Министерство образования Российской Федерации

Балтийский государственный технический университет «Военмех»

С.Н. КОРОЛЕВ

МАРКОВСКИЕ МОДЕЛИ МАССОВОГО ОБСЛУЖИВАНИЯ

Учебное пособие

Санкт-Петербург
2002

УДК 519.872(075.8)
К68

Королев С.Н. Марковские модели массового обслуживания:
К68 Учеб. пособие / Балт. гос. техн. ун-т. СПб., 2002. 79 с.

В пособии, соответствующем курсу "Системы массового обслуживания", рассматриваются общие понятия и определения теории массового обслуживания. В рамках теории марковских случайных процессов приведены формальное описание и вывод формул для характеристик основных и специальных типов систем массового обслуживания. Описаны аналитические приемы, которые приводят реальные процессы функционирования систем к марковским.

На примере марковских моделей показаны приложения теории массового обслуживания к телефонии, вычислительной технике, организации производства, торговле, транспорту и другим областям практической деятельности.

Предназначено для студентов технических и экономических направлений, специализирующихся по прикладной математике и автоматизированным системам управления.

УДК 519.872 (075.8)

Р е ц е н з е н т ы : кафедра ‹‹Авиационные приборы и измерительно-вычислительные комплексы›› Санкт-Петербургского государственного университета аэрокосмического приборостроения (зав. каф. д-р техн. наук, проф. *А.Н. Синяков*); заведующий кафедрой прикладной математики БГТУ, д-р физ.-мат. наук, проф. *С.Д. Шапорев*

Утверждено
редакционно-издательским
советом университета

ПРЕДИСЛОВИЕ

Теория массового обслуживания – это математическое исследование таких систем, в которых в случайные моменты времени поступают требования (извне или изнутри системы). Они должны быть обслужены системой на некотором приборе, причем длительность обслуживания в общем случае случайна.

Природа требований и их обслуживания зависит от конкретного вида системы. Первые задачи теории массового обслуживания были рассмотрены сотрудником Копенгагенской телефонной компании датским ученым А.К. Эрлангом (1878 – 1929) в период между 1908 и 1922 годами. Эти задачи были вызваны к жизни стремлением упорядочить работу телефонной сети и разработать методы, позволяющие заранее рассчитать качество обслуживания потребителей в зависимости от числа используемых устройств. Если под требованиями понимать, например, отказы элементов системы или станков и соответственно под обслуживанием их замену или ремонт, то многочисленные задачи надежности можно решать методами теории массового обслуживания.

Очень много задач из разных областей техники и экономики удается сформулировать и решить с помощью теории массового обслуживания методами современной математики. В качестве примеров можно назвать различные технические дисциплины (связь, вычислительная техника, техника измерений, управление и автоматическое регулирование), организацию производства, торговли, транспорта, здравоохранения, военного дела. Выявляются все новые сферы применения быстро развивающейся теории массового обслуживания.

Потребности практики привели к необходимости увеличения выпускаемых высшей школой специалистов, владеющих основами теории вероятностей, математической статистики, теории случайных процессов и некоторыми специальными разделами современной прикладной математики.

Монографии по теории массового обслуживания [1–3, 5–7, 9–11] сложны для первоначального изучения, а в учебной литературе [4, 8, 13] акцент делается на изложение таких разделов теории, как асимптотические методы, полумарковские процессы, статистика систем массового обслуживания. В то же время аналитическое решение задач массового обслуживания [12, 14] обладает рядом положительных качеств, которых не имеют другие методы исследования: оно не привязано к определенным числовым значениям параметров системы обслуживания, позволяет находить оптимальные решения и делать общие заключения.

В связи с этим основная задача предлагаемого учебного пособия состоит в изложении наиболее общих методов аналитического исследования систем массового обслуживания с подробностью, необходимой для первоначального изучения.

В первом разделе учебного пособия даются основные понятия и определения теории массового обслуживания, вводятся показатели эффективности систем рассматриваемого класса. Второй раздел посвящен формальному описанию и исследованию основных типов систем массового обслуживания с использованием математического аппарата марковских случайных процессов с непрерывным временем и дискретными состояниями. Анализ специальных систем массового обслуживания на основе марковских процессов составляет содержание третьего раздела пособия. В четвертом разделе излагаются аналитические приемы, которые приводят реальные процессы функционирования систем массового обслуживания к марковским.

Настоящее пособие предназначено, прежде всего, для студентов, обучающихся по специальностям "Автоматизированные системы обработки информации и управления" и "Информационные системы и технологии". В то же время для понимания и применения изложенного материала достаточно знаний, полученных при изучении дисциплин общеинженерного цикла (дискретной математики, теории вероятностей, математической статистики и теории случайных процессов). Поэтому оно может быть рекомендовано также студентам старших курсов широкого круга технических и экономических специальностей и направлений подготовки.

1. ОБЩАЯ ХАРАКТЕРИСТИКА СИСТЕМ МАССОВОГО ОБСЛУЖИВАНИЯ

1.1. Основные элементы систем массового обслуживания

Любая система массового обслуживания (СМО) может включать в себя следующие элементы (рис.1): входящий поток требований (или заявок) на обслуживание, очередь, обслуживающее устройство, состоящее из приборов (или каналов) обслуживания, выходящий поток. Это поток требований, покидающих систему. Требования этого потока могут быть обслужены приборами системы и не обслужены.

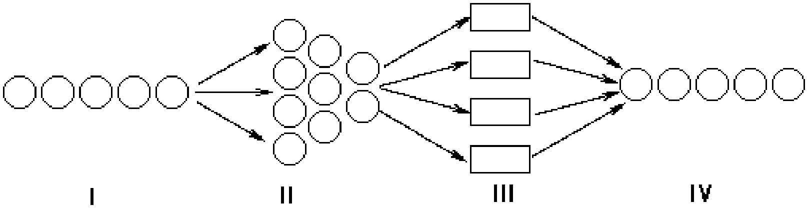

Рис. 1

Входящий поток (один или несколько потоков) представляет собой совокупность требований, которые поступают в систему одно за другим в какие-то моменты времени и нуждаются в обслуживании.

Если с точки зрения обслуживания все требования оказываются равноправными, то такой поток характеризуется только моментами времени поступления требований и называется **однородным**.

Однородный поток называется **стационарным**, если его характеристики не зависят от времени. В частности, интенсивность (плотность) стационарного потока λ (среднее число требований, поступающих в СМО в единицу времени) должна быть постоянной.

Если требования поступают на обслуживание поодиночке, то такой поток требований называется **ординарным**.

Если требования поступают на обслуживание через равные промежутки времени, то поток требований называется **регулярным**. Регулярные потоки сравнительно редко используются в теории СМО. Это объясняется их свойством “неограниченного последействия”, которое заключается в том, что по моменту поступления требования потока можно предсказать будущее и восстановить прошлое этого потока.

Если требования поступают в случайные моменты времени, то, очевидно, промежутки времени между этими моментами являются случайными величинами.

Поток требований называется **рекуррентным** (или простым процессом восстановления), если он стационарен, ординарен, а интервалы времени между моментами поступления требований представляют собой независимые случайные величины с одинаковым произвольным распределением.

Приведем пример рекуррентного потока.

Технический элемент (скажем, микросхема) работает непрерывно до своего отказа (выхода из строя); отказавший элемент мгновенно заменяется новым. Если отдельные экземпляры элемента выходят из строя независимо друг от друга, то поток отказов (он же “поток замен” или ”восстановлений”) будет рекуррентным.

Рекуррентный поток является частным случаем потока с **ограниченным последействием**.

Потоком с ограниченным последействием называется любой поток, у которого интервалы времени между моментами поступления требований – независимые случайные величины. Из этого определения следует, что в момент поступления требования будущее поведение процесса (поступления требований) в вероятностном смысле не зависит от прошлого.

Очевидно, что, если интервалы времени между моментами поступления требований имеют показательное (экспоненциальное) распределение, то такой поток представляет собой частный случай рекуррентного потока. Его называют **пуассоновским**, поскольку число требований, поступающих на обслуживание за фиксированный промежуток времени, в этом случае имеет дискретное распределение Пуассона.

Целую гамму рекуррентных потоков можно получить “просеиванием” пуассоновского потока.

Пусть, например, в какое-то учреждение поступает пуассоновский поток посетителей, а у входа стоит “страж”, направляющий первого посетителя – к первому столу, второго – ко второму столу и т.д. Если столов n, то на каждый из них поступает так называемый “поток **Эрланга n-го порядка**”.

Таким образом, поток Эрланга n-го порядка получается из пуассоновского, если сохранить в потоке каждое n-е требование, а промежуточные – выбрасывать.

Очевидно, что пуассоновский поток есть ни что иное, как поток Эрланга первого порядка.

Практически наиболее часто входящий поток требований предполагается пуассоновским. Он играет в теории СМО примерно такую же роль, как и нормальное распределение в теории вероятностей. Дело в том, что суммирование (наложение) достаточно большого числа случайных потоков при некоторых условиях приводит к пуассоновскому потоку.

В наиболее общем случае входящий поток требований характеризуется:

- числом требований (конечная или бесконечная совокупность требований);
- распределением входящего потока (распределением промежутков времени между моментами поступления требований);
- характером требований (одиночные – групповые, однородные – неоднородные);
- алгоритмом поведения требований, заставших СМО занятой (ожидание обслуживания, ожидание с ограничениями, неприсоединение к очереди, уход из системы).

Очередь (одна или несколько очередей) характеризуется: количеством мест ожидания (длиной очереди): от 0 до ∞; порядком, по которому ожидающие требования направляются на обслуживание (дисциплиной очереди).

Основные виды дисциплины:

- направление требований по определенным приборам;
- очередь в порядке поступления требований;
- требование, поступившее последним, обслуживается первым;
- случайный выбор на обслуживание;
- приоритет.

Различают два важных приоритетных правила:

1. Абсолютный (прерывающий) приоритет: при поступлении требования высокого приоритета прерывается обслуживание требования низкого приоритета, если такое имеется. Рассматривают случаи абсолютного приоритета с потерями, абсолютного приоритета с дообслуживанием и абсолютного приоритета с обслуживанием заново.
2. Относительный приоритет: требование высокого приоритета занимает первое место в очереди, и не происходит никаких прерываний.

Выделяют также: статические приоритетные правила, при которых приоритет (абсолютный или относительный) назначается до начала обслуживания (заранее); динамические приоритетные правила, при которых выбор приоритета зависит от текущего состояния системы, например, от количества уже имеющихся или ожидающих обслуживания требований.

Обслуживающее устройство – их число может быть любым: от 0 до ∞.

Основной характеристикой функционирования прибора является время обслуживания требования этим прибором. Заметим, что этот показатель характеризует не качество обслуживания, а лишь пропускную способность. Время обслуживания требования прибором в общем случае – случайная величина. При этом считают, что продолжительности обслуживания разных требований данным прибором есть случайные величины с одним и тем же законом распределения. Если время обслуживания подавляющей части требований мало и только для сравнительно небольшой части требований велико, то этот закон распределения является показательным.

В случае показательного времени обслуживания τ плотность распределения вероятностей

$$f(t) = \mu e^{-\mu t}, \ \mu > 0, \ 0 \le t < \infty, \tag{1.1}$$

а функция распределения

$$F(t) = P(\tau < t) = \int_0^t \mu e^{-\mu t} dt = 1 - e^{-\mu t}.$$

При этом математическое ожидание

$$M\tau = \mu \int_0^\infty t e^{-\mu t} dt = {}^1\!/\!{}_\mu.$$

И, следовательно, параметр μ в выражении (1.1) имеет простой физический смысл: обратная его величина равна среднему времени обслуживания требования прибором. Поэтому в литературе μ часто называют производительностью прибора.

При показательном распределении времени обслуживания все теоретические рассуждения существенно упрощаются. Многие окончательные результаты оказываются справедливыми и для произвольного закона распределения, но с тем же средним временем обслуживания. Поэтому в теории массового обслуживания обычно делают допущение, что время обслуживания распределено по показательному закону.

В наиболее общем случае обслуживающее устройство характеризуется:

- числом обслуживающих приборов;
- распределениями длительностей обслуживания требований этими приборами;
- особенностями организации обслуживания (обслуживание одиночных требований или групповое обслуживание; приборы обслуживания с одинаковым или различным распределением времени обслуживания; переменное или постоянное число обслуживающих приборов; специальные и общие приборы обслуживания);
- структурой (параллельные приборы, последовательные приборы, специальные комбинации последовательных и параллельных приборов обслуживания).

Выходящий поток (один или несколько потоков) одной системы массового обслуживания может служить полностью или частично входящим потоком требований для другой системы. Существуют также системы массового обслуживания, в которых входящий поток требований формируется из выходящего. В этих случаях имеет большое значение исследование структуры выходящего потока.

1.2. Пуассоновский поток требований

Обозначим через $X(t)$ число требований, поступающих в систему массового обслуживания за промежуток времени $(0, t)$. Если требования поступают на обслуживание в случайные моменты времени, то $X(t)$ – случайная функция, одна из возможных реализаций которой показана на рис. 2.

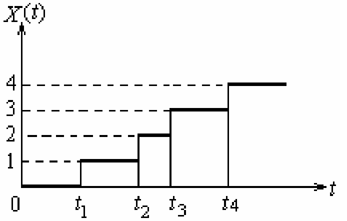

Рис. 2

Функция $X(t)$, очевидно, принимает только целые неотрицательные значения и является неубывающей при возрастании времени t. Она определяет некоторую временную последовательность событий, или, как принято говорить в теории массового обслуживания, поток требований.

Далее, если не оговорено противное, предполагается, что входящий поток требований обладает следующими тремя свойствами:

1) вероятность поступления k требований в промежутке времени $(0, t)$ равна вероятности поступления k требований в любом другом промежутке времени $(\alpha, \alpha+t)$ той же продолжительности. Это свойство потока называется стационарностью. Оно выражает неизменность вероятностного режима потока во времени;

2) вероятность поступления k требований в систему после произвольного момента времени t_0 не зависит от того, когда и сколько поступило требований до этого момента времени. Такое свойство потока называется «отсутствием последействия». Из него следует взаимная независимость поступления того или иного числа требований на обслуживание в непересекающиеся промежутки времени;

3) вероятность поступления более одного требования за малый промежуток времени Δt есть величина более высокого порядка малости, чем Δt. Это свойство потока называется ординарностью и выражает практическую невозможность одновременного поступления двух или более требований.

Поток требований, удовлетворяющий двум последним перечисленным свойствам (отсутствие последействия, ординарность), называется пуассоновским. Если поток обладает всеми тремя свойствами, то он называется простейшим.

Обозначим через $P_k(t)$ вероятность поступления в систему массового обслуживания ровно k требований за промежуток времени t, т. е. положим $P_k(t)=P(X(t)=k)$. Если λ – плотность (интенсивность) потока требований, т. е. среднее число требований, поступающих на обслуживание в единицу времени, то для пуассоновского потока при малом h можно записать:

$$P_k(h)\approx 0,\ k=2,3,\ldots;\ P_1(h)\approx\lambda h,\quad P_0(h)\approx 1-\lambda h.$$

Отсюда, воспользовавшись теоремой умножения вероятностей независимых событий, получим:

$$P_0(t+h)=P_0(t)P_0(h)\approx P_0(t)(1-\lambda h);$$

$$P_k(t+h)=P_k(t)P_0(h)+P_{k-1}(t)P_1(h)+$$
$$+P_{k-2}(t)P_2(h)+\ldots\approx P_k(t)(1-\lambda h)+P_{k-1}(t)\lambda h, k\ge 1$$

или

$$\frac{P_0(t+h)-P_0(t)}{h}\approx-\lambda P_0(t);$$

$$\frac{P_k(t+h)-P_k(t)}{h}\approx-\lambda P_k(t)+\lambda P_{k-1}(t), k\ge 1.$$

Переходя к пределу при $h\to 0$, получаем:

$$P_0'(t)=-\lambda P_0(t);\quad P_k'(t)=-\lambda P_k(t)+\lambda P_{k-1}(t), k\ge 1. \tag{1.2}$$

Допустим, что к начальному моменту $t=0$ не поступило ни одного требования, т. е. $P_0(0)=1$, $P_k(0)=0$, $k=1, 2, \ldots$ Тогда решение системы дифференциальных уравнений (1.2) можно записать в виде

$$P_0(t)=e^{-\lambda t}; \tag{1.3}$$

$$P_k(t)=\lambda e^{-\lambda t}\int_0^t e^{\lambda t}P_{k-1}(t)dt. \tag{1.4}$$

Решая рекуррентное соотношение (1.4) последовательно для $k=1,2, \ldots$ и учитывая при этом (1.3), находим:

$$P_1(t)=\lambda te^{-\lambda t};$$

$$P_2(t)=\frac{(\lambda t)^2e^{-\lambda t}}{2!};$$

. . . .

$$P_k(t)=\frac{(\lambda t)^k e^{-\lambda t}}{k!}. \tag{1.5}$$

По определению математического ожидания из (1.5)

$$MX(t)=\sum_{k=1}^{\infty}kP_k(t)=\sum_{k=1}^{\infty}k\frac{(\lambda t)^k e^{-\lambda t}}{k!}=$$
$$=\lambda t e^{-\lambda t}\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}=\lambda t,$$

т. е. среднее число требований, поступающих на обслуживание за время t, равно λt.

Формула (1.5) дает распределение вероятностей числа требований, поступающих на обслуживание за данный промежуток времени t. Вместо этого можно пользоваться распределением вероятностей промежутка времени T между двумя последовательными моментами поступления требований. Это распределение получается автоматически из формулы (1.5) при k=0. В самом деле, величина $P_0(t)$ по определению равна вероятности $P(T>t)$ того, что случайная величина T превзойдет t, поэтому можно записать

$$F(t)=P(T<t)=1-P(T>t)=1-e^{-\lambda t}. \tag{1.6}$$

Отсюда для соответствующей плотности вероятностей имеем

$$f(t)=F'(t)=\lambda e^{-\lambda t}, \tag{1.7}$$

т. е. длительность интервала времени между двумя последовательными моментами поступлений требований пуассоновского потока имеет показательное распределение. При этом

$$MT=\int_0^{\infty}t\lambda e^{-\lambda t}dt=\frac{1}{\lambda};\ DT=\int_0^{\infty}\left(t-\frac{1}{\lambda}\right)^2\lambda e^{-\lambda t}dt=\frac{1}{\lambda^2}.$$

Таким образом, математическое ожидание промежутка времени между двумя последовательными моментами поступлений требований равно обратной величине интенсивности потока требований. Этого и следовало ожидать, так как ясно, что средняя продолжительность времени между двумя последовательными моментами поступления требований должна быть равна обратной величине от среднего числа требований, поступающих в единицу времени.

Формулы (1.6) и (1.7) представляют закон распределения вероятностей для интервала времени не только между двумя последовательными моментами поступления требований, но и между произвольным фиксированным моментом времени и ближайшим моментом поступления требования. Действительно, если T – интервал времени между двумя последовательными моментами поступления требований, а τ – интервал времени после поступления последнего требования в систему, то функция распределения $F_1(t)$ оставшегося интервала времени $T-\tau$ в силу (1.6)

$$F_1(t) = P(T - \tau < t \mid T > \tau) = \frac{P(T - \tau < t, T > \tau)}{P(T > \tau)} =$$
$$= \frac{P(\tau < T < \tau + t)}{P(T > \tau)} = \frac{F(\tau + t) - F(\tau)}{1 - F(\tau)} =$$
$$= \frac{e^{-\lambda t} - e^{-\lambda(t+\tau)}}{e^{-\lambda\tau}} = 1 - e^{-\lambda t} = F(t),$$

следовательно, функция распределения оставшегося интервала времени T–τ совпадает с функцией распределения всего интервала времени T.

Таким образом, для пуассоновского потока можно утверждать, что закон распределения интервала времени от произвольного момента t_0 до момента поступления ближайшего требования в систему массового обслуживания не зависит от того, поступило или нет требование в момент t_0. Это следует из свойства отсутствия последействия, являющегося основным для пуассоновского потока.

1.3. Типы систем обслуживания. Краткая символика

Существуют следующие типы систем массового обслуживания.

Системы **с отказами (потерями)** – требования, которые при поступлении не находят ни одного свободного прибора, теряются.

Системы **с ожиданием** – возможно ожидание для любого числа требований, которые не могут быть обслужены сразу.

Комбинированные системы с ожиданием и потерями (системы ожидания с ограничениями). Например, ожидать может только конечное число требований, определяемое конечным числом мест ожидания. Требование может теряться и тогда, когда время ожидания или пребывания в системе превышает заданные границы.

Приоритетные системы – поступающие требования имеют различные личные приоритеты.

Многофазные системы – обслуживание состоит из нескольких последовательных этапов или “фаз”.

Другой важный класс систем – так называемые **замкнутые** СМО. В отличие от незамкнутых (говорят “открытых”) систем массового обслуживания в замкнутых системах входящий поток требований формируется из выходящего. К таким системам относятся и так называемые **сети массового обслуживания**.

Сеть массового обслуживания представляет собой совокупность конечного числа M обслуживающих центров, в которых циркулируют сообщения, переходящие в соответствии с маршрутной матрицей из одного центра в другой. Под центром обслуживания понимают систему массового обслуживания, состоящую из A одинаковых приборов ($1 \le A \le \infty$) и буфера объемом C ($0 \le C \le \infty$). При $A = 1$ центр называется однолинейным, при $A = \infty$ – бесконечнолинейным.

Рассмотрим пример сети массового обслуживания, являющейся простейшей моделью мультипрограммной ЭВМ (рис.3).

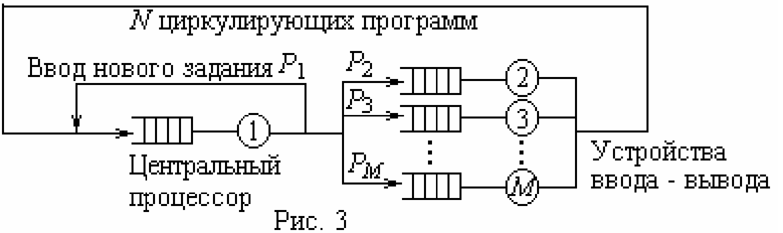

Рис. 3

Конечное число N программ (сообщений), соответствующих уровню мультипрограммирования, согласно вероятностям P_i ($i = \overline{1, M}$) поочередно обращается к одному из M центров обслуживания.

Обслуживающий центр *1* моделирует работу центрального процессора, а центры *2*, *3*, …, *M* представляют группу внешних запоминающих устройств.

Для краткой характеристики систем массового обслуживания Кендалл впервые ввел символику, которая оправдала себя. Изложим ее в несколько расширенном виде.

Символика использует пять разрядов: первый для характеристики входящего потока однородных событий, второй для характеристики обслуживания, третий определяет особенности структуры системы, в четвертом разряде фиксируются особенности очереди, пятый разряд вводится для описания приоритетных систем массового обслуживания.

