Глава 15
ЭЛЕМЕНТЫ ТЕОРИИ МАССОВОГО ОБСЛУЖИВАНИЯ

15.1. Основные понятия. Классификация СМО

При исследовании операций часто приходится сталкиваться с системами, предназначенными для многоразового использования при решении однотипных задач. Возникающие при этом процессы получили название *процессов обслуживания*, а системы — *систем массового обслуживания* (*СМО*). Примерами таких систем являются телефонные системы, ремонтные мастерские, вычислительные комплексы, билетные кассы, магазины, парикмахерские и т.п.

Каждая СМО состоит из определенного числа обслуживающих единиц (приборов, устройств, пунктов, станций), которые будем называть *каналами* обслуживания. Каналами могут быть линии связи, рабочие точки, вычислительные машины, продавцы и др. По числу каналов СМО подразделяют на *одноканальные* и *многоканальные*.

Заявки поступают в СМО обычно не регулярно, а случайно, образуя так называемый *случайный поток заявок (требований)*. Обслуживание заявок, вообще говоря, также продолжается какое-то случайное время. Случайный характер потока заявок и времени обслуживания приводит к тому, что СМО оказывается загруженной неравномерно: в какие-то периоды времени скапливается очень большое количество заявок (они либо становятся в очередь, либо покидают СМО необслуженными), в другие же периоды СМО работает с недогрузкой или простаивает.

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

В качестве *показателей эффективности* СМО используются: среднее[1] число заявок, обслуживаемых в единицу времени, среднее число заявок в очереди; среднее время ожидания обслуживания; вероятность отказа в обслуживании без ожидания; вероятность того, что число заявок в очереди превысит определенное значение и т.п.

СМО делят на два основных типа (класса): СМО *с отказами* и СМО *с ожиданием* (*очередью*). В СМО с отказами заявка, поступившая в момент, когда все каналы заняты, получает отказ, покидает СМО и в дальнейшем процессе обслуживания не участвует (например, заявка на телефонный разговор в момент, когда все каналы заняты, получает отказ и покидает СМО необслуженной). В СМО с ожиданием заявка, пришедшая в момент, когда все каналы заняты, не уходит, а становится в очередь на обслуживание.

СМО с ожиданием подразделяются на разные виды в зависимости от того, как организована очередь: с ограниченной или неограниченной длиной очереди, с ограниченным временем ожидания и т.п.

Для классификации СМО важное значение имеет *дисциплина обслуживания*, определяющая порядок выбора заявок из числа поступивших и порядок распределения их между свободными каналами. По этому признаку обслуживание заявки может быть организовано по принципу *"первая пришла — первая обслужена"*, *"последняя пришла — первая обслужена"* (такой порядок может применяться, например, при извлечении для обслуживания изделий со склада, ибо последние из них оказываются часто более доступными) или *обслуживание с приоритетом* (когда в первую очередь обслуживаются наиболее важные заявки). Приоритет может быть как *абсолютным*, когда более важная заявка "вытесняет" из-под обслуживания обычную заявку (например, в случае аварийной ситуации плановые работы ремонтных бригад прерываются до ликвидации аварии), так и *относительным*, когда более важная заявка получает лишь "лучшее" место в очереди.

---

[1] Здесь и в дальнейшем средние величины понимаются как математические ожидания соответствующих случайных величин.

15.2. Понятие марковского случайного процесса

Процесс работы СМО представляет собой *случайный процесс*.

Под *случайным (вероятностным или стохастическим) процессом* понимается процесс изменения во времени состояния какой-либо системы в соответствии с вероятностными закономерностями.

Процесс называется *процессом с дискретными состояниями*, если его возможные состояния S_1, S_2, S_3... можно заранее перечислить, а переход системы из состояния в состояние происходит мгновенно (скачком). Процесс называется *процессом с непрерывным временем*, если моменты возможных переходов системы из состояния в состояние не фиксированы заранее, а случайны.

Процесс работы СМО представляет собой случайный процесс с дискретными состояниями и непрерывным временем. Это означает, что состояние СМО меняется скачком в случайные моменты появления каких-то событий (например, прихода новой заявки, окончания обслуживания и т.п.).

Математический анализ работы СМО существенно упрощается, если процесс этой работы — марковский. Случайный процесс называется *марковским* или *случайным процессом без последствия*, если для любого момента времени t_0 вероятностные характеристики процесса в будущем зависят только от его состояния в данный момент t_0 и не зависят от того, когда и как система пришла в это состояние.

Пример марковского процесса: система S — счетчик в такси. Состояние системы в момент t характеризуется числом километров (десятых долей километров), пройденных автомобилем до данного момента. Пусть в момент t_0 счетчик показывает S_0. Вероятность того, что в момент $t > t_0$ счетчик покажет то или иное число километров (точнее, соответствующее число рублей) S_1, зависит от S_0, но не зависит от того, в какие моменты времени изменялись показания счетчика до момента t_0.

Многие процессы можно приближенно считать марковскими. Например, процесс игры в шахматы; система S — группа шахматных фигур. Состояние системы характеризуется числом фигур противника, сохранившихся на доске в момент t_0. Вероятность того, что в момент $t > t_0$ материальный перевес будет на стороне одного из противников, зависит в первую очередь от того, в ка-

ком состоянии находится система в данный момент t_0, а не от того, когда и в какой последовательности исчезли фигуры с доски до момента t_0.

В ряде случаев предысторией рассматриваемых процессов можно просто пренебречь и применять для их изучения марковские модели.

При анализе случайных процессов с дискретными состояниями удобно пользоваться геометрической схемой — так называемым *графом состояний*. Обычно состояния системы изображаются прямоугольниками (кружками), а возможные переходы из состояния в состояние — стрелками (ориентированными дугами), соединяющими состояния.

15.1. Построить граф состояний следующего случайного процесса: устройство S состоит из двух узлов, каждый из которых в случайный момент времени может выйти из строя, после чего мгновенно начинается ремонт узла, продолжающийся заранее неизвестное случайное время.

Решение. Возможные состояния системы: S_0 — оба узла исправны; S_1 — первый узел ремонтируется, второй исправен; S_2 — второй узел ремонтируется, первый исправен; S_3 — оба узла ремонтируются. Граф системы приведен на рис. 15.1.

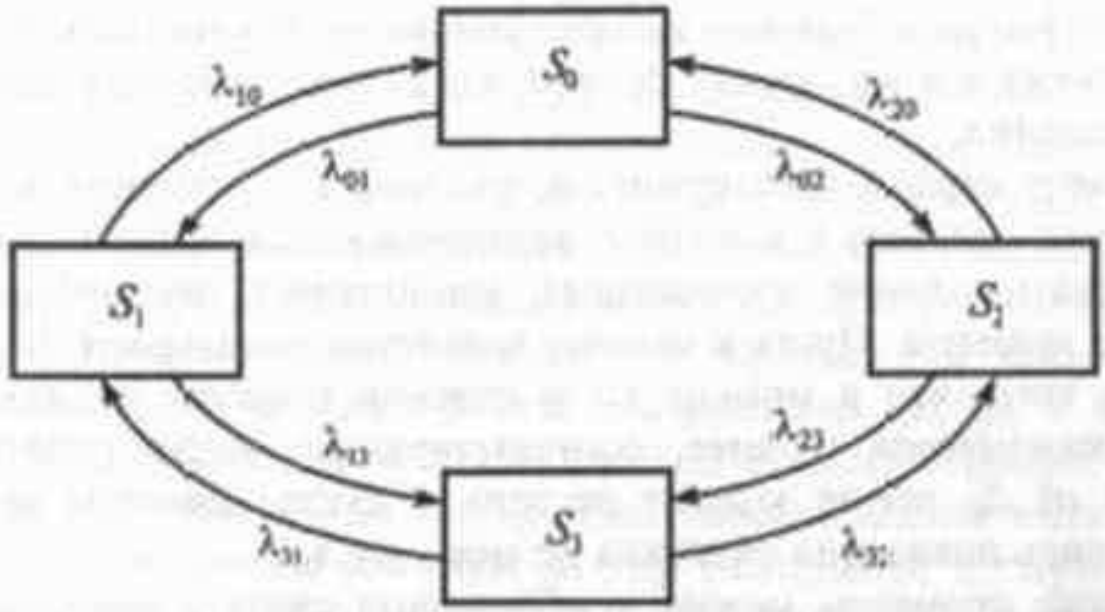

Рис. 15.1

Стрелка, направленная, например, из S_0 в S_1, означает переход системы в момент отказа первого узла, из S_1 в S_0 — переход в момент окончания ремонта этого узла.

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

Для математического описания марковского случайного процесса с дискретными состояниями и непрерывным временем, протекающего в СМО, познакомимся с одним из важных понятий теории вероятностей — понятием потока событий.

15.3. Потоки событий

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

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

Поток событий называется *регулярным*, если события следуют одно за другим через определенные равные промежутки времени. Например, поток изделий на конвейере сборочного цеха (с постоянной скоростью движения) является регулярным.

Поток событий называется *стационарным*, если его вероятностные характеристики не зависят от времени. В частности, интенсивность стационарного потока есть величина постоянная: $\lambda(t)=\lambda$. Например, поток автомобилей на городском проспекте не является стационарным в течение суток, но этот поток можно считать стационарным в течение суток, скажем, в часы пик. Обращаем внимание на то, что в последнем случае фактическое число проходящих автомобилей в единицу времени (например, в каждую минуту) может заметно отличаться друг от друга, но среднее их число будет постоянно и не будет зависеть от времени.

Поток событий называется *потоком без последействия*, если для любых двух непересекающихся участков времени τ_1 и τ_2 — число событий, попадающих на один из них, не зависит от числа событий, попадающих на другие. Например, поток пассажиров, входящих в метро, практически не имеет последействия. А, скажем,

поток покупателей, отходящих с покупками от прилавка, уже имеет последействие (хотя бы потому, что интервал времени между отдельными покупателями не может быть меньше, чем минимальное время обслуживания каждого из них).

