МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

С. М. НАМЕСТНИКОВ, М. Н. СЛУЖИВЫЙ, Ю. Д. УКРАИНЦЕВ

ОСНОВЫ ТЕОРИИ ТЕЛЕТРАФИКА

Ульяновск
2016

УДК 621.391 (075)
ББК 32я7
Н 24

Рецензенты:
доктор технических наук, заведующий кафедрой «Информационные технологии и сети» Ульяновского государственного университета, профессор Смагин А. А.
генеральный директор ООО «АйПиТелеком» Половов А. А.

Наместников, С. М.
Н 24 Основы теории телетрафика : учебное пособие / С. М. Наместников, М. Н. Служивый, Ю. Д. Украинцев. – Ульяновск : УлГТУ, 2016. – 154 с.

ISBN 978-5-9795-1000-0

Рассмотрены основы теории систем массового обслуживания в применении к анализу трафика в современных телекоммуникационных системах, а также основные сведения о математических моделях потоков заявок и систем обработки пакетов в сетях связи. Описаны методы решения задач оценки рабочих характеристик и параметров качества обслуживания в сетях связи.

Для студентов направления «Инфокоммуникационные технологии и системы связи», а также аспирантов, специализирующихся в области анализа и моделирования инфокоммуникационных систем.

УДК 621.391 (075)
ББК 32я7

ISBN 978-5-9795-1000-0

ВВЕДЕНИЕ

В течение последних двух десятилетий значительно ускорилось развитие телекоммуникационных систем (ТС). При этом одной из важнейших задач при проектировании ТС является задача обеспечения требуемого качества (уровня) обслуживания, определяемого посредством совокупности таких параметров, как средняя задержка передачи сообщения, вероятность потери пакета и др., или задача обеспечения минимума потерь при заданном объеме телекоммуникационного оборудования, необходимого для построения ТС. Подобные задачи могут быть решены с помощью методов теории телетрафика.

Основным математическим аппаратом, положенным в основу теории телетрафика, является *теория массового обслуживания*, а также ряд других разделов теории вероятностей. Поэтому *теория телетрафика* может быть определена как приложение теории вероятностей к решению проблем планирования и оценки рабочих характеристик, качества работы и обслуживания ТС [8].

Когда известны будущий трафик и емкость системных элементов, практическая работа, основанная на теории телетрафика, состоит в том, чтобы проектировать ТС наиболее рентабельно при заранее заданном уровне обслуживания. Теория телетрафика позволяет определить потребности (например, на основе объемов трафика) и вычислить емкость системы и спецификации количественных характеристик для обеспечения заданного уровня обслуживания. Приложение теории телетрафика к проектированию новых систем позволяет сравнить различные решения на основе соответствующих моделей и избежать неоптимальных решений при построении сетей.

Основоположником теории телетрафика считается датский ученый А.К. Эрланг, в работах которого (1908–1918 гг.) были получены выражения для вычисления вероятностей потерь для некоторых схем коммутации, в которых поступающий поток заявок рассматривался как случайный процесс. Поэтому *предметом теории телетрафика* являются процессы обслуживания системами распределения информации поступающих потоков сообщений и их численные характеристики.

Понятие «поток сообщений» включает в себя информацию о параметрах и свойствах потока вызовов, а также информацию о виде передаваемых сообщений и форме их представления. При математической записи потока сообщений важной характеристикой является

время между соседними вызовами, которое может быть фиксированным или случайным. В первом случае оно задается длительностью, а во втором – вероятностным законом распределения.

В теории телетрафика все рассматриваемые объекты могут быть объединены под общим названием *системы распределения информации*, включающей в себя коммутационные устройства, маршрутизаторы и линии связи, которые обслуживают по определенному алгоритму пакеты данных, телефонные или другие сообщения. Система распределения информации характеризуется в свою очередь структурой построения: полнодоступная или неполнодоступная, однозвенная или многозвенная и т. п., а также набором структурных параметров: число входов и выходов, число промежуточных линий, коммутаторов и т. д.

В первой главе учебного пособия представлены основные определения и понятия теории телетрафика, изложены методологические основы теории систем массового обслуживания. В главе также представлены основы теории непрерывных и дискретных марковских процессов, где введено понятие о процессах «размножения-гибели», а также об уравнениях равновесия (баланса).

Во второй главе описаны математические вероятностные модели потоков заявок в виде законов распределения параметров потоков. Анализируемые системы и случайные потоки в них представлены в виде марковских моделей. Подробно проанализированы характеристики простейшего потока. Дано представление о нестационарном пуассоновском потоке и потоках с ограниченным последействием (потоках Пальма) на примере потоков Эрланга. Определено понятие примитивного потока заявок на примере нагрузки на телефонный коммутатор. Рассмотрены самоподобные (фрактальные) модели трафика, такие как фрактальное броуновское движение и фрактальный гауссовский шум. Описан R/S-метод анализа характеристик самоподобных случайных процессов.

Третья глава посвящена подробному анализу процесса обслуживания простейшего потока системами с отказами и ожиданием. Значительное внимание уделяется выводу уравнений Эрланга, характеризующих параметры обслуживания простейшего потока заявок, а также формул Эрланга для установившегося режима обслуживания, из которых получены зависимости вероятностей занятости заданного количества каналов от параметров потока и системы обслуживания.

В четвертой главе проанализированы распределения Эрланга, Энгсета, Бернулли и Пуассона, на основе чего получены зависимости

вероятностей занятости заданного количества каналов от параметров потока и системы обслуживания. Представлен подробный анализ распределения Энгсета и характеристики системы обслуживания примитивного потока. Описано распределение вероятностей занятия фиксированных выходов в системе.

В пятой главе представлена классификация Кендалла и анализ основных характеристик различных систем массового обслуживания (СМО) с использованием временных диаграмм. Подробно рассмотрены СМО с приоритетным обслуживанием. Введено понятие о законах сохранения в СМО. Представлен подробный анализ функционирования СМО с несколькими серверами, а также СМО с ограниченной очередью. Описаны вероятностные характеристики времени ожидания заявки в очереди. Рассмотрена СМО с пуассоновским потоком на входе и произвольным распределением времени обслуживания, в процессе анализа которой получена формула Поллачека-Хинчина. Описаны СМО с самоподобным входным потоком и детерминированным временем обслуживания, а также СМО с самоподобным характером времени обслуживания. Даны практические примеры анализа трафика в элементах сетей связи.

В шестой главе рассмотрены основные методы моделирования потоков в сетях, таких как пуассоновский поток, а также самоподобные потоки с распределениями Вейбулла, Парето и логнормальным. Описано моделирование фрактального броуновского движения посредством RMD-метода. Дано краткое описание моделирования СМО посредством сетей Петри. Представлены наиболее важные параметры трафика, которые обычно измеряются в сетях. Кратко описаны основные средства для измерения трафика в ТС.

Седьмая глава посвящена методам анализа сетей массового обслуживания на примере марковских сетей без потерь. Рассмотрены графовые модели сетей с блокировками в виде параллельно-последовательных схем. Уделено значительное внимание анализу и оптимизации коммутационных схем с описанием комбинаторного метода Якобеуса на примере двухзвенной коммутационной системы. Проведен анализ потерь в двухзвенных схемах без сжатия и расширения, а также при наличии сжатия и расширения. Дано понятие о многозвенных коммутационных схемах на примере трехзвенной коммутационной схемы. Методы управления доступом к среде в системах радиодоступа представлены протоколами «классическая ALOHA» и «тактированная ALOHA».

1. ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ТЕЛЕТРАФИКА

1.1. Основные термины и определения в теории телетрафика

Теория телетрафика представляет собой научную дисциплину о закономерностях и количественном описании процессов движения сообщений (информационных потоков) в телекоммуникационных сетях и системах (ТС).

Методы теории телетрафика позволяют рассчитывать характеристики качества обслуживания (QoS – Quality of Service) в ТС, управлять основными параметрами QoS реальных сетей и измерять их с целью оптимизации проектирования новых ТС.

Вопросы построения сетей с гарантированным QoS являются предметом внимания Международного союза электросвязи (МСЭ) (International Telecommunication Union, ITU) и представлены в рекомендациях ITU (E.600), основанных на методах теории телетрафика (Traffic Engineering) [34].

При построении ТС необходим поиск компромисса между обеспечением заданного уровня QoS и технико-экономической эффективностью. Основными количественными характеристиками QoS являются среднее время задержки в предоставлении услуги, а также вероятность отказа в услуге [36].

Методы теории телетрафика основаны на принципах теории систем массового обслуживания (СМО). При этом коммутируемая ТС может рассматриваться как СМО, в которой происходит передача и обработка **сообщений** – одномерных информационных элементов с выделенным началом и концом. Появление в системе нового сообщения назовем **заявкой (требованием)** на его передачу или обработку, которые занимают некоторое конечное **время обслуживания**. Часть системы, осуществляющая передачу или обработку не более одного сообщения в заданный момент времени, называется **сервером**. Система функционирует во времени, т. е. преобразует входной поток заявок в поток обслуженных заявок и, таким образом, относится к классу динамических систем, которые могут быть описаны с помощью концепции **состояния**. Состояние динамической системы в данный момент времени включает в себя число заявок, находящихся в системе в данный момент времени, а также интервал времени, прошедшего с момента поступления заявок на обработку в каждый сервер.

Заявка, поступившая на такую систему, занятую в течение интервала времени обслуживания, получает **отказ** в обслуживании (ресурсный конфликт) – сервер не успевает обслужить заявки за время между их поступлениями. В системах со многими серверами поступающие заявки распределяются для обслуживания между всеми серверами в соответствии с некоторым алгоритмом. Если все серверы окажутся занятыми в данный момент, также возникает ресурсный конфликт. Поступившая заявка будет отброшена (система заблокирована), что определяется **вероятностью блокировки (отказа)**.

Для снижения или исключения потерь заявок в системе должен быть предусмотрен специальный **буфер** памяти, содержащий заявки, которые не могут быть обслужены немедленно при поступлении из-за занятости всех серверов. В системе организуется **очередь** заявок и рассматривается **система с очередью**. В очереди может оказаться несколько заявок, если количество поступающих заявок за некоторый интервал времени превысит число освободившихся за это время серверов. Если очередь не будет бесконечно нарастать, все заявки рано или поздно будут обслужены, однако время их пребывания в очереди будет случайной величиной, вероятностное распределение которой является важной характеристикой СМО. Для оценки **QoS** часто используется **среднее время ожидания обслуживания**. Недостаточность ресурсов в ТС может приводить либо к потере заявок, либо к задержке их обслуживания.

Трафик представляет собой поток информации, передаваемый по сети, линии или каналу связи. Кроме полезной абонентской нагрузки в трафик входит поток служебной (вспомогательной и управляющей) информации.

Связь (коммуникация) – это процесс передачи информации в соответствии с некоторыми правилами. **Соединение** – это некоторая ассоциация двух или более устройств внутри сети или посредством сети для осуществления связи между ними. **Пользователь** (абонент, терминал) – это общий термин для всех внешних по отношению к сети объектов, использующих соединения через сеть для коммуникации. Под **ресурсом** обычно понимается сервер или набор серверов.

1.2. Понятие о системах массового обслуживания

Во многих системах обслуживания потоков **заявок** (требований на обслуживание или вызовов), таких как конвейеры на фабрике, оче-

реди в кассах, движение автомобилей по мосту или шоссе и т. п., имеют место конфликты обслуживания, в виде задержек реагирования или отказов в обслуживании, зависящих от интенсивности прихода заявок на предоставление услуги с учетом ограниченности ресурсов исполнения. Такие системы могут быть отнесены к общему классу динамических (функционирующих во времени) систем. Наука, занимающаяся исследованием подобных систем, получила название **теории массового обслуживания**, а в зарубежной литературе – **теории очередей** (Queuing Theory) [6, 10, 19, 20].

Каждая СМО состоит из некоторого количества обслуживающих единиц – «каналов» обслуживания (ресурсов - в частности, серверов) [6, 10, 19, 20]. Работа любой СМО состоит в обслуживании поступающего на нее потока **заявок**. В общем случае заявки поступают в случайные моменты времени. Обслуживание поступившей заявки продолжается некоторое время, после чего канал освобождается и снова готов к приему следующей заявки.

Предметом теории массового обслуживания является установление зависимости между характером потока заявок, производительностью отдельного канала, числом каналов и параметрами качества обслуживания.

СМО делятся на два основных типа:

1) СМО **с отказами**, где заявка, поступившая в момент, когда все каналы обслуживания заняты, немедленно получает отказ, покидает систему и в дальнейшем процессе обслуживания не участвует;

2) СМО **с ожиданием (очередью)**, где заявка, заставшая все каналы занятыми, не покидает систему, а становится в очередь и ожидает, пока не освободится какой-нибудь канал.

Пропускная способность канала (capacity) – фундаментальное теоретическое понятие, определяющее потенциальные возможности данного канала: максимальное количество единиц информации, которое можно передать по каналу или через систему в единицу времени.

Пропускная способность системы (throughput) – техническая характеристика, определяющая максимально возможную скорость передачи с учетом сложности реализации и стоимости (производительность оборудования): 1) число бит, знаков, пакетов, сообщений, вызовов, которые могут быть переданы через систему в единицу времени в условиях максимальной загрузки; 2) максимальное количество устройств, которые могут одновременно обслуживать поступающие на них запросы без снижения их номинальной скорости.

Пропускная способность СМО зависит от числа каналов и их производительности и обычно представляет собой среднее число заявок, обслуживаемое системой в единицу времени. Под заявками здесь могут подразумеваться обычные телефонные вызовы (требующие, в общем случае, различного времени соединения) или пакеты различной длины. **Относительная пропускная способность** представляет собой среднее отношение числа обслуженных заявок к числу поданных. Снижение пропускной способности СМО происходит из-за наличия случайных сгущений и разрежений в потоке заявок, которые нельзя предвидеть заранее.

Качество обслуживания в СМО, в зависимости от конкретной задачи, может определяться следующими параметрами:

– средним процентом заявок, получающих отказ и покидающих систему необслуженными;

– средним временем «простоя» отдельных каналов и системы в целом;

– средним временем ожидания в очереди;

– вероятностью потери пакета;

– законом распределения (ЗР) длины очереди и т. д.

Сгущения заявок приводят либо к отказам, либо к образованию очередей. Разрежения могут привести к непроизводительным простоям канала (системы). На эти случайности, связанные с неоднородностью потока заявок, накладываются еще случайности, связанные с задержками обслуживания отдельных заявок. Процесс функционирования СМО в общем случае представляет собой случайный процесс – наличие случайного потока заявок и системы со случайной длительностью обслуживания.

Типичная модель СМО может быть представлена в виде физической дискретной системы с конечным (счетным) множеством состояний, в которой переход системы из одного состояния в другое происходит скачком, в дискретный момент времени, когда происходит какое-то **событие** (приход новой заявки, освобождение канала, уход заявки из очереди и т. п.) [6]. Переходы из состояния в состояние обратимы: занятый канал может освободиться, очередь может как увеличиваться, так и уменьшаться.

В одноканальной СМО заявка, заставшая канал занятым, не становится в очередь, а покидает систему (получает «отказ»). На рис. 1.1 представлена модель одноканальной СМО (например, одна телефонная линия) в виде диаграммы состояний, описывающей дискретную

систему с непрерывным временем и двумя возможными состояниями: x_0 – канал свободен; x_1 – канал занят [5].