В первом и во втором разрядах символами M, D, E_k, G обозначаются соответственно экспоненциальное, регулярное, эрланговское *k*-го порядка и произвольное распределения. Стрелка над буквой указывает, что исследуется многомерный случай (входящий поток есть сумма некоторого числа потоков различных требований и распределения длительностей обслуживания требований разных потоков различны).

Если в первом разряде стоит один из перечисленных выше символов, то это означает, что входящий поток рекуррентный с соответствующим распределением интервалов между моментами поступления требований. Каждый из введенных выше символов во втором разряде означает, что распределение длительности обслуживания имеет указанный вид.

Третий разряд содержит число обслуживающих приборов (1, 2 и т. д.). Если в этом разряде написана латинская буква (*m, c* и т. д.), то это означает, что число приборов может быть произвольным и рассматривается общий случай.

В четвертом разряде указывается число мест для ожидания (максимальная длина очереди): 0 – система с потерями без ожидания; 1, 2 и т. д. – система с ограниченным числом мест для ожидания. В случае системы с бесконечным числом мест ожидания ставят символ "∞" или, что чаще, опускают в символике четвертый разряд.

В пятом разряде (для приоритетных систем) фиксируется символ f_i^j; $i = 0,1,2$; $j = 0,1$. Если $i = 0$, то осуществляется обслуживание без приоритета, при $i = 1$ в системе имеется относительный приоритет, $i = 2$ означает наличие абсолютного приоритета (т.е. нижний индекс указывает на характер приоритета); $j = 0$ указывает, что требование, заставшее все места занятыми, теряется, $j = 1$ – вновь прибывшее требование вытесняет требование с низшим приоритетом.

Описанная символика не позволяет учесть все мыслимые особенности СМО, поэтому она часто дополняется словесными описаниями. Например, говорят: ”система $D/M/1$ с ненадежным обслуживающим прибором”. Это означает, что входящий поток требований регулярный, обслуживание производится по показательному закону, система одноканальная (однолинейная), каждое требование, проходящее через обслуживающий прибор (он один), обслуживается не достоверно (с некоторой вероятностью), а раз символ в четвертом разряде опущен и пятого разряда нет, то система неприоритетная с ожиданием.

1.4. Показатели эффективности систем массового обслуживания

Работа систем массового обслуживания может рассматриваться как с точки зрения организаторов (владельцев) СМО так и с точки зрения обслуживаемых клиентов. С первой точки зрения желательно “выжать из системы все, что возможно” и добиться того, чтобы ее каналы были предельно загружены. С точки зрения клиентов желательно всемерное уменьшение очередей.

В связи с этим при решении задач оптимизации в теории массового обслуживания необходимы “системный подход”, полное и комплексное рассмотрение всех “за” и ”против”. Поэтому в задачах массового обслуживания не выделяют какого-либо одного показателя эффективности, а сразу ставят эти задачи как многокритериальные.

Показатели эффективности систем массового обслуживания зависят от трех групп факторов:

- характеристик качества и надежности системы;

- экономических показателей, характеризующих работу системы (ее стоимости, трудовых затрат обслуживающего персонала, убытков, связанных с несвоевременным обслуживанием и т.д.);
- особенностей ситуации, в которой эксплуатируется система (параметров потока требований, ограничений на длину очереди и т.д.).

К числу наиболее часто применяемых показателей эффективности функционирования СМО относятся следующие:

1) вероятность P_k того, что обслуживанием требований в системе занято k приборов. Это наиболее полная характеристика, частными случаями которой являются: вероятность P_0 того, что все приборы свободны; вероятность P_n того, что все приборы заняты, где n – число приборов в СМО. Эта характеристика в системе с отказами называется "вероятностью отказа" $P_{\text{отк}}$, а в СМО с ожиданием – "вероятностью ожидания" π;

2) средняя длина очереди M_1, т.е. математическое ожидание числа требований, ожидающих начала обслуживания;

3) среднее число требований, находящихся в системе (как обслуживаемых, так и стоящих в очереди) M_2;

4) среднее число свободных от обслуживания приборов M_3;

5) коэффициент простоя обслуживающих приборов $\xi_3 = M_3 / n$;

6) среднее число приборов, занятых обслуживанием M_4;

7) коэффициент загруженности приборов $\xi_4 = M_4 / n$;

8) среднее время $M\theta$ ожидания в очереди, т.е. математическое ожидание времени θ ожидания начала обслуживания;

9) вероятность того, что время пребывания требования в очереди не продлится больше определенной величины, $P(\theta < t)$, $t > 0$. Очевидно, что эта характеристика является значением функции распределения $F(t)$ времени ожидания в фиксированный момент времени t;

10) вероятность того, что число требований в очереди, ожидающих начала обслуживания, больше некоторого числа, $P_{>m} = \sum_{k=m+1}^{\infty} P_k$. Этот показатель особенно необходим при оценке возможностей размещения требований при ограниченности времени для ожидания.

Кроме перечисленных критериев при оценке эффективности системы массового обслуживания могут быть использованы стоимостные показатели:

$C_{\text{об}}$ – стоимость обслуживания каждого требования в системе;

$C_{\text{ож}}$ – стоимость потерь, связанных с простаиванием требований в очереди в единицу времени;

$C_{\text{у}}$ – стоимость убытков, связанных с уходом из системы требований;

C_k – стоимость эксплуатации каждого прибора системы в единицу времени;

C_{nk} – стоимость единицы времени простоя прибора системы.

Для выбора оптимальных параметров СМО по экономическим показателям можно использовать функцию стоимости потерь в системе:

– для системы с ожиданием $\dot{\varphi}(n) = (C_{\text{ож}} M_1 + C_{nk} M_3 + C_k M_4)T$,

где T – интервал времени;

– для системы с отказами $\dot{\varphi}(n) = (C_k M_4 + C_{\text{у}} p_n \lambda)T$;

– для смешанных систем массового обслуживания

$$\dot{\varphi}(n) = (C_{nk} M_3 + C_{\text{ож}} M_1 + C_{\text{у}} p_n \lambda + C_k M_4)T .$$

При решении некоторых задач целесообразно пользоваться критерием экономической эффективности системы массового обслуживания $E = p_{\text{об}} \lambda C T - \varphi(n)$, где C – экономический эффект, полученный при обслуживании каждого требования; $p_{\text{об}}$ – вероятность обслуживания требования.

2. ОСНОВНЫЕ ТИПЫ СИСТЕМ МАССОВОГО ОБСЛУЖИВАНИЯ

В разделе рассматриваются системы массового обслуживания в предположении пуассоновского входящего потока и показательного закона распределения как времени обслуживания требований, так и других случайных факторов работы СМО. Такие системы массового обслуживания называют марковскими, поскольку процессы их функционирования являются марковскими случайными процессами с непрерывным временем и дискретными состояниями. Для них хорошо разработан математический аппарат и получены аналитические решения.

2.1. Системы массового обслуживания с отказами

Пусть система массового обслуживания состоит из n однотипных обслуживающих приборов и в эту систему поступает пуассоновский поток однородных требований с интенсивностью λ. Предположим далее, что время обслуживания каждого требования каждым прибором подчиняется показательному закону с параметром μ. Если оказывается, что все обслуживающие приборы заняты, то требование получает отказ и покидает систему необслуженным. Такие системы массового обслуживания называются системами с отказами (или системами с потерями).

Прежде всего выведем систему дифференциальных уравнений для вероятности $p_k(t)$ того, что в системе в момент времени t находится на обслуживании k требований (k = 0, 1, ..., n). Сразу же заметим, что в любой момент времени t должно выполняться условие нормировки $\sum_{k=0}^{n} p_k(t) = 1$.

Это непосредственно следует из того, что в любой момент времени t в системе находится на обслуживании либо 0, либо 1, либо, наконец, n требований.

По формуле полной вероятности

$$p_k(t+h) = \sum_{i=0}^{n} p_i(t) p_{ik}(h), \tag{2.1}$$

где $p_{ik}(h)$ – условная вероятность того, что через промежуток времени h будет обслуживаться k требований, если вначале обслуживалось i требований.

Будем считать h достаточно малым. Так как рассматриваемый поток требований – пуассоновский и его плотность равна λ, то можно написать:

$$P_{\geq 1}(h) = \lambda h + \mathrm{O}(h); \tag{2.2}$$

$$P_{\geq 2}(h) = \mathrm{O}(h), \tag{2.3}$$

где $P_{\geq m}(h)$ – вероятность того, что в течение промежутка времени h в систему на обслуживание поступит не менее m требований, $\mathrm{O}(h)$ – величина более высокого порядка малости, чем h. Вероятность поступления в течение промежутка времени h в систему точно одного требования равна $\lambda h + \mathrm{O}(h)$ (см. формулы (2.2), (2.3)).

Далее, используя функцию распределения $F(t) = 1 - e^{-\mu t}$ времени обслуживания τ, подчиняющегося показательному закону с параметром μ, для вероятности P ($\tau < h$) того, что требование будет обслужено за время, меньшее h, находим

$$P(\tau \langle h) = 1 - e^{-\mu h} = \mu h + \mathrm{O}(h), \tag{2.4}$$

откуда

$$P(\tau \rangle h) = e^{-\mu h}. \tag{2.5}$$

Правая часть соотношения (2.5) равна вероятности того, что время обслуживания превзойдет h, т. е. вероятности того, что ни один занятый прибор не освободится за время h. Следовательно, если на обслуживании находится k требований, то вероятность того, что ни один из k обслуживающих приборов не освободится за время h, по теореме умножения вероятностей независимых событий равна $e^{-\mu kh}$. Поэтому для вероятности того, что за время h освободится хотя бы один из k приборов, получаем $1-e^{-\mu kh}=\mu kh+O(h)$. А так как вероятность освобождения двух и более приборов за время h равна $O(h)$, то вероятность освобождения за время h точно одного из k приборов равна

$$\mu kh+O(h). \tag{2.6}$$

Найдем формулы для условных вероятностей $p_{ik}(h)$. Пусть $p_{ik}(h) = P(A)$, где событие A означает, что в системе через время h не будет ни одного требования при условии, что в начальный момент в ней также не было требований. Очевидно, событие A можно представить в виде суммы несовместных событий:

$$A = A_1 + A_2A_3 + A_4A_5\,, \tag{2.7}$$

где A_1 означает, что за время h не поступило ни одного требования; A_2 – за время h поступило одно требование; A_3— это требование будет обслужено до окончания рассматриваемого промежутка времени h; A_4 – за время h поступило более одного требования; A_5 – эти требования будут обслужены до окончания рассматриваемого промежутка времени h.

Но из формул (2.2) – (2.4) следует:

$$P(A_1)=1-\lambda h+O(h);\ P(A_2)=\lambda h+O(h);\ P(A_3)=\mu h+O(h);\ P(A_4)=O(h);$$
$$P(A_5)=O(h).$$

Отсюда и из (2.7), используя теоремы сложения вероятностей несовместных событий и умножения вероятностей независимых событий, находим

$$\begin{aligned}&p_{00}(h)=P(A)=P(A_1)+P(A_2)P(A_3)+P(A_4)P(A_5)=\\&=[1-\lambda h+O(h)]+[\lambda h+O(h)]\cdot[\mu h+O(h)]+O(h)O(h)=1-\lambda h+O(h).\end{aligned} \tag{2.8}$$

Таким образом, вероятность поступления в промежуток времени $(t,\ t + h)$ m требований (m =1, 2, ...) и их обслуживания в этом же промежутке имеет более высокий порядок малости, чем h. По этой причине эту вероятностную компоненту мы далее явно выписывать не будем, считая ее содержащейся в символе $O(h)$.

Пусть $k - i \geq 2$. Тогда $p_{ik}(h)$ равна вероятности того, что в систему за время h поступит на обслуживание не менее двух требований, и, следовательно, по формуле (2.3)

$$p_{ik}(h)=O(h), i\leq k-2\,. \tag{2.9}$$

Если же $i - k \geq 2$, то $p_{ik}(h)$ равна вероятности того, что за время h систему покинут не менее двух обслуженных требований, и, следовательно, по формуле (2.4)

$$p_{ik}(h)=[\mu h+\mathrm{O}(h)]^{i-k}=\mathrm{O}(h), i\geq k+2\ . \qquad (2.10)$$

Величина $p_{k,k+1}(h)$ с точностью до O(h) равна вероятности того, что в систему за время h поступит на обслуживание одно требование, т. е.

$$p_{k,k+1}(h)=\lambda h+\mathrm{O}(h), 0\leq k\langle n\ . \qquad (2.11)$$

Аналогично величина $p_{k,k-1}(h)$ с точностью до O(h) равна вероятности того, что за время h освободится точно один прибор из k, и, следовательно, согласно (2.6)

$$p_{k,k-1}(h)=\mu kh+\mathrm{O}(h), 0< k\leq n\ . \qquad (2.12)$$

Если в системе в некоторый момент времени t находится на обслуживании k требований, то в момент $t+h$ в этой системе может оказаться любое число требований от 0 до n, и поэтому можно записать $\sum\limits_{i=0} p_{ik}(h)=1$.

Отсюда и из формул (2.9) и (2.10)

$$p_{kk}(h)=1-p_{k,k-1}(h)-p_{k,k+1}(h)+\mathrm{O}(h), 0\leq k\leq n \qquad (2.13)$$

или, принимая во внимание соотношения (2.11) и (2.12), получаем

$$p_{kk}(h)=1-\lambda h-\mu kh+\mathrm{O}(h), 1\leq k\leq n-1\ . \qquad (2.14)$$

Наконец, учитывая, что $p_{n,n+1}(h)$=0 (в системе не может находиться n+1 требований, так как рассматривается система с отказами и в ней имеется только n приборов), из соотношений (2.13) и (2.12) получаем

$$p_{nn}(h)=1-\mu kh+\mathrm{O}(h). \qquad (2.15)$$

Подставим теперь оценки (2.8) – (2.15) для $p_{ik}(h)$ в формулу (2.1), получим:

$$\begin{cases} p_0(t+h)=p_0(t)(1-\lambda h)+p_1(t)\mu h+\mathrm{O}(h); \\ p_k(t+h)=p_{k-1}(t)\lambda h+p_k(t)(1-\lambda h-\mu kh)+ \\ \quad +p_{k+1}(t)\mu(k+1)h+\mathrm{O}(h), 1\leq k\leq n-1; \\ p_n(t+h)=p_{n-1}(t)\lambda h+p_n(t)(1-\mu nh)+\mathrm{O}(h). \end{cases} \qquad (2.16)$$

Здесь в каждом уравнении все члены, содержащие множителем величину O(h), объединены вместе и обозначены тем же символом O(h).

Запишем систему (2.16) в виде

$$\begin{cases} \dfrac{p_0(t+h)-p_0(t)}{h}=-\lambda p_0(t)+\mu p_1(t)+\mathrm{O}(1); \\ \dfrac{p_k(t+h)-p_k(t)}{h}=\lambda p_{k-1}(t)-(\lambda+k\mu)p_k(t)+ \\ \quad +\mu(k+1)p_{k+1}(t)+\mathrm{O}(1), 1\leq k\leq n-1; \\ \dfrac{p_n(t+h)-p_n(t)}{h}=\lambda p_{n-1}(t)-\mu np_n(t)+\mathrm{O}(1), \end{cases} \qquad (2.17)$$

где символ O(1) означает величину, стремящуюся к нулю при $h\to 0$.

Система разностных уравнений (2.17) в пределе при $h \to 0$ превращается в систему дифференциальных уравнений:

$$\begin{cases} p_0'(t) = -\lambda p_0(t) + \mu p_1(t); \\ p_k'(t) = \lambda p_{k-1}(t) - (\lambda + k\mu) p_k(t) + \\ + \mu(k+1) p_{k+1}(t), 1 \le k \le n-1; \\ p_n'(t) = \lambda p_{n-1}(t) - \mu n p_n(t). \end{cases} \tag{2.18}$$

Если предположить, что в начальный момент t=0 в системе нет требований, то получим начальные условия:

$$p_0(0) = 1; p_1(0) = 0; \ldots; p_n(0) = 0. \tag{2.19}$$

Хотя система обыкновенных дифференциальных уравнений (2.18) с начальными условиями (2.19) и является линейной, причем с постоянными коэффициентами, найти ее точное, т.е. аналитическое, решение в явном виде при m>2 в общем случае невозможно. Конечно, при желании систему (2.18), (2.19) всегда можно решить с любой степенью точности численно, например методом Рунге – Кутта. Однако численное решение с точки зрения качественного анализа представляет известные неудобства. Вместе с тем качественное исследование решений систем дифференциальных уравнений, возникающих в теории массового обслуживания, весьма желательно.

Рассмотрим более подробно частный случай системы (2.18), (2.19) при n=2, μ=l, λ=2. В этом случае система (2.18) принимает вид

$$\begin{cases} p_0' = -2p_0 + p_1; \\ p_1' = 2p_0 - 3p_1 + 2p_2; \\ p_2' = 2p_1 - 2p_2. \end{cases} \tag{2.20}$$

Используя равенство $p_2(t) = 1 - p_0(t) - p_1(t)$, исключим из второго уравнения системы (2.20) функцию $p_2(t)$, тогда, отбрасывая третье уравнение, получаем:

$$p_0' = -2p_0 + p_1; p_1' = -5p_1 + 2.$$

Решение этой системы, как легко проверить, имеет вид

$$p_0(t) = \frac{1}{5} + \frac{2}{3}e^{-2t} + \frac{2}{15}e^{-5t}; \; p_1(t) = \frac{2}{5} - \frac{2}{5}e^{-5t}.$$

Кроме того,

$$p_2(t) = \frac{2}{5} - \frac{2}{3}e^{-2t} + \frac{4}{15}e^{-5t}.$$

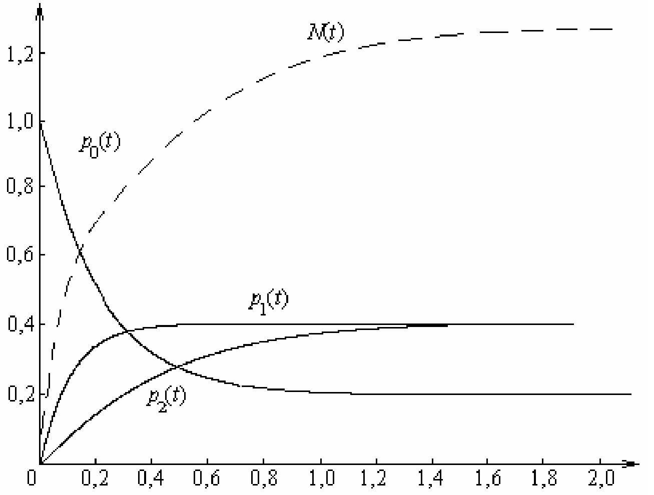

Рис. 4

На рис. 4 функция $N(t) = p_1(t) + 2p_2(t)$ равна среднему числу занятых обслуживающих приборов. В рассматриваемом случае, начиная с момента t =1,5, практически устанавливается стационарный режим, определяемый по формулам

$$p_0 = \lim_{t\to\infty} p_0(t) = 0{,}2; \quad p_1 = \lim_{t\to\infty} p_1(t) = 0{,}4; \quad p_2 = \lim_{t\to\infty} p_2(t) = 0{,}4. \qquad (2.21)$$

Решение (2.21) называется стационарным решением системы (2.20). Оно соответствует так называемому состоянию равновесия, характеризуемому постоянством во времени всех вероятностных показателей системы массового обслуживания. При этом, что весьма примечательно, это решение не зависит от исходного распределения вероятностей числа требований в системе, т.е. от величин $p_0(0)$, $p_1(0)$, $p_2(0)$. Этот факт имеет для теории массового обслуживания неоценимое значение.

Как уже отмечалось, найти явное аналитическое решение системы (2.18) при произвольном n в общем случае невозможно. Однако стационарное решение этой системы, если оно существует, найти легко. Действительно, полагая $\lim\limits_{t\to\infty} p_k(t) = p_k,\ k = 0, 1, ..., n$ и переходя к пределу в дифференциальной системе (2.18) при $t \to \infty$, получаем уже алгебраическую систему:

$$\begin{cases} 0 = -\lambda p_0 + \mu p_1; \\ 0 = \lambda p_{k-1} - (\lambda + k\mu) p_k + \mu(k+1) p_{k+1}, 1 \le k \le n-1; \\ 0 = \lambda p_{n-1} - \mu n p_n. \end{cases} \qquad (2.22)$$

Для упрощения записи системы (2.22) разделим каждое ее уравнение на μ, и введем обозначение $\alpha=\lambda/\mu$. В результате будем иметь:

$$\begin{cases} 0 = -\alpha p_0 + p_1; \\ 0 = \alpha p_{k-1} - (\alpha + k) p_k + (k+1) p_{k+1}, 1 \le k \le n-1; \\ 0 = \alpha p_{n-1} - n p_n. \end{cases} \tag{2.23}$$

Предположим, что p_0 известно. Тогда, используя последовательно уравнения системы (2.23), получаем:

$$\begin{cases} p_1 = \alpha p_0; \\ p_2 = \dfrac{1}{2}\left[-\alpha p_0 + (\alpha+1) p_1\right] = \dfrac{\alpha^2 p_0}{2}; \\ p_2 = \dfrac{1}{3}\left[-\alpha p_1 + (\alpha+2) p_2\right] = \dfrac{\alpha^3 p_0}{3!} \end{cases}$$

и вообще для любого $k \le n$

$$p_k = \alpha^k p_0 / k!. \tag{2.24}$$

Чтобы избавиться в формуле (2.24) от p_0, а заодно и определить p_0, используем условие нормировки $\sum\limits_{k=0}^{n} p_k = 1$, справедливое также и при $t \to \infty$. Подставляя в это условие вместо p_k правую часть равенства (2.24), получаем $p_0 \sum\limits_{k=0}^{n} \dfrac{\alpha^k}{k!} = 1$, откуда $p_0 = \dfrac{1}{\sum\limits_{k=0}^{n} \dfrac{\alpha^k}{k!}}$. (2.25)

Наконец, из формул (2.24) и (2.25) окончательно находим

$$p_k = \frac{\alpha^k / k!}{\sum\limits_{k=0}^{n} \dfrac{\alpha^k}{k!}}, k = 0, 1, ..., n. \tag{2.26}$$

Формулы (2.26) называются формулами Эрланга. Они дают явное аналитическое решение для стационарного распределения вероятностей числа требований, находящихся на обслуживании в системе с отказами.

Заметим, что прямой путь получения стационарного решения из нестационарного при $t \to \infty$ (как это было сделано для случая n=2, μ=l, λ=2) практически исключается из-за технических трудностей, возникающих при нахождении нестационарного решения системы (2.18) в явном виде. Вместе с тем на практике обычно интересуются именно установившимися во времени (т. е. стационарными) вероятностными характеристиками функционирования системы массового обслуживания.