Поток событий называется *ординарным*, если вероятность попадания на малый (элементарный) участок времени Δt двух и более событий пренебрежимо мала по сравнению с вероятностью попадания одного события. Другими словами, поток событий ординарен, если события появляются в нем поодиночке, а не группами. Например, поток поездов, подходящих к станции, ординарен, а поток вагонов не ординарен.

Поток событий называется ***простейшим*** *(или* ***стационарным пуассоновским****), если он одновременно стационарен, ординарен и не имеет последействия.* Название "простейший" объясняется тем, что СМО с простейшими потоками имеет наиболее простое математическое описание. Заметим, что регулярный поток не является "простейшим", так как он обладает последействием: моменты появления событий в таком потоке жестко зафиксированы.

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

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

Рассмотрим на оси времени Ot (рис. 15.2) простейший поток событий как неограниченную последовательность случайных точек.

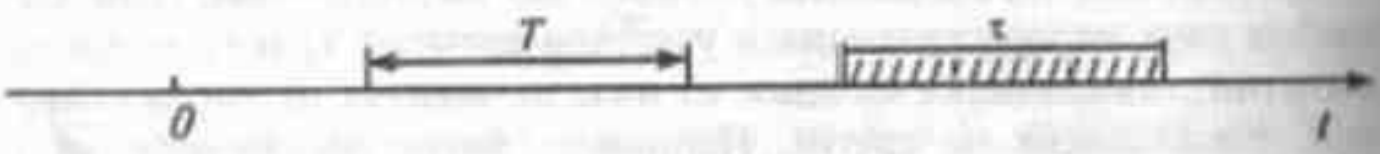

Рис. 15.2

Можно показать (*см.*, например, [3]), что для простейшего потока число m событий (точек), попадающих на произвольный участок времени τ, распределено по *закону Пуассона*

$$P_m(\tau) = \frac{(\lambda\tau)^m}{m!} e^{-\lambda\tau}, \qquad (15.1)$$

для которого математическое ожидание случайной величины равно ее дисперсии: $a = \sigma^2 = \lambda\tau$.

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

$$P_0(\tau) = e^{-\lambda\tau}. \qquad (15.2)$$

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

В соответствии с (15.2) вероятность того, что на участке времени длиной t не появится ни одного из последующих событий, равна

$$P(T \geq t) = e^{-\lambda t}, \qquad (15.3)$$

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

$$F(t) = P(T < t) = 1 - e^{-\lambda t}. \qquad (15.4)$$

Плотность вероятности случайной величины есть производная ее функции распределения (рис. 15.3), т.е.

$$\varphi(t) = F'(t) = \lambda e^{-\lambda t}. \qquad (15.5)$$

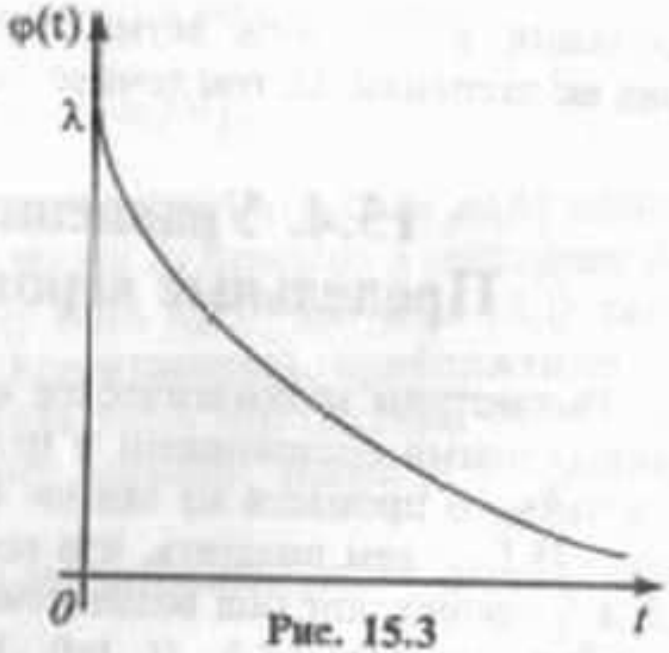

Рис. 15.3

Распределение, задаваемое плотностью вероятности (15.5) или функцией распределения (15.4), называется *показательным* (или *экспоненциальным*). Таким образом, интервал времени между двумя соседними произвольными событиями имеет показательное распределение, для которого матема-

тическое ожидание равно среднему квадратическому отклонению случайной величины

$$a = \sigma = \frac{1}{\lambda} \qquad (15.6)$$

и обратно по величине интенсивности потока λ.

Важнейшее свойство показательного распределения (присущее только показательному распределению) состоит в следующем: если промежуток времени, распределенный по показательному закону, уже длился некоторое время τ, то это никак не влияет на закон распределения оставшейся части промежутка $(T-\tau)$: он будет таким же, как и закон распределения всего промежутка T.

Другими словами, для интервала времени T между двумя последовательными соседними событиями потока, имеющего показательное распределение, любые сведения о том, сколько времени протекал этот интервал, не влияют на закон распределения оставшейся части. Это свойство показательного закона представляет собой, в сущности, другую формулировку для "отсутствия последействия" — основного свойства простейшего потока.

Для простейшего потока с интенсивностью λ вероятность попадания на *элементарный* (*малый*) отрезок времени Δt хотя бы одного события потока равна согласно (15.4)

$$P_{\Delta t} = P(T < \Delta t) = 1 - e^{-\lambda \Delta t} \approx \lambda \Delta t\,. \qquad (15.7)$$

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

15.4. Уравнения Колмогорова. Предельные вероятности состояний

Рассмотрим математическое описание марковского процесса с дискретными состояниями и непрерывным временем на примере случайного процесса из задачи **15.1**, граф которого изображен на рис. 15.1. Будем полагать, что все переходы системы из состояния S_i в S_j происходят под воздействием простейших потоков событий с интенсивностями λ_{ij} (i, j=0, 1, 2, 3); так, переход системы из состояния S_0 в S_1 будет происходить под воздействием потока отка-

зов первого узла, а обратный переход из состояния S_1 в S_0 — под воздействием потока "окончаний ремонтов" первого узла и т.п.

Граф состояний системы с проставленными у стрелок интенсивностями будем называть *размеченным* (*см.* рис. 15.1). Рассматриваемая система S имеет четыре возможных состояния: S_0, S_1, S_2, S_3.

Вероятностью i-го состояния называется вероятность $p_i(t)$ того, что в момент t система будет находиться в состоянии S_i. Очевидно, что для любого момента t сумма вероятностей всех состояний равна единице:

$$\sum_{i=0}^{3} p_i(t) = 1. \qquad \textbf{(15.8)}$$

Рассмотрим систему в момент t и, задав малый промежуток Δt, найдем вероятность $p_0(t+\Delta t)$ того, что система в момент $t+\Delta t$ будет находиться в состоянии S_0. Это достигается разными способами.

1. Система в момент t с вероятностью $p_0(t)$ находилась в состоянии S_0, а за время Δt не вышла из него.

Вывести систему из этого состояния (*см.* граф на рис. 15.1) можно суммарным простейшим потоком с интенсивностью $(\lambda_{01}+\lambda_{02})$, т.е. в соответствии с (15.7), с вероятностью, приближенно равной $(\lambda_{01}+\lambda_{02})\Delta t$. А вероятность того, что система не выйдет из состояния S_0, равна $[1-(\lambda_{01}+\lambda_{02})\Delta t]$. Вероятность того, что система будет находиться в состоянии S_0 по первому способу (т.е. того, что находилась в состоянии S_0 и не выйдет из него за время Δt), равна по теореме умножения вероятностей:

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

2. Система в момент t с вероятностями $p_1(t)$ (или $p_2(t)$) находилась в состоянии S_1 или S_2 и за время Δt перешла в состояние S_0.

Потоком интенсивностью λ_{10} (или λ_{20} — *см.* рис. 15.1) система перейдет в состояние S_0 с вероятностью, приближенно равной $\lambda_{10}\Delta t$ (или $\lambda_{20}\Delta t$). Вероятность того, что система будет находиться в состоянии S_0 по этому способу, равна $p_1(t)\lambda_{10}\Delta t$ (или $p_2(t)\lambda_{20}\Delta t$).

Применяя теорему сложения вероятностей, получим

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

откуда

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

Переходя к пределу при $\Delta t \to 0$ (приближенные равенства, с занные с применением формулы (15.7), перейдут в точные), по чим в левой части уравнения производную $p_0'(t)$ (обозначим для простоты: p_0'):

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

Получили дифференциальное уравнение первого порядка, уравнение, содержащее как саму неизвестную функцию, так и производную первого порядка.

Рассуждая аналогично для других состояний системы S, мож получить систему дифференциальных уравнений Колмогорова д вероятностей состояний:

$$\begin{cases} p_0' = \lambda_{10}p_2 + \lambda_{20}p_2 - (\lambda_{01} + \lambda_{02})p_0, \\ p_1' = \lambda_{01}p_0 + \lambda_{31}p_3 - (\lambda_{10} + \lambda_{13})p_1, \\ p_2' = \lambda_{02}p_0 + \lambda_{32}p_3 - (\lambda_{20} + \lambda_{23})p_2, \\ p_3' = \lambda_{13}p_1 + \lambda_{23}p_2 - (\lambda_{31} + \lambda_{32})p_3. \end{cases} \quad (15.$$

Сформулируем правило составления уравнен Колмогорова. *В левой части каждого из них стоит произ ная вероятности i-го состояния. В правой части — сумма прои дений вероятностей всех состояний (из которых идут стрел данное состояние) на интенсивности соответствующих пото событий, минус суммарная интенсивность всех потоков, выводя систему из данного состояния, умноженная на вероятность дан (i-го состояния).*

В системе (15.9) независимых уравнений на единицу мен общего числа уравнений. Поэтому для решения системы необ димо добавить уравнение (15.8).