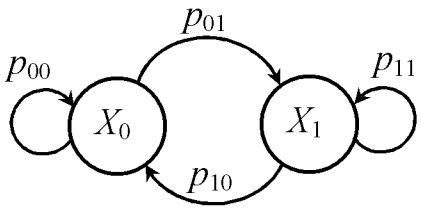


Рис. 1.1. Диаграмма состояний для одноканальной СМО

1.3. Цепи Маркова как основа для моделирования трафика

1.3.1. Дискретные цепи Маркова

Математические модели СМО наиболее часто описываются **вероятностными (стохастическими)** уравнениями, основанными на теории дискретных и непрерывных цепей Маркова [10, 24]. **Дискретная цепь Маркова** считается заданной (рис. 1.2), если для последовательности случайных величин (СВ) выполняется равенство

$$\mathrm{P}\left\{X_n = j \middle| X_1 = i_1,\, X_2 = i_2, \ldots, X_k = i_k, \ldots, X_{n-1} = i_{n-1}\right\} = \mathrm{P}\left\{X_n = j \middle| X_{n-1} = i_{n-1}\right\}.$$

При этом поток СВ определяется только вероятностью перехода от предыдущего значения СВ к последующему. Зная начальное распределение вероятностей, можно найти распределение на любом шаге.

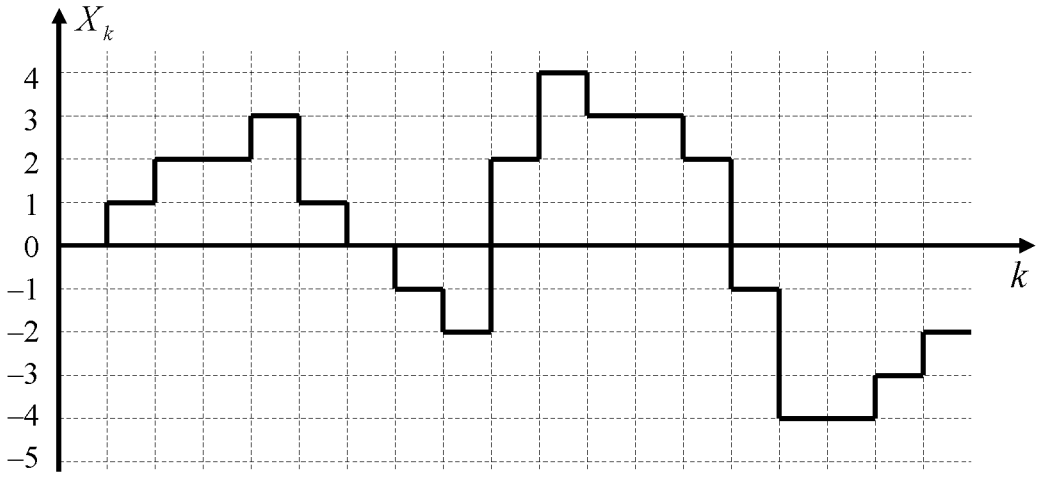


Рис. 1.2. Дискретная цепь Маркова

Величины i_n можно интерпретировать как номера состояний некоторой динамической системы с дискретным множеством состояний (типа конечного автомата), k – номер шага на временной сетке.

Для **однородной** цепи Маркова вероятности переходов не зависят от номера шага

$$p_{ij} = \mathrm{P}\left\{X_n = j \middle| X_{n-1} = i\right\}.$$

Вероятности перехода из состояния i в состояние j за m шагов вычисляются как

$$p_{ij}^{(m)} = \mathrm{P}\left\{X_{n+m} = j \middle| X_n = i\right\} = \sum_k p_{ik}^{(m-1)} p_{kj}\,,\ m = 2, 3, \ldots.$$

Цепь Маркова называется **неприводимой**, если каждое ее состояние может быть достигнуто из любого другого состояния. Состояние i называется **поглощающим**, если для него $p_{ii} = 1$.

Состояние называется **возвратным**, если вероятность попадания в него за конечное число шагов равна 1, в противном случае – **невозвратное** состояние. Возвратное состояние может быть **периодическим** и **апериодическим** в зависимости от наличия кратных шагов возврата. Среднее число шагов (среднее время) возврата в состояние j

$$M_j = \sum_{n=1}^{\infty} n f_j^{(n)}\,,$$

где $f_j^{(n)}$ – вероятность возврата в состояние j через n шагов после ухода из этого состояния.

Состояние называется **возвратным нулевым**, если среднее время возвращения в него равно бесконечности, и **возвратным ненулевым**, если это время конечно. Если все состояния возвратные ненулевые, то существует стационарное распределение вероятностей:

$$\{\pi_i\},\quad \pi_i = \frac{1}{M_i},\quad \pi_j = \sum_i \pi_i p_{ij};\quad \sum_i \pi_i = 1.$$

Возвратное ненулевое состояние, которое к тому же является апериодическим, называется **эргодическим**. Если все состояния цепи Маркова эргодические, то и вся цепь называется эргодической. Предельные вероятности эргодической цепи Маркова называют **вероятностями состояния равновесия**. При этом нет зависимости от начального распределения вероятностей.

Цепь Маркова с конечным числом состояний (конечная цепь) изображается в виде ориентированного графа, называемого **диаграммой переходов**. Вершины графа ассоциируются с состояниями, а ребра с вероятностями переходов [14].

Введем матрицу вероятностей переходов и вектор-строку вероятностей на шаге n:

$$\mathbf{P} = \left[p_{ij} \right]; \quad \mathbf{P}^{(n)} = \left[\pi_1, \pi_2, \ldots \right]^{(n)}.$$

Распределение вероятностей на произвольном шаге определяется как:

$$\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(n-1)} \mathbf{P},$$

и позволяет рекуррентно вычислять все вероятности состояний.

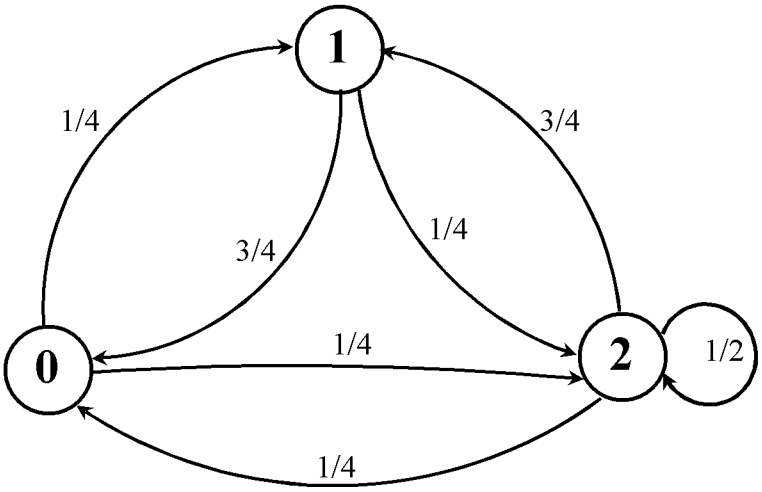


Рис. 1.3. Диаграмма переходов дискретной цепи Маркова

Для нахождения предельного (стационарного) распределения нужно решить уравнение $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$, которое раскрывается в систему линейных уравнений (разрешимых, если цепь конечна)

$$\pi_0 = p_{00}\pi_0 + p_{01}\pi_1 + p_{02}\pi_2 + \ldots$$
$$\pi_1 = p_{10}\pi_0 + p_{11}\pi_1 + p_{12}\pi_2 + \ldots$$
$$\pi_n = p_{n0}\pi_0 + p_{n1}\pi_1 + p_{n2}\pi_2 + \ldots$$

Для примера, показанного на диаграмме переходов (рис. 1.3), имеем

$$\boldsymbol{\pi} = [\pi_0 \ \pi_1 \ \pi_2], \quad \mathbf{P} = \begin{bmatrix} p_{00} & p_{10} & p_{20} \\ p_{01} & p_{11} & p_{21} \\ p_{02} & p_{12} & p_{22} \end{bmatrix} = \begin{bmatrix} 0 & 3/4 & 1/4 \\ 1/4 & 0 & 3/4 \\ 1/4 & 1/4 & 1/2 \end{bmatrix},$$

и решение матричного уравнения сводится к решению системы трех уравнений:

$$\pi_0 = 0 \cdot \pi_0 + (1/4)\pi_1 + (1/4)\pi_2,$$
$$\pi_1 = (3/4)\pi_0 + 0 \cdot \pi_1 + (1/4)\pi_2,$$
$$\pi_2 = (1/4)\pi_0 + (3/4)\pi_1 + (1/2)\pi_2.$$

Так как эти уравнения линейно зависимы, то для решения системы нужно ввести дополнительное нормирующее условие: $1 = \pi_0 + \pi_1 + \pi_2$. Решая систему полученных уравнений, имеем:

$$\pi_0 = 1/5 = 0.2;\ \pi_1 = 7/25 = 0.28;\ \pi_2 = 13/25 = 0.52\ .$$

Вероятности перехода системы из состояния i на m-м шаге в состояние j на n-м шаге для $n > m$ связаны между собой **уравнением Колмогорова-Чепмена:**

$$p_{ij}(m, n) = \sum_k p_{ik}(m, n-1)\, p_{kj}(n-1, n).$$

Для однородных цепей Маркова $p_{ij}(m, n) = p_{ij}^{m-n}$.

1.3.2. Непрерывные цепи Маркова

Случайный процесс $X(t)$ с дискретным множеством значений образует **непрерывную цепь Маркова**, если

$$P\{X(t) = j \mid X(\tau)\ \forall\ \tau_1 \le \tau \le \tau_2 < t\} = P\{X(t) = j \mid X(\tau_2)\}.$$

При этом данная цепь Маркова является непрерывной во времени и дискретной по множеству состояний.

Для однородной непрерывной цепи Маркова уравнение Колмогорова-Чепмена имеет вид

$$\mathbf{P}(T + t) = \mathbf{P}(T)\mathbf{P}(t),$$

где $\mathbf{P}(t)$ – матрица вероятностей переходов из состояния i в состояние j за время t с элементами

$$p_{ij}(t) = P\{X(t + \tau) = j \mid X(\tau) = i\}.$$

Вектор вероятностей состояний марковской цепи в момент времени t определяется матричным уравнением

$$\boldsymbol{\pi}(t) = \boldsymbol{\pi}(0)\, \mathbf{P}(t).$$

Матрица вероятностей переходов $\mathbf{P}(t)$ для непрерывной цепи Маркова, как правило, неизвестна. Значения элементов матрицы часто удается найти, решая систему дифференциальных уравнений (ДУ), которая в матричной форме имеет вид [14]

$$\frac{d\,\mathbf{P}(t)}{dt} = \mathbf{P}(t)\,\mathbf{Q},$$

где $\mathbf{Q}$ - матрица интенсивностей переходов, которая полагается известной. Ее элементы имеют следующий смысл: если в момент времени t система находилась в состоянии x_i, то вероятность перехода в

течение промежутка времени $(t, t+\Delta t)$ в произвольное состояние x_j задается величиной $q_{ij}(t)+o(\Delta t)$, а вероятность сохранить состояние x_i – величиной $1-\sum q_{ij}(t)+o(\Delta t)$.

Рассмотрим непрерывную цепь Маркова применительно к модели СМО с одним сервером и очередью, а также ординарным и стационарным входным потоком заявок (п. 3.2). При этом ординарность потока подразумевает, что вероятность появления заявки на малом интервале $(t, t+\Delta t)$ пропорциональна длине этого интервала и может зависеть только от состояния системы в момент *t*. Обозначим эту вероятность $p_+(k)=\lambda_k\, dt+o(dt)$. Коэффициент λ_k перед dt определяет интенсивность входного потока (количество заявок в единицу времени), а второе слагаемое обозначает бесконечно малую более высокого порядка, чем dt.

Вероятность того, что заявка покинет систему в течение малого интервала времени $(t, t+\Delta t)$ представляет собой величину, прямо пропорциональную этому интервалу времени с точностью до бесконечно малых более высокого порядка малости. Предположив, что интенсивность обслуживания μ_k также может зависеть от состояния *k* системы в данный момент, получим формулу для вероятности ухода заявки с сервера $p_-(k)=\mu_k\, dt+o(dt)$.

При сделанных предположениях можно определить состояние системы в момент *t* как суммарное число заявок *k* в системе и записать уравнение состояния в виде

$$P(k, t+\Delta t)=\Phi(k,\, p_+,\, p_-,\, t).$$

Рассматриваемая система может быть интерпретирована как биологическая популяция численностью *k*, в которой в течение каждого малого интервала времени *dt* с заданной вероятностью может либо появиться одна новая особь (рождение), либо исчезнуть одна особь (гибель). Альтернативой может быть сохранение неизменной численности на этом интервале. Данный процесс принято называть **процессом «размножения-гибели»** [10, 14].

По формуле полной вероятности

$$\begin{aligned} P(k,t+\Delta t)=P(k-1,t)\,p_+(k-1)+P(k+1,t)\,p_-(k+1)+\\ +P(k,t)\left[1-p_+(k)-p_-(k)\right] \end{aligned}, \qquad (1.1)$$

где $P(k,t)$ – вероятность k-го состояния в момент t; $p_-(k+1)=P_{x_{k+1}\to x_k}(\Delta t)$, $p_+(k-1)=P_{x_{k-1}\to x_k}(\Delta t)$ – вероятности переходов между состояниями x_i; $1-p_+(k)-p_-(k)$ – вероятность того, что состояние x_k не изменится за время Δt, откуда следует, что $p_+(k)+p_-(k)$ – вероятность того, что состояние x_k обязательно изменится за время Δt либо на x_{k+1} либо на x_{k-1} (рис. 1.4).

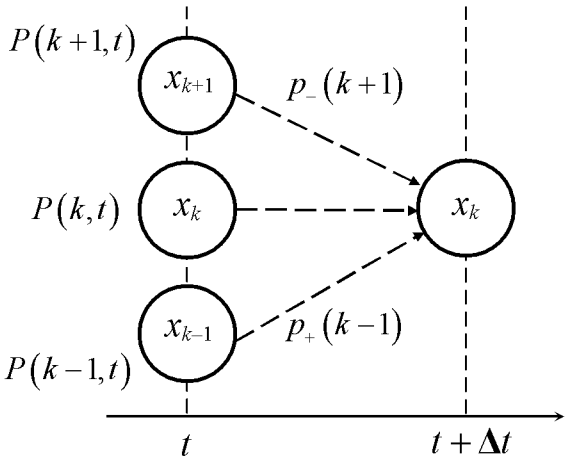


Рис. 1.4. Процесс переходов в непрерывной цепи Маркова

Найдем решение уравнения (1.1) путем его преобразования и перегруппировки слагаемых:

$$P(k,t+\Delta t)-P(k,t)=\left[p_+(k)+p_-(k)\right]P(k,t)+$$
$$+P(k-1,t)\,p_+(k-1)+P(k+1,t)\,p_-(k+1).$$

Подставив введенные ранее выражения для вероятностей прихода и ухода заявок и разделив обе части на dt, получим:

$$\frac{P(k,t+\Delta t)-P(k,t)}{dt}=-(\lambda_k+\mu_k)P(k,t)+\lambda_{k-1}P(k-1,t)+\mu_{k+1}P(k+1,t)+\frac{o(dt)}{dt}.$$

Поскольку отрицательного числа заявок в системе быть не может, и в системе без заявок не могут освобождаться серверы, то следует принять следующие соотношения

$$\lambda_{-1}=\lambda_{-2}=\lambda_{-3}=\ldots=0 \quad \text{и} \quad \mu_0=\mu_{-1}=\mu_{-2}=\ldots=0. \qquad (1.2)$$

В силу этого при $k=0$ будет иметь место уравнение

$$\frac{P(0,t+\Delta t)-P(0,t)}{dt}=-\lambda_0 P(0,t)+\mu_1 P(1,\,t).$$

В пределе мы получим следующую систему ДУ:

$$\frac{dP(k,t)}{dt} = -(\lambda_k + \mu_k)P(k,t) + \lambda_{k-1}P(k-1,t) + \mu_{k+1}P(k+1,t), \quad k \geq 1, \qquad (1.3)$$

$$\frac{dP(0,t)}{dt} = -\lambda_0 P(0,t) + \mu_1 P(1,\, t).$$

В соответствие данной системе уравнений можно поставить диаграмму интенсивностей переходов (рис. 1.5), в которой окружностям соответствуют дискретные состояния, а дуги определяют **интенсивности потоков вероятности переходов** от одного состояния к другому.

