$M_3 = \sum_{k=0}^{n-1} (n-k)p_k = \sum_{k=0}^{n-1} (n-k)p_0 \frac{(n\rho)^k}{k!} = p_0 \sum_{k=0}^{n-1} (n-k) \frac{(n\rho)^k}{k!}.$$

Найдем среднее время ожидания в очереди, рассматривая очередь как отдельную систему массового обслуживания и воспользовавшись формулой Литтла, тогда:

$$M_4 = \frac{M_1}{\lambda} = p_0 \frac{(n\rho)^n \rho}{n!(1-\rho)^2 \lambda} = \pi \frac{\rho}{(1-\rho)\lambda} = \frac{\pi \lambda}{n\mu(1-\rho)\lambda} = \frac{\pi}{n\mu - \lambda}.$$

Отсюда видно, что M_4 пропорционально вероятности того, что все приборы заняты, и пропорционально среднему времени обслуживания $\frac{1}{\mu}$.

Учитывая последний полученный результат, можно найти M_5 – среднее время пребывания заявки в системе:

$$M_5 = M_4 + \frac{1}{\mu} = \frac{\pi}{n\mu - \lambda} + \frac{1}{\mu}.$$

Пример.

В мастерской по ремонту часов есть n мастеров, работающих с одинаковой производительностью. В течение семичасового рабочего дня от населения в среднем поступает на ремонт 30 часов, причем каждый мастер за один рабочий день ремонтирует в среднем 10 часов. рассматриваемый поток требований будем считать пуассоновским с параметром $\lambda = 30$. Время ремонта подавляющей части часов невелико. В капитальном ремонте часы нуждаются сравнительно редко, т.е. предполагается, что обслуживание подчиняется показательному закону с параметром $\mu = 10$ штук в рабочий день. Проанализировать работу мастерской с точки зрения теории массового обслуживания.

Ясно, что в мастерской должно быть, по крайней мере, 4 мастера, так как должно выполняться: $\rho = \frac{30}{10n} < 1$, значит, $n > 3$.

Пусть работают 4 мастера, тогда вероятность того, что все они не заняты в момент поступления очередных часов, равна:

$$p_0 = \left[1 + 3 + \frac{3^2}{2!} + \frac{3^3}{3!} + \frac{3^4}{4!(1 - \frac{3}{4})} \right]^{-1} \approx 0,038;$$

при подсчете мы учитывали, что $\rho = \frac{3}{4}$. Таким образом, в течение семичасового рабочего дня в среднем $0,038 \cdot 7 \cdot 60 \approx 15,96$ минут все 4 мастера свободны.

Найдем вероятность того, что все 4 мастера заняты в момент поступления заявки:

$$\pi = \frac{0,038 \cdot 3^4}{4! \left(1 - \frac{3}{4}\right)} \approx 0,51,$$

т.е. половину рабочего дня все мастера заняты одновременно.

Вычислим среднее время ожидания:

$$M_4 = \frac{0,51}{40 - 30} = 0,051,$$

что составляет примерно 21 минуту.

Учитывая, что $p_4 = \frac{3^4 \cdot 0,038}{4!} \approx 0,13$, найдем среднее число заявок в очереди:

$$M_1 = \frac{3 \cdot 0,13}{4 \cdot \left(1 - \frac{3}{4}\right)^2} = 1,56.$$

Вычислим среднее число экземпляров часов, ожидающих ремонта и ремонтируемых:

$$M_2 = 0,038 \left(3 + \frac{3^2}{1} + \frac{3^3}{2!} \right) + \frac{0,13 \cdot 4}{1 - \frac{3}{4}} + 1,56 = 4,53.$$

Среднее число мастеров, свободных от работы, равно:

$$M_3 = 0,038 \left(\frac{4-0}{0!} \cdot 1 + \frac{4-1}{1!} \cdot 3 + \frac{4-2}{2!} \cdot 3^2 + \frac{4-3}{3!} \cdot 3^3 \right) = 1,007,$$

т.е. каждый мастер свободен в среднем одну четверть рабочего времени, т.е. 1,75 часа в день.

Если $n = 5$, то $p_0 = 0,047$, $\pi = 0,24$, $M_4 = 0,012$ (≈ 5 минут), $M_1 = 0,36$, $M_2 = 3,37$, $M_3 = 2$.

Оптимальное число мастеров равно 4. Если $n < 4$, то число неисправных часов, ожидающих ремонта, возрастает, и мастерская не справляется с ремонтом. Если $n = 5$, то выигрыш, получаемый клиентами, незначителен, по сравнению с экономическими затратами, связанными с привлечением дополнительных мастеров: среднее время ожидания ремонта уменьшится с 21 до 5 минут, но это не даст преимуществ клиентам, так как ремонт одних часов в среднем длится $\frac{7 \cdot 60}{10} = 42$ минуты и, следовательно, производится не в присутствии клиента. Вместе с тем, при $n = 5$ только 24% рабочего времени все мастера полностью загружены, причем, в среднем, 2 мастера все время свободны.