Особенность решения дифференциальных уравнений вооб состоит в том, что требуется задать так называемые начальн условия, т.е. в данном случае вероятности состояний систем начальный момент $t = 0$. Так, например, систему уравнений (1 естественно решать при условии, что в начальный момент узла исправны и система находилась в состоянии S_0, т.е. при чальных условиях $p_0(0)=1$, $p_1(0)=p_2(0)=p_3(0)=0$.

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

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

Предельная вероятность состояния S_i имеет четкий смысл: она показывает *среднее относительное время пребывания системы в этом состоянии*. Например, если предельная вероятность состояния S_0, т.е. $p_0=0,5$, то это означает, что в среднем половину времени система находится в состоянии S_0.

Так как предельные вероятности постоянны, то, заменяя в уравнениях Колмогорова их производные нулевыми значениями, получим систему линейных алгебраических уравнений, описывающих стационарный режим. Для системы S с графом состояний, изображенном на рис. 15.1), такая система уравнений имеет вид:

$$\begin{cases} (\lambda_{01} + \lambda_{02})p_0 = \lambda_{10}p_1 + \lambda_{20}p_2, \\ (\lambda_{10} + \lambda_{13})p_1 = \lambda_{01}p_0 + \lambda_{31}p_3, \\ (\lambda_{20} + \lambda_{23})p_2 = \lambda_{02}p_0 + \lambda_{32}p_3, \\ (\lambda_{31} + \lambda_{32})p_3 = \lambda_{13}p_1 + \lambda_{23}p_2. \end{cases} \qquad \textbf{(15.10)}$$

Систему (15.10) можно составить непосредственно по размеченному графу состояний, если руководствоваться правилом, согласно которому *слева в уравнениях стоит предельная вероятность данного состояния p_i, умноженная на суммарную интенсивность всех потоков, ведущих из данного состояния, а справа — сумма произведений интенсивностей всех потоков, входящих в i-е состояние, на вероятности тех состояний, из которых эти потоки исходят.*

15.2. Найти предельные вероятности для системы S из задачи **15.1**, граф состояний которой приведен на рис. 15.1, при $\lambda_{01}=1$, $\lambda_{02}=2$, $\lambda_{10}=2$, $\lambda_{13}=2$, $\lambda_{20}=3$, $\lambda_{23}=1$, $\lambda_{31}=3$, $\lambda_{32}=2$.

Решение. Система алгебраических уравнений, описывающих стационарный режим для данной системы, имеет вид (15.10) или

$$\begin{cases} 3p_0 = 2p_1 + 3p_2, \\ 4p_1 = p_0 + 3p_3, \\ 4p_2 = 2p_0 + 2p_3, \\ p_0 + p_1 + p_2 + p_3 = 1. \end{cases} \qquad (15.11)$$

(Здесь мы вместо одного "лишнего" уравнения системы (15.10) записали нормировочное условие (15.8))

Решив систему (15.11), получим p_0=0,40, p_1=0,20, p_2=0,27, p_3=0,13, т.е. в предельном, стационарном режиме система S в среднем 40% времени будет находиться в состоянии S_0 (оба узла исправны), 20% — в состоянии S_1 (первый узел ремонтируется, второй работает), 27% — в состоянии S_2 (второй узел ремонтируется, первый работает) и 13% времени — в состоянии S_3 (оба узла ремонтируются).▶

▷ **15.3.** Найти средний чистый доход от эксплуатации в стационарном режиме системы S в условиях задач **15.1** и **15.2**, если известно, что в единицу времени исправная работа первого и второго узлов приносит доход соответственно в 10 и 6 ден.ед., а их ремонт требует затрат соответственно в 4 и 2 ден.ед. Оценить экономическую эффективность имеющейся возможности уменьшения вдвое среднего времени ремонта каждого из двух узлов, если при этом придется вдвое увеличить затраты на ремонт каждого узла (в единицу времени).

Решение. Из задачи **15.2** следует, что в среднем первый узел исправно работает долю времени, равную p_0+p_3=0,40+0,27=0,67, а второй узел — p_0+p_1=0,40+0,20=0,60. В то же время первый узел находится в ремонте в среднем долю времени, равную p_1+p_3= =0,20+0,13=0,33, а второй узел — p_2+p_3=0,27+0,13=0,40. Поэтому средний чистый доход в единицу времени от эксплуатации системы, т.е. разность между доходами и затратами, равен

$$Д=0{,}67\cdot10+0{,}60\cdot6-0{,}33\cdot4-0{,}40\cdot2=8{,}18 \text{ ден.ед.}$$

Уменьшение вдвое среднего времени ремонта каждого из узлов в соответствии с (15.6) будет означать увеличение вдвое интенсивностей потока "окончаний ремонтов" каждого узла, т.е. теперь λ_{10}=4, λ_{20}=6, λ_{31}=6, λ_{32}=4 и система линейных алгебраических

уравнений (15.10), описывающая стационарный режим системы S, вместе с нормировочным условием (15.8) примет вид[1]:

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

Решив систему, получим p_0=0,60, p_1=0,15, p_2=0,20, p_3=0,05.

Учитывая, что p_0+p_2=0,60+0,20=0,80, p_0+p_1=0,60+0,15=0,75, p_1+p_3=0,15+0,05=0,20, p_2+p_3=0,20+0,05=0,25, а затраты на ремонт первого и второго узла составляют теперь соответственно 8 и 4 ден. ед., вычислим средний чистый доход в единицу времени:

$$Д_1=0{,}80\cdot10+0{,}75\cdot6-0{,}20\cdot8-0{,}25\cdot4=9{,}9 \text{ ден.ед.}$$

Так как $Д_1$ больше Д (примерно на 20%), то экономическая целесообразность ускорения ремонтов узлов очевидна.▶

15.5. Процесс гибели и размножения

В теории массового обслуживания широкое распространение имеет специальный класс случайных процессов — так называемый *процесс гибели и размножения*. Название этого процесса связано с рядом биологических задач, где он является математической моделью изменения численности биологических популяций.

Граф состояний процесса гибели и размножения имеет вид, показанный на рис. 15.4.

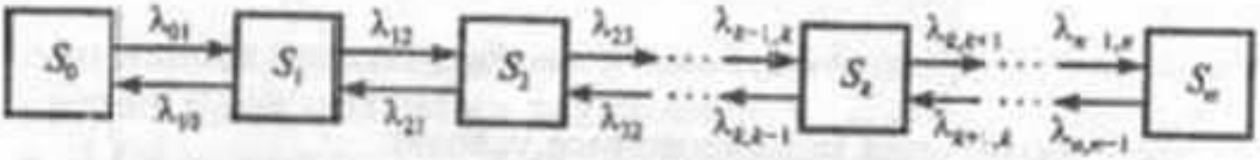

Рис. 15.4

Рассмотрим упорядоченное множество состояний системы S_0, S_1, S_2, ..., S_k. Переходы могут осуществляться из любого со-

[1] При записи системы (15.10) одно "лишнее" уравнение мы исключили.

стояния только в состояния с соседними номерами, т.е. из состояния S_k возможны переходы только либо в состояние S_{k-1}, либо в состояние S_{k+1}[1].

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

По графу, представленному на рис. 15.4, составим и решим алгебраические уравнения для предельных вероятностей состояний (их существование вытекает из возможности перехода из каждого состояния в каждое другое и конечности числа состояний).

В соответствии с правилом составления таких уравнений (см. 15.13) получим: для состояния S_0

$$\lambda_{01}p_0 = \lambda_{10}p_1, \tag{15.12}$$

для состояния S_1 — $(\lambda_{12}+\lambda_{10})p_1 = \lambda_{01}p_0+\lambda_{21}p_2$, которое с учетом (15.12) приводится к виду

$$\lambda_{12}p_1=\lambda_{21}p_2. \tag{15.13}$$

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

$$\begin{cases} \lambda_{01}p_0 = \lambda_{10}p_1, \\ \lambda_{12}p_1 = \lambda_{21}p_2, \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ \lambda_{k-1,k}p_{k-1} = \lambda_{k,k+1}p_k, \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ \lambda_{n-1,n}p_{n-1} = \lambda_{n,n+1}p_n, \end{cases} \tag{15.14}$$

к которой добавляется нормировочное условие

$$p_0 + p_1 + p_2+\ldots+p_n = 1. \tag{15.15}$$

---

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

Решая систему (15.14), (15.15), можно получить

$$p_0=\left(1+\frac{\lambda_{01}}{\lambda_{10}}+\frac{\lambda_{12}\lambda_{01}}{\lambda_{21}\lambda_{10}}+\ldots+\frac{\lambda_{n-1,n}\ldots\lambda_{12}\lambda_{01}}{\lambda_{n,n-1}\ldots\lambda_{21}\lambda_{10}}\right)^{-1}, \qquad (15.16)$$

$$p_1=\frac{\lambda_{01}}{\lambda_{10}}p_0,\ p_2=\frac{\lambda_{12}\lambda_{01}}{\lambda_{21}\lambda_{10}}p_0,\ \ldots,\ p_n=\frac{\lambda_{n-1,n}\ldots\lambda_{12}\lambda_{01}}{\lambda_{n,n-1}\ldots\lambda_{21}\lambda_{10}}p_0. \qquad (15.17)$$

Легко заметить, что в формулах (15.17) для p_1, p_2, ..., p_n коэффициенты при p_0 есть слагаемые, стоящие после единицы в формуле (15.16). Числители этих коэффициентов представляют произведение всех интенсивностей, стоящих у стрелок, ведущих слева направо до данного состояния S_k ($k=1, 2, \ldots, n$), а знаменатели — произведение всех интенсивностей, стоящих у стрелок, ведущих справа налево до состояния S_k.