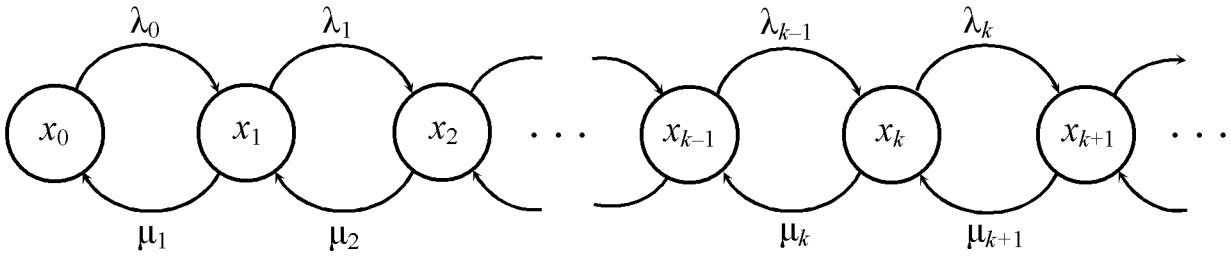


Рис. 1.5. Диаграмма интенсивностей переходов непрерывной цепи Маркова

В соответствии с ДУ (1.3) для k-го состояния (x_k) можно сформулировать «**закон сохранения**»: разность между интенсивностью, с которой система попадает в k-е состояние, и интенсивностью, с которой система покидает это состояние, должна равняться интенсивности изменения потока в этом состоянии (производной по времени).

Рассмотрим стационарный (установившийся) режим, когда можно полагать, что вероятности в произвольный, достаточно отдаленный момент времени, остаются постоянными. Приравнивая производную к нулю $dP(k,t)/dt = 0$, получаем систему разностных уравнений:

$$-(\lambda_k + \mu_k)\, p_k + \lambda_{k-1}\, p_{k-1} + \mu_{k+1}\, p_{k+1} = 0, \quad k \geq 1, \qquad (1.4)$$

$$-\lambda_0 p_0 + \mu_1 p_1 = 0, \quad k = 0.$$

Учитывая соотношения (1.2), уравнение для состояния x_0 ($k = 0$) выписывать отдельно далее не потребуется; необходимо только добавить условие нормировки

$$\sum_{k=0}^{m} p_k = 1, \qquad (1.5)$$

где m – общее число возможных состояний системы.

По закону сохранения в стационарном режиме разность входящих и выходящих потоков в состоянии k равна нулю и уравнение (1.4) приобретает смысл **уравнения равновесия** или **баланса**:

$$\lambda_{k-1} p_{k-1} + \mu_{k+1} p_{k+1} = (\lambda_k + \mu_k) p_k ,$$

т. е. интенсивность потока вероятностей в состояние k равна интенсивности потока из этого состояния (рис. 1.6).

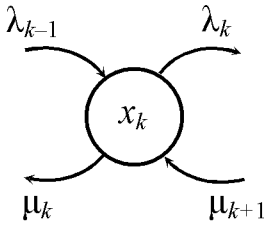


Рис. 1.6. Участок диаграммы интенсивностей переходов для состояния x_k

Решив уравнение баланса (1.4) для $k = 0$, получим [10]

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

Затем, построив систему уравнений для $k = 1$, можно получить

$$p_2 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} p_0 .$$

В общем случае имеем

$$p_k = \frac{\lambda_0 \lambda_1 \lambda_2 \ldots \lambda_{k-1}}{\mu_1 \mu_2 \mu_3 \ldots \mu_k} p_0 = p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}, \qquad p_{k+1} = \frac{\lambda_k}{\mu_{k+1}} p_k . \tag{1.6}$$

С учетом условия нормировки (1.5) имеем

$$p_0 = \frac{1}{1 + \sum_{k=1}^{m} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}} . \tag{1.7}$$

Вероятности состояний будут стационарными, если система уравнений равновесия является эргодической, т. е. если существует некоторое значение k, начиная с которого выполняется неравенство

$$\frac{\lambda_k}{\mu_{k+1}} < C < 1 .$$

Частным случаем выражения (1.7) является формула, выражающая вероятность незанятости СМО с m каналами, полученная при $\mu_{i+1} = (i+1)\mu$ и $\lambda_i = \lambda$ [12]

$$p_0 = \frac{1}{1 + \sum_{k=1}^{m} \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu}} = \frac{1}{\sum_{k=0}^{m} \frac{\lambda^k}{k!\mu^k}} = \left[\sum_{k=0}^{m} \frac{\rho^k}{k!} \right]^{-1}. \qquad (1.8)$$

1.4. Контрольные вопросы

1. Поясните понятие «состояние динамической системы».
2. Приведите примеры конкретных СМО из реальной жизни. Поясните их основные параметры.
3. В чем принципиальное отличие системы с ожиданием от системы с отказами?
4. Какую функцию в СМО выполняет сервер?
5. Дайте определение пропускной способности канала? В каких единицах она измеряется?
6. Какие параметры качества обслуживания в СМО вы знаете? Назовите их.
7. Что представляет собой *событие* в дискретной системе?
8. Что представляет собой дискретная цепь Маркова?
9. Какая цепь Маркова называется однородной?
10. Какая цепь Маркова называется неприводимой?
11. Какое состояние в цепи Маркова называется возвратным?
12. Какое состояние в цепи Маркова называется невозвратным?
13. Какое состояние в цепи Маркова называется поглощающим?
14. Приведите пример диаграммы переходов дискретной цепи Маркова.
15. Запишите уравнение Колмогорова-Чепмена для однородной цепи Маркова?
16. Приведите пример процесса «размножения-гибели».
17. Сформулируйте закон сохранения для непрерывной цепи Маркова.
18. Поясните уравнение равновесия (баланса) для конкретного состояния СМО.
19. Поясните смысл вероятности состояния СМО.
20. Что представляет собой условие нормировки при составлении системы уравнений состояния?

2. ПОТОКИ ЗАЯВОК И ИХ ХАРАКТЕРИСТИКИ

2.1. Простейший поток и его свойства

Под **потоком событий** (заявок) в теории вероятностей понимается последовательность событий, происходящих одно за другим в случайные моменты времени. События, образующие поток, в общем случае могут быть различного типа и формировать неоднородный поток (например, заявки с приоритетом). Рассмотрим поток однородных событий, различающихся только моментами появления (рис. 2.1 – последовательность точек $t_1, t_2, \ldots, t_k, \ldots$ на оси времени Ot).

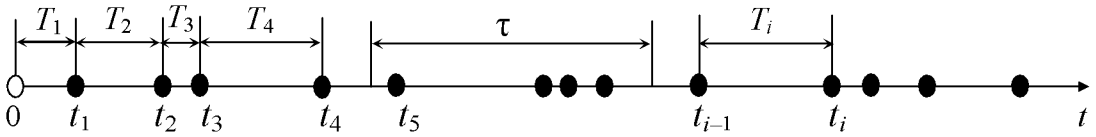


Рис. 2.1. Случайный поток событий

Поток называется **регулярным**, если события следуют одно за другим через одинаковые промежутки времени T (рис. 2.2). На практике такой поток встречается редко. Типичным для СМО является случайный поток заявок, в котором промежутки T_i представляют собой случайные величины (СВ).

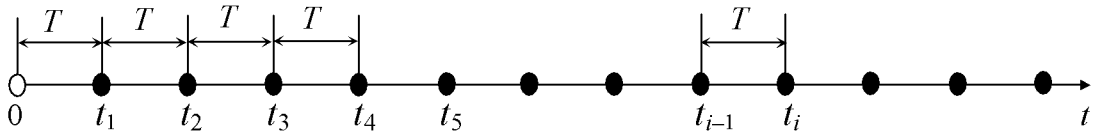


Рис. 2.2. Регулярный поток событий

Наиболее распространенной в приложениях моделью случайного потока событий является **простейший** (или стационарный пуассоновский) поток, в котором число событий, попадающих на любой фиксированный интервал времени, распределено по закону Пуассона.

Простейший поток обладает следующими тремя свойствами:

1. В **стационарном** потоке событий вероятность попадания некоторого числа событий на участок времени длиной τ зависит только от длины участка и не зависит от местоположения этого участка на оси Ot (рис. 2.1). Для стационарного потока характерна постоянная плотность заявок.

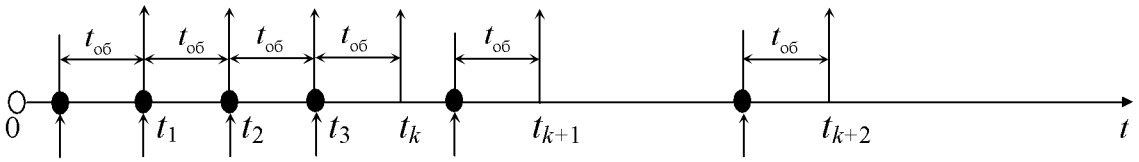


Рис. 2.3. СМО с постоянным временем обслуживания

2. В **потоке без последействия** для любых, не перекрывающихся участков времени число событий, попадающих на один из них, не зависит от числа событий, попадающих на другие. Примером системы с последействием является одноканальная СМО с отказами, для которой время обслуживания одной заявки постоянно и равно $t_{об}$ (рис. 2.3). Тогда в потоке обслуженных заявок минимальный интервал времени между заявками, покидающими систему, будет равен $t_{об}$. Наличие такого минимального интервала неизбежно приводит к последействию. Действительно, пусть известно, что в какой-то момент времени t_k систему покинула обслуженная заявка. Тогда на любом участке времени τ, лежащем в пределах $(t_k, t_k + t_{об})$, обслуженной заявки (на рис. 2.3 – стрелка над осью Ot) не появится – очевидна зависимость между числами событий на не перекрывающихся участках. Регулярный поток (рис. 2.2) является потоком с последействием, т. к. моменты появления событий связаны жесткой функциональной зависимостью.

3. В **ординарном** потоке вероятность попадания на элементарный участок Δt двух или более событий пренебрежимо мала по сравнению с вероятностью попадания одного события (заявки приходят поодиночке, а не парами, тройками и т. д.)

Пуассоновские потоки обладают **свойством аддитивности**: если взять N источников пуассоновских потоков событий с интенсивностями λ_k и рассмотреть новый поток как суммарный поток всех событий от этих источников, то результирующий поток окажется также пуассоновским с интенсивностью

$$\lambda_\Sigma = \sum_{k=1}^{N} \lambda_k .$$

Можно также расщеплять пуассоновский поток с интенсивностью λ на m пуассоновских потоков с интенсивностями λ_k (рис. 2.4). Для этого нужно определить полную систему из m случайных событий ν_k с вероятностями p_k, наступление каждого из которых будем

считать фактом переключения на k-е направление. При поступлении каждого нового события C_i в пуассоновском потоке будем находить совместное событие (декартово произведение) $[C_i \times \nu_k]$ и относить его к k-му потоку. Каждый из m потоков будет пуассоновским с интенсивностью $\lambda_k = \lambda p_k$.

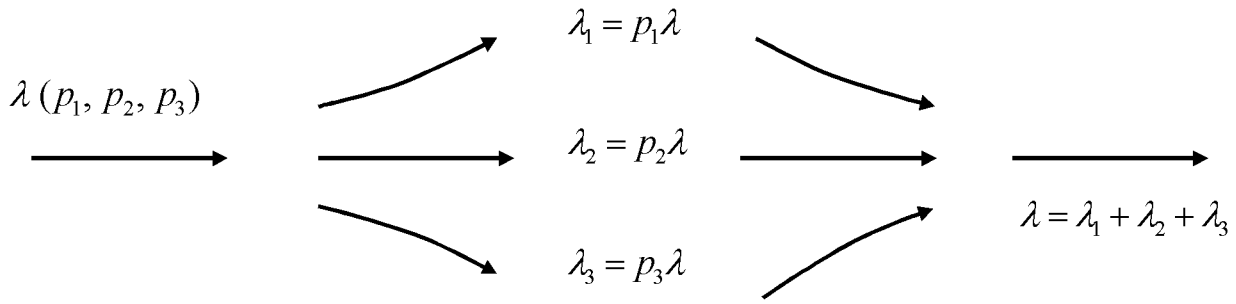


Рис. 2.4. Расщепление и слияние пуассоновских потоков при $m = 3$

Простейший поток играет среди потоков событий особую роль, в некоторой степени аналогичную роли нормального закона среди других законов распределения в соответствии с центральной предельной теоремой (ЦПТ) теории вероятностей [5]: при суммировании (взаимном наложении) большого числа ординарных, стационарных потоков с практически любым последействием получается поток, сколь угодно близкий к простейшему. При этом условия, которые должны соблюдаться, аналогичны условиям ЦПТ – количество складываемых потоков $n \to \infty$.

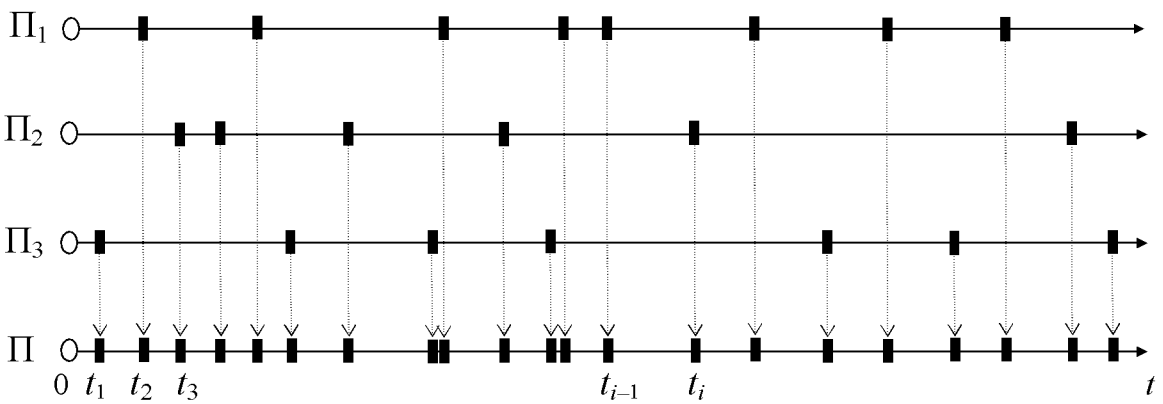


Рис. 2.5. Формирование суммарного потока П из независимых потоков Π_i

Пусть имеется ряд независимых потоков $\Pi_1, \Pi_2, \ldots, \Pi_n$, имеющих плотности одного порядка, а их количество n достаточно велико [5]. «Суммирование» потоков состоит в том, что все моменты появления событий сносятся на одну и ту же ось Ot (рис. 2.5). При этом суммарный поток

$$\Pi = \sum_{k=1}^{n} \Pi_k \qquad (2.1)$$

также является стационарным и ординарным. При увеличении числа слагаемых n последействие в суммарном потоке (2.1), даже если оно значительно в отдельных потоках, постепенно ослабевает.