Для рассматриваемой системы с отказами наибольший интерес представляет вероятность отказа, т. е. вероятность того, что в момент поступления очередного требования на обслуживание все приборы будут заняты:

$$p_{\text{отк}} = p_n = \frac{\alpha^n / n!}{\sum_{k=0}^{n} \frac{\alpha^k}{k!}} = p_0 \frac{\alpha^n}{n!}, \tag{2.27}$$

и среднее число приборов, занятых обслуживанием:

$$M_4 = \sum_{k=1}^{n} k p_k = p_0 \sum_{k=1}^{n} \frac{\alpha^k}{(k-1)!} = \alpha(1 - p_n). \tag{2.28}$$

p_n характеризует полноту обслуживания входящего потока, M_4 – степень загрузки системы массового обслуживания.

Пример. Автоматическая телефонная станция (АТС) имеет пять линий связи (пять обслуживающих приборов). Очевидно, что моменты поступления на АТС требований, т. е. заявок на соединение с абонентом для ведения разговоров, равно как и продолжительность этих разговоров, являются случайными. Предположим, что поток требований – пуассоновский с интенсивностью λ=2 вызова в минуту и время каждого разговора имеет показательное распределение с математическим ожиданием $1/\mu$ равным 1,5 мин. Предполагается, что требование получает отказ, если в момент его поступления на АТС все пять линий связи заняты. В этом случае n=5, $\alpha=\lambda/\mu=2\cdot1{,}5=3$ и, следовательно, для вероятности того, что телефонный разговор с первого вызова не состоится, согласно формуле (2.27)

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

Среднее число занятых линий во время работы АТС определим по формуле (2.28): N=3 (1—0,11) =2,67 линии. Поэтому коэффициент загрузки линий связи данной АТС $\xi_4 = M_4 / n = 2{,}67/5 = 0{,}53$, а коэффициент простоя $\xi_3 = 1 - 0{,}53 = 0{,}47$.

Работу рассмотренной АТС вряд ли можно считать удовлетворительной, так как станция не соединяет с абонентом в среднем в 11% случаев (p_5 = 0,11). Очевидно, что 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 АТС в 10 раз, т.е., чтобы вероятность отказа в соединении с абонентом не превосходила 0,011. Для этого, используя формулу (2.27) с $\alpha = 3$, составим таблицу:

n	1	2	3	4	5	6	7	8
p_n	0,750	0,529	0,346	0,206	0,110	0,052	0,022	0,008

Отсюда следует, что если проектировать АТС для данных значений λ и μ так, что она сможет одновременно обслуживать восемь разговоров (n=8), то только в 0,8% случаев будет получен отказ из-за занятости всех восьми линий связи.

2.2. Системы с бесконечным числом приборов

Рассмотрим систему массового обслуживания, отличающуюся от рассмотренной в предыдущем подразделе только тем, что число обслуживающих приборов неограниченно, т.е. $n = \infty$. Естественно считать, что при любой интенсивности входящего потока все требования, поступающие в систему, начинают немедленно обслуживаться, ибо в рассматриваемой системе всегда имеются незанятые приборы. Таким образом, если в системе с отказами очередь из требований не образуется просто из-за того, что требования покидают систему, когда в ней все приборы заняты, то здесь очередь не возникает по другой причине.

Поскольку предположения относительно входящего потока требований и времени обслуживания этих требований остаются те же, что и для случая, рассмотренного в предыдущем подразделе (промежутки времени между моментами поступления требований независимы и имеют показательное распределение с параметром λ, промежутки времени обслуживания этих требований также независимы и имеют показательное распределение, но с параметром μ), то для нахождения $p_k(t)$ здесь может быть использована система (2.18), за исключением последнего ее уравнения, имеющего специальный вид из-за ограниченного числа обслуживающих приборов.

Отбрасывая это уравнение и полагая в (2.18) $k \ge 1$ вместо $1 \le k \le n-1$, получаем

$$\begin{cases} p_0'(t) = -\lambda p_0(t) + \mu p_1(t); \\ p_k'(t) = \lambda p_{k-1}(t) - (\lambda + k\mu) p_k(t) + \mu(k+1) p_{k+1}(t), \; k \ge 1. \end{cases} \tag{2.29}$$

Система (2.29), в отличие от аналогичной (2.18), имеет бесконечный порядок. Однако отсутствие последнего замыкающего уравнения специального вида позволяет найти ее явное аналитическое решение. Но мы этого делать не будем, поскольку нас интересует только стационарное решение. Его можно получить предельным переходом при $n \to \infty$ в (2.26):

$$p_k = \frac{\alpha^k / k!}{\sum_{k=0}^{n} \frac{\alpha^k}{k!}}, k = 0, 1, \ldots (\alpha = \lambda / \mu).$$

В результате

$$p_k = \lim_{t \to \infty} p_k(t) = \frac{\alpha^k e^{-\alpha}}{k!}, k = 0, 1, \ldots \quad (2.30)$$

Пример. Пусть на некоторую станцию неотложной медицинской помощи поступают за час в среднем два вызова врачей. Моменты поступления этих вызовов, очевидно, являются случайными величинами. Предположим, что поток вызовов врачей является пуассоновским и, следовательно, характеризуется только одним параметром. В нашем случае этот параметр λ=2 вызовам в час. Будем считать, что после получения каждого вызова станция немедленно высылает врача к больному, а время, затрачиваемое на оказание медицинской помощи и дорогу в оба конца, имеет показательное распределение с математическим ожиданием $1/\mu$ =1,5 ч.

Состояние данной системы массового обслуживания определяется количеством врачей, находящихся на обслуживании (включая дорогу). Обозначим через p_k вероятность того, что в установившемся режиме на обслуживании находятся k врачей. Так как в нашем случае λ/μ=2·1,5=3, то согласно формуле (2.30) $p_k = \frac{1}{k!} 3^k e^{-3}, k = 0, 1, \ldots$ Ниже приведены значения p_k найденные по этой формуле:

k	0	1	2	3	4	5	6	7	8	9
p_k	0,050	0,149	0,224	0,224	0,168	0,101	0,050	0,022	0,008	0,003

Отсюда видно, что наиболее вероятными являются состояния, при которых заняты одновременно два или три врача. Вероятность того, что ни один врач не будет занят, равна 0,05. Следовательно, вероятность того, что хотя бы один врач находится на обслуживании, равна 0,95.

Если при той же средней продолжительности обслуживания вызовов в единицу времени поступает меньше (например, в среднем один вызов за час), то получаются следующие значения p_k:

k	0	1	2	3	4	5	6
p_k	0,223	0,335	0,251	0,126	0,047	0,014	0,003

Здесь уже наиболее вероятным является состояние, при котором занят только один врач. Вероятность того, что все врачи будут не заняты, равна 0,223.

2.3. Системы массового обслуживания с ожиданием

Рассмотрим опять, как и в подразд. 2.1, систему массового обслуживания, состоящую из конечного числа n однотипных обслуживающих приборов. Но требование, заставшее все обслуживающие приборы занятыми, не покидает систему, а «становится» в очередь и ждет, пока его не обслужит один из освободившихся приборов. Разумеется, в системе с бесконечным числом обслуживающих приборов такая ситуация возникнуть не может.

Как обычно, будем предполагать, что входящий поток требований является пуассоновским с интенсивностью λ, а продолжительность каждого обслуживания каждым прибором подчиняется показательному закону с параметром μ. При этом продолжительности обслуживания, рассматриваемые априорно, считаются стохастически независимыми как друг от друга, так и от моментов поступления требований в систему.

Обозначим через $p_k(t)$ вероятность того, что система находится в состоянии k, т. е. вероятность того, что в системе в момент времени t находятся на обслуживании и в очереди (если таковая имеется) k требований. Если $k \le n$, то в системе нет очереди и число свободных от обслуживания приборов равно n-k. Легко видеть, что при $k<n$ функционирование данной системы совпадает с функционированием системы с отказами, рассмотренной в подразд 2.1, и, следовательно, величины $p_0(t)$, $p_1(t)$,..., $p_{n-1}(t)$ удовлетворяют первым уравнениям системы (2.18):

$$p_0'(t) = -\lambda p_0(t) + \mu p_1(t); \tag{2.31}$$

$$p_k'(t) = \lambda p_{k-1}(t) - (\lambda + k\mu) p_k(t) + \mu(k+1) p_{k+1}(t), 1 \le k \le n-1. \tag{2.32}$$

Последнее уравнение системы (2.18) здесь не используется, так как оно составлено в предположении, что переход системы из состояния n в состояние n+1 невозможен. С другой стороны, рассматриваемая система с ожиданием может находиться в любом состоянии k = 0, 1, ...

Обозначим через $p_{ik}(h)$ вероятность перехода системы из состояния i в состояние k за время h, т. е. условную вероятность того, что через время h в системе на обслуживании и в ожидании его будут находиться k требований, если вначале на обслуживании и в ожидании его находилось i требований. Тогда, рассуждая так же, как в подразд. 2.1, для любого $k \ge n$ получаем:

$$p_{ik}(h) = \mathrm{O}(h), |i-k| \ge 2; \tag{2.33}$$

$$p_{k-1,k}(h) = \lambda h + \mathrm{O}(h); \tag{2.34}$$

$$p_{k+1,k}(h) = n\mu h + \mathrm{O}(h); \tag{2.35}$$

где O(h) – бесконечно малая более высокого порядка малости, чем h.

Формулы (2.33) и (2.34) совпадают с соответствующими формулами подразд. 2.1, (2.35) заменяет формулу $p_{k+1,k}(h)=(k+1)\mu h+O(h)$, полученную в подразд. 2.1 для СМО с отказами и отражает тот факт, что при $k \ge n$ в системе с ожиданием, находящейся в состоянии k+1, имеется не k+1,а n занятых приборов.

Далее из условия нормировки $\sum\limits_{i=0}^{\infty} p_{ik}(h)=1$ и равенств (2.33) – (2.35) находим

$$p_{kk}(h)=1-p_{k,k-1}(h)-p_{k,k+1}(h)+O(h)=1-n\mu h-\lambda h+O(h). \qquad (2.36)$$

Подставляя теперь равенства (2.33) – (2.36) в формулу полной вероятности $p_k(t+h)=\sum\limits_{i=0}^{\infty} p_i(t)p_{ik}(h)$, имеем $p_k(t+h)=p_{k+1}(t)\lambda h+$ $+p_k(t)(1-n\mu h-\lambda h)+p_{k+1}(t)n\mu h+O(h)$, откуда при $h \to 0$

$$p_k'(t)=\lambda p_{k-1}(t)-(\lambda+n\mu)p_k(t)+n\mu p_{k+1}(t), k \ge n. \qquad (2.37)$$

Полагая $\lim\limits_{t\to\infty} p_k(t)=p_k$ и переходя к пределу при $t \to \infty$ в уравнениях (2.31), (2.32), (2.37), получаем бесконечную систему уже алгебраических уравнений:

$$\begin{cases} 0=-\lambda p_0+\mu p_1; \\ 0=\lambda p_{k-1}-(\lambda+k\mu)p_k+\mu(k+1)p_{k+1}, 1 \le k \le n-1; \\ 0=\lambda p_{k-1}-(\lambda+n\mu)p_k+n\mu p_{k+1}, k \ge n. \end{cases} \qquad (2.38)$$

Для упрощения записи разделим каждое уравнение системы (2.38) на μ и введем обозначение α=λ/μ, в результате получим:

$$\begin{cases} 0=-\alpha p_0+p_1; \\ 0=\alpha p_{k-1}-(\alpha+k)p_k+(k+1)p_{k+1}, 1 \le k \le n-1; \\ 0=\alpha p_{k-1}-(\alpha+n)p_k+np_{k+1}, k \ge n. \end{cases} \qquad (2.39)$$

Если положить $z_k=\begin{cases} \alpha p_{k-1}-kp_k, 1 \le k \le n; \\ \alpha p_{k-1}-np_k, k \ge n, \end{cases}$ то система (2.39) примет вид $z_1=0; z_k-z_{k+1}=0, k \ge 1,$ откуда $z_k=0, k \ge 1$. А это означает, что

$$\alpha p_{k-1}=\begin{cases} kp_k, 1 \le k \le n; \\ np_k, k \ge n \end{cases} \text{ или } p_k=\begin{cases} \dfrac{\alpha}{k}p_{k-1}, 1 \le k \le n; \\ \dfrac{\alpha}{n}p_{k-1}, k \ge n. \end{cases}$$

Решая последовательно эти рекуррентные соотношения, получаем

$$p_k = \begin{cases} \dfrac{\alpha^k}{k!} p_0, & 1 \le k \le n; \\ \left(\dfrac{\alpha}{n}\right)^{k-n} p_n = \dfrac{\alpha^k}{n! n^{k-n}} p_0, k \ge n. \end{cases} \tag{2.40}$$

Для определения p_0 используем условие нормировки

$$\sum_{k=0}^{\infty} p_k = 1, \tag{2.41}$$

означающее, что обслуживающая система и в установившемся режиме находится в одном и только в одном из возможных состояний 0, 1, ...

Из (2.41) и (2.40)

$$p_0 \left[\sum_{k=0}^{n-1} \frac{\alpha^k}{k!} + \frac{\alpha^n}{n!} \sum_{k=n}^{\infty} \left(\frac{\alpha}{n}\right)^{k-n} \right] = 1. \tag{2.42}$$

Если выполняется условие

$$\alpha < n , \tag{2.43}$$

то $\sum_{k=n}^{\infty} \left(\frac{\alpha}{n}\right)^{k-n} = \sum_{k=0}^{\infty} \left(\frac{\alpha}{n}\right)^{k} = \frac{1}{1-\frac{\alpha}{n}} = \frac{n}{n-\alpha}$ и, следовательно,

$$p_0 \left(\sum_{k=0}^{n-1} \frac{\alpha^k}{k!} + \frac{\alpha^n}{n!} \frac{n}{n-\alpha} \right) = 1 .$$

Таким образом, вероятность того, что в системе с ожиданием для стационарного режима не будет находиться на обслуживании ни одно требование, можно найти по формуле

$$p_0 = \left[\sum_{k=0}^{n-1} \frac{\alpha^k}{k!} + \frac{\alpha^n}{(n-1)!(n-\alpha)} \right]^{-1}. \tag{2.44}$$

Вероятность же того, что в этой системе на обслуживании будет находиться $k(k \ge 1)$ требований, можно найти по формулам (2.40), (2.44).

При $n = 1$, т. е. в случае, когда система массового обслуживания состоит только из одного прибора, из формул (2.40) и (2.44) $p_k = \alpha^k (1-\alpha), k = 0, 1, ...$

Если условие (2.43) не выполняется, то геометрическая прогрессия в квадратных скобках в (2.42) расходится. Это означает, что p_0=0, а также в силу (2.40) p_k=0, k =1, 2, ... Таким образом, если условие (2.43) не выполняется, то вероятность того, что в системе находится на обслуживании любое конечное число требований, при $t \to \infty$ равна нулю. Иными словами, при нарушении условия (2.43) длина очереди при $t \to \infty$ становится бесконечной. Система явно не справляется с обслуживанием.

Заметим, что (2.43) выполняется, если имеет место по крайней мере одно из следующих трех условий: интенсивность λ входящего потока достаточно мала; среднее время обслуживания $1/\mu$ каждого требования каждым прибором достаточно мало; число приборов n достаточно велико. Далее будем предполагать, что условие (2.43) выполнено.

Формулы (2.40) позволяют легко найти следующие средние характеристики функционирования системы массового обслуживания с ожиданием: M_1 – среднюю длину очереди, т. е. математическое ожидание числа требований, ожидающих начала обслуживания; M_2 – среднее число требований, находящихся в системе (как обслуживаемых, так и стоящих в очереди); M_3 – среднее число свободных от обслуживания приборов. Действительно, используя формулу $\sum_{i=0}^{\infty} i\alpha^i = \frac{\alpha}{(1-\alpha)^2}, 0 \le \alpha \le 1,$ получающуюся в результате дифференцирования равенства $\sum_{i=0}^{\infty} \alpha^i = 1/(1-\alpha)$ и умножения результата на α, и формулы (2.40), находим:

$$M_1 = \sum_{k=n+1}^{\infty}(k-n)p_k = \frac{p_0 n^n}{n!}\sum_{k=n+1}^{\infty}(k-n)\left(\frac{\alpha}{n}\right)^k =$$

$$= \frac{p_0 n^n}{n!}\sum_{i=0}^{\infty} i\left(\frac{\alpha}{n}\right)^{n+i} = p_n \sum_{i=0}^{\infty} i\left(\frac{\alpha}{n}\right)^i = \frac{\alpha p_n}{n\left(1-\frac{\alpha}{n}\right)^2}; \qquad (2.45)$$

$$M_2 = \sum_{k=1}^{\infty} k p_k = p_0 \sum_{k=1}^{n-1}\frac{k}{k!}\alpha^k + p_n \sum_{k+n}^{\infty} k\left(\frac{\alpha}{n}\right)^{k-n} =$$

$$= p_0 \sum_{k=1}^{n-1}\frac{\alpha^k}{(k-1)!} + p_n \sum_{i=0}^{\infty}(n+i)\left(\frac{\alpha}{n}\right)^i =$$

$$= p_0 \sum_{k=1}^{n-1}\frac{\alpha^k}{(k-1)!} + \frac{p_n n}{1-\frac{\alpha}{n}} + M_1; \qquad (2.46)$$

$$M_3 = \sum_{k=0}^{n-1}(n-k)p_k = p_0 \sum_{k=0}^{n-1}\frac{n-k}{k!}\alpha^k. \qquad (2.47)$$

Эти средние характеристики можно использовать для выбора лучшего варианта при проектировании системы массового обслуживания. При этом в общем случае исходят из условия минимизации полных потерь за счет ожидания в очереди и простоя приборов. Например, число n приборов с экономической точки зрения естественно определять из условия минимума функции $\varphi(n) = C_{\text{ож}} M_1 + C_{nk} M_3$, где $C_{\text{ож}}$ —удельные издержки простоя требования в очереди; C_{nk} – удельные издержки простоя прибора.

Формулы (2.40) позволяют легко вычислить вероятность π того, что все n обслуживающих приборов заняты. Действительно, используя теорему сложения вероятностей несовместных событий, получаем

$$\pi = \sum_{k=n}^{\infty} p_k = \frac{n^n p_0}{n!} \sum_{k=n}^{\infty} \left(\frac{\alpha}{n}\right)^k = \frac{p_0 \alpha^n}{(n-1)!\,(n-\alpha)}. \qquad (2.48)$$

Таким образом, вероятность того, что все приборы окажутся занятыми и, следовательно, требованию придется встать в очередь, равна правой части равенства (2.48). Иначе эту вероятность называют вероятностью ожидания. Для сравнения напомним, что в случае системы с отказами, рассмотренной в подразд. 2.1, вероятность того, что все n обслуживающих приборов окажутся занятыми («вероятность отказа»), определяется по формуле: $p_{\text{отк}} = p_0 \alpha^n / n!$.

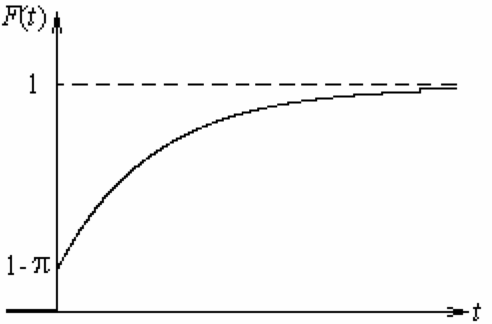

Рис. 5

Однако для систем с ожиданием вероятность ожидания π играет меньшую роль, чем аналогичная величина – вероятность отказа $p_{\text{отк}}$ для систем с отказами. Дело в том, что даже при значениях π, близких к единице, функционирование системы массового обслуживания может быть признано удовлетворительным, если только время ожидания θ начала обслуживания для подавляющей части требований мало. В связи с этим для системы с ожиданием важное значение имеет функция распределения $F(t)$ времени ожидания, или, что то же, времени пребывания в очереди. Эта функция имеет вид (рис. 5):

$$F(t)=P(\theta<t)=1-P(\theta>t)-P(\theta=t)=\begin{cases}1-\pi e^{-(n\mu-\lambda)t};\ t\rangle 0;\\ 0, \qquad t\le 0.\end{cases} \tag{2.49}$$

Наличие разрыва непрерывности и интервала возрастания функции распределения $F(t)$ свидетельствует о том, что время θ ожидания в очереди является случайной величиной смешанного типа, из (2.49) ее математическое ожидание

$$M\theta=\int_0^\infty tdF(t)=0\cdot(1-\pi)+\int_0^\infty tF'(t)dt=\pi\int_0^\infty t(n\mu-\lambda)e^{-(n\mu-\lambda)}dt.$$

Отсюда, интегрируя по частям, для среднего времени ожидания начала обслуживания, получаем

$$M\theta=\pi\left[-te^{-(n\mu-\lambda)t}\Big|_0^\infty+\int_0^\infty e^{-(n\mu-\lambda)t}dt\right]=-\frac{\pi e^{-(n\mu-\lambda)t}}{n\mu-\lambda}\Big|_0^\infty=\frac{\pi}{n\mu-\lambda} \text{ или}$$

$$M\theta=\pi/\mu(n-\alpha). \tag{2.50}$$

Как и следовало ожидать, среднее время ожидания начала обслуживания оказалось пропорциональным среднему времени обслуживания $1/\mu$ требования в системе и вероятности π того, что все приборы заняты. Заметим, что если уже найдена средняя длина M_1 очереди, то среднее время $M\theta$ ожидания в очереди легко можно найти по формуле $M\theta = M_1/\lambda$. Действительно, при установившемся режиме число покидающих в единицу времени систему обслуженных требований равно числу поступающих в единицу времени в систему на обслуживание требований. А поступает в единицу времени λ требований.

Пример. В некоторой мастерской по ремонту часов имеется n мастеров, работающих с одинаковой производительностью. Наблюдения показали, что в течение рабочего дня от населения поступает на ремонт в среднем 30 шт. часов разных марок, причем каждый мастер за один рабочий день ремонтирует в среднем 10 часов. Общее число часов у населения велико и они выходят из строя в случайные моменты времени. По всей вероятности, клиенты приносят часы на ремонт более или менее равномерно в течение рабочего дня (стационарность), независимо друг от друга (отсутствие последействия) и не группами в два и более человека (ординарность). Значит, рассматриваемый поток требований можно считать пуассоновским, причем с интенсивностью λ = 30 шт. часов в рабочий день.

Время ремонта подавляющей части неисправных часов сравнительно мало (замена разбитого стекла, стрелки и др.). В капитальном же ремонте, требующем большого времени (изготовление нестандартных деталей для импортных часов, точка осей и пр.), часы нуждаются сравнительно редко. В связи с этим предположим, что время обслуживания (время ремонта часов) подчиняется показательному закону, причем с параметром μ =10 шт. часов в рабочий день.