▷ **15.4.** Процесс гибели и размножения представлен графом (рис. 15.5). Найти предельные вероятности состояний.

Рис. 15.5

Решение. По формуле (15.16) найдем

$$p_0=\left(1+\frac{1}{4}+\frac{2\cdot 1}{3\cdot 4}\right)^{-1}=0{,}706,$$

по (15.17) — $p_1=\frac{1}{4}0{,}706=0{,}176$, $p_2=\frac{2\cdot 1}{3\cdot 4}0{,}706=0{,}118$, т.е. в установившемся, стационарном режиме в среднем 70,6% времени система будет находиться в состоянии S_0, 17,6% — в состоянии S_1 и 11,8% — в состоянии S_2.▶

15.6. СМО с отказами

В качестве показателей эффективности СМО с отказами будем рассматривать:

A — абсолютную пропускную способность СМО, т.е. среднее число заявок, обслуживаемых в единицу времени;

Q — *относительную пропускную способность*, т.е. среднюю долю пришедших заявок, обслуживаемых системой;

$P_{отк.}$ — *вероятность отказа*, т.е. того, что заявка покинет СМО необслуженной;

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

Одноканальная система с отказами. Рассмотрим задачу.

Имеется один канал, на который поступает поток заявок с интенсивностью λ. Поток обслуживаний имеет интенсивность μ[1]. Найти предельные вероятности состояний системы и показатели ее эффективности.

Система S (СМО) имеет два состояния: S_0 — канал свободен, S_1 — канал занят. Размеченный граф состояний представлен на рис. 15.6.

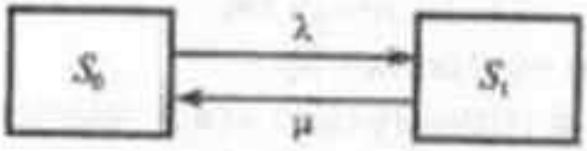

Рис. 15.6

В предельном, стационарном режиме система алгебраических уравнений для вероятностей состояний имеет вид (*см.* правило составления таких уравнений на с. 343).

$$\begin{cases} \lambda p_0 = \mu p_1, \\ \mu p_1 = \lambda p_0, \end{cases} \qquad (15.18)$$

т.е. система вырождается в одно уравнение. Учитывая нормировочное условие $p_0 + p_1 = 1$, найдем из (15.18) предельные вероятности состояний

$$p_0 = \frac{\mu}{\lambda + \mu}, \; p_1 = \frac{\lambda}{\lambda + \mu}, \qquad (15.19)$$

которые выражают среднее относительное время пребывания системы в состоянии S_0 (когда канал свободен) и S_1 (когда канал

---

[1] Здесь и в дальнейшем предполагается, что все потоки событий, переводящие СМО из состояния в состояние, будут простейшими. К ним относится и поток обслуживаний — поток заявок, обслуживаемых одним непрерывно занятым каналом. Среднее время обслуживания $\bar{t}_{об.}$ обратно по величине интенсивности μ, т.е. $\bar{t}_{об.} = 1/\mu$.

заняты, т.е. определяют соответственно относительную пропускную способность Q системы и вероятность отказа $P_{отк}$:

$$Q=\frac{\mu}{\lambda+\mu}, \tag{15.20}$$

$$P_{отк}=\frac{\lambda}{\lambda+\mu}. \tag{15.21}$$

Абсолютную пропускную способность найдем, умножив относительную пропускную способность Q на интенсивность потока отказов

$$A=\frac{\lambda\mu}{\lambda+\mu}. \tag{15.22}$$

▷ **15.5.** Известно, что заявки на телефонные переговоры в телевизионном ателье поступают с интенсивностью λ, равной 90 заявок в час, а средняя продолжительность разговора по телефону $\bar{t}_{об.}=2$ мин. Определить показатели эффективности работы СМО (телефонной связи) при наличии одного телефонного номера.

Решение. Имеем $\lambda=90$ (1/ч), $\bar{t}_{об.}=2$ мин. Интенсивность потока обслуживаний $\mu=1/\bar{t}_{об.}=1/2=0,5$ (1/мин)=30 (1/ч). По (15.20) относительная пропускная способность СМО $Q=30/(90+30)=0,25$, т.е. в среднем только 25% поступающих заявок осуществят переговоры по телефону. Соответственно вероятность отказа в обслуживании составит $P_{отк.}=0,75$ (*см.* (15.21)). Абсолютная пропускная способность СМО по (15.29) $A=90\cdot0,25=22,5$, т.е. в среднем в час будут обслужены 22,5 заявки на переговоры. Очевидно, что при наличии только одного телефонного номера СМО будет плохо справляться с потоком заявок. ▶

Многоканальная система с отказами. Рассмотрим классическую *задачу Эрланга.*

Имеется n каналов, на которые поступает поток заявок с интенсивностью λ. Поток обслуживаний имеет интенсивность μ. Найти предельные вероятности состояний системы и показатели ее эффективности.

Система S (СМО) имеет следующие состояния (нумеруем их по числу заявок, находящихся в системе): S_0, S_1, S_2, ..., S_k, ... S_n, где S_k — состояние системы, когда в ней находится k заявок, т.е. занято k каналов.

Граф состояний СМО соответствует процессу гибели и размножения и показан на рис. 15.7.

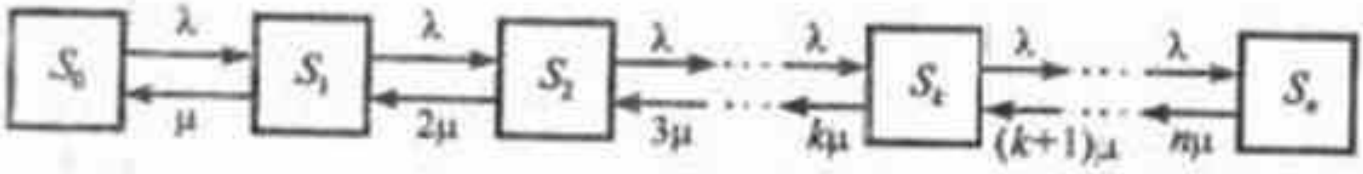

Рис. 15.7

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

В формуле (15.16) для схемы гибели и размножения получим для предельной вероятности состояния

$$p_0=\left(1+\frac{\lambda}{\mu}+\frac{\lambda^2}{2!\mu^2}+\ldots+\frac{\lambda^k}{k!\mu^k}+\ldots+\frac{\lambda^n}{n!\mu^n}\right)^{-1}, \qquad (15.23)$$

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

$$\rho=\frac{\lambda}{\mu} \qquad (15.24)$$

называется *приведенной интенсивностью потока заявок* или *интенсивностью нагрузки канала*. Она выражает среднее число заявок, приходящее за среднее время обслуживания одной заявки. Теперь

$$p_0=\left(1+\rho+\frac{\rho^2}{2!}+\ldots+\frac{\rho^k}{k!}+\ldots+\frac{\rho^n}{n!}\right)^{-1}, \qquad (15.25)$$

$$p_1=\rho p_0,\ p_2=\frac{\rho^2}{2!}p_0,\ \ldots,\ p_k=\frac{\rho^k}{k!}p_0,\ \ldots,\ p_n=\frac{\rho^n}{n!}p_0. \qquad (15.26)$$

Формулы (15.25) и (15.26) для предельных вероятностей получили названия *формул Эрланга*[1] в честь основателя теории массового обслуживания.

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

$$P_{\text{отк}}=\frac{\rho^n}{n!}p_0. \qquad (15.27)$$

Относительная пропускная способность — вероятность того, что заявка будет обслужена:

$$Q=1-P_{\text{отк}}=1-\frac{\rho^n}{n!}p_0. \qquad (15.28)$$

Абсолютная пропускная способность:

$$A=\lambda Q=\lambda\left(1-\frac{\rho^n}{n!}p_0\right). \qquad (15.29)$$

Среднее число занятых каналов $\bar{k}$ есть математическое ожидание числа занятых каналов:

$$\bar{k}=\sum_{k=0}^{n}kp_k,$$

где p_k — предельные вероятности состояний, определяемых по формулам (15.25), (15.26).

Однако среднее число занятых каналов можно найти проще, если учесть, что абсолютная пропускная способность системы A есть не что иное, как интенсивность потока *обслуженных* системой заявок (в единицу времени). Так как каждый занятый канал обслуживает в среднем μ заявок (в единицу времени), то среднее число занятых каналов

$$\bar{k}=\frac{A}{\mu} \qquad (15.30)$$

---

[1] Эрланг А.К. (конец XIX в. — начало XX в.) — датский инженер, математик.

или, учитывая (15.29), (15.24):

$$\bar{k} = \rho\left(1 - \frac{\rho^n}{n!} p_0\right). \qquad \textbf{(15.31)}$$

▷ **15.6.** В условиях задачи **15.5** определить оптимальное число телефонных номеров в телевизионном ателье, если условием оптимальности считать удовлетворение в среднем из каждых 100 заявок не менее 90 заявок на переговоры.

Решение. Интенсивность нагрузки канала по формуле (15.25) $\rho=90/30=3$, т.е. за время среднего (по продолжительности) телефонного разговора $\bar{t}_{об.}=2$ мин. поступает в среднем 3 заявки на переговоры.