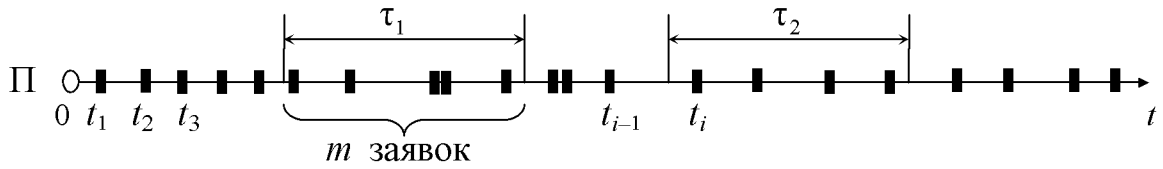


Рис. 2.6. Не перекрывающиеся участки на простейшем потоке

Рассмотрим на оси Ot два не перекрывающихся отрезка τ_1 и τ_2 (рис. 2.6). Каждая из заявок, попадающих в эти отрезки, случайным образом может оказаться принадлежащей тому или иному потоку. По мере увеличения n удельный вес заявок, принадлежащих одному и тому же потоку и, следовательно, зависимых, должен уменьшаться. Остальные заявки принадлежат разным потокам и появляются на отрезках τ_1 и τ_2 независимо друг от друга. При увеличении n суммарный поток будет терять последействие и приближаться к простейшему. На практике достаточно сложить 4-5 потоков, чтобы получить поток, достаточно близкий к простейшему.

Рассмотрим на оси Ot простейший поток событий Π как неограниченную последовательность случайных заявок. Выделим произвольный участок времени длиной τ. Число заявок, попадающих на участок τ, распределено по закону Пуассона с математическим ожиданием

$$a = \lambda\tau,$$

где λ – плотность потока (среднее число событий в единицу времени). Пуассоновский ЗР определяет вероятность того, что за время τ произойдет ровно m событий (рис. 2.6):

$$P_m(\tau) = \frac{(\lambda\tau)^m}{m!} e^{-\lambda\tau}. \qquad (2.2)$$

В частности, вероятность того, что участок окажется пустым (не произойдет ни одного события), будет

$$P_0(\tau) = e^{-\lambda\tau}. \qquad (2.3)$$

Важной характеристикой потока является ЗР длины промежутка между соседними событиями. Рассмотрим случайную величину T – промежуток времени между двумя произвольными соседними событиями в простейшем потоке (рис. 2.1) и найдем ее функцию распределения (вероятность того, что интервал T между соседними заявками не превысит величины t)

$$F(t) = P(T < t).$$

Вероятность противоположного события

$$1 - F(t) = P(T > t)$$

представляет собой вероятность того, что на участке времени длиной t, начинающемся в момент времени t_k появления одного из событий потока, не появится ни одного из последующих событий. Так как простейший поток не обладает последействием, то наличие в начале участка (в точке t_k) какого-то события никак не влияет на вероятность появления тех или других событий в дальнейшем. Поэтому вероятность $P(T > t)$ можно вычислить по формуле (2.3)

$$P_0(t) = e^{-\lambda t},$$

откуда

$$F(t) = 1 - e^{-\lambda t} \quad (t > 0). \qquad (2.4)$$

Дифференцируя $F(t)$, получим показательный ЗР интервала между соседними заявками:

$$f(t) = \lambda e^{-\lambda t} \quad (t > 0).$$

Математическое ожидание и дисперсия величины T, распределенной по показательному закону, соответственно: $m_t = M\{T\} = \frac{1}{\lambda}$, $D_t = D\{T\} = \frac{1}{\lambda^2}$.

Пример. На вход СМО поступает пуассоновский поток с плотностью $\lambda = 5$ выз/с. Найти:

1) среднее количество заявок, приходящееся на интервал времени $\tau = 2$ с;
2) вероятность попадания $m = 3$ заявок в интервал времени $\tau = 3$ с;
3) вероятность попадания $m = 1$ заявки в интервал времени $\tau = 1$ с.

Решение.

1) $a = \lambda\tau = 5 \cdot 2 = 10$ заявок;

2) $P_3(3) = \dfrac{(5 \cdot 3)^3}{3!} e^{-5 \cdot 3} = \dfrac{15^3}{6} e^{-15} = 1.72 \cdot 10^{-4}$;

3) $P_1(1) = \dfrac{(5 \cdot 1)^1}{1!} e^{-5 \cdot 1} = \dfrac{5}{1} e^{-5} = 0.337$.

Отсутствие последействия также может выражаться тем, что время, оставшееся до момента поступления новой заявки, не зависит от того, сколько времени прошло после поступления последней заявки [18]. Пусть, например, в систему поступает простейший поток со средним интервалом между заявками 10 секунд. После прихода очередной заявки прошло 9 секунд. Через сколько секунд (в среднем) придет новая заявка? – Средняя продолжительность оставшегося времени ожидания 10 секунд. А если случилось так, что с момента прихода последней заявки прошла минута – когда (в среднем) придет новая заявка? – Опять через 10 секунд! Хотя это и выглядит парадоксально, это можно доказать.

Пусть после прихода последней заявки прошло время t_0 (рис. 2.7). Интервал между приходами последней и ожидаемой заявок – случайная величина T с функцией распределения (2.4). Нужно найти $P\{T - t_0 < t \mid T \geq t_0\}$ – вероятность того, что оставшееся до прихода новой заявки время окажется меньше t при условии, что весь интервал больше t_0.

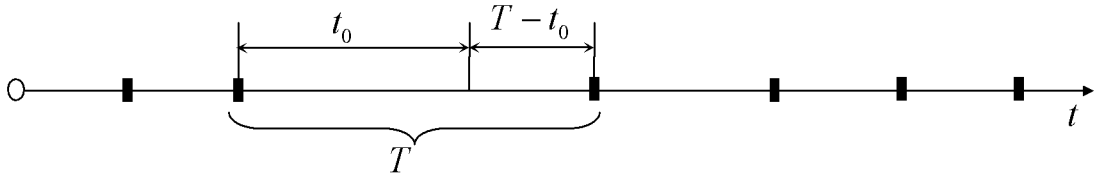


Рис. 2.7. Интервал между моментами прихода соседних заявок

По формуле условной вероятности:

$$P\{T - t_0 < t \mid T \geq t_0\} = \frac{P\{T - t_0 < t, T \geq t_0\}}{P\{T \geq t_0\}} = \frac{P\{t_0 \leq T < t + t_0\}}{P\{T \geq t_0\}} = \frac{P\{T < t + t_0\} - P\{T < t_0\}}{P\{T \geq t_0\}} =$$

$$= \frac{1 - e^{-\lambda(t + t_0)} - \left(1 - e^{-\lambda t_0}\right)}{e^{-\lambda t_0}} = \frac{e^{-\lambda t_0}\left(1 - e^{-\lambda t}\right)}{e^{-\lambda t_0}} = 1 - e^{-\lambda t} = F(t),$$

так как

$$P\{T \geq t_0\} = 1 - P\{T < t_0\} = e^{-\lambda t_0},$$

т. е. условная функция распределения времени, оставшегося до прихода новой заявки, равна функции распределения целого интервала между поступлениями заявок.

Данное замечательное свойство показательного закона можно сформулировать следующим образом: если промежуток времени, распределенный по показательному закону, **уже** длился некоторое время t_0, то это никак не влияет на ЗР **оставшейся** части промежутка – он будет таким же, как и ЗР всего промежутка T. Или, иными словами, интервал времени между произвольным моментом t и моментом поступления очередной заявки имеет такое же распределение, что и интервал между двумя последовательными заявками [28].

Для простейшего потока характерно, что поступление заявок через короткие промежутки времени более вероятно, чем через длинные, что создает более тяжелый режим работы СМО. Если реальный поток отличен от простейшего, то система будет функционировать не хуже, чем это следует из полученных оценок предельных значений характеристик QoS [28].

2.2. Нестационарный пуассоновский поток

Поток однородных событий, ординарный и без последействия, но не стационарный, с переменной плотностью $\lambda(t)$ называется **нестационарным пуассоновским потоком**. Для такого потока число событий, попадающих на участок длины τ, начинающийся в точке t_0, подчиняется закону Пуассона

$$P_m(\tau, t_0) = \frac{a^m}{m!} e^{-a} \quad (m = 0, 1, 2, \ldots),$$

где a – математическое ожидание числа событий на участке от t_0 до $t_0 + \tau$, равное $a = \int_{t_0}^{t_0+\tau} \lambda(t)\,dt$ и зависящее не только от длины τ участка, но и от его положения на оси Ot.

Если поток событий нестационарен, то его основной характеристикой является мгновенная плотность $\lambda(t)$ – предел отношения среднего числа событий, приходящегося на элементарный участок времени $(t, t + \Delta t)$ к длине этого участка $\Delta t \to 0$:

$$\lambda(t)=\lim_{\Delta t\to 0}\frac{m(t+\Delta t)-m(t)}{\Delta t}=m'(t),$$

где $m(t)$ – математическое ожидание числа событий на участке $(0,t)$.

Найдем ЗР промежутка времени T между соседними событиями для нестационарного потока. Поскольку поток нестационарен, ЗР будет зависеть от местоположения первого из событий (t_0) на оси Ot, а также от вида функции $\lambda(t)$. Предположим, что первое из двух событий появилось в момент t_0, и найдем при этом условии ЗР времени T между этим событием и последующим:

$$F_{t_0}(t)=P(T<t)=1-P(T>t).$$

Найдем вероятность $P(T>t)$ того, что на участке (t_0,t_0+t) не появится ни одного события

$$P(T>t)=e^{-a}=e^{-\int_{t_0}^{t_0+t}\lambda(t)dt},\quad \text{откуда}\quad F_{t_0}(t)=1-e^{-\int_{t_0}^{t_0+t}\lambda(t)dt}.$$

Дифференцируя по t, найдем плотность распределения

$$f_{t_0}(t)=\lambda(t_0+t)\,e^{-\int_{t_0}^{t_0+t}\lambda(t)dt}\qquad (t>0).$$

Этот ЗР уже не будет показательным, т. к. в нем имеется зависимость от параметра t_0 и вида функции $\lambda(t)$. Например, при линейном изменении мгновенной плотности $\lambda(t)=a+bt$ плотность распределения вероятностей (ПРВ) будет иметь вид

$$f_{t_0}(t)=\left[a+b(t_0+t)\right]e^{-at-bt_0t-\frac{bt^2}{2}}.$$

Главное свойство простейшего потока – отсутствие последействия, сохраняется и для нестационарного потока, т. е. если зафиксировать на оси Ot произвольную точку t_0, то ПРВ $f_{t_0}(t)$ времени T, отделяющего эту точку от ближайшего по времени будущего события, не зависит от того, что происходило на участке времени, предшествующем t_0, и в самой точке t_0 (т. е. появлялись ли ранее другие события и когда именно).

Модель в виде нестационарного пуассоновского потока в некоторых случаях отражает поведение трафика более адекватно, нежели стационарный пуассоновский поток. Например, наличие часа наибольшей нагрузки (ЧНН) может быть отражено ростом средней интенсивности в определенное время суток.

2.3. Потоки с ограниченным последействием (потоки Пальма)

Ординарный поток однородных событий называется **потоком с ограниченным последействием** (**потоком Пальма**), если промежутки времени между последовательными событиями T_i (рис. 2.1) являются независимыми случайными величинами (в общем случае – с неэкспоненциальной ПРВ). Простейший поток (экспоненциальная ПРВ) является частным случаем потока Пальма [15].

Нестационарный пуассоновский поток не является потоком Пальма, т. к. ЗР промежутка между событиями в нестационарном потоке зависит от того, где этот промежуток начинается, а для двух соседних промежутков T_k и T_{k+1} начало промежутка T_{k+1} совпадает с концом промежутка T_k, значит, длины этих промежутков зависимы [5].

Потоки Пальма часто получаются в виде выходных потоков СМО. Входящий поток заявок разделяется системой СМО на две части: поток обслуженных заявок и поток необслуженных заявок (получивших отказ), который, в свою очередь, поступает на вход какой-либо другой СМО.

Теорема Пальма: пусть на СМО поступает поток заявок типа Пальма, причем заявка, заставшая все каналы занятыми, получает отказ (не обслуживается). Если при этом время обслуживания в СМО имеет показательный ЗР, то поток необслуженных заявок также является потоком Пальма. В частности, если входной поток – простейший, то (просеянный простейший) поток необслуженных заявок будет иметь ограниченное последействие.

Примером потоков с ограниченным последействием могут служить **потоки Эрланга**, которые образуются «просеиванием» простейшего потока (рис. 2.8). Если в простейшем потоке выбросить каждую вторую заявку, то оставшиеся заявки образуют поток Эрланга первого порядка $(Э_1)$. Поскольку независимы промежутки между событиями в простейшем потоке, то независимы и величины T_i', получающиеся суммированием таких промежутков по два в потоке Пальма.

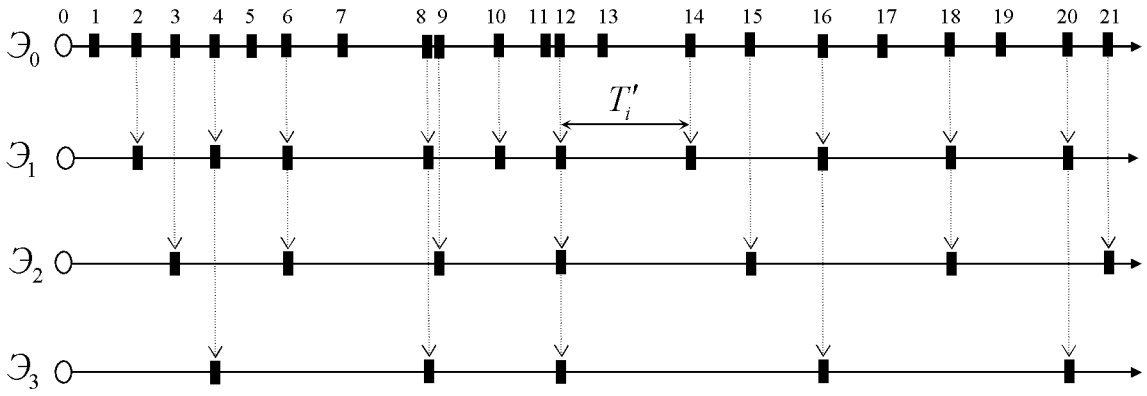


Рис. 2.8. Формирование потоков Эрланга

Поток Эрланга k-го порядка $(Э_k)$ получается из простейшего, если сохранить каждую $(k+1)$-ю заявку, а остальные отбросить (рис. 2.9). При этом простейший поток можно рассматривать как поток Эрланга нулевого порядка $(Э_0)$.