Упражнения

- 1. Выписать систему алгебраических уравнений для вероятностей p_k системы $M/M/n$.
- 2. СМО с тремя приборами и неограниченной очередью работает в стационарном режиме. Входной поток – пуассоновский со средней интенсивностью, равной пяти заявкам в час. Средняя интенсивность обслуживания равна трем заявкам в час. Найти среднюю длину очереди, среднее время ожидания, вероятность того, что в системе нет очереди.
- 3. СМО с двумя приборами, имеющими одинаковую среднюю производительность, равную двум заявкам в минуту, работает в стационарном режиме. Очередь неограничена, интенсивность входного потока равна трем заявкам в минуту. Найти среднее число занятых приборов, среднее число заявок в системе, вероятность того, что заявке не придется стоять в очереди, среднее время ожидания начала обслуживания.

13. Система $M/M/1/V$

Пусть СМО такова, что объем накопителя в ней не превышает $V - 1$ заявки, т.е. максимальное число заявок, которые могут быть в системе, равно V (одна заявка на приборе). Если в системе уже есть V заявок, то любое следующее требование получает отказ и покидает систему необслуженным. Если $V = 1$, то систему называют *системой с удалением заблокированных вызовов*. Поступление заявок происходит по закону Пуассона.

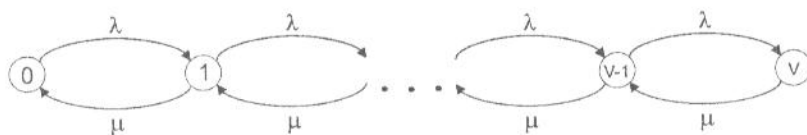
Эта система может быть описана с помощью процесса гибели и размножения.

На время, когда в системе имеются V заявок, будем перекрывать входной поток, т.е. положим:

$$\lambda_k = \begin{cases} \lambda, & k < V, \\ 0, & k \geq V, \end{cases}$$

$$\mu_k = \mu, \quad k = 1, 2, \dots, V.$$

Тогда диаграмма интенсивностей – следующая конечная цепь:



Дифференциальная система превратится (в условиях стационарного режима) в алгебраическую:

$$\begin{cases} 0 = -\lambda p_0 + \mu p_1, & k = 0, \\ \dots\dots\dots \\ 0 = \lambda p_{k-1} - (\lambda + \mu)p_k + \mu p_{k+1}, & 0 < k < V, \\ \dots\dots\dots \\ 0 = \lambda p_{k-1} - \mu p_k, & k = V, \\ \sum_{k=0}^V p_k = 1. \end{cases}$$

Для всех $k > V$ выполняется: $p_k = 0$.

Найдем p_k при $k \leq V$:

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} = p_0 \left(\frac{\lambda}{\mu} \right)^k.$$

Тогда:

$$p_0 = \frac{1}{1 + \sum_{k=1}^V \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}} = \left[1 + \sum_{k=1}^V \left(\frac{\lambda}{\mu} \right)^k \right]^{-1} = \left[1 + \frac{\frac{\lambda}{\mu} \left(1 - \left(\frac{\lambda}{\mu} \right)^V \right)}{1 - \frac{\lambda}{\mu}} \right]^{-1} = \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu} \right)^{V+1}},$$

при подсчете учитывался тот факт, что $\sum_{k=1}^V \left(\frac{\lambda}{\mu} \right)^k$ – сумма V членов геометрической прогрессии.

Итак, для p_k имеем:

$$p_k = \begin{cases} \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{V+1}} \left(\frac{\lambda}{\mu}\right)^k, & 0 \leq k \leq V, \\ 0, & k > V. \end{cases}$$

Для системы с удалением заблокированных вызовов, т.е. системы, в которой отсутствует очередь, и заявка, пришедшая в систему и обнаружившая прибор занятым, покидает систему без обслуживания, имеем:

$$p_k = \begin{cases} p_0 = \frac{1}{1 + \frac{\lambda}{\mu}} = \frac{\mu}{\lambda + \mu}, \\ p_1 = \frac{\frac{\lambda}{\mu}}{1 + \frac{\lambda}{\mu}} = \frac{\lambda}{\lambda + \mu}, \\ 0, \text{ если } k > 1. \end{cases}$$

Рассмотрим теперь систему с удалением заблокированных вызовов не при установившемся процессе обслуживания, когда после момента включения системы прошло достаточно много времени. Тогда дифференциальная система уравнений будет выглядеть следующим образом:

$$\begin{cases} p_0'(t) = -\lambda p_0(t) + \mu p_1(t), \\ p_1'(t) = \lambda p_0(t) - \mu p_1(t), \\ p_0(t) + p_1(t) = 1. \end{cases}$$

Эта система линейно зависима, и достаточно рассмотреть, например, только первое и третье уравнения. Получаем:

$$\begin{aligned} p_0'(t) &= -\lambda p_0(t) + \mu(1 - p_0(t)), \\ p_0'(t) + (\lambda + \mu)p_0(t) - \mu &= 0, \\ p_0(t) &= e^{-(\lambda + \mu)t} \left[\mu \frac{1}{\lambda + \mu} e^{(\lambda + \mu)t} + C \right] = \frac{\mu}{\lambda + \mu} + C e^{-(\lambda + \mu)t}. \end{aligned}$$