Проанализируем работу этой мастерской с точки зрения теории массового обслуживания. Прежде всего заметим, что для выполнения условия (2.43) необходимо, чтобы в мастерской было не менее четырех мастеров. Если $n < 4$, то условие (2.43) не будет выполнено: $\alpha=\lambda/\mu = 3< n$.

Пусть в мастерской работает минимально возможное число мастеров, т.е. $n = 4$. Тогда вероятность p_0 того, что все четыре мастера в момент поступления очередных часов на ремонт не заняты, можно найти по формуле (2.44):

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

Таким образом, в течение семичасового рабочего дня в среднем $0{,}038\cdot 7\cdot 60 \approx 14$ мин все четыре мастера свободны.

С другой стороны, для вероятности π того, что все 4 мастера в момент поступления очередных часов на ремонт заняты, согласно формуле (2.48) имеем

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

т. е. все мастера заняты половину рабочего времени.

Клиента интересует, прежде всего, среднее время $M\theta$ ожидания начала ремонта его часов после того, как он их отнес в мастерскую. Это время можно найти по формуле (2.50):

$$M\theta = \frac{0{,}51}{10(4-3)} = 0{,}051 \text{ раб. дня } \approx 21 \text{ мин.}$$

Руководство мастерской, по-видимому, прежде всего интересует среднее число часов M_1, ожидающих ремонта, и общее среднее число часов M_2, ожидающих ремонта и ремонтируемых. По формулам (2.40), (2.45) и (2.46) находим:

$$p_4 = \frac{3^4}{4!}0{,}038 = 0{,}13;\quad M_1 = \frac{3\cdot 0{,}13}{4\left(1-\frac{3}{4}\right)^2} = 1{,}56 \text{ шт.};$$

$$M_2 = 0{,}038\left(3 + \frac{3^2}{1} + \frac{3^3}{2!}\right) + \frac{0{,}13 \cdot 4}{1 - \dfrac{3}{4}} + 1{,}56 = 4{,}58 \text{ шт.}$$

Представляет, конечно, интерес и среднее число M_3 мастеров, свободных от работы. По формуле (2.47) получаем

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

Это означает, что каждый мастер свободен в среднем одну четверть рабочего времени, т. е. 1,75 ч в день.

Для $n = 5$ аналогичные вычисления дают: $p_0 = 0{,}047$; $\pi = 0{,}24$; $M\theta = 5$ мин; $M_1 = 0{,}35$ шт.; $M_2 = 3{,}37$ шт.; $M_3 = 2$.

Сравнение всех этих результатов показывает, что оптимальное число мастеров равно 4. Если $n < 4$, это означает, что со временем груда неисправных часов, поступающих в мастерскую на ремонт, будет неограниченно увеличиваться и, таким образом, мастерская явно не будет справляться со своей работой. Если же $n > 4$, то выигрыш, получаемый клиентами, будет несущественным по сравнению с экономическими затратами, связанными с привлечением дополнительных мастеров.

Так, если вместо четырех использовать пять мастеров, то среднее время ожидания начала ремонта часов уменьшится с 21 до 5 мин. Однако вряд ли это даст какое-либо преимущество клиентам, так как ремонт в их присутствии практически все равно исключается из-за большого среднего времени ремонта одних часов. Вместе с тем при $n = 5$ только 24% рабочего времени все мастера полностью загружены работой, причем в среднем два мастера все время свободны. Иными словами, загрузка мастеров низкая.

Из графиков функции распределения времени ожидания (см. формулу (2.49)) начала ремонта часов

$$F(t) = \begin{cases} 1 - 0{,}51e^{-10t}, t > 0, n = 4; \\ 1 - 0{,}24e^{-20t}, t > 0, n = 5; \\ 0, \qquad t \le 0, n = 4;5, \end{cases}$$

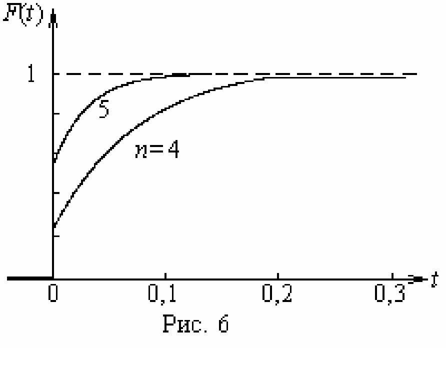

Рис. 6

(рис. 6) следует, что вероятность того, что время ожидания начала ремонта часов окажется меньшим, например 0,1 раб. дня, при $n = 4$ и $n = 5$ соответственно равна 0,81 и 0,97.

2.4. Замкнутые системы массового обслуживания

До сих пор мы рассматривали только такие системы массового обслуживания, для которых интенсивность λ входящего потока требований не зависит от состояния системы. В этом случае источник требований находится вне системы и генерирует неограниченный поток требований. В этом подразделе обратимся к системы массового обслуживания, для которых λ зависит от состояния системы, причем источник требований является «внутренним» и генерирует ограниченный поток требований.

Например, пусть в гараже имеется m автомашин и $n(n<m)$ площадок для ремонта, причем на каждой площадке может находиться только одна автомашина. Здесь автомашины – источники требований, а ремонтные площадки – обслуживающие приборы. Неисправная автомашина после обслуживания (ремонта) используется по своему прямому назначению и вновь становится потенциальным источником возникновения требований на обслуживание. Ясно, что интенсивность поступления требований на обслуживание зависит от того, сколько автомашин в данный момент находится в эксплуатации и сколько автомашин обслуживается или стоит в очереди, ожидая обслуживания.

Итак, рассмотрим систему массового обслуживания, состоящую из приборов, каждый из которых может одновременно обслуживать только одно требование, причем, как обычно, время обслуживания одного требования есть случайная величина с функцией распределения $F(t)=1-e^{-\mu t}$.

Входящий поток требований исходит из $m(m>n)$ обслуживаемых объектов, которые в случайные моменты времени выходят из строя и требуют обслуживания. При этом каждый объект, находящийся в эксплуатации, генерирует независимо от других объектов пуассоновский поток требований с интенсивностью λ, так что если в эксплуатации находится k объектов, то общий (суммарный) входящий поток имеет интенсивность $k\lambda$. Требование, поступившее в систему в момент, когда свободен хотя бы один прибор, немедленно идет на обслуживание. Если требование застает все приборы занятыми обслуживанием других требований, то оно не покидает систему (как в случае систем с отказами), а становится в очередь и ждет до тех пор, пока один из приборов не станет свободным (как в случае систем с ожиданием).

Системы массового обслуживания, в которых входящий поток требований формируется из выходящего, называются замкнутыми.

Как и в подразд. 2.3, будем говорить, что система массового обслуживания находится в состоянии k, если общее число требований, находящихся на обслуживании и в очереди, равно k. Для рассматриваемой замкнутой системы, очевидно, $k = 0, 1, \ldots, m$. При этом если система находится в состоянии k, то число объектов, находящихся в эксплуатации, равно $m - k$. Заметим, что из этих и только этих $m - k$ объектов формируется в данный момент времени входящий поток требований.

Составим систему дифференциальных уравнений для вероятности $p_k(t)$ того, что замкнутая система массового обслуживания в момент времени t находится в состоянии $k = 0, 1, ..., m$. Для этой цели выпишем ряд элементарных, ранее уже неоднократно встречавшихся, формул.

Так как входящий поток требований по предположению является пуассоновским с интенсивностью λ, то вероятность того, что за время h в систему поступит k требований,

$$P_k(h) = \frac{(\lambda h)^k}{k!} e^{-\lambda h}, \tag{2.51}$$

отсюда

$$P_0(h) = e^{-\lambda h} = 1 - \lambda h + \mathrm{O}(h), \tag{2.52}$$

$$P_1(h) = \lambda h e^{-\lambda h} = \lambda h + \mathrm{O}(h). \tag{2.53}$$

Далее, так как время τ обслуживания подчиняется показательному закону с параметром μ, то

$$P(\tau < h) = 1 - e^{-\mu h}; \tag{2.54}$$

$$P(\tau > h) = e^{-\mu h}. \tag{2.55}$$

Далее применим стандартный классический метод, а именно в формуле полной вероятности

$$p_k(t+h) = \sum_{i=0}^{m} p_i(t) p_{ik}(h) \tag{2.56}$$

последовательно оценим переходные вероятности для различных возможных комбинаций индексов i и k, подставим эти оценки в формулу (2.56) и перейдем к пределу при $h \to 0$. В результате получим следующие четыре случая.

Случай 1

$$p_{kk}(h) = \begin{cases} 1 - \lambda(m-k)h - \mu k h + \mathrm{O}(h), 0 \le k < n; \\ 1 - \lambda(m-k)h - \mu n h + \mathrm{O}(h), n \le k \le m. \end{cases} \tag{2.57}$$

Действительно, пусть система в момент времени t находится в состоянии k, $0 \le k < n$. Тогда через время h она останется в том же состоянии k, если за это время не поступит в систему ни одно из $m - k$ возможных требований (событие A_1) и не покинут систему ни одно из k требований, находящихся там в момент времени t на обслуживании (событие A_2). Но, используя теорему умножения вероятностей независимых событий и формулы (2.52) и (2.55), получаем:

$$P(A_1)=\left(e^{-\lambda h}\right)^{m-k}, \tag{2.58}$$

$$P(A_2)=\left(e^{-\mu h}\right)^{k}. \tag{2.59}$$

Следовательно, учитывая независимость событий A_1 и A_2, можно записать

$$p_{kk}(h)=P(A_1A_2)=P(A_1)P(A_2)=\left(e^{-\lambda h}\right)^{m-k}\left(e^{-\mu h}\right)^{k}=$$
$$=[1-(m-k)\lambda h+\mathrm{O}(h)]\cdot[1-\mu kh+\mathrm{O}(h)]=1-(m-k)\lambda h-\mu kh+\mathrm{O}(h)..$$

Случай, когда $n \le k < m$ отличается от рассмотренного $(0 \le k < n)$ только тем, что при $k > n$ обслуживаются не k, а только n требований (так как в системе имеется только n обслуживающих приборов, каждый из которых может обслуживать только одно требование).

Случай 2

$$p_{k,k-1}(h)=\begin{cases}\mu kh+\mathrm{O}(h), 0<k<n;\\ \mu nh+\mathrm{O}(h), n\le k\le m.\end{cases} \tag{2.60}$$

Действительно, система за время h перейдет из состояния $k(0<k<n)$ в состояние $k-1$, если за это время не поступит в систему ни одно из $m-k$ возможных требований (событие A_1) и покинет систему (будучи обслуженным) только одно из каких-либо k обслуживаемых требований (событие A_3). Но по формуле Бернулли

$$P(A_3)=k\left(1-e^{-\mu h}\right)\left(e^{-\mu h}\right)^{k-1}=k\mu h+\mathrm{O}(h). \tag{2.61}$$

Поэтому, учитывая независимость событий A_1 и A_3 и используя формулу (2.58), запишем

$$p_{k,k-1}(h)=P(A_1A_3)=P(A_1)P(A_3)=k\mu h+\mathrm{O}(h).$$

Случай, когда $n\le k\le m$, отличается от рассмотренного $(0<k<n)$ только тем, что при $k\ge n$ вероятность $P(A_3)$ того, что освободится только один какой-либо из обслуживаемых приборов, не зависит от k, и, следовательно, вместо (2.61) справедливо равенство

$$P(A_3)=n\left(1-e^{-\mu h}\right)\left(e^{-\mu h}\right)^{n-1}=n\mu h+\mathrm{O}(h).$$

Случай 3

$$p_{k,k+1}(h)=(m-k)\lambda h+\mathrm{O}(h), 0\le k<m. \tag{2.62}$$

Действительно, система за время h перейдет из состояния $k(0\le k<m)$ в состояние $k+1$, если за это время в систему поступит только одно из $m-k$ возможных требований (событие A_4) и систему не покинет ни одно из k требований (событие A_2). Но, применяя теорему сложения вероятностей несовместных событий и формулу (2.53), получаем $P(A_4)=(m-k)\lambda he^{-\lambda h}$.

Поэтому, учитывая независимость событий A_2 и A_4 и используя (2.59), имеем:

$$p_{k,k+1}(h) = P(A_4 A_2) = P(A_4)P(A_2) =$$
$$= (m-k)\lambda h e^{-\lambda h}\left(e^{-\mu h}\right)^k = (m-k)\lambda h + \mathrm{O}(h).$$

Случай 4

$$p_{ik}(h) = \mathrm{O}(h), |i-k| \geq 2 . \qquad (2.63)$$

Действительно, вероятность того, что за время h в систему поступят два или более требования, согласно (2.51) при $k \geq 2$ равна $\mathrm{O}(h)$. С другой стороны, вероятность того, что за время h систему покинут два или более требования, согласно (2.54) и теореме умножения вероятностей независимых событий также равна $\mathrm{O}(h)$. Поэтому вероятность перехода системы из состояния i в состояние k при $|i-k| \geq 2$ за время h тем более равна $\mathrm{O}(h)$.

Подставим теперь полученные оценки (2.57), (2.60), (2.62), (2.63) для $p_{ik}(h)$ в формулу (2.56). Тогда, например, при $k = 0$ получим

$$p_0(t+h) = p_0(t)p_{00}(h) + p_1(t)p_{10}(h) + \mathrm{O}(h) =$$
$$= p_0(t)[1 - \lambda h + \mathrm{O}(h)] + p_1(t)[\mu h + \mathrm{O}(h)] + \mathrm{O}(h) \text{ или}$$
$$p_0(t+h) - p_0(t) / h = -\lambda m p_0(t) + \mathrm{O}(h).$$

Переходя к пределу при $h \to 0$, имеем $p_0'(t) = -\lambda m p_0(t) + \mu p_1(t)$. А так как нас будут интересовать только стационарные решения, то в последнем равенстве переходим к пределу при $t \to \infty$: $0 = -\lambda m p_0 + \mu p_1$. Проведя аналогичные рассуждения для $k = 1, 2, ..., m$, получим следующую однородную систему алгебраических уравнений:

$$\left\{ \begin{array}{c} 0 = -\alpha m p_0 + p_1; \\ 0 = (m-k+1)\alpha p_{k-1} - [(m-k)\alpha + k]p_k + \\ + (k+1)p_{k+1}; 0 < k < n; \\ 0 = (m-k+1)\alpha p_{k-1} - [(m-k)\alpha + n]p_k + np_{k+1}, n \leq k < m; \\ 0 = \alpha p_{m-1} - np_m. \end{array} \right. \qquad (2.64)$$

Здесь $\alpha = \lambda / \mu, p_k = \lim\limits_{t \to \infty} p_k(t), k = 0, 1, ..., m.$

Для решения системы (2.64) введем новые переменные:

$$z_k = (m-k)\alpha p_k - (k+1)p_{k+1}, 0 \leq k < n;$$
$$z_k = (m-k)\alpha p_k - np_{k+1}, n \leq k \leq m. \qquad (2.65)$$

В новых переменных система (2.64), как легко проверить, принимает очень простой вид

$$z_0 = 0;$$
$$z_{k-1} - z_k = 0, 0 < k < m;$$
$$z_{m-1} = 0$$
и имеет, очевидно, только нулевое решение $z_k = 0,\ \ k = 0, 1, ..., m-1$. Следовательно, равенства (2.65) можно переписать так: $p_{k+1} = \dfrac{(m-k)\alpha}{k+1} p_k, 0 \le k < n;\ p_{k+1} = \dfrac{(m-k)\alpha}{n} p_k, n \le k \le m.$

Отсюда последовательными подстановками легко находим

$$p_k = \begin{cases} \dfrac{m!\alpha^k}{k!(m-k)!} p_0, 1 \le k \le n; \\ \dfrac{m!\alpha^k}{n^{k-n} n!(m-k)!} p_0, n \le k \le m. \end{cases} \qquad (2.66)$$

Величину же p_0, как обычно, определяем из условия нормировки $\sum\limits_{k=0}^{m} p_k = 1$ и полученных формул (2.66) для p_k, $k = 1, 2, ..., m$. Зная вероятности p_k, $k = 1, 2, ..., m$, можно вычислить среднюю длину очереди

$$M_1 = \sum_{k=n}^{m} (k-n) p_k, \qquad (2.67)$$

среднее число обслуживаемых и ожидающих обслуживания требований

$$M_2 = \sum_{k=1}^{m} k p_k, \qquad (2.68)$$

среднее число свободных от обслуживания приборов

$$M_3 = \sum_{k=0}^{n-1} (n-k) p_k \ . \qquad (2.69)$$

Отсюда легко находим коэффициент простоя обслуживаемого объекта

$$\xi_1 = M_1 / m, \qquad (2.70)$$

коэффициент использования объектов

$$\xi_2 = 1 - (M_2 / m) \qquad (2.71)$$

и коэффициент простоя обслуживающих приборов

$$\xi_3 = M_3 / n. \qquad (2.72)$$

Коэффициент ξ_2 – весьма важная интегральная характеристика замкнутой системы массового обслуживания. Он характеризует интенсивность эксплуатации объектов, так как практически равен вероятности того, что данный объект в любой момент времени t будет находиться в эксплуатации, т. е. работать. Для установившегося режима работы СМО его можно вычислить по формуле

$$\xi_2 = t_1/(t_1 + t_2)\,, \tag{2.73}$$

где $t_1 = 1/\lambda$ – среднее время безотказной работы объекта; t_2 – среднее время простоя этого объекта.

Обозначим, как и в подразд. 2.3, время ожидания в очереди через θ. Очевидно, общее время простоя $П$ слагается из времени ожидания в очереди θ и времени обслуживания τ, т. е. $П = \theta + \tau$. Переходя к математическим ожиданиям, учитывая равенства $MП = t_2$; $M\tau = 1/\mu$ и определяя t_2 из формулы (2.73), находим

$$M\theta = MП - M\tau = t_2 - \frac{1}{\mu} = t_1\frac{1-\xi_2}{\xi_2} - \frac{1}{\mu} = \frac{1}{\lambda}\frac{1-\xi_2}{\xi_2} - \frac{1}{\mu}. \tag{2.74}$$

Пример. Пусть для обслуживания (наладки) шести однотипных автоматов выделено два мастера одинаковой производительности. Время от времени эти автоматы останавливаются и требуют наладки. Предполагается, что поток остановок автоматов является пуассоновским, а время их наладки подчиняется показательному закону. Каждый автомат останавливается в среднем один раз в два часа, а среднее время наладки одного автомата одним мастером составляет 0,4 ч. Таким образом, λ= 0,5; μ = 2,5; α=λ/μ=0,2.

Возможны следующие варианты организации обслуживания: 1) оба мастера обслуживают все шесть автоматов, так что при остановке автомата его обслуживает один из свободных мастеров; в этом случае $n = 2$, $m = 6$; 2) каждый из двух мастеров обслуживает по три закрепленных за ним автомата. В этом случае $n = 1$, $m = 3$.

В результате вычислений по формулам (2.66) получаем табл.1 для первого варианта и табл.2 для второго варианта.

Таблица 1

Число неработающих автоматов	Число автоматов, ожидающих наладки	Число свободных мастеров	p_k
0	0	2	0,320
1	0	1	0,384
2	0	0	0,192
3	1	0	0,077
4	2	0	0,023
5	3	0	0,005
6	4	0	0,000

Таблица 2

Число неработающих автоматов	Число автоматов, ожидающих наладки	Число свободных мастеров	p_k
0	0	1	0,531
1	0	0	0,318
2	1	0	0,127
3	2	0	0,025

Используя значения p_k в табл. 1 и 2, вычислим для обоих вариантов коэффициент простоя автоматов ξ_1, коэффициент использования автоматов ξ_2 и коэффициент простоя мастеров ξ_3. Начнем с первого варианта. Среднее число автоматов, простаивающих в каждый момент в очереди из-за того, что мастера заняты наладкой других автоматов, определяется по формуле (2.67): $M_1 = \sum_{k=2}^{6}(k-2)p_k = 0{,}140$. Иными словами, в среднем из шести автоматов 0,14 автомата будет стоять в очереди до тех пор, пока не освободится мастер.

Для среднего числа неработающих автоматов, т. е. автоматов, которые налаживают и которые стоят в очереди, имеем согласно (2.68) $M_2 = \sum_{k=1}^{6} kp_k = 1{,}109$. Это означает, что в среднем из шести автоматов не работают немного больше одного.

Наконец, среднее число мастеров, не занимающихся в каждый момент времени наладкой автоматов, определяется по формуле (2.69):

$$M_3 = \sum_{k=0}^{1}(2-k)p_k = 1{,}024 .$$

Теперь по формулам (2.70) и (2.72) вычислим коэффициенты ξ_1, ξ_2 и ξ_3. Имеем $\xi_1 = M_1/6 = 0{,}023$, т. е. каждый автомат стоит в очереди в ожидании начала его отладки приблизительно 0,023 части рабочего времени. Далее $\xi_2 = 1-(M_2/6) = 0{,}815$, т. е. вероятность того, что данный автомат в любой момент времени будет работать, равна 0,815. Наконец, $\xi_3 = M_3/2 = 0{,}512$, т. е. каждый мастер свободен от наладки автоматов чуть больше половины рабочего времени. Кроме того, по формуле (2.74) найдем среднее время ожидания начала обслуживания:

$$M\theta = \frac{1}{0{,}5}\frac{1-0{,}815}{0{,}815} - \frac{1}{2{,}5} = 0{,}054 \text{ ч} \approx 32\text{мин}.$$

Вычислим аналогично коэффициенты ξ_1 ξ_2, ξ_3 и величину $M\theta$ для второго варианта и соберем полученные результаты в табл.3.

Т а б л и ц а 3

Показатели функционирования	Вариант	
	Первый	Второй
ξ_1	0,023	0,059
ξ_2	0,815	0,784
ξ_3	0,512	0,531
$M\theta$, мин	3,2	9,1

Из табл. 3 заключаем, что по всем показателям первый вариант лучше второго. Например, простой автоматов в первом варианте на 0,059—0,023 = 0,036 рабочего времени меньше, т. е. производительность труда на один автомат в первом варианте на 3,6% больше, чем во втором.

2.5. Смешанные системы с ожиданием

В подразд. 2.3 и 2.4 рассмотрены системы массового обслуживания с ожиданием, для которых типично следующее: требования, поступающие в систему на обслуживание, покидают ее только после обслуживания. Однако на практике часто встречаются системы массового обслуживания с ожиданием, в которых имеются различные ограничения, например на время пребывания в очереди, на время пребывания требования в системе, на длину очереди. Возможны также различные комбинации подобных ограничений. Такие системы массового обслуживания называются смешанными системами с ожиданием. В связи с этим системы массового обслуживания с ожиданием, рассмотренные в подразд. 2.3, иногда называют чистыми системами с ожиданием.