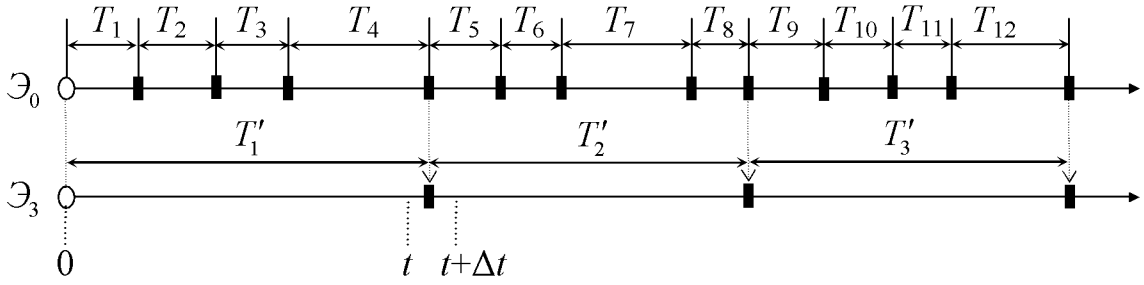


Рис. 2.9. Формирование потока Эрланга третьего порядка

Найдем ЗР промежутка времени T' между соседними событиями в потоке Эрланга k-го порядка $(Э_k)$. Рассмотрим на оси Ot простейший поток с интервалами T_i . Величина T' представляет собой сумму $(k+1)$ независимых случайных величин

$$T' = \sum_{i=1}^{k+1} T_i ,$$

где $T_1, T_2, \ldots, T_{k+1}$ – независимые случайные величины, подчиненные одному и тому же показательному ЗР $f(t) = \lambda e^{-\lambda t}$ $(t > 0)$.

Обозначим $f_k(t)$ плотность распределения величины T' для потока $(Э_k)$; следовательно, $f_k(t)dt$ есть вероятность того, что величи-

на T' примет значение между t и $t+\Delta t$. Это означает, что последняя заявка промежутка $(0, t+\Delta t)$ должна попасть на элементарный участок $(t, t+\Delta t)$, а предыдущие k заявок простейшего потока – на участок $(0, t)$ (рис. 2.9). Вероятность первого события равна $P_1(dt) = \lambda\, dt\, e^{-\lambda\, dt} \underset{dt\to 0}{=} \lambda\, dt$; вероятность второго события $P_k(t) = \dfrac{(\lambda t)^k}{k!} e^{-\lambda t}$. Вероятность одновременного выполнения этих событий получается как произведение:

$$P_1(dt)\, P_k(t) = f_k(t)\, dt = \frac{\lambda(\lambda t)^k}{k!} e^{-\lambda t}\, dt,$$

откуда $f_k(t) = \dfrac{\lambda(\lambda t)^k}{k!} e^{-\lambda t}$ $(t>0)$ – закон Эрланга k-го порядка. При $k=0$ имеем показательный ЗР $f_0(t) = \lambda e^{-\lambda t}$ $(t>0)$.

Характеристики закона Эрланга $f_k(t)$ (ЗР промежутка времени между соседними событиями):

– математическое ожидание $m_k = \sum_{i=1}^{k+1} m_0 = (k+1) m_0$, где $m_0 = 1/\lambda$ – математическое ожидание промежутка между событиями в простейшем потоке, откуда $m_k = (k+1)/\lambda$;

– дисперсия: $D_k = (k+1)/\lambda^2$;

– плотность Λ_k потока $(Э_k)$ будет обратной величине m_k: $\Lambda_k = \dfrac{\lambda}{k+1}$.

Таким образом, при увеличении порядка потока увеличиваются как математическое ожидание, так и дисперсия промежутка времени между событиями, а плотность потока уменьшается.

Интересно выяснить, как будут изменяться характеристики потока Эрланга при $k \to \infty$, если его плотность будет сохраняться постоянной. Для этого необходимо пронормировать величину T' так, чтобы ее математическое ожидание и, следовательно, плотность потока, оставались неизменными. Для этого изменим масштаб по оси времени и вместо T' рассмотрим величину $\tilde{T} = \dfrac{T'}{k+1}$. Назовем такой поток нормированным потоком Эрланга k-го порядка. При этом ЗР промежутка $\tilde{T}$ между событиями этого потока будет

$$\tilde{f}_k(t) = \frac{\theta_k(\theta_k t)^k}{k!} e^{-\theta_k t} \quad (t>0),$$

где $\theta_k = \lambda(k+1)$ – плотность нормированного («сжатого») потока Эрланга k-го порядка.

Математическое ожидание величины $\tilde{T}$ с ПРВ $\tilde{f}_k(t)$ не зависит от k и равно $\tilde{m}_k = m_0 = 1/\lambda$, где λ – плотность потока, совпадающая при любом k с плотностью исходного простейшего потока. Дисперсия величины $\tilde{T}$ равна $\tilde{D}_k = \dfrac{D_k}{(k+1)^2} = \dfrac{1}{\lambda^2(k+1)}$ и неограниченно убывает с возрастанием k.

При неограниченном увеличении k нормированный поток Эрланга приближается к регулярному потоку с постоянными интервалами (рис. 2.2), равными $T = 1/\lambda$. Задавая различные k, можно получить любую степень последействия: от полного отсутствия последействия $(k=0)$ до жесткой функциональной связи между моментами появления событий $(k=\infty)$. Порядок потока Эрланга может служить «мерой последействия», имеющегося в потоке.

2.4. Характеристики времени обслуживания

Производительность СМО характеризуется числом каналов n и быстродействием каждого канала, определяемым временем обслуживания одной заявки $T_{об}$ (далее также обозначается как x).

Рассмотрим случайную величину $T_{об}$ с функцией распределения вероятностей (ФРВ) $G(t) = P(T_{об} < t)$ и ПРВ $g(t) = G'(t)$. На практике обычно величина $T_{об}$ имеет показательное распределение

$$g(t) = \mu e^{-\mu t}, \ (t > 0),$$

где параметр μ обратно пропорционален среднему времени обслуживания одной заявки: $\mu = 1/m_{t_{об}}$, $m_{t_{об}} = M\{T_{об}\}$ – среднее значение времени обслуживания.

Если закон распределения величины $T_{об}$ показательный, то имеет место **отсутствие последействия**: если в некоторый момент t_0 происходит обслуживание заявки, то закон распределения *оставшегося* времени обслуживания не зависит от того, сколько времени обслуживание *уже продолжалось* (рис. 2.7).

2.5. Примитивный поток

Примитивный поток – это нестационарный пуассоновский поток с параметром λ_k, зависящим от состояния СМО, на которую он поступает. Интенсивность примитивного потока

$$\lambda_k = \alpha(N - k), \qquad (2.5)$$

где α – интенсивность парциального потока (рис. 2.10); k – состояние системы $(0 \le k \le N)$; N – общее количество источников, создающих нагрузку на СМО. При этом k-е состояние СМО представляет собой количество источников, обслуживаемых в данный момент (число занятых серверов).

Среднее значение параметра примитивного потока может быть найдено из распределения вероятностей состояний СМО $\{p_k\}$:

$$\overline{\lambda} = \sum_{k=0}^{N} \lambda_k p_k ,$$

где p_k – вероятность того, что занято k источников.

Модель примитивного потока описывает обслуживание нескольких независимых одинаковых пуассоновских источников одной системой. При поступлении на обслуживание заявки от какого-либо из этих источников (занятии одного из серверов), данный источник исключается из числа создающих нагрузку на систему. Величину λ_k можно представить как интенсивность потока, порожденного источниками, не получившими обслуживания (в данный момент не обслуживаемыми).

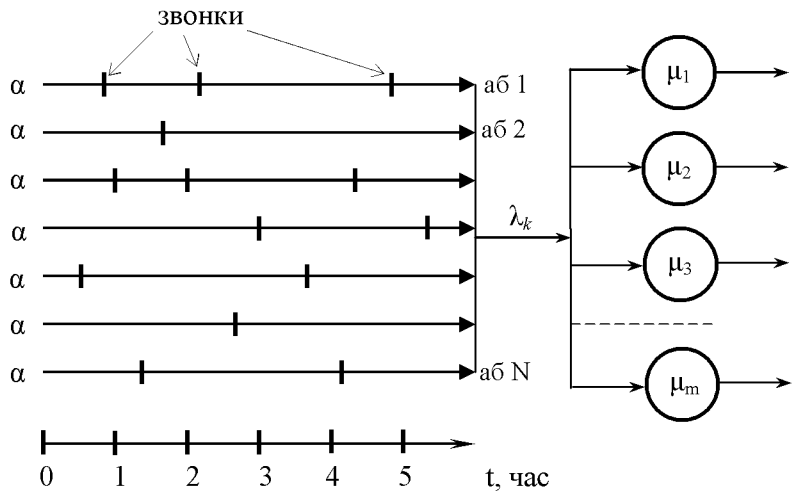


Рис. 2.10. Модель СМО с примитивным потоком на входе

Модель примитивного потока удобна для представления абонентской нагрузки на телефонный коммутатор. Каждый абонент является потенциальным источником независимого пуассоновского потока заявок-звонков. Сумма таких потоков представляет собой общую нагрузку на коммутатор. Когда абонент получает обслуживание своего звонка, его поток исчезает из общего входного потока на время длительности обслуживания данного звонка, т. е. в течение времени собственного обслуживания абонент не является потенциальным отправителем заявки (разговаривая по телефону, одновременно не позвонит по нему), и интенсивность входного потока уменьшается скачком, что является причиной нестационарности потока.

Примитивный поток является обобщением, частным случаем которого является простейший поток [12]. С ростом числа источников N и соответствующим уменьшением величины α последействие примитивного потока сокращается. В пределе при $N \to \infty$ и $\alpha \to 0$, но так, что $\alpha N = \text{const}$ и k ограничено сверху, примитивный поток переходит в простейший с параметром $\lambda = \alpha N$. Практически уже при $N = 300 \ldots 500$ (в зависимости от величины α и максимального значения k) можно пользоваться моделью простейшего потока (п. 4.4). Вносимая при этом погрешность невелика.

2.6. Самоподобные (фрактальные) модели трафика

Рассмотренные выше модели случайных пуассоновских потоков, имеющих показательный ЗР интервала времени между соседними заявками и обладающих свойством марковости (п.1.3), а также модели потоков с последействием (памятью), такие как модели Пальма и Эрланга, оказываются неадекватными при анализе трафика в сетях пакетной коммутации, таких как Ethernet, Internet, Telnet и др. [26-30]. В подобных сетях обнаруживается **долгосрочная зависимость**, или **самоподобие**. На интуитивном уровне это означает, что число событий на заданном временном интервале может зависеть от числа событий, поступивших в весьма отдаленные от него интервалы времени. При этом часто процесс носит «пачечный» характер, т. е. имеют место сгущения и разрежения поступающих заявок. Проведенные в последние годы исследования показали, что поведение потока данных в сетях с пакетной коммутацией хорошо описывается с помощью моделей, основанных на теории фракталов [13, 26, 27, 30].

Фрактал (лат. fractus – «состоящий из фрагментов»; Б. Мандельброт, 1975 г.) – это геометрическая структура, состоящая из частей, которые в каком-то смысле подобны целому или, иными словами, обладающая свойством самоподобия [13]. В самом простом случае небольшая часть фрактала содержит информацию обо всем фрактале. Простейший фрактал имеет вид, представленный на рис. 2.11. За один шаг алгоритма каждый из отрезков, составляющих ломаную, заменяется на ломаную в соответствующем масштабе. В результате бесконечного повторения этой процедуры получается геометрический фрактал. Фрактальная кривая образует фигуру конечной площади, однако периметр этой кривой бесконечен (в случае кривой Коха $(4/3)^n$, $n \to \infty$). Наглядно увидеть рост фракталов можно на сайте: http:// www.shodor.org/master/fractal/software/Snowflake.html.

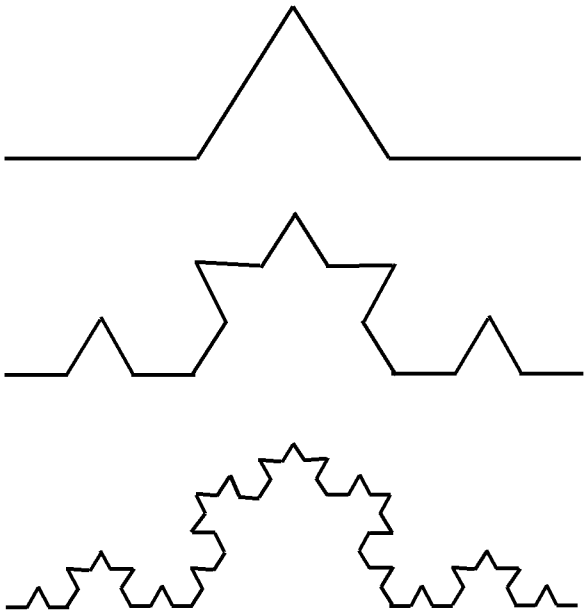


Рис. 2.11. Кривая Коха

Долгосрочная зависимость у случайных процессов (СП) может быть наглядно описана корреляционной функцией (КФ). Будем рассматривать в качестве значения СП число событий, поступающих в систему за фиксированный интервал времени (рис. 2.12). Тогда СП является дискретной последовательностью таких случайных величин, т. е. аргументом является порядковый номер такого интервала времени:

$$X = \{X_i = X(t_i)\},\ i = 1, 2, \ldots.$$

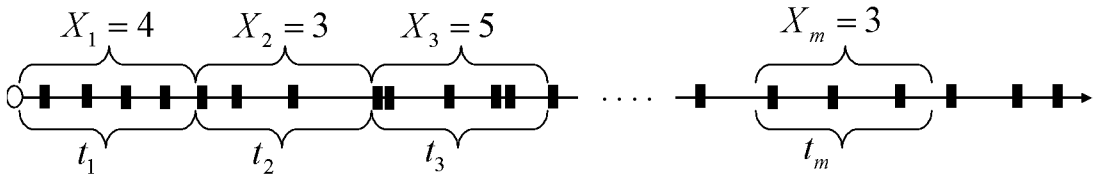


Рис. 2.12. Первичное агрегирование случайного процесса X_i

Будем рассматривать только стационарные СП с ограниченной ковариацией

$$\operatorname{cov}\{X_i, X_{i+k}\} = M\{(X_i - \bar{X})(X_{i+k} - \bar{X})\} < +\infty,$$

дисперсией

$$D\{X\} = M\{(X_i - \bar{X})^2\} = \sigma^2$$

и автокорреляционной функцией (АКФ)

$$r(k) = \frac{\operatorname{cov}\{X_i, X_{i+k}\}}{D\{X\}}.$$

Для того чтобы выявить у СП самоподобие, необходимо рассмотреть агрегированные из него процессы, построенные с помощью усреднения m значений исходного процесса на непересекающихся временных интервалах:

$$X^{(m)} = \{X_k^{(m)}\}, \quad k = 1, 2, \ldots; \qquad X_k^{(m)} = \frac{1}{m}\sum_{i=1}^{m} X_{(k-1)m+i}. \tag{2.6}$$