Будем постепенно увеличивать число каналов (телефонных номеров) $n=2, 3, 4, \ldots$ и определим по формулам (15.25), (15.28), (15.29) для получаемой n-канальной СМО характеристики обслуживания. Например, при $n = 2$ $p_0 = \left(1+3+3^2/2!\right)^{-1} = 0{,}118 \approx 0{,}12$; $Q = 1 - \left(3^2/2!\right)\cdot 0{,}118 = 0{,}471 \approx 0{,}47$; $A=90\cdot 0{,}471=42{,}4$ и т.д. Значение характеристик СМО сведем в табл. 15.1.

Таблица 15.1

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

По условию оптимальности $Q \geq 0{,}9$, следовательно, в телевизионном ателье необходимо установить 5 телефонных номеров (в этом случае $Q = 0{,}90$ — *см.* табл. 15.1). При этом в час будут обслуживаться в среднем 80 заявок ($A = 80{,}1$), а среднее число занятых телефонных номеров (каналов) по формуле (15.30) $\bar{k} = 80{,}1/30 = 2{,}67$. ▶

▷ **15.7.** В вычислительный центр коллективного пользования с тремя ЭВМ поступают заказы от предприятий на вычислительные работы. Если работают все три ЭВМ, то вновь поступающий заказ

не принимается, и предприятие вынуждено обратиться в другой вычислительный центр. Среднее время работы с одним заказом составляет 3 ч. Интенсивность потока заявок 0,25 (1/ч). Найти предельные вероятности состояний и показатели эффективности работы вычислительного центра.

Решение. По условию $n=3$, $\lambda=0{,}25$ (1/ч), $\bar{t}_{об.}=3$ (ч). Интенсивность потока обслуживаний $\mu=1/\bar{t}_{об.}=1/3=0{,}33$. Интенсивность нагрузки ЭВМ по формуле (15.24) $\rho=0{,}25/0{,}33=0{,}75$. Найдем предельные вероятности состояний:

по формуле (15.25) $p_0=(1+0{,}75+0{,}75^2/2!+0{,}75^3/3!)^{-1}=0{,}476$;

по формуле (15.26) $p_1=0{,}75\cdot0{,}476=0{,}357$; $p_2=(0{,}75^2/2!)\cdot0{,}476=$ $=0{,}134$; $p_3=(0{,}75^3/3!)\cdot0{,}476=0{,}033$, т.е. в стационарном режиме работы вычислительного центра в среднем 47,6% времени нет ни одной заявки, 35,7% — имеется одна заявка (занята одна ЭВМ), 13,4% — две заявки (две ЭВМ), 3,3% времени — три заявки (заняты три ЭВМ).

Вероятность отказа (когда заняты все три ЭВМ), таким образом, $P_{отк.}=p_3=0{,}033$.

По формуле (15.28) относительная пропускная способность центра $Q = 1-0{,}033 = 0{,}967$, т.е. в среднем из каждых 100 заявок вычислительный центр обслуживает 96,7 заявок.

По формуле (15.29) абсолютная пропускная способность центра $A = 0{,}25\cdot0{,}967 = 0{,}242$, т.е. в один час в среднем обслуживается 0,242 заявки.

По формуле (15.30) среднее число занятых ЭВМ $\bar{k} =$ $= 0{,}242/0{,}33 = 0{,}725$, т.е. каждая из трех ЭВМ будет занята обслуживанием заявок в среднем лишь на $72{,}5/3 = 24{,}2\%$.

При оценке эффективности работы вычислительного центра необходимо сопоставить доходы от выполнения заявок с потерями от простоя дорогостоящих ЭВМ (с одной стороны, у нас высокая пропускная способность СМО, а с другой стороны — значительный простой каналов обслуживания) и выбрать компромиссное решение.▶

15.7. СМО с ожиданием (очередью)

В качестве показателей эффективности СМО с ожиданием, кроме уже известных показателей — абсолютной A и относительной Q пропускной способности, вероятности отказа $P_{отк.}$, средне-

го числа занятых каналов $\bar{k}$ (для многоканальной системы) будем рассматривать также следующие: $L_{\text{сист.}}$ — *среднее число заявок в системе*; $T_{\text{сист.}}$ — *среднее время пребывания заявки в системе*; $L_{\text{оч.}}$ — *среднее число заявок в очереди (длина очереди)*; $T_{\text{оч.}}$ — *среднее время пребывания заявки в очереди*; $P_{\text{зан.}}$ — *вероятность того, что канал занят (степень загрузки канала)*.

Одноканальная система с неограниченной очередью. На практике часто встречаются одноканальные СМО с неограниченной очередью (например, телефон-автомат с одной будкой). Рассмотрим задачу.

Имеется одноканальная СМО с очередью, на которую не наложены никакие ограничения (ни по длине очереди, ни по времени ожидания). Поток заявок, поступающих в СМО, имеет интенсивность λ, а поток обслуживаний — интенсивность μ. Необходимо найти предельные вероятности состояний и показатели эффективности СМО.

Система может находиться в одном из состояний $S_0, S_1, S_2, \ldots, S_k$, по числу заявок, находящихся в СМО: S_0 — канал свободен; S_1 — канал занят (обслуживает заявку), очереди нет; S_2 — канал занят, одна заявка стоит в очереди; ... S_k — канал занят, $(k-1)$ заявок стоят в очереди и т.д.

Граф состояний СМО представлен на рис. 15.8.

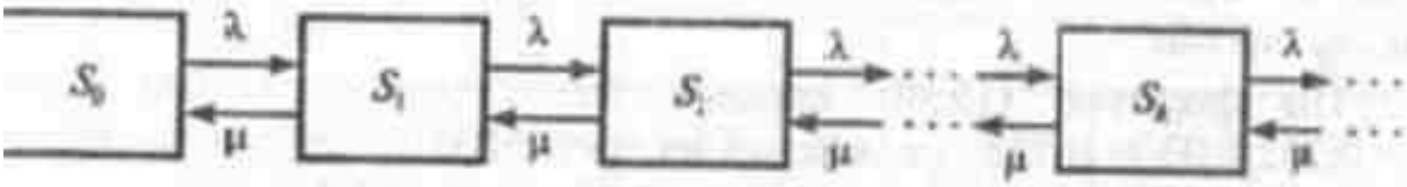

Рис. 15.8

Это процесс гибели и размножения, но с бесконечным числом состояний, в котором интенсивность потока заявок равна λ, а интенсивность потока обслуживаний μ.

Прежде чем записать формулы предельных вероятностей, необходимо быть уверенным в их существовании, ведь в случае, когда время $t \to \infty$, очередь может неограниченно возрастать. Доказано, что *если* $\rho < 1$, *т.е. среднее число приходящих заявок меньше среднего числа обслуженных заявок (в единицу времени), то предельные вероятности существуют. Если* $\rho \geq 1$, *очередь растет до бесконечности.*

Для определения предельных вероятностей состояний воспользуемся формулами (15.16), (15.17) для процесса гибели и размножения (здесь мы допускаем известную нестрогость, так как ранее эти формулы были получены для случая конечного числа состояний системы). Получим:

$$p_0=\left[1+\frac{\lambda}{\mu}+\left(\frac{\lambda}{\mu}\right)^2+\ldots+\left(\frac{\lambda}{\mu}\right)^k+\ldots\right]^{-1}=$$

$$=\left(1+\rho+\rho^2+\ldots+\rho^k+\ldots\right)^{-1}. \qquad \textbf{(15.32)}$$

Так как предельные вероятности существуют лишь при $\rho < 1$, то геометрический ряд со знаменателем $\rho < 1$, записанный в скобках в формуле (15.32), сходится к сумме, равной $\frac{1}{1-\rho}$. Поэтому

$$p_0=1-\rho, \qquad \textbf{(15.33)}$$

и с учетом соотношений (15.17)

$$p_1=\rho p_0,\ p_2=\rho^2 p_0,\ \ldots,\ p_k=\rho^k p_0,\ \ldots$$

найдем предельные вероятности других состояний

$$p_1=\rho(1-\rho),\ p_2=\rho^2(1-\rho),\ \ldots,\ p_k=\rho^k(1-\rho),\ \ldots. \qquad \textbf{(15.34)}$$

Предельные вероятности $p_0, p_1, p_2, \ldots, p_k, \ldots$ образуют убывающую геометрическую прогрессию со знаменателем $\rho < 1$, следовательно, вероятность p_0 — наибольшая. Это означает, что если СМО справляется с потоком заявок (при $\rho < 1$), то наиболее вероятным будет отсутствие заявок в системе.

Среднее число заявок в системе $L_{\text{сист.}}$ определим по формуле математического ожидания, которая с учетом (15.34) примет вид

$$L_{\text{сист.}}=\sum_{k=1}^{\infty} kp_k=(1-\rho)\sum_{k=1}^{\infty} k\rho^k \qquad \textbf{(15.35)}$$

(суммирование от 1 до ∞, так как нулевой член $0p_0=0$).

Можно показать, что формула (15.35) преобразуется (при $\rho < 1$) к виду

$$L_{\text{сист.}} = \frac{\rho}{1-\rho}. \qquad (15.36)$$

Найдем среднее число заявок в очереди $L_{\text{оч}}$. Очевидно, что

$$L_{\text{оч.}} = L_{\text{сист.}} - L_{\text{об.}}, \qquad (15.37)$$

где $L_{\text{об.}}$ — среднее число заявок, находящихся под обслуживанием.