В смешанных системах массового обслуживания требование, поступившее в систему на обслуживание и заставшее все приборы занятыми, может покинуть систему и не обслуженным. Так, если имеется ограничение на длину очереди, выражающееся в том, что число требований, ожидающих начала обслуживания, не должно превышать m, то очередное требование, поступившее в систему на обслуживание в момент времени, когда все приборы заняты и m требований ожидают обслуживания, обязано покинуть систему, хотя оно и не обслужено. Такая ситуация возникает, например, в мастерских, предназначенных для ремонта каких-либо машин, с ограниченной площадью для их хранения.

Кстати заметим, что если длина очереди $m = \infty$, то смешанная система с ожиданием вырождается в чистую систему с ожиданием, рассмотренную в подразд. 2.3; в другом крайнем случае, когда длина очереди $m = 0$, рассматриваемая смешанная система вырождается в систему с отказами (см. подразд. 2.1.).

Ограничение на время пребывания в очереди выражается в том, что время ожидания требованием начала обслуживания не должно превышать некоторой величины $t_{\text{ож}}$, где $t_{\text{ож}}$ – постоянная или случайная. Иначе говоря, если за время $t_{\text{ож}}$ с того момента, как требование стало в очередь, не освободится ни один из обслуживающих приборов, то требование покидает систему, хотя оно и не обслужено. Предполагается, что если процесс обслуживания уже начат, то он доводится до конца независимо от времени ожидания начала обслуживания. Такая ситуация может возникнуть, например, в парикмахерской: с одной стороны, деловые люди, знающие цену времени, не могут позволить себе находиться в очереди больше, чем некоторое $t_{\text{ож}}$. С другой стороны, если уж клиент сел в кресло и его начали брить, то он вряд ли уйдет с одной выбритой щекой.

Ограничение на время пребывания требования в системе массового обслуживания означает, что общее время пребывания требования в системе – время ожидания начала обслуживания плюс время собственно обслуживания – не должно превышать некоторой величины $T_{\text{ож}}$. В противном случае требование независимо от того, начато его обслуживание или нет и если начато обслуживание, то окончено оно или нет, покидает систему. Такая ситуация может возникнуть, например, в привокзальном ателье мелких услуг: пассажир, несомненно, предпочтет забрать свой пиджак с еще не всеми выведенными пятнами, чем опоздать на поезд.

Оставляя в стороне все многообразие смешанных систем массового обслуживания, рассмотрим для примера смешанную систему с ограничением на длину очереди и время пребывания в ней. Итак, пусть на вход системы массового обслуживания, состоящей из n однотипных приборов, поступает пуассоновский поток требований с интенсивностью λ. Время обслуживания τ каждого требования подчинено показательному закону $P(\tau < t) = 1 - e^{-\mu t}$. Обозначим максимальное число мест в очереди через m. Если в системе находится $k(k \le n)$ требований, то все они обслуживаются, причем каждое требование одним прибором. Если в системе находится $n + s(s \le m)$ требований, то по понятной причине n из них обслуживаются, а остальные s стоят в очереди и ждут начала обслуживания. Если вновь прибывшее в систему требование застанет в ней $n + m$ требований (n находящихся на обслуживании и m ожидающих начала обслуживания), то оно покидает систему необслуженным.

Предположим далее, что время пребывания требования в очереди, заставшего в системе $n + s(s < m)$ других требований, не должно превосходить $t_{\text{ож}}$, где $t_{\text{ож}}$ – случайная величина, имеющая показательное распределение $P(t_{ож}<t)=1-e^{-\nu t}$.Таким образом, требование, заставшее все приборы уже занятыми, покидает систему необслуженным в следующих двух случаях: 1) если в очереди стоит уже m требований; 2) если требование простаивает в напрасном ожидании начала обслуживания в среднем $1/\nu$ единиц времени (в связи с этим параметру ν иногда придают смысл интенсивности ухода из очереди необслуженных требований).

Обозначим, как обычно, через $p_k(t)$ вероятность того, что система массового обслуживания в момент времени t находится в состоянии k. Напомним, что здесь различные состояния определяются так: 0 – в системе нет ни одного требования и, стало быть, все приборы свободны; k – в системе находится k требований и все они обслуживаются, k =1, 2, ..., n; $n+s$ — в системе находится $n+s$ требований, из них n обслуживаются, а s стоят в очереди в ожидании обслуживания, s = 1, 2, . . ., m. Вероятности для всех этих возможных состояний системы массового обслуживания удовлетворяют дифференциальным уравнениям:

$$\begin{gathered}
p_0'(t) = -\lambda p_0(t) + \mu p_1(t);\\
p_k'(t) = \lambda p_{k-1}(t) - (\lambda + k\mu) p_k(t) + (k+1)\mu p_{k+1}(t), 1 \le k \le n-1;\\
p_k'(t) = \lambda p_{k-1}(t) - [\lambda + n\mu + (k-n)\nu] p_k(t) + \\
+ [n\mu + (k-n+1)] p_{k+1}(t), n \le k \le n+m-1;\\
p_{n+m}'(t) = \lambda p_{n+m-1}(t) - (n\mu + m_\nu) p_{n+m}(t).
\end{gathered}$$

Опуская вывод и решение соответствующей системы уравнений для установившегося режима, приведем окончательный результат:

$$p_k = \begin{cases} \dfrac{\alpha^k}{k!} p_0, 1 \le k \le n; \\[2ex] \dfrac{\alpha^k}{n! \prod\limits_{i=1}^{k-n} (n + i\beta)} p_0, n \le k \le n+m; \end{cases} \tag{2.75}$$

$$p_0 = 1 - \sum_{k=1}^{n+m} p_k, \tag{2.76}$$

где $\alpha = \lambda/\mu,\ \beta = \nu/\mu$.

Знание вероятностей p_k, как и раньше, позволяет вычислить среднее число требований, находящихся в очереди (длину очереди),

$$M_1 = \sum_{k=n+1}^{n+m} (k-n) p_k, \tag{2.77}$$

среднее число требований, находящихся в системе,

$$M_2 = \sum_{k=1}^{n+m} kp_k, \tag{2.78}$$

среднее число свободных от обслуживания приборов

$$M_3 = \sum_{k=0}^{n-1} (n-k)p_k, \tag{2.79}$$

Важнейшей характеристикой качества обслуживания в рассматриваемой системе является вероятность отказа

$$p_{\text{отк}} = 1 - (n - M_3)/\alpha, \tag{2.80}$$

где $n - M_3$ – среднее число приборов, занятых обслуживанием, а $(n - M_3)/\alpha$ – вероятность обслуживания. Другая важнейшая характеристика качества обслуживания – среднее время $M\theta$ пребывания требования в очереди, т. е. средняя длительность ожидания начала обслуживания.

Если $f(t)$ —плотность вероятности времени θ нахождения требования в очереди до начала его обслуживания, то $P(t < \theta < t + dt) \approx f(t)dt.$ Но за время t нахождения требования в очереди за ним образуется очередь, состоящая в среднем из λt требований (по определению входящего потока требований). Поэтому среднее число требований, находящихся в очереди в интервале времени $(t, t + dt)$, равно $\lambda t f(t) dt$.

Отсюда непосредственно следует, что среднее число требований, находящихся в очереди в интервале времени $(0, \infty)$, т.е. математическое ожидание числа требований, ожидающих начала обслуживания

$$M_1 = \int_0^\infty \lambda t f(t) dt = \lambda M\theta \text{ или } M\theta = M_1/\lambda. \tag{2.81}$$

Формула (2.81) позволяет вычислить математическое ожидание времени θ пребывания требования в очереди без знания закона распределения величины θ.

Пример. В крупный магазин поступают из различных отдаленных хозяйств машины с фруктами. Эти машины поступают в случайные моменты времени с плотностью λ = 4 машины в день. В магазине имеется подсобное помещение, позволяющее хранить не более m машин фруктов одновременно, и n бригад продавцов одинаковой производительности. Время продажи (обслуживания) одной машины с фруктами одной бригадой случайно и составляет в среднем 0,5 дня.

Фрукты – скоропортящийся товар и срок их хранения зависит от многих случайных факторов, например вида, времени, прошедшего с момента сбора фруктов, степени их созревания, условия транспортировки. Предположим, что среднее допустимое время хранения этих фруктов составляет

2,5 дня. Формально, чтобы можно было использовать формулы, приведенные в этом подразделе, нужно предположить, что поток поступления автомашин с фруктами пуассоновский, а продолжительности обслуживания автомашин и сохранности содержимого автомашин подчиняются показательному распределению. На практике это, конечно, может не выполняться. Например, продолжительность сохранности свежих фруктов скорее имеет гамма-распределение, чем показательное. Однако, как показали соответствующие расчеты методом статистического моделирования, интегральные характеристики смешанных систем с ожиданием практически мало зависят от вида распределений времени обслуживания требований и времени их пребывания в очереди. Таким образом, $\lambda = 4;$ $\mu = 2;$ $\nu = 1/2{,}5 = 0{,}4;$ $\alpha = 2;$ $\beta = 0{,}2.$

Соответствующие результаты вычислений, полученные по формулам (2.75) – (2.81) для $n = 1; 2$ и $m = 1; 2; 3,$ приведены в табл. 4.

Таблица 4

Показатели функционирования	n=1			
	m=0	m=1	m=2	m=3
P_0	0,333	0,158	0,090	0,059
p_1	0,667	0,316	0,181	0,117
p_2	—	0,526	0,300	0,195
p_3	—	—	0,430	0,281
p_4	—	—	—	0,350
M_1	—	0,53	1,16	1,81
M_2	0,67	1,37	1,97	2,75
M_3	0,33	0,16	0,09	0,06
$p_{отк}$	0,67	0,58	0,54	0,53
$M\theta$, дней	—	0,13	0,29	0,45
	n=2			
p_0	0,2	0,145	0,119	0,105
p_1	0,4	0,290	0,238	0,210
p_2	0,4	0,290	0,238	0,210
p_3	—	0,244	0,225	0,198
p_4	—	—	0,181	0,159
p_5	—	—	—	0,122
M_1	—	0,21	0,59	0,88
M_2	1,2	1,6	2,11	2,47
M_3	0,8	0,58	0,48	0,42
$p_{отк}$	0,4	0,29	0,24	0,21
$M\theta$, дней	—	0,06	0,15	0,22

Из этой таблицы следует, что одна бригада продавцов $(n = 1)$ явно не справится со своей работой, ибо в среднем свыше половины (53—67%) всех поступающих фруктов не будет распродано. Вероятность отказа в обслуживании при наличии двух бригад продавцов $(n = 2)$ уже значительно меньше (21-40%), но еще сравнительно велика. По-видимому, целесообразно использовать три бригады.

С другой стороны, вероятность отказа $p_{отк}$ менее чувствительна к изменению величины m. Поэтому если содержание подсобного помещения связано с относительно большими издержками, то рационально использовать небольшие значения m, например $m = 1$. Вместе с тем полное отсутствие подсобного помещения ($m = 0$) также невыгодно, так как при этом заметно возрастает вероятность отказа.

3. СПЕЦИАЛЬНЫЕ СИСТЕМЫ МАССОВОГО ОБСЛУЖИВАНИЯ

Все системы массового обслуживания, рассмотренные в предыдущем разделе, являются в определенном смысле классическими. Вместе с тем существует большое количество разновидностей СМО, отражающих те или иные закономерности и особенности работы реальных объектов исследования. Для них в рамках марковских ограничений также получены аналитические решения (по крайней мере, для установившихся режимов). Охарактеризуем основные подобные модели, используемые для описания широкого круга производственно-экономических и организационно-технических систем.

3.1. Упорядоченные системы

До сих пор предполагалось, что все обслуживающие приборы равноправны в том смысле, что очередное требование может обслуживать любой из свободных приборов. Однако легко себе представить такую систему массового обслуживания, в которой все приборы пронумерованы, причем очередное требование может обслуживать данный свободный прибор только при условии, если все приборы с меньшим номером уже заняты обслуживанием. В этом случае вероятность того, что k приборов будут заняты одновременно обслуживанием, имеем

$$p_k = \frac{\alpha^k / k!}{\sum\limits_{i=0}^{k} \alpha^i / i!}, \quad k = 1,2,..., \tag{3.1}$$

где $\alpha = \lambda / \mu$, λ – среднее число требований, поступающих в единицу времени, $1/\mu$ – среднее время обслуживания одного требования.

Пример. Пусть в некоторый гараж поступает в течение семичасового рабочего дня в среднем девять заявок на перевозки грузов, причем среднее время перевозок (включая время возвращения в гараж) составляет 0,7 ч.

Предполагается, что все автомашины пронумерованы, а обслуживание очередной заявки по перевозке груза осуществляет свободная автомашина с наименьшим номером. Результаты вычислений по формуле (3.1) следующие ($\lambda = 9$, $\mu = 10$, $\alpha = 0{,}9$):

k	1	2	3	4	5
p_k	0,474	0,176	0,050	0,011	0,002

Отсюда следует, что в среднем только 0,2% рабочего времени загружены одновременно пять автомашин. Это указывает на то, что гаражу можно обойтись меньшим, чем пять, количеством автомашин. Если в гараже имеется всего три автомашины, то вероятность того, что очередному требованию придется встать в очередь, равна 0,05. Так как эта вероятность достаточно мала, то, по-видимому, гаражу достаточно иметь три автомашины. Таким образом, формула (3.1) позволила определить не только степень загруженности каждой автомашины, но и оптимальное количество автомашин, которое нужно иметь данному гаражу.

Существуют различные обобщения рассматриваемой упорядоченной системы массового обслуживания. Например, иногда упорядочивают не отдельные приборы, а группы приборов, так что очередное требование начинает обслуживаться каким-либо свободным прибором *k*-й группы, если только все приборы всех предыдущих групп заняты.

Пусть r_k – число приборов в *k*-й группе, а A_k – событие, состоящее в том, что требование получает отказ на приборах *k*-й группы. Тогда вероятность того, что очередное требование застанет все приборы первой группы занятыми, можно определить по формуле Эрланга:

$$P(A_1) = \frac{\alpha^{r_1} / r_1}{\sum\limits_{i=0}^{r_1} \alpha^i / i!}. \tag{3.2}$$

Вероятность того, что это требование застанет также занятыми все приборы второй группы

$$P(A_1 A_2) = \frac{\alpha^{r_1 + r_2} / (r_1 + r_2)!}{\sum\limits_{i=0}^{r_1 + r_2} \frac{a^i}{i!}} \tag{3.3}$$

Получив отказ в приборах первой и второй групп, требование поступает на обслуживание в приборы третьей группы. Вероятность того, что и здесь оно получит отказ, т. е. застанет все приборы занятыми, можно определить по формуле

$$P(A_1 A_2 A_3) = \frac{\alpha^{r_1+r_2+r_3} / (r_1+r_2+r_3)!}{\sum_{i=0}^{r_1+r_2+r_3} \frac{\alpha^i}{i!}}.$$

Продолжая этот процесс, можно найти вероятность $P(A_1 A_2 ... A_k)$ того, что требование прошло необслуженным последовательно сквозь приборы первой, второй,..., k-й групп. Это позволяет по формуле

$$P(A_k \mid A_1 A_2 ... A_{k-1}) = \frac{P(A_1 A_2 ... A_k)}{P(A_1 A_2 ... A_{k-1})} \tag{3.4}$$

определить условную вероятность отказа на приборах k-й группы при условии, что получен отказ на приборах всех предыдущих k-1 групп. Заметим, что величина $1 - P(A_k \mid A_1 A_2 ... A_{k-1})$ равна вероятности того, что приборы k-й группы обслужат требование при условии, что был получен отказ на приборах всех предыдущих k – 1 групп.

Пример. На заводе имеются два красильных цеха: основной и дополнительный. В основном цехе функционируют четыре красильные установки, а в дополнительном – две. Производительность всех этих шести установок одна и та же и в среднем равна 5 изд./ч. Из многих цехов завода в случайные моменты времени в основной красильный цех поступают на окраску изделия с интенсивностью 10 изд./ч. Если в момент поступления на окраску очередного изделия все четыре красильные установки основного цеха оказываются занятыми, то это изделие поступает в дополнительный цех. Если при этом обе красильные установки дополнительного цеха также оказываются занятыми, то изделие остается неокрашенным.

Здесь $\lambda = 10$, $\mu = 5$, $r_1 = 4$, $r_2 = 2$, $\alpha = 2$. Поэтому по формулам (3.2) и (3.3) получаем $P(A_1) = 0{,}095$; $P(A_1 A_2) = 0{,}012$. Это означает, что 9,5% всех поступающих на окраску в основной цех изделий не будут в этом цехе окрашены, а 1,2% изделий вообще не будут окрашены. Используя эти данные, согласно формуле (3.4) имеем $P(A_2 \mid A_1) = 0{,}012 / 0{,}095 = 0{,}127$; иначе говоря, если изделие не было окрашено в основном цехе, то с вероятностью 0,127 оно не будет окрашено также и в дополнительном красильном цехе.

3.2. Системы с поступлением групповых заявок

До сих пор предполагалось, что требования поступают в систему массового обслуживания по одному, т. е. входящий поток требований – ординарный. Рассмотрим теперь системы, в которых требования поступают группами, причем число требований в каждой группе либо постоянно, либо случайно.

Предположим:

1) поток групп требований является пуассоновским с интенсивностью λ групп в единицу времени;

2) время обслуживания каждого требования в каждой группе подчинено показательному закону с математическим ожиданием $1/\mu$ единиц времени;

3) число требований X в каждой группе случайное и имеет распределение вероятностей

$$P(X = s) = a_s\,,\ a_s \geq 0\,;\ \sum_{s=1}^{\infty} a_s = 1\,;$$

4) каждое требование в поступившей в систему группе начинает обслуживаться с одной и той же вероятностью;

5) число обслуживающих приборов равно n и все они однотипны;

6) рассматриваемая система массового обслуживания принадлежит к типу систем с отказами; это означает, что очередное требование покидает систему необслуженным, если в момент его поступления все приборы заняты.

Для такой системы массового обслуживания справедливы рекуррентные соотношения:

$$\begin{cases} p_1 = \alpha p_0; \\ p_{k+1} = \dfrac{k+\alpha}{k+1} p_k - \dfrac{\alpha}{k+1} \sum\limits_{i=1}^{k} a_i p_{k-i}, 1 \leq k \leq n-2; \\ p_n = \dfrac{\alpha}{n} \sum\limits_{i=1}^{n} p_{n-i} \sum\limits_{s=i}^{\infty} a_s, \end{cases} \tag{3.5}$$

где $\alpha = \lambda/\mu$, p_n – вероятность того, что в системе при установившемся режиме находится на обслуживании k требований (или, что то же, занято k приборов). При этом условие нормировки имеет вид $\sum\limits_{k=0}^{n} p_k = 1$. Основной интегральной характеристикой для систем с отказами является вероятность $p_{\text{отк}}$ того, что требование получит отказ в обслуживании. Для рассматриваемого случая

$$p_{\text{отк}} = 1 - \left(\sum_{k=1}^{n} k p_k \Bigg/ \alpha \sum_{k=1}^{\infty} k a_k \right). \tag{3.6}$$

Пример. Над некоторой зенитной батареей, состоящей из трех однотипных орудий, в случайные моменты времени пролетают группы самолетов противника с интенсивностью 10 групп/ч. Каждая группа состоит из одного, двух или трех самолетов соответственно с вероятностями 0,2; 0,3 и 0,5. Среднее время, необходимое для обстрела одного самолета одним орудием, равно 3 мин. Оценим эффективность этой батареи.

Здесь $\lambda = 10$, $\mu = 20$, $\alpha = 0{,}5$, $n = 3$, $a_1 = 0{,}2$, $a_2 = 0{,}3$, $a_3 = 0{,}5$, $a_s = 0$, $s > 3$. Из формул (3.5) находим:

$$p_1 = 0{,}5 p_0;$$

$$p_2 = \frac{1+0{,}5}{1+1} p_1 - \frac{0{,}5}{1+1} 0{,}2 p_0 = 0{,}75 p_1 - 0{,}05 p_0 = 0{,}325 p_0;$$

$$p_3 = \frac{0{,}5}{3}\left[p_2(0{,}2+0{,}3+0{,}5) + p_1(0{,}3+0{,}5) + p_0 0{,}5\right] = 0{,}204 p_0.$$

Отсюда, используя условие нормировки, получаем: $p_0 = 0{,}494$, $p_1 = 0{,}247$, $p_2 = 0{,}160$, $p_3 = 0{,}101$ и, следовательно, по формуле (3.6)

$$p_{\text{отк}} = 1 - \frac{0{,}247 + 2\cdot 0{,}160 + 3\cdot 0{,}101}{0{,}5(0{,}2 + 2\cdot 0{,}3 + 3\cdot 0{,}5)} = 0{,}24 = 24\%.$$

Это означает, что в среднем 76% самолетов противника будут обстреливаться («обслуживаться») рассматриваемой гипотетической батареей.

Разумеется, требования могут поступать группами также и в системах массового обслуживания с ожиданием (как чистых, так и смешанных). Такая ситуация может возникнуть, например, при разгрузке судов, прибывающих в порт не по одному, а караванами и, естественно, ожидающих начала их разгрузки в случае занятости всех причалов.

Рассмотрим, например, смешанную систему массового обслуживания с ожиданием, у которой имеется ограничение на время пребывания требований в очереди, причем это время подчиняется показательному закону с параметром ν. Предположим далее, что число требований в каждой группе постоянно и равно s, а в остальном эта система с ожиданием совпадает с рассмотренной уже системой с отказами и групповым поступлением требований. Тогда вместо формул (3.5) имеем:

$$\begin{cases} p_1 = \alpha p_0; \\ p_{k+1} = \dfrac{k+\alpha}{k+a} p_k, \quad 1 \le k \le s; \\ p_{k+1} = \dfrac{k+\alpha}{k+1} p_k - \dfrac{\alpha}{k+1} p_{k-s}, \quad s \le k \le n-1; \\ p_{n+1} = \dfrac{n+\alpha}{n+\beta} p_n - \dfrac{\alpha}{n+\beta} p_{n-s}; \\ p_{k+1} = \dfrac{n+\alpha+(k-n)\beta}{n+(k-n+1)\beta} p_k - \dfrac{\alpha}{n+(k-n+1)\beta} p_{k-s}, \quad k \ge n+1, \end{cases} \tag{3.7}$$

где $\alpha = \lambda/\mu$, $\beta = \nu/\mu$; p_0 — вероятность того, что все n приборов свободны; $p_k(1 \le k \le n)$ – вероятность того, что заняты k приборов; $p_k(k \ge n+1)$ – вероятность того, что все n приборов заняты и $k-n$ требований стоят в очереди. При этом условие нормировки здесь уже имеет вид $\sum\limits_{k=0}^{\infty} p_k = 1$, а вероятность отказа в обслуживании можно определить по формуле

$$p_{\text{отк}} = 1 - M_4 / s\alpha, \tag{3.8}$$

где

$$M_4 = \sum_{k=1}^{n} k p_k + n\left(1 - \sum_{k=0}^{n} p_k\right) \tag{3.9}$$

есть среднее число приборов, занятых обслуживанием.