Будем считать, что в начальный момент времени система была пуста, т.е. $p_0(0) = 0$, тогда имеем:

$$1 = \frac{\mu}{\lambda + \mu} + C,$$

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

$$p_0(t) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t}, \quad (13.1)$$

$$p_1(t) = \frac{\lambda}{\lambda + \mu} - \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t}. \quad (13.2)$$

Если начальные условия изменим, например, возьмем $p_0(0) = 0$, т.е. в момент включения системы в ней находилась заявка, то имеем:

$$0 = \frac{\mu}{\lambda + \mu} + C,$$

$$C = -\frac{\mu}{\lambda + \mu},$$

$$p_0(t) = \frac{\mu}{\lambda + \mu} - \frac{\mu}{\lambda + \mu} e^{-(\lambda + \mu)t}, \quad (13.3)$$

$$p_1(t) = \frac{\lambda}{\lambda + \mu} + \frac{\mu}{\lambda + \mu} e^{-(\lambda + \mu)t}. \quad (13.4)$$

Из формул (13.1) – (13.4) следует, что при $t \rightarrow \infty$ существуют одинаковые для обоих случаев пределы:

$$p_0 = \frac{\mu}{\lambda + \mu} = \frac{1}{1 + \rho},$$

$$p_1 = \frac{\lambda}{\lambda + \mu} = \frac{\rho}{1 + \rho}.$$

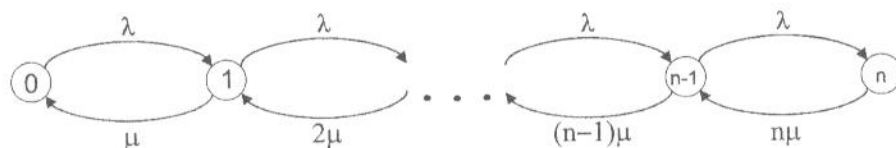
Эти предельные значения получаются при любых начальных условиях, т.е. стационарные вероятности не зависят от начального состояния системы.

Упражнение

- Проверить, будут ли стационарные вероятности p_0, p_1, p_2 зависеть от начального состояния системы $M/M/1/2$.

14. Система $M/M/n$ с n обслуживающими приборами и с потерями

По-прежнему рассматриваем систему с удалением заблокированных вызовов, но с n приборами, т.е., если требование поступает в тот момент, когда все приборы заняты, то оно покидает систему без обслуживания (теряется). Тогда имеем следующую диаграмму интенсивности переходов:



В нашем случае:

$$\begin{aligned} \lambda_k &= \lambda, & k < n, \\ \lambda_k &= 0, & k \geq n, \\ \mu_k &= k\mu, & 1 \leq k \leq n. \end{aligned}$$

Формулы для вероятностей p_k имеют вид:

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} = p_0 \frac{1}{k!} \left(\frac{\lambda}{\mu} \right)^k, \quad k \leq n, \quad (14.1)$$

$$p_k = 0, \quad k > n,$$

$$p_0 = \left[\sum_{k=0}^n \left(\frac{\lambda}{\mu} \right)^k \frac{1}{k!} \right]^{-1}. \quad (14.2)$$

На величины p_k можно смотреть как на вероятности того, что заняты k приборов, а можно – как на отношение длины промежутка времени, когда заняты k приборов, ко всему рассматриваемому отрезку времени, т.е. как на долю времени, когда заняты k приборов. Тогда p_0 – относительная длительность промежутка времени, когда свободны все приборы, а p_n – относительная длительность времени занятости всех n приборов. Но так как при этом вновь поступающие в систему заявки получают отказ, то p_n – это также вероятность потери заявки.

Поскольку даже при малом значении $\frac{\lambda}{\mu}$ величина $p_n = \frac{\left(\frac{\lambda}{\mu} \right)^n}{n! \sum_{k=0}^n \left(\frac{\lambda}{\mu} \right)^k \frac{1}{k!}}$ отлична от

нуля, то даже при малой загрузке имеется вероятность потери заявки из-за занятости всех приборов.

С другой стороны, поскольку даже при значениях $\frac{\lambda}{\mu}$, близких к единице, величина p_0 отлична от нуля, то всегда имеются промежутки времени, когда все приборы свободны.

Найдем математическое ожидание числа занятых приборов:

$$M = \sum_{k=0}^n k p_k = \sum_{k=0}^n k p_0 \frac{\left(\frac{\lambda}{\mu} \right)^k}{k!} = \frac{\lambda}{\mu} \sum_{k=1}^n p_0 \frac{\left(\frac{\lambda}{\mu} \right)^{k-1}}{(k-1)!} = \frac{\lambda}{\mu} \sum_{k=1}^n p_{k-1} = \frac{\lambda}{\mu} \sum_{k=0}^{n-1} p_k = \frac{\lambda}{\mu} (1 - p_n).$$

Пример 1.