Среднее число заявок под обслуживанием определим по формуле математического ожидания числа заявок под обслуживанием, принимающего значения 0 (если канал свободен) либо 1 (если канал занят):

$$L_{\text{об.}} = 0 \cdot p_0 + 1(1-p_0),$$

т.е. среднее число заявок под обслуживанием равно вероятности того, что канал занят:

$$L_{\text{об.}} = P_{\text{кан.}} = 1 - p_0. \qquad (15.38)$$

В силу (15.33)

$$L_{\text{об.}} = P_{\text{зан.}} = \rho. \qquad (15.39)$$

Теперь по формуле (15.37) с учетом (15.36) и (15.39)

$$L_{\text{оч.}} = \frac{\rho^2}{1-\rho}. \qquad (15.40)$$

Доказано, что *при любом характере потока заявок, при любом распределении времени обслуживания, при любой дисциплине обслуживания среднее время пребывания заявки в системе (очереди) равна среднему числу заявок в системе (в очереди), деленному на интенсивность потока заявок*, т.е.

$$T_{\text{сист.}} = \frac{1}{\lambda} L_{\text{сист.}}, \qquad (15.41)$$

$$T_{\text{оч.}} = \frac{1}{\lambda} L_{\text{оч.}}. \qquad (15.42)$$

Формулы (15.41) и (15.42) называются *формулами Литтла*. Они вытекают из того, что *в предельном, стационарном режиме среднее число заявок, прибывающих в систему, равно среднему числу заявок, покидающих ее*: оба потока заявок имеют одну и ту же интенсивность λ.

На основании формул (15.41) и (15.42) с учетом (15.36) и (15.40) среднее время пребывания заявки в системе определится по формуле:

$$T_{\text{сист.}} = \frac{\rho}{\lambda(1-\rho)}, \qquad (15.43)$$

а среднее время пребывания заявки в очереди —

$$T_{\text{оч.}} = \frac{\rho^2}{\lambda(1-\rho)}. \qquad (15.44)$$

▷ **15.8.** В порту имеется один причал для разгрузки судов. Интенсивность потока судов равна 0,4 (судов в сутки). Среднее время разгрузки одного судна составляет 2 суток. Предполагается, что очередь может быть неограниченной длины. Найти показатели эффективности работы причала, а также вероятность того, что ожидают разгрузки не более чем 2 судна.

Решение. Имеем $\rho = \lambda/\mu = \lambda \bar{t}_{\text{об.}} = 0{,}4 \cdot 2 = 0{,}8$. Так как $\rho = 0{,}8 < 1$, то очередь на разгрузку не может бесконечно возрастать и предельные вероятности существуют. Найдем их.

Вероятность того, что причал свободен, по (15.33) $p_0 = 1 - 0{,}8 = 0{,}2$, а вероятность того, что он занят, $P_{\text{зан.}} = 1 - 0{,}2 = 0{,}8$. По формуле (15.34) вероятности того, что у причала находятся 1, 2, 3 судна (т.е. ожидают разгрузки 0, 1, 2 судна), равны: $p_1 = 0{,}8(1-0{,}8) = 0{,}16$; $p_2 = 0{,}8^2 \cdot (1-0{,}8) = 0{,}128$; $p_3 = 0{,}8^3 \cdot (1-0{,}8) = 0{,}1024$.

Вероятность того, что ожидают разгрузку не более чем 2 судна, равна

$$P = p_1 + p_2 + p_3 = 0{,}16 + 0{,}128 + 0{,}1024 = 0{,}3904.$$

По формуле (15.40) среднее число судов, ожидающих разгрузки,

$$L_{\text{оч.}} = 0{,}8^2/(1-0{,}8) = 3{,}2,$$

а среднее время ожидания разгрузки по формуле (15.42)

$$T_{\text{оч.}} = 3{,}2/0{,}8 = 4 \text{ (сутки)}.$$

По формуле (15.36) среднее число судов, находящихся у причала, $L_{сист.} = 0{,}8/(1-0{,}8) = 4$ (сутки) (или проще по (15.37) $L_{сист.} = 3{,}2+0{,}8 = 4$ (сутки), а среднее время пребывания судна у причала по формуле (15.41) $T_{сист.} = 4/0{,}8 = 5$ (сутки).

Очевидно, что эффективность разгрузки судов невысокая. Для ее повышения необходимо уменьшение среднего времени разгрузки судна $\bar{t}_{об}$ либо увеличение числа причалов n.▶

Многоканальная СМО с неограниченной очередью. Рассмотрим задачу. Имеется n-канальная СМО с неограниченной очередью. Поток заявок, поступающих в СМО, имеет интенсивность λ, а поток обслуживаний — интенсивность μ. Необходимо найти предельные вероятности состояний СМО и показатели ее эффективности.

Система может находиться в одном из состояний S_0, S_1, S_2, ..., S_k, ..., S_n, ..., нумеруемых по числу заявок, находящихся в СМО: S_0 — в системе нет заявок (все каналы свободны); S_1 — занят один канал, остальные свободны; S_2 — заняты два канала, остальные свободны; ..., S_k — занято k каналов, остальные свободны; ..., S_n — заняты все n каналов (очереди нет); S_{n+1} — заняты все n каналов, в очереди одна заявка; ..., S_{n+r} — заняты все n каналов, r заявок стоит в очереди, ...

Граф состояний системы показан на рис. 15.9. Обратим внимание на то, что в отличие от предыдущей СМО, интенсивность потока обслуживаний (переводящего систему из одного состояния в другое справа налево) не остается постоянной, а по мере увеличения числа заявок в СМО от 0 до n увеличивается от величины μ до $n\mu$, так как соответственно увеличивается число каналов обслуживания. При числе заявок в СМО большем, чем n, интенсивность потока обслуживаний сохраняется равной $n\mu$.

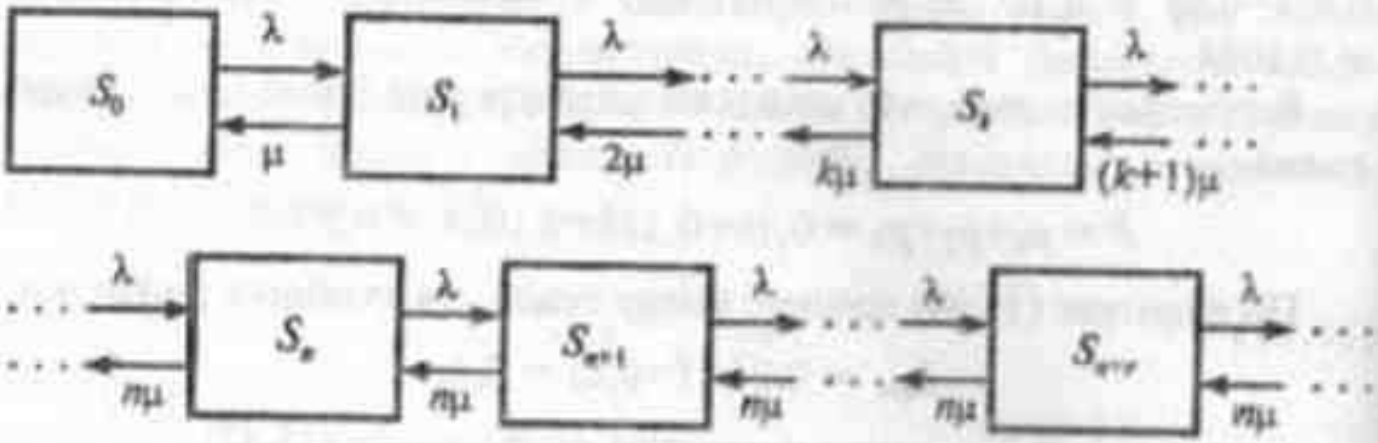

Рис. 15.9

Можно показать, что при $\rho/n < 1$ предельные вероятности существуют. Если $\rho/n \geq 1$, очередь растет до бесконечности. Используя формулы (15.16) и (15.17) для процесса гибели и размножения, можно получить следующие формулы для предельных вероятностей состояний n-канальной СМО с неограниченной очередью

$$p_0 = \left(1 + \frac{\rho}{1!} + \frac{\rho^2}{2!} + \ldots + \frac{\rho^n}{n!} + \frac{\rho^{n+1}}{n!(n-\rho)}\right)^{-1}, \qquad (15.45)$$

$$p_1 = \frac{\rho}{1!} p_0, \ldots, p_k = \frac{\rho^k}{k!} p_0, \ldots, p_n = \frac{\rho^n}{n!} p_0, \qquad (15.46)$$

$$p_{n+1} = \frac{\rho^{n+1}}{n \cdot n!} p_0, \ldots, p_{n+r} = \frac{\rho^{n+r}}{n^r \cdot n!} p_0, \ldots . \qquad (15.47)$$

Вероятность того, что заявка окажется в очереди,

$$P_{\text{оч.}} = \frac{\rho^{n+1}}{n!(n-\rho)} p_0 . \qquad (15.48)$$

Для n-канальной СМО с неограниченной очередью, используя прежние приемы, можно найти:

среднее число занятых каналов

$$\bar{k} = \frac{\lambda}{\mu} = \rho, \qquad (15.49)$$

среднее число заявок в очереди

$$L_{\text{оч.}} = \frac{\rho^{n+1} p_0}{n \cdot n! \left(1 - \frac{\rho}{n}\right)^2}, \qquad (15.50)$$

среднее число заявок в системе

$$L_{\text{сист.}} = L_{\text{оч.}} + \rho. \qquad (15.51)$$

Среднее время пребывания заявки в очереди и среднее время пребывания заявки в системе, как и ранее, находятся по формулам Литтла (15.42) и (15.41).

Замечание. Для СМО с неограниченной очередью при $\rho < 1$ любая заявка, пришедшая в систему, будет обслужена, т.е. ве-

роятность отказа $P_{отк.} = 0$, относительная пропускная способность $Q = 1$, а абсолютная пропускная способность равна интенсивности входящего потока заявок, т.е. $A = \lambda$.

▷ **15.9.** В универсаме к узлу расчета поступает поток покупателей с интенсивностью $\lambda = 81$ чел. в час. Средняя продолжительность обслуживания контролером-кассиром одного покупателя $\bar{t}_{об.} = 2$ мин. Определить:

а. Минимальное количество контролеров-кассиров n_{min}, при котором очередь не будет расти до бесконечности, и соответствующие характеристики обслуживания при $n = n_{min}$.