Пример. Пусть в библиотеку за книгами заходит в среднем 15 человек в час, причем каждый читатель берет две книги. В библиотеке имеются три сотрудника, каждый из которых осуществляет выдачу одной книги в среднем в течение 2 мин. Если в момент прихода читателя в библиотеку все трое сотрудников оказываются занятыми выдачей книг, то он становится в очередь, причем среднее время его пребывания в очереди не должно превышать 5 мин.

Здесь $\lambda = 15$, $\mu = 30$, $\nu = 12$, $n = 3$, $s = 2$, $\alpha = 0{,}5$, $\beta = 0{,}4$. Вероятности p_k состояний, вычисленные по формулам (3.7), следующие (для $k \ge 8$ вероятности p_k не выписаны, так как они меньше 0,001):

k	0	1	2	3	4	5	6	7
p_k	0,468	0,234	0,176	0,068	0,035	0,013	0,005	0,001

Отсюда, пользуясь формулами (3.8) и (3.9), находим: $M_4 = 0{,}952$; $p_{\text{отк}} = 0{,}048$. Таким образом, в среднем 4,8% посетителей покинут библиотеку, не получив в ней книги. Коэффициент загруженности

$\xi_4 = M_4 / n$ здесь равен 0,317, т. е. каждый сотрудник библиотеки загружен в среднем 31,7% рабочего времени. Это сравнительно низкая загруженность. По-видимому, в этой библиотеке целесообразно ограничиться двумя сотрудниками.

Проведя аналогичные вычисления для $n = 2$ при прочих равных условиях, находим $p_{\text{отк}} = 0{,}115$; $\xi_4 = 0{,}442$. Это означает, что при двух сотрудниках их средняя загруженность увеличивается до 44,2% (вместо 31,7% при $n = 3$), что с экономической точки зрения целесообразно. При этом среднее число необслуженных читателей также увеличивается до 11,5% (вместо 4,8% при $n = 3$). Эта величина вероятности отказа сравнительно высокая и вряд ли подобной организацией работы библиотеки будет довольно население.

Однако если увеличить допустимое среднее время пребывания в очереди с 5 мин до, скажем, 20 мин, создав соответствующие условия для чтения свежих журналов и газет, то, для $n = 2$ и $\nu = 3$ при прочих равных условиях, находим $p_{\text{отк}} = 0{,}055$; $\xi_4 = 0{,}472$. Таким образом, два сотрудника могут обеспечить книгами 94,5% всех приходящих в библиотеку читателей. При этом они будут загружены в среднем на 47,2%.

3.3. Системы с приборами разной производительности

Во всех рассмотренных выше системах массового обслуживания предполагалось, что все обслуживающие приборы имеют одинаковую производительность. Рассмотрим систему, состоящую из двух приборов, на которую поступает пуассоновский поток требований с параметром λ. Время обслуживания требований этими приборами подчиняется показательному закону, причем для первого прибора параметр равен μ_1, а для второго – $\mu_2 (\mu_1 > \mu_2)$. Если требование застает один свободный прибор, то оно начинает обслуживаться этим прибором. Если требование застает оба прибора свободными, то оно начинает обслуживаться с равной вероятностью любым из этих приборов. Если требование застает оба прибора занятыми, то оно становится в очередь и ждет, пока не освободится один из двух приборов.

В этом случае для установившегося режима:

$$p_{00} = \frac{1-\alpha}{1 + \dfrac{\alpha}{\gamma(1+2\alpha)}\left[1 + (1+\gamma^2)\alpha - (1-\gamma^2)0{,}5\right]};$$

$$p_{01} = \frac{\alpha(1+\gamma)(1+\alpha-\gamma)}{\gamma(1+2\alpha)} p_{00};$$

$$p_{10} = \frac{\alpha(1+\gamma)(\alpha+0{,}5)}{1+2\alpha} p_{01};$$

$$p_{11} = 1 - p_{01} - p_{10} - p_{00};$$

$$p_k = \frac{\alpha^k(1+\gamma)[1+(1+\gamma)\alpha-(1-\lambda)0{,}5]}{\gamma(1+2\alpha)} p_{00}, k = 2, 3...,$$

где $\alpha = \lambda/(\mu_1 + \mu_2)$; $\gamma = \mu_2/\mu_1$; p_{00}, p_{01}, p_{10}, p_{11}, p_k — вероятности состояний системы соответственно для случаев: 1) оба прибора свободны; 2) первый прибор свободен, второй занят; 3) первый прибор занят, а второй свободен; 4) оба прибора заняты; 5) в системе находится k требований. При этом среднее число требований, находящихся в системе (как обслуживаемых, так и стоящих в очереди), можно вычислить по формуле

$$M_2 = p_{01} + p_{10} + \sum_{k=2}^{\infty} k p_k =$$

$$= \frac{\alpha(1-\gamma)[1+(1+\gamma)\alpha-(1-\gamma)0{,}5]}{\gamma(1+2\alpha)+\alpha\left[1+\left(1+\gamma^2\right)\alpha-\left(1-\gamma^2\right)0{,}5\right]}.$$

Пример. Для заказа лекарств в некоторую аптеку в случайные моменты времени заходят в среднем 24 человека в час. В этой аптеке принимают рецепты на изготовление лекарств два аптекаря. Первый из них оформляет в среднем 18 заказов в час, второй же работает медленнее, оформляя в среднем только 12 заказов в час. При этом никому из них не отдается предпочтение.

Здесь $\lambda = 24$; $\mu_1 = 18$; $\mu_2 = 12$; $\alpha = 24/(18+12) = 0{,}80$; $\gamma = 12/18 = 0{,}60$. Поэтому по предыдущим формулам находим p_{00} = 0,11; p_{01} = 0,11; p_{10} = 0,08; $p_{11} = 0{,}70$; M_2 = 0,18.

Таким образом, 70% рабочего времени оба аптекаря будут оформлять заказы на изготовление лекарств одновременно. При этом очереди либо не будет совсем, либо, что менее вероятно, в очереди будет стоять один, в крайнем случае два человека. Первый аптекарь свободен $p_{00}+p_{01}$=0,22=22% рабочего времени, а второй – $p_{00}+p_{10}$=0,19=19% рабочего времени. Одновременно оба аптекаря свободны p_{00}=0,11 = 11% рабочего времени.

Рассмотрим еще систему массового обслуживания с приборами разной производительности, у которой приборы упорядочены. Для разнообразия исследуем случай системы с отказами. Итак, пусть система массового обслуживания состоит из двух последовательно расположенных приборов, производительность первого из которых характеризуется параметром μ_1 а второго – μ_2. На первый из этих приборов поступает поток требований с параметром λ. Если первый прибор оказывается занятым, то требование поступает на обслуживание на второй прибор. Если и второй прибор занят, то требование получает отказ и покидает систему необслуженным. В этом случае имеем:

$$p_{11} = p_{\text{отк}} = \left[1+\gamma_1+\gamma_2+\frac{\gamma_1\gamma_2(2+\gamma_1+\gamma_2)}{1+\gamma_2}\right]^{-1}; \qquad (3.10)$$

$$p_{01} = \frac{\gamma}{1+\gamma_2}p_{11};\ p_{10} = \frac{\gamma_2(1+\gamma_1+\gamma_2)}{1+\gamma_2}p_{11};\ p_{00} = \frac{\gamma_1\gamma_2(2+\gamma_1+\gamma_2)}{1+\gamma_2}p_{11}, \quad (3.11)$$

где $\gamma_1=\mu_1/\lambda$, $\gamma_2=\mu_2/\lambda$.

Очевидно, наиболее интересной характеристикой является величина $p_{\text{отк}}$, равная вероятности отказа в обслуживании требования. Из формулы (3.10) непосредственно следует, что для уменьшения величины $p_{\text{отк}}$ нужно взять в качестве первого прибор, имеющий большую производительность. В этом случае на второй прибор будет поступать более редкий поток необслуженных требований. Обслуживать же такой поток более целесообразно прибором меньшей производительности.

Пример. В упаковочном цехе имеются два упаковочных автомата, первый из которых упаковывает в среднем 40 изделий в час, а второй – 30 изделий в час. В случайные моменты времени на первый автомат поступает для упаковки в среднем 80 изделий в час. Если этот автомат оказывается занятым, то изделие поступает для упаковки на второй автомат. Если и второй автомат оказывается занятым, то изделие остается неупакованным.

Здесь $\lambda=80$; $\mu_1 = 40$; $\mu_2=30$; $\gamma_1=0{,}5$; $\gamma_2=0{,}375$. Отсюда по формуле (3.10) находим $p_{\text{отк}} = 0{,}44$, т. е. чуть ли не половина всех изделий, поступающих в цех на упаковку, оказывается неупакованной. Улучшения работы упаковочного цеха можно добиться, например, заменой старых упаковочных автоматов новыми, более производительными.

Так, если первый упаковочный автомат заменить на новый, для которого $\mu_1 = 120$, то формула (3.10) при $\gamma_1 = 1{,}5$ и $\gamma_2 = 0{,}375$ дает $p_{\text{отк}} = 0{,}22$, т. е. уже только примерно пятая часть изделий не будет упаковываться. В этом случае из формул (3.11) получаем $p_{01} = 0{,}25$, $p_{10} = 0{,}17$, $p_{00} = 0{,}36$, и, следовательно, первый упаковочный автомат простаивает в среднем $p_{00}+p_{01}=$ $0{,}61=61\%$ рабочего времени, а второй – $p_{00}+p_{10}=0{,}53=63\%$ рабочего времени. Одновременно оба упаковочных автомата простаивают $p_{00}=0{,}36=36\%$ рабочего времени.

На первый взгляд может показаться парадоксальным, что при таких сравнительно больших простоях упаковочных автоматов 22% изделий остаются неупакованными. Дело здесь заключается в том, что изделия поступают на упаковку нерегулярно.

Заметим, что если поменять местами упаковочные автоматы, т. е. первым считать менее производительный автомат, то формула (3.10) при $\gamma_1 = 0{,}375$ и $\gamma_2 = 1{,}5$ дает $p_{\text{отк}} = 0{,}27$. Таким образом, неудачное упорядочение двух упаковочных автоматов приводит к увеличению на 5% числа неупакованных изделий.

3.4. Многофазные системы

Ранее рассматривались только такие системы массового обслуживания, в которых каждое требование обслуживалось только одним прибором. В некоторых случаях, однако, обслуживание одного требования осуществляется последовательно несколькими приборами. Именно сначала на приборах первой группы осуществляется первая фаза обслуживания данного требования, затем на приборах второй группы – вторая фаза обслуживания этого же требования и т. д., так что требование считается обслуженным только в том случае, когда оно пройдет через все фазы обслуживания, начиная с первой. Такие системы массового обслуживания называются многофазными, или многоэтапными. При этом в зависимости от того, покидает требование систему или становится в очередь (если все обслуживающие приборы данной фазы заняты), имеем многофазную систему с отказами или с ожиданием.

В простейшем случае число фаз равно двум и число приборов в каждой фазовой группе равно единице. Рассмотрим подобную систему массового обслуживания с ожиданием. Как обычно, предположим, что входящий поток требований является пуассоновским с параметром λ, а время обслуживания подчиняется показательному закону с параметром μ_1 для прибора первой фазы и с μ_2 для прибора второй фазы. Если очередное требование перед первой или второй фазами обслуживания застает соответствующий прибор занятым, то оно становится в очередь и ждет пока ранее поступившее требование не будет обслужено в данной фазе. В этом случае для установившегося режима

$$p_{n_1,n_2} = \alpha_1^{n_1}\alpha_2^{n_2}(1-\alpha_1)(1-\alpha_2),\ n_1,n_2 \ge 0\,, \tag{3.12}$$

где p_{n_1,n_2} – вероятность того, что в первой фазе находится n_1, а во второй – n_2 требований. При этом предполагаются выполненными неравенства $\alpha_1 = \lambda/\mu_1 < 1$; $\alpha_2 = \lambda/\mu_2 < 1$.

Вероятность простоя одного первого прибора

$$P_1 = \sum_{k=1}^{\infty} p_{0k} = \sum_{k=1}^{\infty}\alpha_2^k(1-\alpha_1)(1-\alpha_2) = \alpha_2(1-\alpha_1). \tag{3.13}$$

Аналогичное выражение получаем для вероятности простоя одного второго прибора:

$$P_2 = \alpha_1(1-\alpha_2). \tag{3.14}$$

Среднее число требований, находящихся в первой фазе (на обслуживании и в очереди), можно определить по формуле

$$N_1 = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} n_1 p_{n_1,n_2} = \frac{\alpha_1}{1-\alpha_1}. \tag{3.15}$$

Аналогичная величина для второй фазы определяется из равенства

$$N_2 = \alpha_2/(1-\alpha_1), \tag{3.16}$$

так что среднее число требований, находящихся в рассматриваемой системе, равно $N_1 + N_2$.

Пример. Из многих цехов для обязательной проверки на контрольный пункт поступают некоторые изделия. Так как цехов много, то естественно предположить, что изделия поступают на контрольный пункт в случайные моменты времени, образуя неограниченный пуассоновский поток требований. Контроль изделий состоит в том, что сначала часть параметров изделий должен проверить контролер А, а затем другую часть параметров должен проверить контролер В. Предполагается, что время проверки параметров априори непредсказуемо (из-за зависимости от многих случайных факторов) и имеет показательное распределение. Статистический анализ показал, что в данных условиях $\lambda = 20$, $\mu_1 = 50$, $\mu_2 = 25$. Здесь $\alpha_1 = 20/50 = 0{,}4$; $\alpha_2 = 20/25 = 0{,}8$ и, следовательно, по формулам (3.12) – (3.16) находим $p_{00} = (1-0{,}4)(1-0{,}8) = 0{,}12$; $P_1 = 0{,}8(1-0{,}4) = 0{,}48$; $P_2 = 0{,}08$; $N_1 = 0{,}67$; $N_2 = 4$.

Таким образом, оба контролера одновременно незаняты в среднем 12% рабочего времени, а в отдельности незаняты в среднем: контролер А – 48% и контролер В – 8% рабочего времени. Очереди у контролера А практически нет, а у контролера В в среднем ожидают проверки три изделия.

Если система массового обслуживания замкнутая и, стало быть, число требований, поступающих на обслуживание, ограничено числом m, то вместо формулы (3.12) имеем

$$p_{n_1,n_2} = \alpha_1^{n_1}\alpha_1^{n_2}\frac{(1-\alpha_1)(1-\alpha_2)(\alpha_1-\alpha_2)}{(\alpha_1-\alpha_2)\left(\alpha_1^{m+2}-\alpha_2^{m+2}\right)+\alpha_1\alpha_2\left(\alpha_1^{m+1}-\alpha_2^{m+1}\right)}.$$

При $m \to \infty$ эта формула, естественно, переходит в формулу (3.14).

Если рассматривается система массового обслуживания с отказами, то вместо формулы (3.12) вероятности всех возможных состояний удовлетворяют следующим формулам:

$$p_{00} = \frac{1}{(1+\alpha_1)(1+\alpha_2)}; \quad p_{11} = \frac{\alpha_1\alpha_2^2}{\alpha_1+\alpha_2} p_{00}; \; p_{01} = \alpha_2 p_{00};$$

$$p_{00} + p_{01} + p_{10} + p_{11} = 1 .$$

При этом вероятность обслуживания каждого требования в обеих фазах (пропускная способность системы) может быть найдена из равенства $p_{обсл} = 1/\alpha_2 (p_{00} + p_{10})$.

Так, если в рассмотренном примере предположить, что изделия, поступающие на контрольный пункт, остаются непроверенными, когда они застают занятым соответствующий прибор, то $p_{00} = 0{,}397$; $p_{11} = 0{,}085$; $p_{01} = 0{,}318$; $p_{10} = 0{,}200$; $p_{обсл} = 0{,}75$, т. е. в этом случае только 75% изделий будут проверены.

3.5. Системы с накопителем требований

На практике встречаются такие системы массового обслуживания, в которых требования поступают сначала в так называемый накопитель, а затем, после окончания обслуживания предыдущей партии накопившихся требований, – на обслуживание. Такая ситуация возникает, например, при обработке деталей, предварительно накапливающихся в бункере.

Рассмотрим систему массового обслуживания, состоящую из накопителя требований и обслуживающего прибора. В накопитель поступает пуассоновский поток требований с параметром λ. Если в накопителе имеется уже m требований, то очередное требование получает отказ. Время обслуживания прибором каждой партии требований, поступивших с накопителя, подчиняется показательному закону с параметром μ. Моменты поступления партий накопившихся требований на обслуживающий прибор являются случайными. В этом случае вероятность p_k того, что в накопителе при установившемся режиме имеется ровно k требований, можно определить по формуле

$$p_k = \begin{cases} \dfrac{\alpha^k}{(\alpha+1)^{k+1}}, k < m; \\ \dfrac{\alpha^m}{(\alpha+1)^m}, k = m, \end{cases} \tag{3.17}$$

где $\alpha = \lambda / \mu$. Очевидно, величина p_m равна вероятности отказа.

Зная вероятности p_k, легко находим среднее число M требований, находящихся в накопителе:

$$M = \sum_{k=1}^{m} kp_k . \qquad (3.18)$$

Если $M << m$, то это значит, что загрузка накопителя слабая.

Пример. В некотором музее имеется только один экскурсовод, который обслуживает группы туристов, состоящие не более чем из 10 человек. Среднее время обслуживания группы туристов этим экскурсоводом составляет 20 мин. Туристы приходят в музей в случайные моменты времени в среднем 30 человек в час и ждут, пока не освободится экскурсовод, если число ожидающих туристов менее 10, или уходят, если нет возможности попасть в музей с ближайшей группой.

Здесь $\alpha = \lambda / \mu = 30/3 = 10$. Следовательно, по формулам (3.17), (3.18) имеем:

k	0	1	2	3	4	5	6	7	8	9	10
p_k	0,091	0,083	0,075	0,068	0,062	0,056	0,052	0,047	0,042	0,038	0,386

и $M = \sum_{k=1}^{10} kp_k = 6{,}14$. Это означает, что в среднем 38,6% туристов не попадают в музей, а среднее число ожидающих туристов равно 6,14. Если среднее время обслуживания группы туристов сократить, скажем, вдвое, то $p_{10} = 0{,}162$, т. е. только 16,2% туристов не попадут в музей.

3.6. Системы со смешанным потоком требований

До сих пор мы рассматривали системы массового обслуживания с единственным входящим потоком требований. Однако на практике весьма распространены системы массового обслуживания с двумя, тремя и более независимыми входящими потоками требований. Такие системы называются системами со смешанным потоком требований. Простейшая система массового обслуживания со смешанным потоком требований, для которой полностью разработан математический аппарат, функционирует следующим образом. В систему, состоящую из одного обслуживающего прибора, поступают независимо друг от друга два пуассоновских потока требований; первый с интенсивностью λ_1, а второй – λ_2. Застав прибор занятым, требование из первого потока становится в очередь, а из второго потока покидает систему необслуженным. Время обслуживания требований из обоих потоков подчиняется показательному закону.

Рассмотрим более интересный случай, когда требования обоих потоков должны быть в конечном счете обслужены, причем требования первого потока имеют абсолютный приоритет в обслуживании. Это означает следующее: если при поступлении требования из любого потока прибор свободен, то это требование начинает немедленно обслуживаться; если же прибор занят, то требование из второго потока становится в очередь в любом случае, а требование из первого, т. е. приоритетного потока, – только при наличии ранее поступивших необслуженных требований из первого потока, включая сюда и требование, находящееся на обслуживании. Если же таковых нет и, стало быть, в очереди стоят только требования из второго потока и на обслуживании находится также требование из второго потока, то это требование снимается с обслуживания независимо от того, сколько времени оно там находилось, и начинает обслуживаться требование из первого (приоритетного) потока. После освобождения прибора на обслуживание поступает то требование приоритетного потока (если таковое имеется), время ожидания которого наибольшее. Требования из второго потока поступают на обслуживание только при условии, если нет ожидающих требований приоритетного потока. При этом среди требований второго потока соблюдается обычная дисциплина очереди: «Первым пришел, первым и поступай на обслуживание». В частности, первым на обслуживание поступает то требование из второго потока, обслуживание которого было прервано появлением требования из приоритетного потока.

Заметим, что система с относительным приоритетом в обслуживании отличается от рассматриваемой системы с абсолютным приоритетом только тем, что требование из приоритетного потока, застав прибор занятым обслуживанием требования из второго (неприоритетного) потока, ждет окончания его обслуживания. Например, хотя в билетной кассе инвалиды войны пропускаются вне очереди, процесс продажи билетов очередному пассажиру не прерывается при появлении инвалида.

Естественно считать, что в общем случае среднее время обслуживания требований из разных потоков неодинаково. Обозначим через $1/\mu_1$ среднее время обслуживания требований, имеющих абсолютный приоритет, а через $1/\mu_2$ – аналогичную величину для требований из второго (неприоритетного) потока.

Для рассматриваемой системы массового обслуживания представляют интерес M_1' – среднее число требований из приоритетного потока, находящихся в очереди; M_2'' – среднее число требований из неприоритетного потока, находящихся в системе, $\theta_{\text{ож}}'$ – среднее время ожидания в очереди требования из приоритетного потока, $\theta_{\text{ож}}''$ – среднее время ожидания в очереди требования из неприоритетного потока.

Если выполняется условие $\alpha_1 + \alpha_2 < 1$, где $\alpha_1 = \lambda_1/\mu_1$, $\alpha_2 = \lambda_2/\mu_2$, то для установившегося режима справедливы формулы

$$M_1' = \frac{\alpha_1^2}{1-\alpha_1}; \;\; M_2'' = \frac{\alpha_2}{1-\alpha_1-\alpha_2}\left(1+\frac{\mu_2}{\mu_1}\frac{\alpha_1}{1-\alpha_1}\right);$$

$$\theta_{\text{ож}}' = \frac{M_1'}{\lambda_1}; \;\; \theta_{\text{ож}}'' = \frac{M_2''}{\lambda_2} - \frac{1}{\mu_2}. \qquad (3.19)$$

Пример. Обувная мастерская с одним мастером осуществляет обычный и срочный (в присутствии заказчика) ремонт обуви. При поступлении заказа на срочный ремонт обычный ремонт прерывается. В среднем в час поступает одна пара обуви на срочный ремонт и три пары обуви на обычный ремонт. На обычный ремонт одной пары мастер затрачивает в среднем 15 мин, а на срочный (обычно мелкий) – 3 мин.