Пусть $n = 1, 2, 4, 6$. Пусть $\frac{\lambda}{n\mu} = 1, 5$, т.е. пусть в среднем каждый прибор одинаково загружен в каждой из четырех систем. Тогда:

n	λ/μ	p_n	M
1	1,5	0,6	0,6
2	3	0,5294	1,4118
4	6	0,4696	3,1824
6	9	0,4405	5,0355

Как видим, с ростом числа приборов вероятность потерь убывает, а растет среднее число занятых приборов.

Отсюда: если в систему поступает некоторый поток, то вероятность потери заявок при сохранении одной и той же средней нагрузки на каждый прибор будет меньше для той системы, где любая заявка может поступать на любой прибор, по сравнению с системой, в которой на каждый прибор поступает определенная часть требований. Именно так и организованы сейчас кассы предварительной продажи билетов (на поезд и самолеты любого направления).

Пример 2.

Автоматическая телефонная станция имеет пять линий связи. Предположим, что поток заявок на соединение с абонентом для ведения разговоров – пуассоновский с интенсивностью два вызова в минуту, и время разговора имеет показательное распределение с математическим ожиданием, равным 1,5 мин. Предполагается, что заявка получает отказ, если в момент ее поступления на АТС все пять линий заняты.

Имеем: $n = 5$, $\frac{\lambda}{\mu} = 2 \cdot 1,5 = 3$, тогда вероятность отказа равна:

$$p_5 = \frac{3^5}{5! \sum_{k=0}^5 3^k \frac{1}{k!}} \approx 0,11.$$

Таким образом, станция не соединяет с абонентом, в среднем, в 11% случаев. Пропускную способность АТС при данных λ и μ можно увеличить только за счет увеличения числа линий связи.

Найдем, сколько нужно использовать линий связи, чтобы увеличить пропускную способность в десять раз, т.е. чтобы вероятность отказа не превосходила 0,011. Для этого с помощью формул (14.1) и (14.2) при условии, что $\frac{\lambda}{\mu} = 3$, составим таблицу:

n	1	2	3	4	5	6	7	8
pn	0,75	0,529	0,346	0,206	0,11	0,052	0,022	0,008

Отсюда можно сделать следующий вывод: если спроектировать АТС при данных значениях λ и μ так, что она сможет одновременно обслужить восемь разговоров ($n = 8$), то только в 0,8% случаев будет получен отказ.

Упражнения

- 1. Имеются две системы, работающие в стационарном режиме. Каждая имеет пуассоновский входной поток со средней интенсивностью, равной двум заявкам в минуту. В каждой системе имеются по два прибора с одинаковой средней производительностью, равной двум заявкам в минуту. В первой системе очередь не может превышать двух заявок, а во второй объем накопителя равен нулю. Для каждой системы найдите вероятность того, что в системе будет не менее двух заявок.
- 2. На телефонной станции имеются две линии. Вызов, поступающий, когда все линии заняты, получает отказ. Поток вызовов является пуассоновским с параметром $\lambda = 0,5$ вызовов в минуту. Время обслуживания распределено по показательному закону. Средняя продолжительность разговора составляет три минуты. Найдите вероятность отказа. При каком наименьшем числе линий доля требований, получающих отказ, будет составлять меньше пяти процентов?

15. Системы $M/M/1/\infty/m$, $M/M/\infty/\infty/m$, $M/M/n/V/m$

15.1. Система $M/M/1/\infty/m$

Когда входной в систему поток пуассоновский, можно было считать, что он обеспечивается бесконечным множеством источников нагрузки (т.е. бесконечным множеством заявок).

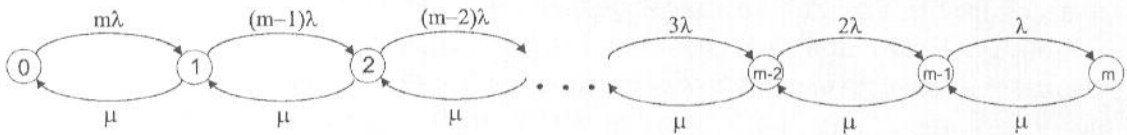
Теперь предположим, что входной поток не является пуассоновским, а создается конечной группой, содержащей m возможных заявок. Знак ∞ в обозначении системы $M/M/1/\infty/m$ указывает на то, что никаких ограничений на объем накопителя нет, хотя понятно, что в данном конкретном случае, когда имеется всего лишь m возможных претендентов на поступление в систему, объем очереди в действительности ограничен. Попутно заметим, что иногда тот факт, что заранее никаких ограничений на объем накопителя нет, обозначают не знаком ∞ , а пробелом, например, вместо $M/M/1/\infty/m$, можно было бы написать $M/M/1/ \quad /m$. Каждая из m заявок либо находится в системе, либо является претендентом на поступление. Момент поступления в систему является случайной величиной, пусть промежутки между поступлениями распределены по показательному закону со средним значением $\frac{1}{\lambda}$ (секунд). Предполагаем, что все заявки поступают независимо друг от друга, тогда если в системе находится k заявок (очередь плюс прибор), то $m - k$ заявок находятся в числе поступающих, и, значит, общая интенсивность поступления равна:

$$\lambda_k = \begin{cases} \lambda(m - k), & 0 \leq k \leq m, \\ 0, & k > m, \end{cases}$$

$$\mu_k = \mu, \quad k = 1, 2, \dots, m.$$

Очевидно, система является саморегулируемой: когда накапливается большая очередь (k близко к m), то интенсивность поступления новых требований уменьшается и предотвращает перегрузку системы.