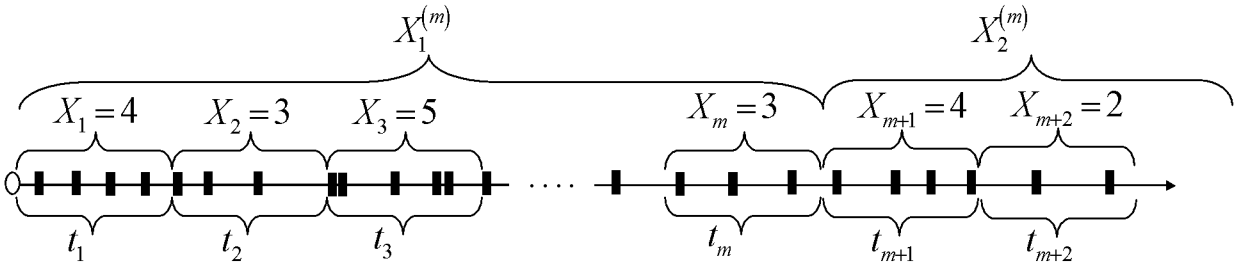


Рис. 2.13. Вторичное агрегирование случайного процесса X_i

Самоподобные СП характеризуются тремя основными свойствами:

1) убыванием дисперсии, которое асимптотически описывается соотношением

$$D\{X^{(m)}\} \approx am^{-\beta}, \quad 0 < \beta < 1, \quad m \to \infty, \tag{2.7}$$

т. е. дисперсия агрегированных процессов уменьшается медленнее, чем величина, обратная размеру выборки ($1/m$);

2) долгосрочной зависимостью, признаком которой является расходимость АКФ СП:

$$\sum_k r(k) = \infty, \quad r(k) \approx k^{-\beta}, \tag{2.8}$$

т. е. происходит медленный спад АКФ по гиперболическому закону;

3) флуктуационным характером спектра мощности, что подразумевает сходство со спектром мощности флуктуаций электронного потока:

$$f(\omega) \approx c\omega^{-\gamma}, \ \omega \to 0, \ \gamma = 1 - \beta, \tag{2.9}$$

где $f(\omega) = \sum_k r(k) e^{-j\omega k}$ – дискретное преобразование Фурье от АКФ.

АКФ самоподобного СП асимптотически совпадает с АКФ агрегированных процессов

$$r(k) \to r^{(m)}(k), \ m \to \infty.$$

Данная формула означает, что корреляционные свойства самоподобного процесса, усредненного на различных временных интервалах, остаются неизменными.

Для выборочного случайного набора X_k, $k = 1, \ldots, N$ можно определить:

1) выборочное среднее $M = \frac{1}{N} \sum_{k=1}^{N} X_k$;

2) выборочную дисперсию $S_N^2 = \frac{1}{N} \sum_{k=1}^{N} (X_k - M)^2$;

3) интегральное отклонение $D_j = \sum_{k=1}^{j} X_k - jM$;

4) изменчивость случайного процесса на интервале N как неубывающую функцию длины интервала

$$R_N = \max_{1<j<N} D_j - \min_{1<j<N} D_j,$$

где N – объем выборки.

Для большинства естественных процессов выполняется отношение [27, 28]

$$\frac{R}{S} \approx \left(\frac{N}{2} \right)^H,$$

или иначе,

$$\log\left(\frac{R}{S} \right) \approx H \log\left(\frac{N}{2} \right)$$

при больших N.

Величина H характеризует «степень самоподобия» СП и называется **параметром Хёрста**, который лежит в интервале $0.5 < H < 1.0$. Для процессов, не обладающих самоподобием $H = 0.5$ (например, простейший поток). Для самоподобных процессов с долгосрочной зависимостью $H \approx 0.7 \ldots 0.9$. При этом $\beta = 2(1 - H)$ [27].

Степень самоподобия процесса можно оценить, определив β из соотношения (2.7):

$$\lim_{m \to \infty} \log D\left\{X^{(m)}\right\} = -\beta \log(m) + \log(a).$$

Построив в логарифмическом масштабе зависимость отношения R/S от логарифма числа выборок или зависимость логарифма дисперсии для агрегированных процессов от логарифма степени агрегирования m, можно оценить H или соответственно β как тангенс угла наклона аппроксимирующей кривой (рис. 2.14). Изложенная выше методика получила название «**R/S-метода**» анализа случайных процессов.

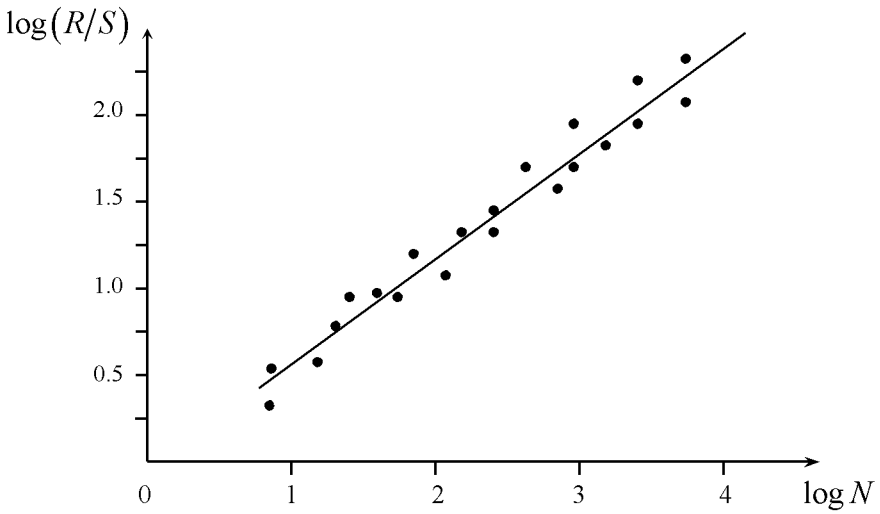


Рис. 2.14. Аппроксимирующая кривая

Эксперименты показали, что распределение числа пакетов в единицу времени в современных пакетных сетях хорошо описывается самоподобным СП с $H \approx 0.65 \ldots 0.8$. При этом процессы, описывающие трафик в различных временных масштабах (шкалах), имеют близкие статистические свойства.

Простейшими самоподобными процессами являются фрактальное броуновское движение и фрактальный гауссовский шум.

Нормированное **фрактальное броуновское движение** (ФБД) (Fractal Brownian Motion, FBM) с параметром Хёрста H – это СП $\{X_t = X(t),\ t > 0\}$, обладающий следующими свойствами [13, с.292-293]:

– $X(t)$ – нормально распределенная СВ для любого $t > 0$, причем приращения $\Delta X(t_k) = X(t_k) - X(t_{k-1})$ также стационарны и нормально распределены;

– $X(t_0) = 0,\ \ M\{\Delta X(t)\} = 0$ для любых t_{k-1}, t_k;

– $M\{\Delta X(t_k)^2\} = \sigma^2 |t_k - t_{k-1}|^{2H}$ для любых t_{k-1}, t_k.

Дисперсия приращений:

$$D\{\Delta X(t_k)\} = M\{\Delta X(t_k)^2\} - [M\{\Delta X(t_k)\}]^2 = \sigma^2 |t_k - t_{k-1}|^{2H}.$$

Приращения $\Delta X(t_k)$ независимы только при $H = 0.5$, и в этом случае ФБД совпадает с обычным броуновским движением.

Последовательность приращений FBM, т. е. его производная, образует **фрактальный гауссовский шум** (ФГШ) (Fractal Gaussian Noise, FGN) или «черный шум» – стационарный гауссовский процесс

$$\Delta X(t_k) = X(t_k) - X(t_{k-1})$$

с заданными параметрами $N(m, \sigma^2)$ и АКФ

$$r(k) = \frac{1}{2}\left[|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\right].$$

При этом ФБД может быть выражено как сумма отсчетов ФГШ

$$X(t_n) = \sum_{i=1}^{n} \Delta X(t_i).$$

Параметр Хёрста сохраняется при суммировании любого конечного числа независимых самоподобных процессов с фиксированным параметром самоподобия, так как для суммы независимых СП КФ равна сумме КФ слагаемых, если каждая из функций имела асимптотическое поведение в соответствии с (2.7). При этом сохраняются все важнейшие характеристики самоподобия, например, характер выбросов самоподобного потока сохраняется при его рассмотрении в различных масштабах времени (при агрегировании) в отличие от пуассоновского потока, у которого на значительных интервалах анализа выбросы практически исчезают [22, с.73]. Это означает, что если записывать нагрузку на каком-либо элементе сети с дискретностью, например, 10 миллисекунд, то, рассматривая график изменения нагрузки во времени на интервалах 10 секунд, 10 минут или 10 часов, не бу-

дет существенных различий в поведении кривой. В этом смысле самоподобие графиков может быть охарактеризовано соотношением

$$y(t) = a^{\alpha}\, y\left(\frac{t}{\alpha}\right),$$

где α – коэффициент самоподобия непрерывных процессов [28, с.294].

Распределение числа событий во времени для трафика, представимого самоподобным процессом, носит характер сложной взаимосвязанной последовательности случайных пачек поступающих заявок. Долгосрочная (или долговременная) зависимость является причиной длительных пульсаций нагрузки в ТС, которые превышают средние уровни нагрузки. Это приводит к быстрому переполнению буферов и вызывает потери или недопустимые задержки в обслуживании.

2.7. Контрольные вопросы

1. Какой поток заявок называется ординарным?
2. Какой поток заявок называется стационарным?
3. Какой поток заявок называется простейшим? Назовите три его основных свойства.
4. Дайте определение термину *последействие*. Поясните на рисунке.
5. Запишите закон распределения Пуассона. Объясните смысл параметров в формуле.
6. Какой случайный процесс называется Марковским?
7. Какому закону распределения подчиняется интервал времени между заявками в пуассоновском потоке?
8. Приведите примеры потоков с сильным последействием.
9. Приведите три примера потоков Пальма.
10. Что такое ограниченное последействие?
11. Можно ли поток с изменяющимися во времени статистическими характеристиками назвать простейшим? Почему?
12. Будет ли поток заявок, получивших отказ в обслуживании при поступлении на одноканальную СМО являться потоком Пальма? Почему?
13. Какой поток заявок называется регулярным?
14. Каким образом изменяются свойства результирующего потока, получаемого путем взаимного наложения большого числа потоков с ограниченным последействием?
15. Запишите выражение для ПРВ длины промежутка времени между соседними событиями в пуассоновском потоке.

16. Какая особенность пуассоновского потока создает более тяжелый режим работы СМО? Как она проявляет себя на графике ПРВ длины промежутка времени между соседними событиями?
17. Что представляет собой «просеивание» потока?
18. Поясните на рисунке, каким образом формируется поток Эрланга 2-го порядка.
19. Что представляет собой нормированный поток Эрланга k-го порядка?
20. Какую характеристику СМО описывает параметр μ?
21. Каким образом формируется примитивный поток?
22. Что характеризует параметр α в примитивном потоке?
23. Что представляет собой k-е состояние СМО с примитивным потоком на входе?
24. При каких значениях N модели простейшего и примитивного потоков практически совпадают?
25. Дайте определение понятию «фрактал»? Каковы его основные свойства?
26. Что представляет собой долгосрочная зависимость в случайном процессе?
27. Какова основная характерная особенность корреляционной функции самоподобного СП?
28. Каким образом осуществляется агрегирование случайного процесса?
29. Каковы три основных свойства самоподобных случайных процессов?
30. Что характеризует изменчивость случайного процесса?
31. Что характеризует интегральное отклонение случайного процесса?
32. Поясните кратко сущность R/S-метода.
33. Чему равен параметр Херста для простейшего потока?
34. Каким образом формируется фрактальное броуновское движение?
35. Каким образом формируется фрактальный гауссовский шум?

3. ОБСЛУЖИВАНИЕ ПРОСТЕЙШЕГО ПОТОКА ЗАЯВОК

3.1. Обслуживание простейшего потока системой с отказами Уравнения Эрланга

Как уже было упомянуто в п. 1.2, в СМО **с отказами** заявка, поступившая в момент, когда все каналы обслуживания заняты, немедленно получает отказ, покидает систему и в дальнейшем процессе обслуживания не участвует.

Рассмотрим n-канальную СМО с отказами как физическую систему X с конечным множеством состояний (рис. 3.1):

x_0 – свободны все каналы; x_1 – занят ровно один канал;

x_k – занято ровно k каналов … x_n – заняты все n каналов.

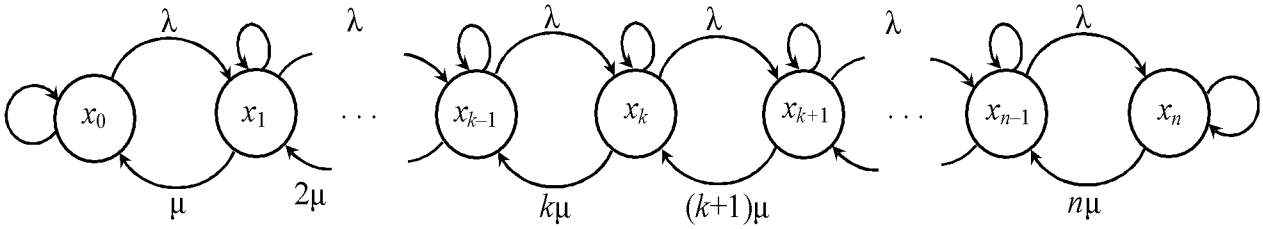


Рис. 3.1. Диаграмма интенсивностей переходов для n-канальной СМО

Определим вероятности состояний системы $p_k(t)$, $k = 0, 1, \ldots, n$ для любого момента времени t при следующих допущениях:

1) поток заявок – простейший с плотностью λ;

2) время обслуживания $T_{об}$ имеет показательный ЗР с параметром μ

$$g(t) = \mu e^{-\mu t},\ t > 0.$$

Аналогично параметру λ («плотности потока заявок»), величину μ можно рассматривать как «плотность потока освобождений» непрерывно занятого канала (бесперебойно снабжаемого заявками), т. е. на выходе канала имеем простейший поток освобождений с плотностью μ.

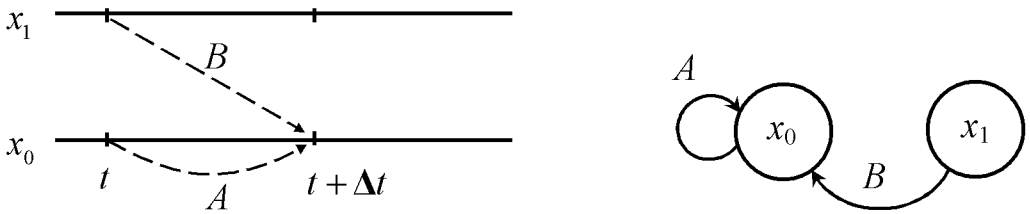
Рассмотрим возможные состояния системы и их вероятности $p_0(t), p_1(t), \ldots, p_n(t)$. В любой момент времени полная вероятность $\sum_{k=0}^{n} p_k(t) = 1$.