Здесь $\lambda_1 = 1; \lambda_2 = 3$; $\mu_1 = 20$; $\mu_2 = 4$ и, следовательно, $\alpha_1 = 1/20 = 0{,}05; \; \alpha_2 = 3/4 = 0{,}75$. Так как $\alpha_1 + \alpha_2 = 0{,}05 + 0{,}75 < 1$, то по формулам (3.19) находим:

$$M_1' = \frac{0{,}05^2}{1-0{,}05} = 0{,}0026; \;\; M_2'' = \frac{0{,}75}{1-0{,}05-0{,}75}\left(1+\frac{4}{20}\frac{0{,}05}{1-0{,}05}\right) = 3{,}8;$$

$$\theta_{\text{ож}}' = \frac{0{,}0026}{1}\text{ ч} \approx 0{,}16 \text{ мин}; \;\; \theta_{\text{ож}}'' = \frac{3{,}8}{3} - \frac{1}{4} = 1{,}02\text{ч} \approx 61 \text{ мин}.$$

Таким образом, для срочного ремонта среднее время ожидания в очереди составляет 0,16 мин (т. е. практически никакой очереди нет), а для обычного – 61 мин.

3.7. Системы с ненадежными обслуживающими приборами

На практике встречаются системы массового обслуживания, в которых каждое требование, проходящее через обслуживающий прибор, обслуживается не достоверно, а с некоторой вероятностью p. Если подобная система есть система с отказами, состоящая из n однотипных приборов со средним временем обслуживания $1/\mu$, на которую поступает поток требований с интенсивностью λ, то вместо классической формулы Эрланга

$$p_{\text{обсл}} = 1 - p_{\text{отк}}, \qquad (3.20)$$

где $p_{\text{отк}} = \frac{\alpha^n}{n!}\left[\sum_{k=0}^{n}\frac{a^k}{k!}\right]^{-1}$, $\alpha = \lambda/\mu$, имеем

$$p_{\text{обсл}}' = p_{\text{обсл}} \cdot p. \qquad (3.21)$$

Действительно, вероятность $p_{\text{обсл}}'$ того, что требование будет обслужено по теореме умножения вероятностей независимых событий, равна произведению вероятности $p_{\text{обсл}}$ того, что требование будет принято к обслуживанию, на вероятность p обслуживания этого требования.

Таким образом, при $p < 1$ вероятность обслуживания и, следовательно, пропускная способность системы уменьшаются по сравнению с рассмотренным в подразд. 2.1 классическим случаем $p = 1$.

Пример. В ателье проката имеется восемь полотеров, каждый из которых берется на прокат в среднем на два дня. В среднем в ателье за полотерами обращаются пять человек в день. Вероятность того, что взятый на прокат полотер будет исправно работать, равна 0,95.

Здесь $n = 8$; $\lambda = 5$; $\mu = 0{,}5$; $p = 0{,}95$, откуда по формулам (3.20), (3.21) находим $p'_{\text{обсл}} = 0{,}76$; $p_{\text{обсл}} = 0{,}76 \cdot 0{,}95 = 0{,}73$. Таким образом, ателье проката в среднем на 73% обеспечивает население полотерами. Если бы эти полотеры время от времени не выходили из строя, то население было бы обеспечено ими на 76%.

Рассмотрим теперь такие системы массового обслуживания, в которых отказавшие приборы ремонтируются в процессе функционирования системы. Пусть в систему массового обслуживания, состоящую из n однотипных приборов, поступает пуассоновский поток требований с интенсивностью λ. Время обслуживания требований каждым прибором подчиняется показательному закону с параметром μ. Приборы в случайные моменты времени выходят из строя как тогда, когда они свободны, так и тогда, когда они заняты обслуживанием требования. Неисправные приборы сразу же поступают в ремонт, так что вместо того, чтобы обслуживать, эти приборы сами начинают обслуживаться. Время работы каждого прибора случайно и подчиняется показательному закону с параметром γ. Одновременно в ремонте может находиться только один прибор, так что если имеются другие неисправные приборы, то они стоят в очереди. Отремонтированные приборы могут также выходить из строя. Суммарный поток отказов еще не ремонтированных и ремонтированных (возможно многократно) приборов предполагается пуассоновским с параметром q. Если прибор отказал во время обслуживания требования, то это требование теряется (даже если имеются свободные исправные приборы).

Рассмотрим сначала систему с отказами. В этом случае требование теряется, если в момент его поступления в систему нет ни одного свободного исправного прибора. Если имеется только один прибор ($n = 1$), то для такой системы в случае установившегося режима справедливы формулы:

$$\pi_1 = \frac{q}{q+\gamma};\ \pi_0 = 1 - \pi_1;\ p_1 = \frac{\lambda_\lambda}{(\lambda+q)(\mu+q)+\gamma\lambda};\ p_0 = 1 - p_1,\ p' = p_1 + \pi_1;$$

$$p'' = \frac{qp_1}{\lambda}; \; p_{\text{отк}} = p' + p'', \tag{3.22}$$

где π_1 – вероятность того, что прибор неисправен; π_0 – вероятность того, что прибор исправен; p_1 – вероятность того, что прибор занят обслуживанием; p_0 – вероятность того, что прибор свободен; p' – вероятность того, что прибор либо неисправен, либо занят обслуживанием; p'' – вероятность того, что требование покинет систему не полностью обслуженным из-за выхода из строя прибора; $p_{\text{отк}}$ – вероятность отказа в обслуживании.

Пример. В случайные моменты времени к моечной установке подъезжают в среднем четыре автомашины в час. Средняя производительность моечной установки составляет 10 автомашин в час. В случайные моменты времени моечная установка выходит из строя в среднем два раза за 100 ч работы. Иными словами, среднее время наработки моечной установки на один отказ равно 50 ч. Время, необходимое для ремонта моечной установки, случайное и составляет в среднем 1 ч. Если автомашина подъезжает к моечной установке в тот момент, когда установка занята или находится в ремонте, то она уезжает невымытой. Здесь $\lambda = 4$; $\mu = 10$; q=0,02; $\gamma = 1$, из формул (3.22) находим:

$$\pi_1 = \frac{0{,}02}{1 + 0{,}02} = 0{,}020; \quad p_1 = \frac{4 \cdot 1}{(1 + 0{,}02)(10 + 0{,}02) + 1 \cdot 4} = 0{,}281;$$

$$p' = 0{,}281 + 0{,}020 = 0{,}301; \; p'' = \frac{0{,}02 \cdot 0{,}281}{4} = 0{,}001;$$

$$p_{\text{отк}} = 0{,}301 + 0{,}001 = 0{,}302.$$

Следовательно, в среднем 30,2% автомашин не будут вымыты моечной установкой. Из них только 0,1% – по причине выхода из строя моечной установки во время мойки. Это указывает на то, что узким местом является не плохая надежность моечной установки, а ее перегруженность.

Перейдем теперь к другому варианту рассматриваемой системы с ненадежными приборами – к системе с ожиданием. В этом случае требование не теряется, а становится в очередь, если в момент его поступления в систему нет ни одного исправного прибора. Очевидно, эта система может быть как с ограничением на длину очереди, время пребывания в очереди или системе, так и без ограничений. Для определенности рассмотрим случай с ограничением на длину очереди, когда в очереди имеется только m мест. Так как формулы для произвольных n и m громоздки, то приведем формулы только для частного случая $n = 1$, $m = 2$ и установившегося режима:

$$p_0 = \frac{1}{1 + \omega + \omega^2 + \omega^3}; \; \omega = \frac{\lambda}{(\mu + q)\pi_0}; \; \pi_0 = \frac{\gamma}{(q + \gamma)};$$

$$p_1 = \omega p_0\,;\; p_2 = \omega^2 p_0\,;\; p_3 = \omega^3 p_0\,;$$
$$M = (p_1 + p_2 + p_3)\pi_0\,;\; M_2 = M + p_2 + 2p_3\,; \qquad (3.23)$$
$$p'' = qM/\lambda\,;\; p_{\text{отк}} = p_3 + p''.$$

Здесь p_0 – вероятность того, что прибор свободен; p_1 – вероятность того, что прибор занят и очередь отсутствует; p_2 – вероятность того, что прибор занят и имеется одно требование в очереди; p_3 – вероятность того, что прибор занят и имеются два требования в очереди; M – среднее число требований, находящихся на обслуживании; M_2 – среднее число требований, находящихся в системе; π_0, $p_{\text{отк}}$, p'' имеют тот же смысл, что и для системы с отказами.

Пример. В ЭВМ в случайные моменты времени поступают для обработки блоки информации некоторой структуры с плотностью два блока в час. Время обработки одного блока информации на ЭВМ в среднем равно 10 мин. Иногда ЭВМ выходит из строя, причем среднее время ее наработки на один отказ равно 3 ч. Среднее же время ремонта ЭВМ составляет 15 мин. Во внешней памяти ЭВМ могут одновременно храниться не более двух блоков информации, так что если очередной блок информации поступает в тот момент, когда в ЭВМ уже два блока информации ожидают обработки, то он теряется. Блок информации теряется также и в том случае, когда во время его обработки ЭВМ дает сбой.

Здесь $\lambda = 2$, $\mu = 6$, $q = 1/3$, $\gamma = 4$. Поэтому согласно (3.23) $p_0 = 0{,}668$; $p_1 = 0{,}228$; $p_2 = 0{,}078$; $p_3 = 0{,}026$; $M = 0{,}307$; $M_2 = 0{,}440$; $p'' = 0{,}051$; $p_{\text{отк}} = 0{,}077$.

Таким образом, в среднем 7,7% блоков информации не будет обработано: из них 5,1% за счет сбоя ЭВМ во время обработки и 2,6% за счет отсутствия места во внешней памяти ЭВМ для хранения свыше двух блоков информации. Среднее число блоков информации, находящихся либо на обработке (во внутренней памяти ЭВМ), либо на хранении (во внешней памяти ЭВМ), равно 0,44.

3.8. Системы с групповым обслуживанием

В некоторых системах массового обслуживания поступившее требование начинают обслуживать сразу несколько приборов. В простейшем случае поступившее требование начинают обслуживать независимо друг от друга сразу все n имеющихся в системе приборов, и обслуживание этого требования считается законченным, как только оканчивается его обслуживание одним из приборов. Затем осуществляется аналогичное групповое обслуживание следующего требования (если таковое имеется) и т. д. Если время обслуживания каждого требования каждым прибором подчиняется показательному закону с параметром μ, то такая система эквивалентна системе с одним прибором, обслуживающим каждое требование в течение времени, подчиняющегося показательному закону с параметром $n\mu$. Иначе говоря, среднее время группового обслуживания уменьшается в n раз. Поэтому для систем с подобной организацией группового обслуживания применимы соответствующие ранее рассмотренные формулы, если в них подставить единицу вместо n, и $n\mu$ вместо μ. В частности, для классической пуассоновской системы с отказами, имеющей параметры (в обычных обозначениях) λ, $n\mu$, μ, вместо формулы Эрланга

$$p_{\text{отк}} = \frac{(\lambda/\mu)^n 1/n!}{\sum_{k=0}^{n} (\lambda/\mu)^k 1/k!} \tag{3.24}$$

имеем

$$p_{\text{отк}} = \frac{\lambda/n\mu}{1+\lambda/\mu}. \tag{3.25}$$

Заметим, что описанная организация группового обслуживания ведет к уменьшению пропускной способности системы массового обслуживания. Это следует из того, что правая часть формулы (3.24), как легко проверить, всегда меньше правой части формулы (3.25).

Рассмотрим другой вариант организации группового обслуживания требований, при котором увеличивается пропускная способность системы массового обслуживания. Пусть это будет опять пуассоновская система с отказами, имеющая параметры λ, μ, n. Обслуживание производится сразу всеми n приборами, но, в отличие от только что рассмотренного случая, здесь вновь поступившее требование не получает отказа, если в системе уже находится требование на групповом обслуживании.

Обслуживание в этой системе организовано так. Если поступившее в систему требование застает все приборы свободными, то оно начинает обслуживаться, причем всеми n приборами одновременно. При этом, как и в только что рассмотренном случае, среднее время обслуживания уменьшается в n раз. После окончания обслуживания все n приборов становятся одновременно свободными. Если поступившее в систему требование застает там ранее поступившее другое требование, то оно также принимается на обслуживание. При этом часть приборов (безразлично каких) продолжает обслуживать ранее поступившее требование, а другая часть начинает обслуживать вновь прибывшее требование. Таким образом, уже два требования обслуживаются всеми n приборами.

В общем случае, если новое требование застает в системе k $(k = 0, 1, ..., n-1)$ ранее прибывших требований, оно принимается на обслуживание. Это означает, что $k+1$ требований обслуживаются всеми n приборами, причем распределение n приборов по $k+1$ требованиям произвольно. При этом важно, чтобы все n приборов были заняты. Если новое требование застает в системе n ранее прибывших требований, то оно получает отказ и покидает систему необслуженным. Такие системы иногда называются системами с полной взаимопомощью между приборами. В них освободившаяся после обслуживания какого-либо требования группа приборов присоединяется к обслуживанию других требований, находящихся в системе. Это означает, что если в данный момент в системе имеется хотя бы одно требование, то все приборы будут заняты его обслуживанием. В этом случае вместо формулы (3.25) имеем:

$$p_{\text{отк}} = \begin{cases} \dfrac{n\mu/\lambda - 1}{(n\mu/\lambda)^{n+1} - 1}, & \text{если} \quad \lambda \neq m\mu; \\ \dfrac{1}{n+1}, & \text{если} \quad \lambda = m\mu \quad . \end{cases} \tag{3.26}$$

Предположим теперь, что в рассматриваемой системе массового обслуживания с полной взаимопомощью между приборами имеется m мест для ожидания. Тогда если очередное требование застает в системе n ранее прибывших требований, то оно уже не получает отказа, а становится в очередь и ожидает освобождения хотя бы одного из приборов.

В общем случае, если новое требование застает в системе $n+k$ $(k = 0, 1, ..., m-1)$ ранее прибывших требований, оно становится в очередь. Наконец, если новое требование застает в системе $n+m$ требований, то оно получает отказ и покидает систему необслуженным.

Для такой смешанной системы с ограничением на длину очереди и полной взаимопомощью между обслуживающими приборами вместо (3.26) имеем:

$$p_{\text{отк}} = \begin{cases} \dfrac{n\mu/\lambda - 1}{(n\mu/\lambda)^{m+n+1} - 1}, & \text{если} \quad \lambda \neq m\mu; \\ \dfrac{1}{n+m+1}, & \text{если} \quad \lambda = m\mu \quad . \end{cases} \tag{3.27}$$

При этом среднее время пребывания требования в очереди, например для случая $\lambda = n\mu$, можно определить по формуле

$$M\theta = \frac{m(m+1)}{2\lambda(n+m+1)}. \tag{3.28}$$

Пример. В случайные моменты времени в порт для разгрузки причаливают баржи; в среднем шесть барж в час. В порту имеются три бригады грузчиков, каждая из которых разгружает баржу в среднем за 0,5 ч. Очередная баржа начинает разгружаться, если в порту менее трех барж, становится в очередь, если в порту менее (3+4) барж и, наконец, покидает порт неразгруженной, если в порту находится (3+4) баржи, из которых три разгружаются и четыре ожидают разгрузки.

Если разгрузка барж организована по принципу полной взаимопомощи между бригадами, то по формуле (3.27) имеем ($\lambda = 6$, $n = 3$, $\mu = 2$, $m = 4$) $p_{\text{отк}} =$ 1/(3+4+1) = 0,125, т. е. в среднем 12,5% барж не будет разгружаться в этом порту. При этом согласно (3.28) среднее время нахождения в очереди барж $M\theta = \dfrac{4(4+1)}{2 \cdot 6(3+4+1)} = \dfrac{5}{24}$ ч $\approx 12{,}5$ мин.

Заметим, что если каждая баржа будет разгружаться только одной бригадой, то $p_{\text{отк}} = 0{,}145$, т. е. на 2,5% больше отказов.

4. МАРКОВИЗИРОВАНИЕ МОДЕЛЕЙ МАССОВОГО ОБСЛУЖИВАНИЯ

В большинстве практически интересных задач потоки событий, переводящие систему из одного состояния в другое, могут быть с достаточной точностью аппроксимированы потоками Пуассона. В этом случае процесс, протекающий в системе, благодаря замене реальных потоков пуассоновскими сводится к марковскому, т.е. становится процессом без последействия.

Представление процесса функционирования системы в виде марковского случайного процесса с непрерывным временем позволяет применить для описания процесса ее функционирования аппарат обыкновенных дифференциальных уравнений (примеры марковских моделей массового обслуживания приведены в предыдущих разделах).

Сведение эмпирического процесса функционирования СМО к марковскому путем замены реальных потоков пуассоновскими приводит, естественно, к некоторым погрешностям, величина которых зависит от степени последействия реальных потоков. В большинстве практических случаев эти погрешности невелики и не превышают погрешностей в исходных данных, возникающих при статистической обработке реальных потоков. Однако иногда имеют место случаи, когда потоки событий, переводящие систему из одного состояния в другое, являются потоками с ограниченным последействием, заменить которые пуассоновскими возможно лишь в качестве первого приближения.

Поэтому необходимо рассмотреть вопрос о том, каким образом учесть последействие потоков так, чтобы процесс функционирования систем можно было описать по-прежнему в виде марковского процесса с непрерывным временем и дискретным числом состояний.

В данном разделе рассматривается простой метод, который в некоторых случаях может быть полезен в ситуациях такого рода. В этом методе используются два свойства потока Эрланга, состоящих в том, что, во-первых, поток Эрланга – поток с ограниченным последействием и, во-вторых, закон распределения промежутка времени между событиями в потоке Эрланга является композицией показательных законов распределения.

4.1. Потоки Эрланга и их свойства

Потоком Эрланга k-го порядка называется ординарный стационарный поток с ограниченным последействием, у которого промежутки времени T между событиями в потоке являются суммами случайных величин, распределенных по показательному закону с параметром λ.

Поясним это с помощью рис. 7. Обозначим через E события, соответствующие пуассоновскому потоку, и предположим, что они пронумерованы в порядке их появления, начиная с некоторого исходного момента. Пусть E^* – события, определяемые следующим образом: E^* наступает в момент появления событий E с номером, кратным k. Тогда поток, состоящий из последовательности событий E^*, и будет потоком Эрланга k-го порядка. В частности, на рисунке приведен поток Эрланга при $k = 3$. Очевидно, при $k = 1$ имеем исходный пуассоновский поток, для которого промежутки времени T между событиями в потоке распределены по закону с плотностью

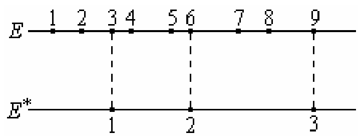

Рис. 7

$$f(t) = \lambda e^{-\lambda t}, \tag{4.1}$$

где λ – параметр потока.

Промежуток времени между n-м и $(n+k)$-м событиями в потоке Эрланга равен сумме k промежутков времени между событиями в пуассоновском потоке. Следовательно, искомый закон является k-кратной композицией закона (4.1), плотность распределения его имеет вид

$$f_k(t) = \frac{\lambda(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t}, \tag{4.2}$$

где k – порядок потока Эрланга, λ – плотность исходного пуассоновского потока.

Закон распределения с плотностью (4.2) называется законом Эрланга k-го порядка.

В качестве примера на рис. 8 приведены законы распределения $f_k(t)$ для $k = 1, 2, 3, 6, 11, 26$ при $\lambda = 1$.

Известно, что поток называется потоком с ограниченным последействием, если промежутки времени между последовательными событиями представляют собой независимые случайные величины. Следовательно, потоки Эрланга – потоки с ограниченным последействием, так как промежутки времени T между событиями, являясь суммами независимых случайных величин, независимы между собой.

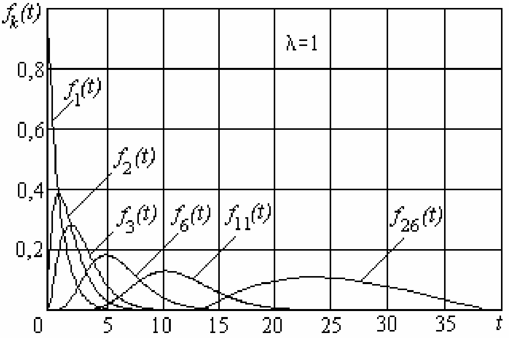

Рис. 8

Математическое ожидание и дисперсия промежутка времени между событиями в потоке Эрланга k-го порядка определяются формулами

$$m_k = k / \lambda \,,\ D_k = k / \lambda^2 \,. \tag{4.3}$$

Из этих формул получаем, что для закона Эрланга любого порядка справедливо соотношение $m_k = \lambda D_k$.

Плотность потока Эрланга Λ_k обратна величине математического ожидания m_k:

$$\Lambda_k = \lambda / k \,, \tag{4.4}$$

где λ – параметр показательного закона.

Из (4.3) и (4.4)

$$m_k = 1 / \Lambda_k \,,\ D_k = 1 / k\Lambda_k^2 \,. \tag{4.5}$$

С учетом (4.4) выразим $f_k(t)$ через плотность Λ_k :

$$f_k(t) = \frac{[k\Lambda_k]^k t^{k-1}}{(k-1)!} e^{-k\Lambda_k t} \,.$$

Законы $f_k(t)$ при $\Lambda_k = 1$ для $k = 1, 2, 3, 6, 11, 26$ показаны на рис. 9.

При постоянной плотности потока Λ_k математическое ожидание не зависит от порядка потока k, а дисперсия с возрастанием k неограниченно убывает, при $k \to \infty$ и $D_k \to 0$ (см. (4.5) и рис. 9).

Таким образом, поток Эрланга обладает ценным свойством, которое состоит в следующем: при неограниченном увеличении порядка потока k и при постоянной плотности Λ_k он приближается к регулярному потоку с постоянными интервалами времени между событиями, равными $1/\Lambda_k$, а плотность распределения $f_k(t)$ при $k \to \infty$ обращается в δ-функцию в точке $t = 1/\Lambda_k$.

Это свойство потоков Эрланга позволяет при различных k получать практически любую степень последействия потока – от полного отсутствия последействия (k = =1) до жесткой функциональной связи между моментами появления событий (при $k \to \infty$). Таким образом, порядок потока k может служить как бы мерой последействия, имеющегося в потоке. Заметим, что даже при $k \to \infty$ поток Эрланга остается потоком с ограниченным последействием.

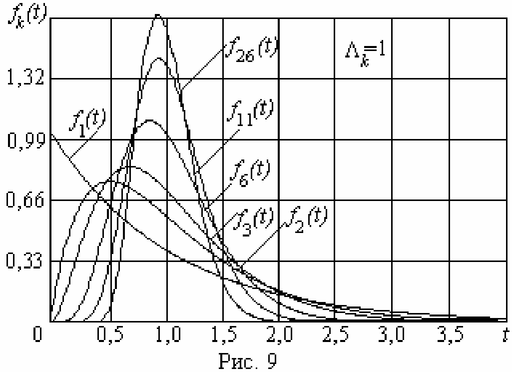

Рис. 9

Определим асимметрию S_k закона Эрланга:

$$S_k = \mu_{k_3} / \left(\sqrt{D_k}\right)^3, \tag{4.6}$$

где μ_{k_3} – третий центральный момент распределения $f_k(t)$.