Диаграмма интенсивностей переходов является конечной цепью:



Учитывая то, что система работает в стационарном режиме, получаем:

$$\begin{aligned} p_k &= p_0 \prod_{i=0}^{k-1} \frac{(m-i)\lambda}{\mu} = p_0 \frac{m \cdot (m-1) \cdot \dots \cdot [m-(k-1)] (m-k) \cdot \dots \cdot 1}{1 \cdot 2 \cdot \dots \cdot (m-k)} \left(\frac{\lambda}{\mu} \right)^k = \\ &= p_0 \left(\frac{\lambda}{\mu} \right)^k \frac{m!}{(m-k)!}, \quad 0 \leq k \leq m, \\ p_k &= 0, \quad k > m, \\ p_0 &= \left[\sum_{k=0}^m \left(\frac{\lambda}{\mu} \right)^k \frac{m!}{(m-k)!} \right]^{-1}. \end{aligned}$$

Заметим, что при $m = 1$ система превращается в систему $M/M/1/1$ – систему с удалением заблокированных вызовов.

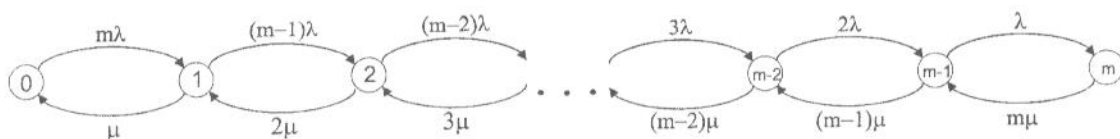
15.2. Система $M/M/\infty/\infty/m$

Пусть имеется конечное число источников нагрузки m (т.е. конечное число заявок) и „бесконечное“ число обслуживающих приборов. Тогда модель может быть описана следующим образом:

$$\lambda_k = \begin{cases} \lambda(m-k), & 0 \leq k \leq m, \\ 0, & k > m, \end{cases}$$

$$\mu_k = k\mu, \quad k = 1, 2, \dots, m.$$

Диаграмма интенсивностей имеет вид:



Воспользовавшись формулами для выражения вероятностей p_k , в случае когда система работает в стационарном режиме, получаем:

$$\begin{aligned} p_k &= p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} = p_0 \prod_{i=0}^{k-1} \frac{(m-i)\lambda}{(i+1)\mu} = p_0 \frac{m!}{(m-k)!k!} \left(\frac{\lambda}{\mu}\right)^k = \\ &= p_0 \left(\frac{\lambda}{\mu}\right)^k C_m^k, \quad 0 \leq k \leq m, \\ p_0 &= \left[\sum_{k=0}^m \left(\frac{\lambda}{\mu}\right)^k C_m^k \right]^{-1} = \frac{1}{\left(1 + \frac{\lambda}{\mu}\right)^m}. \end{aligned}$$

Тогда:

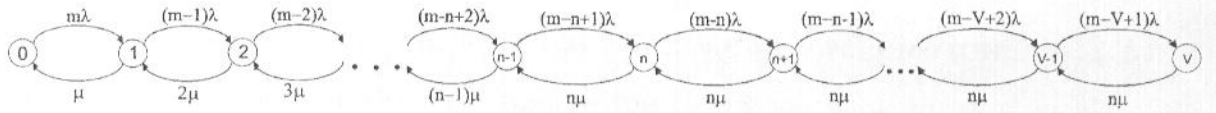
$$\begin{aligned} p_k &= \frac{\left(\frac{\lambda}{\mu}\right)^k C_m^k}{\left(1 + \frac{\lambda}{\mu}\right)^m}, \quad 0 \leq k \leq m, \\ p_k &= 0, \quad k > m. \end{aligned}$$

15.3. Система $M/M/n/V/m$

В системе n обслуживающих приборов, каждый из которых характеризуется параметром обслуживания μ , имеется накопитель такого объема, что суммарное число заявок, находящихся в системе, не превосходит V , если в системе уже имеются V заявок, то следующие поступающие заявки покидают систему необслуженными. Имеется m источников нагрузки, т.е. всего имеется m претендентов на поступление в систему. Пусть средний промежуток времени между заявками равен $\frac{1}{\lambda}$, т.е. поток имеет среднюю интенсивность, равную λ . Заметим, что эта система является обобщением всех ранее рассмотренных систем.

Предполагаем, что $n \leq V \leq m$.