Составим дифференциальные уравнения (ДУ) для всех вероятностей $p_i(t)$, начиная с $p_0(t)$. Зафиксируем момент времени t и найдем вероятность $p_0(t + \Delta t)$ того, что в момент $t + \Delta t$ система будет нахо-

диться в состоянии x_0 (все каналы свободны). Это может произойти двумя способами (рис. 3.2):

А – в момент t система была в состоянии x_0, а за время Δt не перешла из него в x_1 (не пришло ни одной заявки);

В – в момент t система была в состоянии x_1, а за время Δt канал освободился, и система перешла в состояние x_0.



а) б)

Рис. 3.2. Временная диаграмма (а) и участок диаграммы переходов (б) для состояния x_0

Возможностью «перескока» системы через состояние (например, из x_2 в x_0 через x_1) за малый промежуток времени Δt можно пренебречь (ординарный поток), как величиной высшего порядка малости по сравнению с $P(A)$ и $P(B)$. По теореме сложения вероятностей имеем

$$p_0(t+\Delta t) \approx P(A) + P(B). \qquad (3.1)$$

Найдем вероятность события A по теореме умножения. Вероятность того, что в момент t система была в состоянии x_0, обозначим как $p_0(t)$. Вероятность того, что за время Δt не придет ни одной заявки (2.3), равна $P_{x_0 \to x_0}(\Delta t) = P_0^{(\lambda)}(\Delta t) = e^{-\lambda \Delta t}$, где верхний символ (λ) означает, что рассматривается поток поступающих на вход СМО заявок с интенсивностью λ. Учитывая, что $e^{-\lambda \Delta t} \approx 1 - \lambda \Delta t$ с точностью до величин высшего порядка малости получаем $P(A) \approx p_0(t) P_{x_0 \to x_0}(\Delta t) = p_0(t)(1 - \lambda \Delta t)$.

Найдем вероятность события B по теореме умножения. Вероятность того, что в момент t система была в состоянии x_1, равна $p_1(t)$. Вероятность того, что за время Δt один канал освободится, равна $P_{x_1 \to x_0}(\Delta t) = 1 - P_{x_1 \to x_1}(\Delta t) = 1 - e^{-\mu \Delta t}$, где $P_{x_1 \to x_1}(\Delta t) = P_0^{(\mu)}(\Delta t) = e^{-\mu \Delta t}$ – вероятность того, что за время Δt не появится ни одного освобожденного канала (2.3), где верхний символ (μ) означает, что рассматривается поток освобождений на выходе СМО с интенсивностью μ.

Или, с другой стороны, по закону Пуассона (2.2) для параметра μ с учетом того, что при $\Delta t \to 0$ выполняется $1 - e^{-\mu \Delta t} \approx \mu \Delta t$, а также с учетом того, что $(\Delta t)^2 << \Delta t$, получим

$$P_{x_1 \to x_0}(\Delta t) = P_1^{(\mu)}(\Delta t) = \mu \Delta t \, e^{-\mu \Delta t} \approx \mu \Delta t (1 - \mu \Delta t) \approx \mu \Delta t .$$

Следовательно, $P(B) \approx p_1(t) P_{x_1 \to x_0}(\Delta t) = p_1(t) \, \mu \Delta t$ и из (3.1) имеем

$$p_0(t + \Delta t) \approx p_0(t)(1 - \lambda \Delta t) + \mu p_1(t) \Delta t .$$

В результате переноса $p_0(t)$ в левую часть, деления на Δt и перехода к пределу при $\Delta t \to 0$, получим ДУ для $p_0(t)$:

$$\lim_{\Delta t \to 0} \frac{p_0(t + \Delta t) - p_0(t)}{\Delta t} = \frac{d\, p_0(t)}{d\, t} = -\lambda \, p_0(t) + \mu \, p_1(t) . \qquad (3.2)$$

Аналогичные ДУ могут быть составлены и для других вероятностей состояний. Возьмем любое $k \, (0 < k < n)$ и найдем вероятность $p_k(t + \Delta t)$ того, что в момент времени $t + \Delta t$ система будет в состоянии x_k. Эта вероятность вычисляется как вероятность суммы трех событий $p_k(t + \Delta t) \approx P(A) + P(B) + P(C)$ (рис. 3.3):

A – в момент t система была в состоянии x_k (занято k каналов), а за время Δt не перешла из него ни в x_{k+1}, ни в x_{k-1} (ни одна заявка не поступила, ни один канал не освободился);

B – в момент t система была в состоянии x_{k-1} (занято $k-1$ каналов), а за время Δt перешла в x_k (пришла одна заявка);

C – в момент t система была в состоянии x_{k+1} (занято $k+1$ каналов), а за время Δt перешла в x_k (один из каналов освободился).

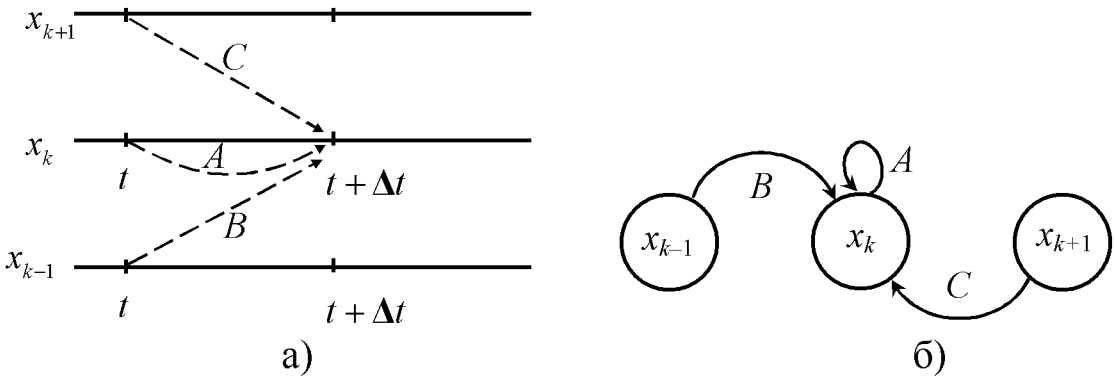


Рис. 3.3. Временная диаграмма (а) и участок диаграммы переходов (б) для состояния x_k

Для нахождения $P(A)$ вычислим сначала вероятность того, что за время Δt не придет ни одна заявка и не освободится ни один из k каналов:

$$P_{x_k \to x_k}(\Delta t) = P_0^{(\lambda)} \left[P_0^{(\mu)} \right]^k = e^{-\lambda \Delta t} \left(e^{-\mu \Delta t} \right)^k = e^{-(\lambda + k\mu)\Delta t}.$$

Учитывая, что при $\Delta t \to 0$ выполняется $e^{-(\lambda+k\mu)\Delta t} \approx 1 - (\lambda + k\mu)\Delta t$, получим

$$P(A) \approx p_k(t) P_{x_k \to x_k}(\Delta t) = p_k(t) \left[1 - (\lambda + k\mu)\Delta t \right].$$

Аналогично, с учетом того, что $(\Delta t)^2 << \Delta t$, имеем

$$P(B) = p_{k-1}(t) P_1^{(\lambda)}(\Delta t) = p_{k-1}(t) \lambda \Delta t \, e^{-\lambda \Delta t} \approx p_{k-1}(t) \lambda \Delta t (1 - \lambda \Delta t) \approx p_{k-1}(t) \lambda \Delta t,$$

$$P(C) = p_{k+1}(t) (k+1) P_1^{(\mu)} = p_{k+1}(t) (k+1) \mu \Delta t \, e^{-\mu \Delta t} \approx$$

$$\approx p_{k+1}(t) (k+1) \mu \Delta t (1 - \mu \Delta t) \approx p_{k+1}(t) (k+1) \mu \Delta t,$$

где $(k+1) P_1^{(\mu)}$ – вероятность освобождения хотя бы одного из $(k+1)$ каналов.

В итоге получаем

$$p_k(t + \Delta t) \approx p_k(t) \left[1 - (\lambda + k\mu)\Delta t \right] + p_{k-1}(t) \lambda \Delta t + p_{k+1}(t)(k+1)\mu \Delta t.$$

Отсюда по аналогии с (3.2) получаем ДУ для $p_k(t)$ $(0 < k < n)$:

$$\frac{d\, p_k(t)}{d\, t} = \lambda p_{k-1}(t) - (\lambda + k\mu) p_k(t) + (k+1)\mu p_{k+1}(t). \qquad (3.3)$$

Составим уравнение для последней вероятности $p_n(t)$ как вероятности суммы двух событий A и B $p_n(t + \Delta t) \approx P(A) + P(B)$ (рис. 3.4):

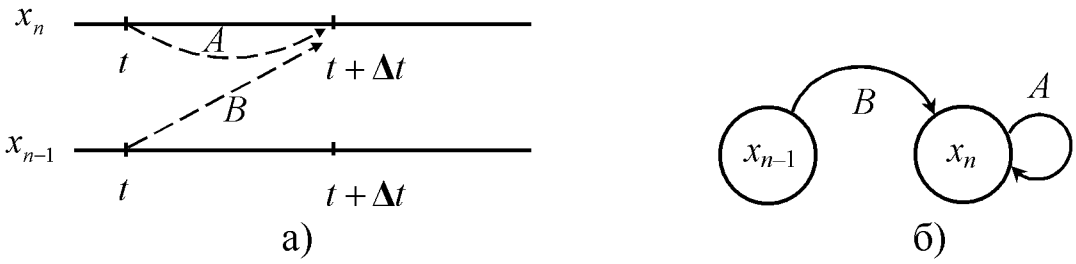


Рис. 3.4. Временная диаграмма (а) и участок диаграммы переходов (б) для состояния x_n

A – в момент t система была в состоянии x_n (заняты все n каналов) и за время Δt осталась в нем же (ни один канал не освободился) $P_{x_n \to x_n}(\Delta t) = \left[P_0^{(\mu)}(\Delta t) \right]^n$, откуда

$$P(A)=p_n(t)\left[P_0^{(\mu)}(\Delta t)\right]^n=p_n(t)\,e^{-n\mu\Delta t}\approx p_n(t)(1-n\mu\Delta t);$$

В – в момент t система была в состоянии x_{n-1} (занято $n-1$ каналов), а за время Δt перешла в x_n (пришла одна заявка) $P_{x_{n-1}\to x_n}(\Delta t)=P_1^{(\lambda)}(\Delta t)$,

$$P(B)=p_{n-1}(t)P_1^{(\lambda)}(\Delta t)=p_{n-1}(t)\lambda\Delta t\,e^{-\lambda\Delta t}\approx p_{n-1}(t)\lambda\Delta t(1-\lambda\Delta t)\approx p_{n-1}(t)\lambda\Delta t\,.$$

В итоге имеем

$$p_n(t+\Delta t)\approx p_n(t)(1-n\mu\Delta t)+p_{n-1}(t)\lambda\Delta t\,,$$

где $\left[P_0^{(\mu)}(\Delta t)\right]^n=e^{-n\mu\Delta t}\approx 1-n\mu\Delta t$, $\Delta t\to 0$ – вероятность того, что за время Δt не освободится ни один из n каналов; $P_1^{(\lambda)}(\Delta t)\approx\lambda\Delta t$ – вероятность того, что за время Δt придет одна заявка.

По аналогии с (3.2) можно получить ДУ для $p_n(t)$:

$$\frac{d\,p_n(t)}{dt}=\lambda p_{n-1}(t)-n\mu p_n(t)\,. \qquad (3.4)$$

Объединив уравнения (3.2), (3.3) и (3.4), получаем систему ДУ для вероятностей $p_0(t)$, $p_1(t)$, $p_2(t),\ldots,p_n(t)$, которая называются системой **уравнений Эрланга**.

$$\begin{cases}\dfrac{d\,p_0(t)}{dt}=-\lambda p_0(t)+\mu p_1(t)\\ \ldots\ldots\ldots\ldots\\ \dfrac{d\,p_k(t)}{dt}=\lambda p_{k-1}(t)-(\lambda+k\mu)\,p_k(t)+(k+1)\mu p_{k+1}(t),\ 0<k<n\\ \ldots\ldots\ldots\ldots\\ \dfrac{d\,p_n(t)}{dt}=\lambda p_{n-1}(t)-n\mu p_n(t)\end{cases} \qquad (3.5)$$

Интегрирование системы уравнений (3.5) при начальных условиях $p_0(0)=1$, $p_1(0)=\ldots=p_n(0)=0$ (в начальный момент все каналы свободны) дает зависимость $p_k(t)$ для любого k. Вероятности $p_k(t)$ характеризуют изменение средней загрузки системы во времени. Очевидно, вероятность занятости всех n каналов является вероятностью отказа $P_{отк}=p_n(t)$. При этом величина $q(t)=1-p_n(t)=\dfrac{\overline{N}_{\text{обсл}}}{\overline{N}_{\text{пост}}}$ представляет собой **относительную пропускную способность** системы – от-

ношение среднего числа обслуженных $\overline{N}_{\text{обсл}}$ за единицу времени заявок к среднему числу заявок поступивших $\overline{N}_{\text{пост}}$ на вход СМО в данный момент t.

3.2. Установившийся режим обслуживания. Формулы Эрланга

При включении СМО в работу вначале имеет место переходный (неустановившийся) процесс, а затем система переходит в установившийся режим работы, вероятностные характеристики которого уже не зависят от времени, т. е. для любой системы с отказами существует предельный режим: при $t \to \infty$ все вероятности $p_0(t), p_1(t), p_2(t), \ldots, p_n(t)$ стремятся к постоянным значениям $p_0, p_1, \ldots, p_n$, а все их производные стремятся к нулю [5]. Из системы ДУ (3.5) получим систему алгебраических уравнений:

$$\begin{cases} -\lambda p_0 + \mu p_1 = 0 \\ \lambda p_0 - (\lambda + \mu) p_1 + 2\mu p_2 = 0 \\ \ldots\ldots\ldots\ldots\ldots \\ \lambda p_{k-1} - (\lambda + k\mu) p_k + (k+1)\mu p_{k+1} = 0\ ,\ 0 < k < n \\ \ldots\ldots\ldots\ldots\ldots \\ \lambda p_{n-2} - [\lambda + (n-1)\mu] p_{n-1} + n\mu p_n = 0 \\ \lambda p_{n-1} - n\mu p_n = 0 \end{cases} \qquad (3.6)$$

К уравнениям (3.6) необходимо добавить условие нормировки (1.5)

$$\sum_{k=0}^{n} p_k = 1. \qquad (3.7)$$

Решив систему (3.6) относительно неизвестных $p_0, p_1, \ldots, p_n$, получим вероятность занятости k каналов

$$p_k = \frac{\lambda^k}{k!\mu^k} p_0\ ,\ k \le n\ .$$

Отношение $\rho = \lambda/\mu = \lambda m_{t_{\text{об}}} < 1$ называется **приведенной плотностью заявок (нагрузкой)**: среднее число заявок, приходящееся на среднее время обслуживания одной заявки, где $m_{t_{\text{об}}} = M\{T_{\text{об}}\}$ – среднее время обслуживания одной заявки. Таким образом, формула

$p_k = \frac{\rho^k}{k!} p_0$ выражает все вероятности p_k через p_0. Используя условие (3.7), получим $\sum_{k=0}^{n} p_k = p_0 \sum_{k=0}^{n} \frac{\rho^k}{k!} = 1$, откуда

$$p_0 = \left(\sum_{k=0}^{n} \frac{\rho^k}{k!} \right)^{-1}. \qquad (3.8)$$

Формула

$$p_k = \frac{\rho^k / k!}{\sum_{i=0}^{n} \frac{\rho^i}{i!}}, \quad 0 \le k \le n \qquad (3.9)$$

называется **формулой Эрланга В**, которая описывает закон распределения числа занятых каналов в зависимости от характеристик потока заявок и производительности СМО (более общими выражениями являются (1.6) и (1.7)) и табулирована в [37]. При $k = n$ (все каналы заняты), получим вероятность отказа (**первая формула Эрланга**) [31, с.59] или вероятность потерь по времени [12]:

$$P_{\text{отк}} = p_n = \frac{\rho^n / n!}{\sum_{k=0}^{n} \frac{\rho^k}{k!}} = E_n(\rho). \qquad (3.10)$$

При отсутствии таблиц вычисление функции $E_n(\rho)$ можно производить по следующей рекуррентной формуле [12]:

$$E_n(\rho) = \rho E_{n-1}(\rho) \big/ \left[n + \rho E_{n-1}(\rho) \right], \quad E_0(\rho) = 1. \qquad (3.11)$$

При больших значениях n и особенно при $\rho > n$ точнее и экономичнее другой алгоритм: $E_n(\rho) = A_m^{-1}$, $A_k = A_{k-1} + B_k$, $B_k = (n - k + 1) B_{k-1} / \rho$, $k = 1 \ldots m$, $m \le n$, $A_0 = B_0 = 1$.