б. Оптимальное количество $n_{опт.}$ контролеров-кассиров, при котором относительная величина затрат $C_{отн.}$, связанная с издержками на содержание каналов обслуживания и с пребыванием в очереди покупателей, задаваемая, например, как $C_{отн.} = \frac{1}{\lambda} n + 3T_{оч.}$, будет минимальна, и сравнить характеристики обслуживания при $n = n_{min}$ и $n = n_{опт.}$.

в. Вероятность того, что в очереди будет не более трех покупателей.

Решение. **а.** По условию $\lambda = 81(1/ч) = 81/60 = 1{,}35$ (1/мин.). По формуле (15.24) $\rho = \lambda/\mu = \lambda \bar{t}_{об.} = 1{,}35 \cdot 2 = 2{,}7$. Очередь не будет возрастать до бесконечности при условии $\rho/n < 1$, т.е. при $n > \rho = 2{,}7$. Таким образом, минимальное количество контролеров-кассиров $n_{min} = 3$.

Найдем характеристики обслуживания СМО при $n = 3$.

Вероятность того, что в узле расчета отсутствуют покупатели, по формуле (15.45) $p_0 = (1+2{,}7+2{,}7^2/2!+2{,}7^3/3!+2{,}7^4/3!(3-2{,}7))^{-1} = 0{,}025$, т.е. в среднем 2,5% времени контролеры-кассиры будут простаивать.

Вероятность того, что в узле расчета будет очередь, по (15.48)

$$P_{оч.} = (2{,}7^4/3!(3-2{,}7))0{,}025 = 0{,}735.$$

Среднее число покупателей, находящихся в очереди, по (15.50)

$$L_{оч.} = (2{,}7^4/3 \cdot 3!(1-2{,}7/3)^2)0{,}025 = 7{,}35.$$

Среднее время ожидания в очереди по (15.42)

$$T_{оч.} = 7{,}35/1{,}35 = 5{,}44 \text{ (мин)}.$$

Среднее число покупателей в узле расчета по (15.51)

$$L_{\text{сист.}} = 7{,}35+2{,}7 = 10{,}05.$$

Среднее время нахождения покупателей в узле расчета по (15.41)

$$T_{\text{сист.}} = 10{,}05/1{,}35 = 7{,}44 \text{ (мин)}.$$

Среднее число контролеров-кассиров, занятых обслуживанием покупателей, по (15.49) $\bar{k} = 2{,}7$.

Коэффициент (доля) занятых обслуживанием контролеров-кассиров

$$\bar{k} = \rho/n = 2{,}7/3 = 0{,}9.$$

Абсолютная пропускная способность узла расчета $A = 1{,}35$ (1/мин), или 81 (1/ч), т.е. 81 покупатель в час.

Анализ характеристик обслуживания свидетельствует о значительной перегрузке узла расчета при наличии трех контролеров-кассиров.

б. Относительная величина затрат при $n = 3$

$$C_{\text{отн.}} = \frac{1}{\lambda} n + 3T_{\text{оч.}} = 3/1{,}35+3\cdot 5{,}44 = 18{,}54.$$

Рассчитаем относительную величину затрат при других значениях n (табл. 15.2).

Таблица 15.2

Характеристика обслуживания	Число контролеров-кассиров				
	3	4	5	6	7
Вероятность простоя контролеров-кассиров p_0	0,025	0,057	0,065	0,067	0,067
Среднее число покупателей в очереди $T_{\text{оч.}}$	5,44	0,60	0,15	0,03	0,01
Относительная величина затрат $C_{\text{отн.}}$	18,54	4,77	4,14	4,53	5,22

Как видно из табл. 15.2, минимальные затраты получены при $n = n_{\text{опт.}} = 5$ контролерах-кассирах.

Определим характеристики обслуживания узла расчета при $n = n_{опт.} = 5$. Получим $P_{оч.} = 0{,}091$; $L_{оч.} = 0{,}198$; $T_{оч.} = 0{,}146$ (мин); $L_{сист.} = 2{,}90$; $T_{сист.} = 2{,}15$ (мин); $\bar{k} = 2{,}7$; $k_3 = 0{,}54$.

Как видим, при $n = 5$ по сравнению с $n = 3$ существенно уменьшились вероятность возникновения очереди $P_{оч.}$, длина очереди $L_{оч.}$ и среднее время пребывания в очереди $T_{оч.}$ и соответственно среднее число покупателей $L_{сист.}$ и среднее время нахождения в узле расчета $T_{сист.}$, а также доля занятых обслуживанием контролеров k_3. Но среднее число занятых обслуживанием контролеров-кассиров $\bar{k}$ и абсолютная пропускная способность узла расчета A естественно не изменились.

в. Вероятность того, что в очереди будет не более 3 покупателей, определится как

$$P(r \le 3) = \underbrace{p_1+p_2+p_3+p_4+p_5}_{\text{(когда заняты от 1 до 5 контролеров-кассиров)}} + \underbrace{p_{5+1}+p_{5+2}+p_{5+3}}_{\text{(когда в очереди стоят от 1 до 3 покупателей)}} =$$

$= 1 - P_{оч.} + p_{5+1} + p_{5+2} + p_{5+3}$, где каждое слагаемое найдем по формулам (15.45)—(15.48). Получим при $n = 5$:

$$P(r \le 3) = 1 - \frac{2{,}7^6}{5!(5-2{,}3)}0{,}065 + \frac{2{,}7^6}{5 \cdot 5!}0{,}065 +$$

$$+ \frac{2{,}7^7}{5^2 \cdot 5!}0{,}065 + \frac{2{,}7^8}{5^3 \cdot 5!}0{,}065 = 0{,}986.$$

(Заметим, что в случае $n=3$ контролеров-кассиров та же вероятность существенно меньше: $P(r \le 3)=0{,}464$).▶

15.10. Железнодорожная касса с двумя окошками продает билеты в два пункта A и B. Интенсивность потока пассажиров, желающих купить билеты, для обоих пунктов одинакова: $\lambda_A = \lambda_B = 0{,}45$ (пассажиров в минуту). На обслуживание пассажиров кассир тратит в среднем 2 мин. Рассматриваются два варианта продажи билетов: первый — билеты продаются в одной кассе с двумя окошками одновременно в оба пункта A и B; второй — билеты продаются в двух специализированных кассах (по одному окошку в каждой), одна только в пункт A, другая — только в пункт B. Необходимо:

а. Сравнить два варианта продажи билетов по основным характеристикам обслуживания.

б. Определить, как надо изменить среднее время обслуживания одного пассажира, чтобы по второму варианту продажи пассажиры затрачивали на приобретение билетов в среднем меньше времени, чем по первому варианту.

Решение. **а. По первому варианту** имеем двухканальную СМО, на которую поступает поток заявок интенсивностью $\lambda = 0{,}45+0{,}45 = 0{,}9$; интенсивность потока обслуживаний $\mu = 1/2 = 0{,}5$; $\rho = \lambda/\mu = 1{,}8$. Так как $\rho/n = 1{,}8/2 = 0{,}9 < 1$, то предельные вероятности существуют.

Вероятность простоя двух кассиров по (15.45)

$$p_0 = \left(1 + \frac{1{,}8}{1!} + \frac{1{,}8^2}{2!} + \frac{1{,}8^3}{2!(2-1{,}8)}\right)^{-1} = 0{,}0526.$$

Среднее число пассажиров в очереди по (15.50)

$$L_{\text{оч.}} = 1{,}8^3/2\cdot 2!(1-1{,}8/2)\cdot 0{,}0526 = 7{,}67.$$

Среднее число пассажиров у кассы по (15.51)

$$L_{\text{сист.}} = 7{,}67+1{,}8 = 9{,}47.$$

Среднее время на ожидание в очереди и покупку билетов равно соответственно (по формулам (15.42) и (15.41)): $T_{\text{оч.}} = 7{,}67/0{,}9 = 8{,}52$ (мин) и $T_{\text{сист.}} = 9{,}47/0{,}9 = 10{,}5$ (мин).

По второму варианту имеем две одноканальные СМО (два специализированных окошка); на каждую поступает поток заявок с интенсивностью $\lambda = 0{,}45$. По-прежнему $\mu = 0{,}5$; $\rho = \lambda/\mu = 0{,}9 < 1$, предельные вероятности существуют. По формулам (15.40), (15.36), (15.42), (15.41)

$$L_{\text{оч.}} = 0{,}9^2/(1-0{,}9) = 8{,}1;\ L_{\text{сист.}} = 0{,}9/(1-0{,}9) = 9{,}0;$$

$$T_{\text{оч.}} = 8{,}1/0{,}45 = 18{,}0\ (\text{мин}),\ T_{\text{сист.}} = 9{,}0/0{,}45 = 20{,}0\ (\text{мин}).$$

Итак, по второму варианту увеличились и длина очереди, и среднее время ожидания в ней и в целом на покупку билетов. Такое различие объясняется тем, что в первом варианте (двухканальная СМО) меньше средняя доля времени, которую простаивает каждый из двух кассиров: если он не занят обслуживанием пассажира, покупающего билет в пункт *A*, он может заняться обслуживанием пассажира, покупающего билет в пункт *B*, и наоборот. Во втором варианте такой взаимозаменяемости нет.

Можно заметить, что среднее время на покупку билетов по второму варианту увеличилось более чем в 2 раза. Такое значительное увеличение связано с тем, что СМО работает на пределе своих возможностей ($\rho = 0{,}9$): достаточно незначительно увеличить среднее время обслуживания $\bar{t}_{об.}$, т.е. уменьшить μ, и ρ превзойдет 1, т.е. очередь начнет неограниченно возрастать.