Диаграмма интенсивностей переходов имеет вид:



Параметры процесса гибели и размножения имеют вид:

$$\lambda_k = \begin{cases} (m-k)\lambda, & 0 \leq k < V, \\ 0, & k \geq V, \end{cases}$$

$$\mu_k = \begin{cases} k\mu, & 0 \leq k < n, \\ n\mu, & n \leq k \leq V, \\ 0, & k > V. \end{cases}$$

Рассмотрим значения p_k для случая $1 \leq k < n$:

$$\begin{aligned} p_k &= p_0 \prod_{i=0}^{k-1} \frac{(m-i)\lambda}{(i+1)\mu} = p_0 \left(\frac{\lambda}{\mu} \right)^k \frac{m(m-1) \cdot \dots \cdot [m-(k-1)](m-k)(m-k-1) \cdot \dots \cdot 1}{1 \cdot 2 \cdot \dots \cdot k (m-k) \cdot \dots \cdot 1} = \\ &= p_0 \left(\frac{\lambda}{\mu} \right)^k C_m^k, \end{aligned}$$

для случая $n \leq k \leq V$:

$$\begin{aligned} p_k &= p_0 \prod_{i=0}^{n-1} \frac{(m-i)\lambda}{(i+1)\mu} \prod_{i=n}^{k-1} \frac{(m-i)\lambda}{n\mu} = \\ &= p_0 \left(\frac{\lambda}{\mu} \right)^k \frac{m(m-1) \cdot \dots \cdot [m-(n-1)][(m-n) \cdot \dots \cdot m-(k-1)]}{n^{k-n} 1 \cdot 2 \cdot \dots \cdot n} = \\ &= p_0 \left(\frac{\lambda}{\mu} \right)^k \frac{m!k!}{n^{k-n} n! (m-k)! k!} = p_0 \left(\frac{\lambda}{\mu} \right)^k C_m^k \frac{k!}{n!} n^{n-k}. \end{aligned}$$

Тогда:

$$\begin{aligned} p_0 &= \left[1 + \sum_{k=1}^{n-1} \left(\frac{\lambda}{\mu} \right)^k C_m^k + \sum_{k=n}^V \left(\frac{\lambda}{\mu} \right)^k C_m^k \frac{k!}{n!} n^{n-k} \right]^{-1} = \\ &= \left[\sum_{k=0}^{n-1} \left(\frac{\lambda}{\mu} \right)^k C_m^k + \sum_{k=n}^V \left(\frac{\lambda}{\mu} \right)^k C_m^k \frac{k!}{n!} n^{n-k} \right]^{-1}. \end{aligned}$$

Упражнение

- Охарактеризуйте каждую из рассмотренных в пункте СМО, изобразите диаграммы интенсивностей и выпишите формулы для вычисления вероятностей p_k .

16. Метод этапов. Эрланговское распределение

Будем рассматривать СМО общего вида, которую нельзя описать процессом гибели и размножения.

Распределение времени обслуживания при непоказательном законе может быть разложено в набор составляющих показательных распределений.

Рассмотрим обслуживающий прибор, время обслуживания которого имеет показательное распределение с плотностью распределения: $b(x) = \mu e^{-\mu x}$, где $x \geq 0$.

Очевидно, что $M(x) = \frac{1}{\mu}$, $D(x) = \frac{1}{\mu^2}$.

Теперь рассмотрим систему, в которой прибор содержит, в свою очередь, два прибора, причем каждый из них имеет показательное распределение с параметром 2μ . Плотность для каждого внутреннего прибора или этапа равна: $h(y) = 2\mu e^{-2\mu y}$, где $y \geq 0$. Тогда $M(y) = \frac{1}{2\mu}$, $D(y) = \frac{1}{(2\mu)^2}$. Тогда заявка, находясь в приборе, обслуживается одним из этапов, и новая заявка поступает на обслуживание только в момент завершения обслуживания второго этапа для предыдущей заявки, т.е. один из этих внутренних приборов всегда обязательно пустует. Это означает, что описываемый процесс обслуживания аналогичен процессу входного потока с просеиванием одной заявки (как бы забывается заявка, виртуально находящаяся на пустом этапе). Тогда для нахождения плотности вероятности времени обслуживания на всем приборе можно воспользоваться распределением Эрланга. В общем случае, если прибор имеет l этапов, то для описания состояния прибора для обслуживаемой заявки достаточно рассмотреть двумерный вектор, первая координата которого указывает на число заявок в очереди, а вторая – число этапов, через которые уже прошло обслуживаемое требование.

В двухэтапном приборе заявка какое-то случайное время находится на одном этапе и какое-то случайное время – на другом. Тогда общее время обслуживания есть случайная величина, равная сумме независимых и одинаково распределенных случайных величин. Следовательно, как известно из курса „Теория вероятностей“, для нахождения плотности данной случайной величины надо взять свертку плотностей распределения каждой из суммируемых случайных величин. Свертка функций относительно преобразования Лапласа обладает следующим свойством: преобразование Лапласа от свертки двух функций есть произведение преобразований Лапласа от каждой из рассматриваемых функций. Таким образом, преобразование Лапласа от плотности распределения времени обслуживания в двухэтапном приборе равно произведению преобразований Лапласа от плотностей времени обслуживания на каждом этапе:

$$L(b(x)) = \left(\frac{2\mu}{2\mu + s} \right)^2 \quad (\text{см. пункт 8}),$$

по данному образцу $L(b(x))$, используя таблицы преобразования Лапласа, находим оригинал. Тогда:

$$b(x) = 2\mu(2\mu x)e^{-2\mu x}, \quad x \geq 0.$$

Зная плотность распределения вероятностей времени обслуживания, можем найти математическое ожидание и дисперсию для этой двухэтапной системы. Поскольку время, проведенное в приборе, является суммой двух случайных величин, то математическое ожидание времени обслуживания является суммой математических ожиданий слагаемых, и дисперсия суммы независимых случайных величин равна сумме