В свою очередь, $\mu_{k_3} = \alpha_{k_3} - 3m_k\left(D_k - m_k^2\right) + 2m_k^3$, где α_{k_3} – третий начальный момент распределения $f_k(t)$,

$$\alpha_{k_3} = \int_0^\infty t^3 f_k(t)dt = \int_0^\infty t^3 \frac{\lambda(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t} \; . \tag{4.7}$$

Берем интеграл по частям, после чего получаем $\alpha_{k_3} = k(k+1)(k+2)/\lambda^3$. Подставляя полученные выражения в (4.6) и производя несложные преобразования, получаем $S_k = 2/\sqrt{k}$. Для показательного закона (k = 1) S_k = 2. При $k \to \infty$ $\lim\limits_{k\to\infty} S_k = \lim\limits_{k\to\infty} 2/\sqrt{k} = 0$.

Таким образом, при увеличении порядка потока k скошенность распределения постепенно исчезает, распределение становится более симметричным. Закон Эрланга любого порядка имеет положительную асимметрию.

Определим эксцесс распределения $f_k(t)$

$$\varepsilon_{x_k} = \left[\mu_{k_4} / \left(\sqrt{D_k}\right)\right] - 3, \qquad (4.8)$$

где μ_{k_4} – четвертый центральный момент распределения $f_k(t)$,

$$\mu_{k_4} = \alpha_{k_4} - 4m_k\alpha_{k_3} + 6m_k^2\alpha_{k_2} - 3m_k^4. \qquad (4.9)$$

Начальные моменты определим аналогично (4.7):

$$\alpha_{k_4} = k(k+1)(k+2)(k+3)/\lambda^4,\ \alpha_{k_3} = k(k+1)(k+2)/\lambda^3,\ \alpha_{k_2} = k(k+1)/\lambda^2.$$

Подставляя полученные выражения в (4.9) и (4.8) и производя несложные преобразования, получаем $\varepsilon_{x_k} = 6/k$.

Для показательного закона (k = 1) ε_{x_k} = 6. С увеличением k ε_{x_k} уменьшается и в пределе при $k \to \infty$ $\varepsilon_{x_k} \to 0$ (так как закон Эрланга при $k \to \infty$ асимптотически стремится к нормальному закону). Пределы изменения ε_{x_k} = 6÷0. Следовательно, закон Эрланга любого порядка всегда имеет положительный эксцесс.

Используя зависимость числовых характеристик и степени последействия потока Эрланга от порядка потока k, рассмотрим способ замены реальных потоков с ограниченным последействием потоками Эрланга различного порядка.

4.2. Замена реальных потоков потоками Эрланга

Произвольные потоки с ограниченным последействием, встречающиеся на практике, можно заменить потоками Эрланга с теми же математическим ожиданием и дисперсией промежутка времени между событиями в потоке.

Пусть, например, в результате статистической обработки промежутков времени между событиями в произвольном потоке с ограниченным последействием получены оценки для математического ожидания и дисперсии T: $m_c = 2$, $D_c = 1$.

Заменим этот поток потоком Эрланга с теми же характеристиками. Из (4.3) получаем $k = m_c^2 / D_c$, откуда $k = 4$.

Таким образом, произвольный поток с ограниченным последействием с математическим ожиданием промежутка времени между событиями $m_c =$ $= 2$ и дисперсией $D_c = 1$ можно заменить потоком Эрланга 4-го порядка с плотностью $\lambda = 2$.

Следует заметить, что такая замена возможна при условии

$$D_c \le m_c^2 \,. \tag{4.10}$$

Ввиду функциональной зависимости между математическим ожиданием и дисперсией в потоке Эрланга рассмотренная замена возможна не для любых значений математического ожидания m_c и дисперсии D_c произвольного потока. Например, при $m_c = 2$, $D_c = 1{,}6$ получаем $k = 2{,}5$, т.е. дробное значение k. Таким образом, используя лишь рассмотренные потоки Эрланга, не всегда удается реальный поток с ограниченным последействием заменить рассмотренными потоками Эрланга. Однако существует класс потоков, так называемые обобщенные потоки Эрланга, с помощью которых можно заменить любой реальный поток с ограниченным последействием, с любым математическим ожиданием и дисперсией, если соблюдено условие (4.10).

Обобщенным потоком Эрланга называется поток, у которого промежутки времени T между событиями являются суммами случайных величин T_i, подчиняющихся показательному закону распределения с различными параметрами λ_i.

Для обобщенных законов Эрланга введем обозначение $f^*{}_k(t)$. Тогда закон распределения с плотностью

$$f^*{}_k(t) = \left(\prod_{i=1}^{k} \lambda_i\right) \sum_{i=1}^{k} e^{-\lambda_i t} \Big/ \prod_{j=1}^{k} \left(\lambda_j - \lambda_i\right)$$

будем называть обобщенным законом Эрланга k-го порядка.

Определим числовые характеристики закона $f^*{}_k(t)$ – математическое ожидание $m^*{}_k$, дисперсию $D^*{}_k$, асимметрию $S^*{}_k$ и эксцесс $\varepsilon^*{}_{x_k}$.

В общем случае для обобщенного закона Эрланга k-го порядка $f^*{}_k(t)$ с параметрами λ_i ($i = 1, 2, \ldots, k$) математическое ожидание

$$m^*{}_k = \sum_{j=1}^{k} \prod_{\substack{i=1 \\ i \ne j}}^{k} \lambda_i \Big/ \prod_{i=1}^{k} \lambda_i \,, \tag{4.11}$$

дисперсия

$$D^*{}_k = \sum_{j=1}^{k} \prod_{\substack{i=1 \\ i \ne j}}^{k} \lambda_i^2 \Big/ \prod_{i=1}^{k} \lambda_i^2 \,, \tag{4.12}$$

асимметрия

$$S^*_k = 2\sum_{j=1}^{k}\prod_{\substack{i=1\\ i\neq j}}^{k}\lambda_i^3 \Bigg/ \sqrt{\left(\sum_{j=1}^{k}\prod_{\substack{i=1\\ i\neq j}}^{k}\lambda_i^2\right)^3},$$

эксцесс

$$\varepsilon^*_{x_k} = 6\sum_{j=1}^{k}\prod_{\substack{i=1\\ i\neq j}}^{k}\lambda_i^4 \Bigg/ \left(\sum_{j=1}^{k}\prod_{\substack{i=1\\ i\neq j}}^{k}\lambda_i^2\right)^2.$$

Рассмотрим некоторые свойства обобщенных законов Эрланга. Для простоты возьмем обобщенный закон Эрланга 2-го порядка

$$f_2^*(t) = \frac{\lambda_1\lambda_2}{\lambda_2-\lambda_1}\left(e^{-\lambda_1 t} - e^{-\lambda_2 t}\right).$$

Очевидно, при $\lambda_2 = \lambda_1 = \lambda$ $f_2^*(t) = f_2(t)$, где $f_2(t)$ – закон Эрланга 2-го порядка.

При $\lambda_2 \to \infty$

$$\lim_{\lambda_2\to\infty} f_2^*(t) = \lim_{\lambda_2\to\infty}\frac{\lambda_2\lambda_1}{\lambda_2-\lambda_1}\left(e^{-\lambda_1 t} - e^{-\lambda_2 t}\right) = \lambda_1 e^{-\lambda_1 t} = f_1(t),$$

где $f_1(t)$ – показательный закон с параметром λ_1.

Естественно, при λ_1 или λ_2, стремящихся к нулю, $f_2^*(t) \to 0$.

При рассмотрении обобщенных законов Эрланга более высокого порядка выявляются некоторые дополнительные свойства. Рассмотрим изменение обобщенного закона Эрланга третьего порядка $f_3^*(t)$ при изменении параметров λ_1, λ_2 и λ_3:

$$f_3^*(t) = \lambda_1\lambda_2\lambda_3\left[\frac{e^{-\lambda_1 t}}{(\lambda_2-\lambda_1)(\lambda_3-\lambda_1)} + \frac{e^{-\lambda_2 t}}{(\lambda_1-\lambda_2)(\lambda_3-\lambda_2)} + \frac{e^{-\lambda_3 t}}{(\lambda_1-\lambda_3)(\lambda_2-\lambda_3)}\right].$$

Очевидно, при $\lambda_1 = \lambda_2 = \lambda_3$ $f_3^*(t) = f_3(t)$. Если $\lambda_1 \neq \lambda_2$, а $\lambda_3 \to \infty$, то $\lim\limits_{\lambda_3\to\infty} f_3^*(t) = f_2^*(t)$. Если же $\lambda_3 = \lambda_2$, но $\lambda_2 \neq \lambda_1$, то $\lim\limits_{\lambda_3\to\lambda_2} f_3^*(t) = f_{3(1)}^*(t)$, где $f^*_{3(1)}(t)$ – обобщенный закон Эрланга 3-го порядка, у которого два параметра из трех равны между собой.

Если для $f_{3(1)}^{*}(t)$ положить $\lambda_1 = \lambda_2$, то $\lim\limits_{\lambda_1 \to \lambda_2} f_{3(1)}^{*}(t) = f_3(t)$, если же $\lambda_1 \to \infty$, то $\lim\limits_{\lambda_1 \to \infty} f_{3(1)}^{*}(t) = f_2(t)$.

Примем для обобщенных законов Эрланга с несколькими равными параметрами λ_0 следующие обозначения: $f^*_k(t)$ – все k параметров λ_i разные; $f^*_{k(k\text{-}2)}(t)$ – два параметра равны λ_0, остальные k-2 параметров λ_i разные; $f^*_{k(k\text{-}3)}(t)$ – три параметра равны λ_0, остальные k-3 параметров λ_i разные; $f^*_{k(1)}(t)$ – k-1 параметр равны λ_0, один параметр λ_1 отличен от остальных; $f_k(t)$ – все параметры равны между собой, т.е. λ_i ($i = 1, 2, \ldots, k$) равны λ_0.

Используя эти обозначения, все превращения обобщенных законов Эрланга при изменении параметров λ_i можно представить в виде рис. 10. Переходы слева направо, обозначенные на схеме горизонтальными стрелками, происходят при одном из $\lambda_i \to \lambda_0$. Переходы слева направо, обозначенные на схеме вертикальными стрелками, происходят, когда один из параметров $\lambda_i \to \infty$, т.е. при понижении порядка законов Эрланга, при этом количество равных параметров λ_0 остается прежним. Переходы справа налево, обозначенные вертикальными стрелками, происходят, когда один из параметров $\lambda_0 \to \infty$,при этом также получается понижение порядка закона Эрланга.

Из рассмотрения свойств обобщенных законов Эрланга можно сделать важный для практики вывод: при постоянной плотности потока Эрланга Λ^*_k, а следовательно, при постоянном математическом ожидании m^*_k, изменяя параметры λ_i от λ_0 до ∞, можно получить различные значения дисперсии D^*_k. При этом наиболее удобным в практическом отношении является обобщенный закон Эрланга $f^*_{k(1)}(t)$, у которого все параметры равны λ_0, кроме одного λ_1. Варьируя эти параметры при постоянном математическом ожидании m^*_k, можно получить различные значения дисперсии D^*_k.

Используя это свойство, можно реальные потоки с ограниченным последействием заменять потоками Эрланга с теми же математическим ожиданием и дисперсией промежутка времени между событиями в потоке. При этом, однако, необходимо соблюдать условие $m_k^2 / k \le D_k \le m_k^2$.

Из выражений (4.11) и (4.12) нетрудно получить общее выражение для математического ожидания и дисперсии закона $f^*_{k(1)}(t)$:

$$m^*_{k(1)} = \frac{(k-1)\lambda_1 + \lambda_0}{\lambda_1 \lambda_0}, \tag{4.13}$$

$$D^{*}{}_{k(1)} = \frac{(k-1)\lambda_1^2 + \lambda_0^2}{\lambda_1^2 \lambda_0^2}. \qquad (4.14)$$

Тогда порядок замены реального потока с ограниченным последействием с математическим ожиданием m_c и дисперсией D_c обобщенным потоком Эрланга $f^*_{k(1)}(t)$ с теми же характеристиками сводится к следующему.

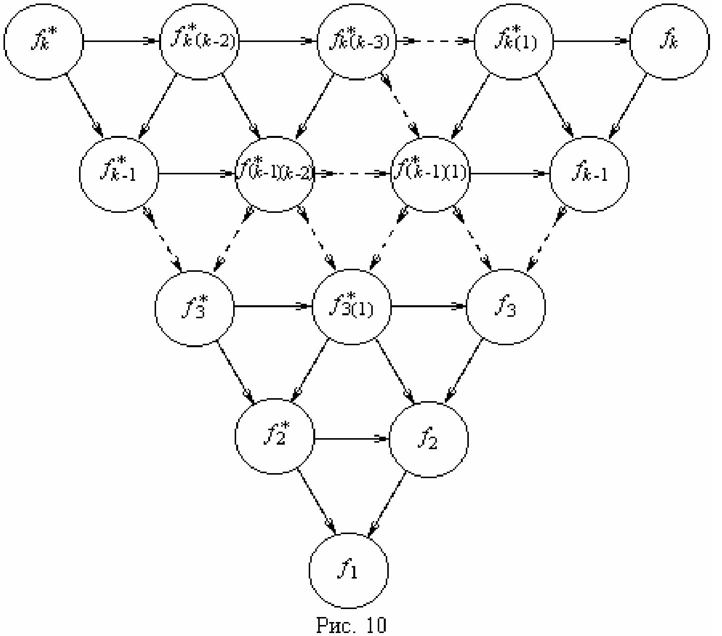

Рис. 10

Определяется порядок потока Эрланга $k = m_c^2 / D_c$. Дробное значение k округляется в большую сторону.

Из выражения (4.13) имеем

$$\lambda_1 = \lambda_0 / (1 - k + \lambda_0 m_c). \qquad (4.15)$$

Подставляя (4.15) в (4.14) и производя несложные преобразования, получаем $(m_c^2 - D_c)\lambda_0^2 - 2(1-k)m_c\lambda_0 + k(k-1) = 0$, откуда

$$\lambda_{0_{1,2}} = \frac{m_0(k-1) \pm \sqrt{(k-1)^2 m_c^2 - k(k-1)(m_c^2 - D_c)}}{(m_c^2 - D_c)}.$$

Подставив λ_{0_1} и λ_{0_2} в (4.15), получим два значения λ_{1_1} и λ_{1_2}. Таким образом, для одних и тех же значений m_c и D_c получаем закон $f^*_{k(1)}(t)$ с двумя различными сочетаниями параметров λ_1 и λ_0: (λ_{1_1}, λ_{0_1}), (λ_{1_2}, λ_{0_2}), при этом $m^*_k = m_c$, $D^*_k = D_c$.

Метод, с помощью которого получено распределение Эрланга, подсказывает способ моделирования процессов с ограниченным последействием, при котором потоки, переводящие систему из одного состояния в другое, являются потоками Эрланга (см. рис. 7).

Предположим для конкретности, что события E обозначают процесс переходов системы из состояния в состояние, число состояний системы равно k и процесс протекает следующим образом: 1) в каждый момент времени система может находиться только в одном из k состояний; 2) система последовательно проходит состояния 1, 2, …, k. Время пребывания системы в каждом состоянии распределено по показательному закону с параметром $\lambda = k\Lambda_k$.

Длительность пребывания системы во всех k состояниях соответствует промежутку времени между событиями E^*, эти события и образуют поток Эрланга.

Такое разбиение каждого состояния системы на k состояний позволяет продемонстрировать важный факт. Пусть $n^*(t)$ и $n(t)$ – соответственно числа появления событий E^* и E за время от 0 до t. Пуассоновский процесс $n(t)$ является марковским, процесс $n^*(t)$, напротив, немарковский: вероятноять того, что $n^*(t)dt = j$ при условии $n^*(t) = i$, зависит от момента появления последнего события в потоке E^*, следовательно, от того, в каком из k состояний находится система [от значения $n(t)$]. Следовательно, разбиение каждого состояния системы на k состояний позволяет свести исследуемый процесс к марковскому $n(t)$, который более удобен для исследования.

4.3. Марковские модели процессов с ограниченным последействием

Рассмотрим способ, позволяющий в необходимых случаях описывать процессы функционирования систем уравнениями для марковского процесса с непрерывным временем и дискретным числом состояний при потоках событий, имеющих ограниченное последействие. В качестве потоков с ограниченным последействием будем рассматривать только потоки Эрланга, считая, что произвольный стационарный поток с ограниченным последействием всегда можно заменить обобщенным потоком Эрланга способом, рассмотренным в подразд. 4.2.

Рассмотрим дискретный марковский процесс (рис. 11), состояния которого составляют множество $X = \{x_0, x_1, \ldots, x_n\}$. Введем в рассмотрение подмножество $Z = \{Y_{\beta,} Y_\gamma\}$ и транзитивное подмножество состояний $Y = \{x_1, x_2, \ldots, x_{n-1}\}$.

$$x_0 \xrightarrow{\lambda_0} x_1 \xrightarrow{\lambda_1} x_2 \xrightarrow{\lambda_2} \ldots \longrightarrow x_{n-1} \xrightarrow{\lambda_{n-1}} x_n$$

Рис. 11

Особенность данного процесса состоит в том, что каждое из подмножеств Y_β и Y_γ имеет единственное состояние, и для определенности будем считать, что $Y_\beta = \{x_0\}$, $Y_\gamma = \{x_n\}$. Состояние x_0 является входным, а x_n – выходным состоянием для подмножества Y.

Введем следующее ограничение: процесс блуждания по состояниям подмножества Y возможен только в одном направлении: $x_1 \rightarrow x_n$.

В первую очередь, нас будет интересовать закон распределения времени T пребывания в транзитивном подмножестве состояний Y при условии, что время отсчитывается от момента начала блуждания процесса из состояния x_0 до момента выхода поцесса из состояний подмножества Y, т.е. до момента перехода процесса из состояния x_{n-1} в состояние x_n подмножества Y_γ.

Система дифференциальных уравнений для определения закона распределения времени T имеет вид

$$
\begin{gathered}
\dot{p}_0(t) = -\lambda_0 p_0(t), \\
\dot{p}_1(t) = -\lambda_1 p_1(t) + \lambda_0 p_0(t), \\
\cdots\cdots\cdots\cdots \\
\dot{p}_n(t) = \lambda_{n-1} p_{n-1}(t).
\end{gathered}
\tag{4.16}
$$

Начальное условие для интегрирования этой системы дифференциальных уравнений: $p_0(0) = 0$, $p_i(0) = 0$ $(i = 1, 2, \ldots, n)$.

Решив систему дифференциальных уравнений (4.16) и найдя вероятность $p_n(t)$, можно определить плотность распределения случайной величины T $f(t) = \dot{p}_n(t)$.

Рассмотрим частный случай, когда процесс блуждания по состояниям простейший, тогда все интенсивности λ_i потоков являются постоянными величинами.

Здесь случайная величина T – время пребывания в транзитивном подмножестве состояний Y – является суммой случайных величин τ_i, распределенных по показательному закону с параметрами λ_i. Очевидно, закон распределения случайной величины T – композиция показательных законов распределения с параметрами λ_i, следовательно, является обобщенным законом Эрланга. Порядок закона Эрланга определяется количеством состояний подмножества Y. Для n-1 состояний подмножества Y имеем обобщенный закон Эрланга (n-1)-го порядка $f^*_{n-1}(t)$.

Для рассматриваемого частного случая, таким образом, закон распределения случайной величины T можно найти, не интегрируя систему уравнений (4.16).

Отметим то важное для практики обстоятельство, о котором речь шла раньше. Оно состоит в том, что если транзитивное подмножество состояний Y представить в виде одного обобщенного состояния $\tilde{x}$ (рис. 12), время пребывания в T котором определяется законом Эрланга, то процесс, протекающий в этом преобразованном множестве состояний $\tilde{X} = \{x_0, \tilde{x}, x_n\}$, будет уже не марковским, а процессом с ограниченным последействием.

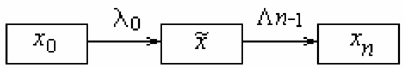

Рис. 12

Справедливо обратное утверждение. Если имеем процесс с ограниченным последействием, состояния которого составляет множество $\tilde{X} = \{x_0, \tilde{x}, x_n\}$, то данный процесс с ограниченным последействием можно представить в виде марковского случайного процесса с множеством состояний $X = \{x_0, x_1, \ldots, x_{n-1}, x_n\}$ путем введения дополнительных состояний x_i ($i = 1, 2, \ldots, n$-1). При этом произвольные потоки переходов процесса из состояния в состояние заменяются потоками Эрланга соответствующего порядка, а состояние $\tilde{x}$ представляется в виде транзитивного множества состояний $Y = \{x_1, x_2, \ldots, x_{n-1}\}$. В этом случае процесс с ограниченным последействием может быть описан дифференциальными уравнениями (4.16).

Библиографический список

1. *Гнеденко Б.В.* и др. Приоритетные модели обслуживания. М.: Изд-во МГУ, 1973.
2. *Гнеденко Б.В., Коваленко И.Н.* Введение в теорию массового обслуживания. М.: Наука, 1987.
3. *Джейсуол Н.* Очереди с приоритетом. М.: Мир, 1973.
4. *Ивченко Г.И., Каштанов В.А., Коваленко И.Н.* Теория массового обслуживания. М.: Высшая школа, 1982.
5. *Кениг Д., Штойян Д.* Методы теории массового обслуживания. М.: Радио и связь, 1981.
6. *Климов Г.П.* Стохастические системы обслуживания. М.: Наука, 1966.
7. *Кокс Д., Смит В.* Теория очередей. М.: Мир, 1966.
8. *Коршунов Ю.М.* Математические основы кибернетики. М.: Энергоатомиздат, 1987.
9. *Кофман А., Крюон Р.* Массовое обслуживание. М.: Мир, 1965.
10. *Риордан Дж.* Вероятностные системы обслуживания. М.: Связь. 1966.
11. *Саати Т.* Элементы теории массового обслуживания и ее приложения. М.: Сов. радио, 1971.
12. *Саульев В.К.* Математические модели теории массового обслуживания. М.: Статистика, 1979.
13. *Советов Б.Я., Яковлев С.А.* Моделирование систем. М.: Высшая школа, 2001.
14. *Тараканов К.Е., Овчаров Л.А., Тырышкин А.Н.* Аналитические методы исследования систем. М.: Сов. радио, 1974.

Оглавление

Королев Сергей Николаевич

Марковские модели массового обслуживания

Редактор *Г.В. Никитина*

Корректор *А.А. Баутдинова*

Подписано в печать 29.12.2002 Формат 60х84/16. Бумага документная.
Печать трафаретная. Усл.-печ. 4,9. Уч.-изд.5,5. Тираж 200 экз. Заказ №
Балтийский государственный технический университет
Типография БГТУ
190005, С.-Петербург, 1-я Красноармейская ул., д.1