б. Выше было получено, что по первому варианту продажи билетов при среднем времени обслуживания одного пассажира $\bar{t}_{об.} = 2$ (мин) среднее время на покупку билетов составит $T_{сист._1} = 10{,}5$ (мин). По условию для второго варианта продажи $T_{сист._2} < T_{сист._1}$ или с учетом (15.36) и (15.41): $\frac{1}{\lambda}\frac{\rho}{1-\rho} < T_{сист._1}$.

Полагая $\rho = \lambda/\mu = \lambda\bar{t}_{об.}$, получим $-\frac{\bar{t}_{об.}}{1-\lambda\bar{t}_{об.}} < T_{сист._1}$, откуда найдем $\bar{t}_{об.} < \frac{T_{сист._1}}{1+\lambda T_{сист._1}}$ или $\bar{t}_{об} < \frac{10{,}5}{1+0{,}45\cdot 10{,}5} = 1{,}83$ (мин).

Итак, средние затраты времени на покупку билетов по второму варианту продажи уменьшатся, если среднее время обслуживания одного пассажира уменьшится более чем на 0,17 мин, или более чем на 8,5%.▶

СМО с ограниченной очередью. СМО с ограниченной очередью отличаются от рассмотренных выше задач лишь тем, что число заявок в очереди ограничено (не может превосходить некоторого заданного m). Если новая заявка поступает в момент, когда все места в очереди заняты, она покидает СМО необслуженной, т.е. получает отказ.

Очевидно: для вычисления предельных вероятностей состояний и показателей эффективности таких СМО может быть использован тот же подход, что и выше, с той разницей, что суммировать надо не бесконечную прогрессию (как, например, мы делали при выводе формулы (15.33)), а конечную. Соответствующие формулы сведем в табл. 15.3.

Среднее время пребывания заявки в очереди и в системе, как и ранее, определяем по формулам Литтла (15.44) и (15.43).

▶ **15.11.** По условию задачи **15.8** найти показатели эффективности работы причала. Известно, что приходящее судно покидает причал (без разгрузки), если в очереди на разгрузку стоит более 3 судов.

Решение. По условию $m = 3$. Используем формулы, приведенные во второй графе табл. 15.3.

Вероятность того, что причал свободен:

$$p_0 = \frac{1-0{,}6}{1-0{,}8^{3+2}} = 0{,}297.$$

Вероятность того, что приходящее судно покинет причал без разгрузки:

$$P_{отк.} = 0{,}8^{3+1} \cdot 0{,}297 = 0{,}122.$$

Относительная пропускная способность причала:

$$Q = 1 - 0{,}122 = 0{,}878.$$

Абсолютная пропускная способность причала $A = 0{,}4 \cdot 0{,}878 = 0{,}351$, т.е. в среднем в сутки разгружается 0,35 судна.

Среднее число судов, ожидающих разгрузку

$$L_{оч.} = \frac{0{,}8^2\left[1-0{,}8^3(3+1-3\cdot 0{,}8)\right]}{(1-0{,}8^{3+2})(1-0{,}8)} = 0{,}861,$$

а среднее время ожидания разгрузки по (15.42)

$$T_{оч.} = \frac{0{,}861}{0{,}8} = 1{,}076 \text{ (сутки)}.$$

Среднее число судов, находящихся у причала

$$L_{сист.} = 0{,}861 + (1 - 0{,}297) = 1{,}564,$$

а среднее время пребывания судна у причала по (15.41):

$$T_{сист.} = \frac{1{,}564}{0{,}8} = 1{,}955 \text{ (сутки)}.\blacktriangleright$$

СМО с ограниченным временем ожидания. На практике часто встречаются СМО с так называемыми "нетерпеливыми" заявками. Такие заявки могут уйти из очереди, если время ожидания превышает некоторую величину. В частности, такого рода заявки возникают в различных технологических системах, в которых задержка с началом обслуживания может привести к потере качества продукции, в системах оперативного управления, когда срочные сообщения теряют ценность (или даже смысл), если они не поступают на обслуживание в течение определенного времени.

Таблица 15.3

Показатели	Одноканальная СМО с ограниченной очередью	Многоканальная СМО с ограниченной очередью
Предельные вероятности	$p_0 = \frac{1-\rho}{1-\rho^{m+2}}$ $p_1 = \rho p_0,\ p_2 = \rho^2 p_0,\ \ldots,\ p_k = \rho^k p_0$	$p_0 = \left[1 + \frac{\rho}{1!} + \ldots + \frac{\rho^n}{n!} + \ldots + \frac{\rho^{n+1}(1-(\rho/n)^m)}{n \cdot n!(1-\rho/n)}\right]^{-1}$ $p_1 = \frac{\rho}{1!} p_0, \ldots, p_k = \frac{\rho^k}{k!} p_0, \ldots, p_n = \frac{\rho^n}{n!} p_0,$ $p_{n+1} = \frac{\rho^{n+1}}{n \cdot n!} p_0, \ldots, p_{n+r} = \frac{\rho^{n+r}}{n^r \cdot n!} p_0\ (r = 1, \ldots, m)$
Вероятность отказа	$P_{\text{отк.}} = P_{m+1} = \rho^{m+1} p_0$	$P_{\text{отк.}} = P_{n+m} = \frac{\rho^{n+m}}{n^m \cdot n!} p_0$
Абсолютная пропускная способность	$A = \lambda Q = \lambda(1-\rho^{m+1} p_0)$	$A = \lambda Q = \lambda(1 - \frac{\rho^{n+m}}{n^m \cdot n!} p_0)$

Продолжение

Относительная пропускная способность	$Q = 1 - P_{отк.} = 1 - \rho^{m+1} p_0$	$Q = 1 - P_{отк.} = 1 - \frac{\rho^{n+m}}{n^m \cdot n!} p_0$
Среднее число заявок в очереди	$L_{оч.} = \rho^2 \frac{\left[1 - \rho^m (m + 1 - m\rho)\right]}{(1 - \rho^{m+2})(1 = \rho)}$	$L_{оч.} = \frac{\rho^{n+1} p_0 \left[1 - \left(m + 1 - m\frac{\rho}{n}\right)\left(\frac{\rho}{n}\right)^m\right]}{n \cdot n! \left(1 - \frac{\rho}{n}\right)^2}$
Среднее число заявок под обслуживанием (среднее число занятых каналов)	$L_{об.} = 1 - p_0$	$\bar{k} = \rho\left(1 - \frac{\rho^{n+m}}{n^m \cdot n!} p_0\right)$
Среднее число заявок в системе	$L_{сист.} = L_{оч.} + L_{об.}$	$L_{сист.} = L_{оч.} + \bar{k}$

В простейших математических моделях таких систем предполагается, что заявка может находиться в очереди случайное время, распределенное по показательному закону с некоторым параметром v, т.е. можно условно считать, что каждая заявка, стоящая в очереди на обслуживание, может покинуть систему с интенсивностью v.

Соответствующие показатели эффективности СМО с ограниченным временем получаются на базе результатов, полученных для процесса гибели и размножения.

В заключение отметим, что на практике часто встречаются *замкнутые* системы обслуживания, у которых входящий поток заявок существенным образом зависит от состояния самой СМО. В качестве примера можно привести ситуацию, когда на ремонтную базу поступают с мест эксплуатации некоторые машины: понятно, что чем больше машин находится в состоянии ремонта, тем меньше их продолжает эксплуатироваться и тем меньше интенсивность потока вновь поступающих на ремонт машин. Для замкнутых СМО характерным является ограниченное число источников заявок, причем каждый источник "блокируется" на время обслуживания его заявки (т.е. он не выдает новых заявок). В подобных системах при конечном числе состояний СМО предельные вероятности будут существовать при любых значениях интенсивностей потоков заявок и обслуживаний. Они могут быть вычислены, если вновь обратиться к процессу гибели и размножения.

15.8. Понятие о статистическом моделировании СМО (методе Монте-Карло)

Основное допущение, при котором анализировались рассмотренные выше СМО, состоит в том, что все потоки событий, переводящие их из состояния в состояние, были простейшими. При нарушении этого требования общих аналитических методов для таких систем не существует. Имеются лишь отдельные результаты, позволяющие выразить в аналитическом виде характеристики СМО через параметры задачи.

В случаях, когда для анализа работы СМО аналитические методы не применимы (или же требуется проверить их точность), используют универсальный *метод статистического моделирования*, или, как его называют, *метод Монте-Карло*.

Идея метода Монте-Карло состоит в том, что вместо аналитического описания СМО производится "розыгрыш" случайного процесса, проходящего в СМО, с помощью специально организованной процедуры. В результате такого "розыгрыша" получается каждый раз новая, отличная от других реализация случайного процесса. Это множество реализаций можно использовать как некий искусственно полученный статистический материал, который обрабатывается обычными методами математической статистики. После такой обработки могут быть получены приближенно любые характеристики обслуживания.

Например, необходимо проанализировать очереди, возникающие в магазине, для решения вопроса о расширении магазина. Время подхода покупателей и время их обслуживания носят случайный характер, и их распределения могут быть установлены по имеющейся информации. В результате взаимодействия этих случайных процессов создается очередь.

Согласно методу Монте-Карло перебирают (с помощью ЭВМ) все возможные состояния системы с различным числом покупателей в час, временем их обслуживания и т.п., сохраняя те же характеристики распределения. В результате многократного искусственного воссоздания работы магазина рассчитывают характеристики обслуживания, как если бы они были получены при наблюдении над реальным потоком покупателей.

При моделировании случайных явлений методом Монте-Карло мы пользуемся самой случайностью как аппаратом исследования. Заметим, что для сложных систем обслуживания с немарковским случайным процессом метод статистического моделирования, как правило, оказывается проще аналитического.