дисперсий, т.е.:

$$M(x) = \frac{1}{2\mu} + \frac{1}{2\mu} = \frac{1}{\mu},$$
$$D(x) = \frac{1}{(2\mu)^2} + \frac{1}{(2\mu)^2} = \frac{1}{2\mu^2}.$$

Мы получили, что математическое ожидание времени обслуживания в двухэтапной системе равно математическому ожиданию времени обслуживания в одноэтапной системе с параметром μ , а дисперсия в двухэтапной системе вдвое меньше дисперсии одноэтапной системы.

Заметим, что здесь мы, в силу показательного закона распределения, имели отсутствие последействия, т.е. время, уже проведенное заявкой на данном этапе обслуживания, никак не сказывается на дальнейшем функционировании системы.

Таким образом, в данной системе удалось распределение Эрланга применить не к входному потоку, как было сделано в пункте 8, а к процессу обслуживания.

Упражнение

- Приведите пример СМО, в которой обслуживающий прибор имеет два этапа, причем новая заявка поступает на обслуживание только в момент завершения обслуживания второго этапа для предыдущей заявки.

17. Статистическое моделирование СМО

Анализ реальной системы может основываться на результатах экспериментов или исследованиях моделей рассматриваемой системы. Следует различать физические и математические модели, а также моделирующие алгоритмы для имитационного (статистического, машинного) моделирования на компьютере.

Математические модели реальных систем исследуются аналитическими методами теории вероятностей. Как правило, основные задачи сводятся к рассмотрению соответствующих случайных процессов и определению их свойств, распределений или числовых характеристик.

В ходе компьютерного моделирования для конкретных наборов параметров имитируются траектории случайных процессов, точнее их значения в некоторых точках. По реализациям траекторий статистическими методами оцениваются значения искоемых характеристик.

Статистическое моделирование реальной системы включает в себя: формулировку задачи исследования, выбор или построение модели (обычно выбирают некоторую математическую модель и для нее строят моделирующий алгоритм), составление плана имитационного эксперимента, непосредственное моделирование функционирования системы (имитацию), обработку результатов моделирования, их анализ и интерпретацию.

Остановимся на проблемах, связанных с моделированием конкретной системы обслуживания.

Для произвольной СМО можно дать унифицированное описание, удобное для построения моделирующего алгоритма. При этом учитывается следующая особенность систем обслуживания. Траектории случайных процессов, описывающих изменения

значений характеристик системы, кусочно-линейные, т.е. на конечных интервалах времени некоторые характеристики не меняют своих значений, другие же изменяются пропорционально течению времени. В моменты времени, соответствующие концам этих интервалов (поступление в систему или окончание обслуживания требования и т.п.), характеристики могут скачком изменить свое значение.

В связи с этим будем считать, что СМО состоит из некоторого набора элементов A_k ; $k \geq 1$ (элементы могут представлять источник требований, конкретные требования, очередь, прибор обслуживания, и т.п.). Элемент A_k в каждый момент времени находится в некотором состоянии e_k из заданного набора состояний. Причем все состояния делятся на активные и пассивные (так, например, для прибора: обслуживание – активное состояние, ожидание поступления в свободную систему требования – пассивное состояние). Состояние системы характеризуется набором состояний элементов. Каждое состояние e_k описывается набором характеристик $x_k = \{x_{ki}; i \geq 0\}$, которые принимают для данного состояния некоторое фиксированное значение, либо линейно изменяются с течением времени. У активных состояний имеются активные характеристики, которые с течением времени линейно изменяют свои значения до некоторой фиксированной величины – границы, достижение которой может вызвать изменение состояния элемента и системы в целом.

Каждая активная характеристика определяет момент времени, когда она может достичь границы. Оставшееся время до ближайшего достижения границы одной из активных характеристик состояния e_k задает остаточное время активного состояния и приписывается характеристике x_{k0} . У пассивных состояний значение x_{k0} полагается равным бесконечности (при компьютерном представлении – наибольшим представимым в компьютере числом), $\Delta = \min\{x_{k0}; k \geq 1\}$ определяет остаточное время эволюции системы, по истечении которого возможны изменения состояний элементов, изменения скачком их характеристик и сам набор характеристик.

Моделирующий алгоритм СМО представляется компьютерной программой, и его схему можно представить в виде выполнения следующих этапов.

- **Этап 1.** *Фиксация начального состояния системы.*
- **Этап 2.** *Определение шага моделирования.* Иногда шаг моделирования фиксирован, как один из параметров системы. Более эффективно за шаг моделирования брать остаточное время эволюции системы.
- **Этап 3.** *Определение нового состояния СМО* по истечении времени ее эволюции. Изменения состояний элементов определяются исходя из свойств конкретной СМО. При этом для некоторых элементов бывает необходимо генерировать значение активной характеристики, как значение случайной величины с заданным законом распределения, при помощи датчика случайных чисел.
- **Этап 4.** *Промежуточная обработка данных.* На каждом шагу моделирования мы определяем значения характеристик (траекторий) СМО. Сохранение всех этих значений до конца моделирования нецелесообразно. Поэтому с учетом предстоящей статистической обработки данных моделирования выполняются промежуточные вычисления.
- **Этап 5.** *Проверка условий окончания имитации.* В качестве критериев принимают либо достижение заданного времени моделирования, либо обслуживание определенного числа требований, либо достижение фиксированной точности. Если моделирование не закончено, следует перейти к этапу 2.

• **Этап 6. Статистическая обработка данных. Представление результатов моделирования.**

Рассмотрим более подробно реализацию этапа 3, а именно проблему генерирования значений случайной величины с заданным законом распределения.

Датчики случайных чисел вырабатывают последовательности $\{w_n\}$ независимых случайных величин, равномерно распределенных в интервале $(0, 1)$. Поэтому нужно научиться строить случайную величину с заданным законом распределения, исходя из равномерно распределенной случайной величины. В случае одномерных величин можно использовать следующий прием.

Рассмотрим в плоскости xOy кривую $y = F(x)$, где $F(x)$ – функция распределения случайной величины ξ , которую необходимо построить. В точках разрыва $F(x)$ дополним график отрезками прямых, параллельных оси Oy , так, чтобы в результате получилась непрерывная кривая. Пусть $\phi(y)$ – значение x , для которого $F(\phi(y)) = y$. Положим $\xi = \phi(w)$, где w – равномерно распределенная в интервале $(0, 1)$ случайная величина. Тогда ξ имеет функцию распределения $F(x)$. Действительно, вследствие монотонности $\phi(y)$ выполняется:

$$p(\xi < x) = p(\phi(w) < x) = p(w < F(x)) = F(x).$$

Заметим, что уравнение $F(x) = y$ при некоторых значениях y может иметь неоднозначное решение. Однако таких значений y не более чем счетное множество, следовательно, вероятность неоднозначности решения уравнения $F(x) = y$ равна нулю и $\phi(y)$ можно определить как монотонную функцию.

Для моделирования пуассоновского потока надо генерировать промежутки времени между моментами поступления соседних заявок, т.е. значения показательно распределенной случайной величины с функцией распределения $F(x) = 1 - e^{-\lambda x}$, где λ – интенсивность входного потока. Чтобы получить конечный массив значений x_i , решим уравнение:

$$F(x_i) = w_i,$$

где w_i – значение равномерно распределенной на интервале $(0, 1)$ случайной величины.

Тогда:

$$1 - e^{-\lambda x_i} = w_i,$$

откуда:

$$x_i = -\frac{1}{\lambda} \ln(1 - w_i). \quad (17.1)$$

Чтобы смоделировать обслуживание, осуществляемое по показательному закону, т.е. получить массив длительности обслуживания отдельных заявок, можно воспользоваться этой же формулой (17.1), в которой интенсивность λ будет заменена на интенсивность обслуживания μ .

Заметим, что составление программы моделирующего алгоритма трудоемко и существенно влияет на эффективность моделирования СМО. Помогают решить данную проблему специальные языки моделирования, накапливаемое проверенное программное обеспечение.

Литература

1. Ивченко Г.И., Каптанов В.А., Коваленко И.Н. Теория массового обслуживания: Учеб. пособие для вузов. – М.: Высш. школа, 1982. – 256 с., ил.
2. Клейнрок Л. Теория массового обслуживания. Пер. с англ. / Пер. И.И. Грушко; ред. В.И. Нейман. – М.: Машиностроение, 1979. – 432 с.
3. Матвеев В.Ф., Ушаков В.Г. Системы массового обслуживания. – М.: Изд-во МГУ, 1984. – 240 с.
4. Математика для экономистов: В 6 т. / Под ред. А.Ф. Тарасюка. – М.: ИНФРА-М, 2000. Серия „Высшее образование“. Т. 6: Чернов В.П., Ивановский В.Б. Теория массового обслуживания. – 158 с.
5. Статистическое моделирование систем массового обслуживания / Лифшиц А.Л., Мальц Э.А., М.: Сов. радио, 1978. – 248 с.

Учебное издание

Кузнецова Валентина Анатольевна
Никулина Елена Вячеславовна

Введение в теорию
массового обслуживания

Текст лекций

Редактор, корректор А.А. Антонова
Компьютерная верстка А.Е. Никулин

Подписано в печать 02.12.2005г. Формат 60х84/8.

Бумага тип. Усл. печ. л. 6,97 . Уч.-изд. л. 3,7.

Тираж 75 экз. Заказ 074/05

Оригинал-макет подготовлен
в редакционно-издательском отделе ЯрГУ.

Отпечатано на ризографе.

Ярославский государственный университет
150000 Ярославль, ул. Советская, 14