Для одноканальной системы $(n = 1)$ $P_{\text{отк}} = p_1 = \frac{\rho}{1 + \rho}$, а относительная пропускная способность $q = 1 - P_{\text{отк}} = \frac{1}{1 + \rho}$.

Пример. На АТС (систему с отказами), имеющую 4 линии связи, поступает простейший поток заявок с плотностью $\lambda = 3$ (выз/мин). Средняя длительность разговора 2 минуты. Найти вероятность отказа и среднюю долю времени, в течение которой АТС вообще не загружена.

Решение. Имеем $m_{t_{об}} = 2$ (мин); $\mu = 0.5$ (разг/мин), $\rho = \lambda/\mu = 6$.

По формуле (3.10) для $n = 4$ имеем $P_{отк} = p_4 = \dfrac{\rho^4/4!}{1+\dfrac{\rho^1}{1!}+\dfrac{\rho^2}{2!}+\dfrac{\rho^3}{3!}+\dfrac{\rho^4}{4!}} \approx 0.47$.

Из формулы (3.8) можно получить $p_0 = \dfrac{1}{1+\dfrac{\rho^1}{1!}+\dfrac{\rho^2}{2!}+\dfrac{\rho^3}{3!}+\dfrac{\rho^4}{4!}} \approx 0.0087$.

Рассмотрим другие характеристики качества обслуживания [12, с.85]:

1. Интенсивность обслуженной нагрузки. В соответствии с определением, используя второе соотношение формулы (1.6) при $\mu_k = k\mu$ и $\lambda_k = \lambda$, получим $(k+1)\mu\, p_{k+1} = \lambda\, p_k$ или $(k+1)\, p_{k+1} = \rho p_k$ и интенсивность обслуженной нагрузки

$$Y = \sum_{i=1}^{n} i\, p_i = \rho \sum_{i=1}^{n} p_{i-1} = \rho \sum_{i=0}^{n-1} p_i = \rho\left(1 - p_n\right) = \rho\left[1 - E_n(\rho)\right]. \qquad (3.12)$$

С ростом числа выходов n и величины потерь $p_n = E_n(\rho)$ пропускная способность системы Y увеличивается.

2. Интенсивность потенциальной нагрузки. По аналогии с предыдущим

$$A = \sum_{i=1}^{\infty} i\, p_i = \rho \sum_{i=1}^{\infty} p_{i-1} = \rho \sum_{i=0}^{\infty} p_i = \rho, \qquad (3.13)$$

где вероятность p_i определяется в соответствии с распределением Пуассона (глава 4). Равенство интенсивностей потенциальной и поступающей нагрузок приводит к равенству интенсивностей потерянной и избыточной нагрузок:

$$A_{пот} = R = \rho E_n(\rho), \qquad (3.14)$$

где R – интенсивность потерянных вызовов.

Из (3.14) непосредственно следует равенство потерь по нагрузке и по вызовам. Таким образом, все три вида потерь равны между собой. Обусловлено это двумя свойствами простейшего потока: стационарностью и отсутствием последействия.

3.3. Обслуживание простейшего потока заявок с ожиданием

В **СМО с ожиданием** заявка, заставшая все каналы занятыми, становится в очередь и ждет, пока не освободится какой-нибудь ка-

нал. Если время ожидания заявки в очереди ничем не ограничено, то имеем «**чистую систему с ожиданием**». Если оно ограничено какими-то условиями, то имеем «**систему смешанного типа**».

Ограничения, наложенные на ожидание:

– время ожидания заявки в очереди ограничено сверху некоторым сроком $T_{\text{ож}}$, который может быть как строго определенным, так и случайным;

– ограничение на общее время пребывания заявки в системе;

– ограничение на число заявок в очереди – заявка становится в очередь только в том случае, если длина очереди не превышает некоторого значения.

При наличии **дисциплины очереди** ожидающие заявки могут вызываться на обслуживание как в порядке очереди (раньше прибывший раньше и обслуживается, FCFS – First Came First Served), так и в случайном порядке. Существуют СМО «с приоритетами», где некоторые заявки (VIP-заявки) обслуживаются предпочтительно перед другими («ветераны войны вне очереди»).

Рассмотрим смешанную СМО (с очередью) X с n каналами, на вход которой поступает простейший поток заявок с плотностью λ [5]. Длительность обслуживания одной заявки $T_{\text{об}}$ имеет показательный ЗР с параметром $\mu = 1/m_{t_{\text{об}}}$. Заявка, заставшая все каналы занятыми, становится в очередь и ожидает обслуживания. При этом время ожидания ограничено некоторым сроком $T_{\text{ож}}$, и если до истечения этого срока заявка не будет принята к обслуживанию, то она покидает очередь и остается не обслуженной. Длительность ожидания $T_{\text{ож}}$ – случайная величина, распределенная по показательному закону

$$h(t) = ve^{-vt} \quad (t > 0), \tag{3.15}$$

где v – величина, обратная средней длительности ожидания: $v = 1/m_{t_{\text{ож}}}$, $m_{t_{\text{ож}}} = M\{T_{\text{ож}}\}$.

Параметр v можно интерпретировать как плотность «потока уходов» заявки, стоящей в очереди. При $v \to \infty$ система смешанного типа превращается в чистую систему с отказами; при $v \to 0$ – в чистую систему с ожиданием.

При показательном ЗР длительности ожидания пропускная способность системы не зависит от того, обслуживаются ли заявки в порядке очереди или в случайном порядке: для каждой заявки ЗР остав-

шегося времени ожидания не зависит от того, сколько времени заявка уже стояла в очереди (п. 2.1).

Так как все потоки событий (с параметрами λ, μ и v) – пуассоновские, то процесс изменения состояний системы будет *марковским*. Получим уравнения для вероятностей состояний системы. Пронумеруем состояния не по числу занятых каналов, а по числу связанных с системой заявок. Заявку будем называть *«связанной с системой»*, если она либо находится в состоянии обслуживания, либо ожидает очереди.

Возможные состояния системы (рис. 3.5):

x_0 – ни один канал не занят (очереди нет);

x_1 – занят ровно один канал (очереди нет);

. . . . . .

x_k – занято ровно k каналов (очереди нет);

. . . . . .

x_n – заняты все n каналов (очереди нет);

x_{n+1} – заняты все n каналов, одна заявка стоит в очереди;

. . . . . .

x_{n+s} – заняты все n каналов, s заявок стоят в очереди.

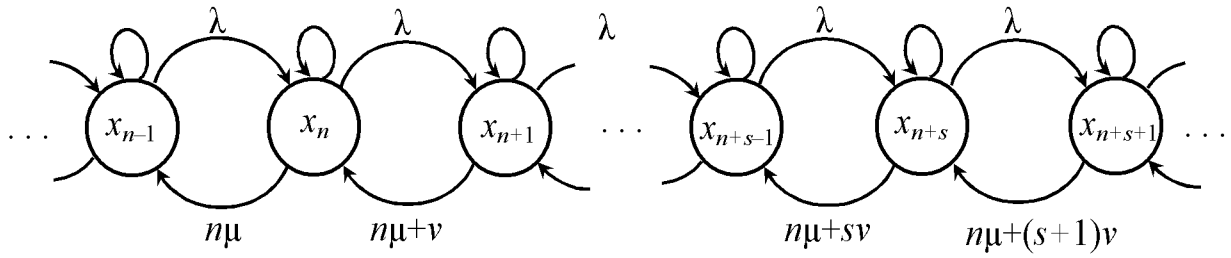


Рис. 3.5. Диаграмма переходов для n-канальной СМО с очередью

В данном случае число заявок s, стоящих в очереди, может быть сколь угодно большим – система X имеет бесконечное (хотя и счетное) множество состояний.

Первые n ДУ совпадают с ДУ (3.5) для системы с отказами:

$$\begin{cases} \dfrac{d\,p_0(t)}{d\,t} = -\lambda p_0(t) + \mu p_1(t), \\ \cdots\cdots\cdots\cdots\cdots \\ \dfrac{d\,p_k(t)}{d\,t} = \lambda p_{k-1}(t) - (\lambda + k\mu)\,p_k(t) + (k+1)\mu p_{k+1}(t) \\ \cdots\cdots\cdots\cdots\cdots \\ \dfrac{d\,p_{n-1}(t)}{d\,t} = \lambda p_{n-2}(t) - \left[\lambda + (n-1)\mu\right] p_{n-1}(t) + n\mu p_n(t). \end{cases},$$

Отличие данной системы уравнений от системы уравнений Эрланга из п. 3.1 начинается при $k = n$, когда система с ожиданием может перейти в состояние x_n не только из x_{n-1}, но и из x_{n+1} (все каналы заняты и одна заявка стоит в очереди) (рис. 3.6).

Составим ДУ для $p_n(t)$. Для момента t найдем $p_n(t+\Delta t)$ – вероятность того, что система в момент $t+\Delta t$ будет в состоянии x_n. Эта вероятность вычисляется как вероятность суммы трех событий $p_n(t+\Delta t) \approx P(A) + P(B) + P(C)$:

A – в момент t система уже была в состоянии x_n и за время Δt не вышла из него (не поступило ни одной заявки и ни один из каналов не освободился);

B – в момент t система была в состоянии x_{n-1} и за время Δt перешла в состояние x_n (поступила одна заявка);

C – в момент t система была в состоянии x_{n+1} (все каналы заняты и одна заявка стоит в очереди) и за время Δt перешла в x_n (либо освободился один канал и стоящая в очереди заявка заняла его, либо стоящая в очереди заявка ушла в связи с окончанием срока ожидания).

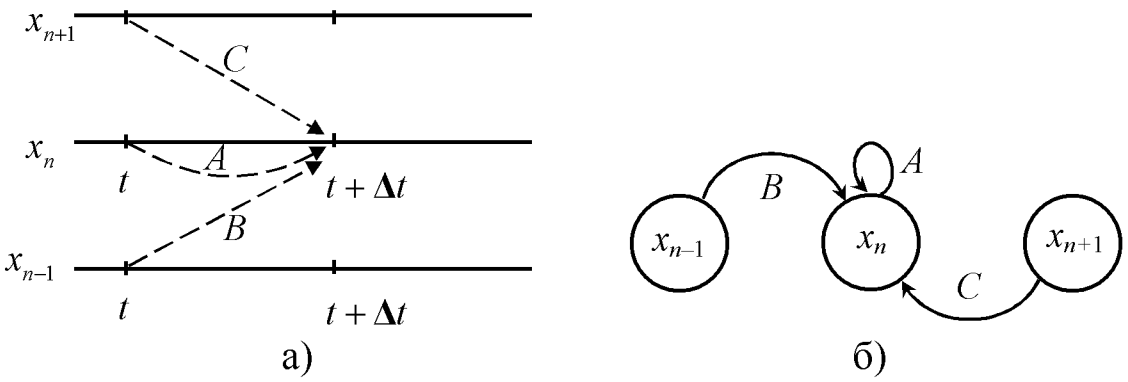


Рис. 3.6. Временная диаграмма (а) и участок диаграммы переходов (б) для состояния x_n

Найдем $P(A)$, вычислив сначала вероятность того, что за время Δt не придет ни одной заявки и не освободится ни один из n каналов:

$$P_{x_n \to x_n}(\Delta t) = P_0^{(\lambda)} \left[P_0^{(\mu)} \right]^n = e^{-\lambda \Delta t} \left(e^{-\mu \Delta t} \right)^n = e^{-(\lambda + n\mu)\Delta t}.$$

Учитывая, что при $\Delta t \to 0$ имеем $e^{-(\lambda+n\mu)\Delta t} \approx 1 - (\lambda + n\mu)\Delta t$, получим

$$P(A) \approx p_n(t) P_{x_n \to x_n}(\Delta t) = p_n(t) \left[1 - (\lambda + n\mu)\Delta t \right].$$

Аналогично, с учетом того, что $(\Delta t)^2 << \Delta t$, имеем

$$P(B) = p_{n-1}(t) P_1^{(\lambda)}(\Delta t) = p_{n-1}(t) \lambda \Delta t \, e^{-\lambda \Delta t} \approx p_{n-1}(t) \lambda \Delta t (1 - \lambda \Delta t) \approx p_{n-1}(t) \lambda \Delta t,$$

$$P(C) = p_{n+1}(t) \left[n P_1^{(\mu)} + P_1^{(v)} \right] = p_{n+1}(t) \left[n\mu \Delta t \, e^{-\mu \Delta t} + v \Delta t \, e^{-v \Delta t} \right] \approx$$

$$\approx p_{n+1}(t) \left[n\mu \Delta t (1 - \mu \Delta t) + v \Delta t (1 - v \Delta t) \right] \approx p_{n+1}(t) (n\mu + v) \Delta t,$$

где $n P_1^{(\mu)}$ – вероятность освобождения хотя бы одного из n каналов; $P_1^{(v)}$ – вероятность ухода из очереди одной заявки.

В итоге получаем

$$p_n(t + \Delta t) \approx p_n(t) \left[1 - \lambda \Delta t - n\mu \Delta t \right] + p_{n-1}(t) \lambda \Delta t + p_{n+1}(t) (n\mu + v) \Delta t.$$

Откуда по аналогии с (3.3) имеем

$$\frac{d\, p_n(t)}{dt} = -(\lambda + n\mu) p_n(t) + \lambda p_{n-1}(t) + (n\mu + v) p_{n+1}(t). \qquad (3.16)$$

Составим ДУ для $p_{n+s}(t + \Delta t)$ при любом $s > 0$ – вероятность того, что в момент $t + \Delta t$ все n каналов будут заняты и s заявок будут в очереди. Эта вероятность вычисляется как вероятность суммы трех событий $p_{n+s}(t + \Delta t) \approx P(A) + P(B) + P(C)$ (рис. 3.7):

A – в момент t система уже была в состоянии x_{n+s}, а за время Δt это состояние не изменилось (не поступило ни одной заявки, ни один канал не освободился и ни одна из s стоящих в очереди заявок не ушла);

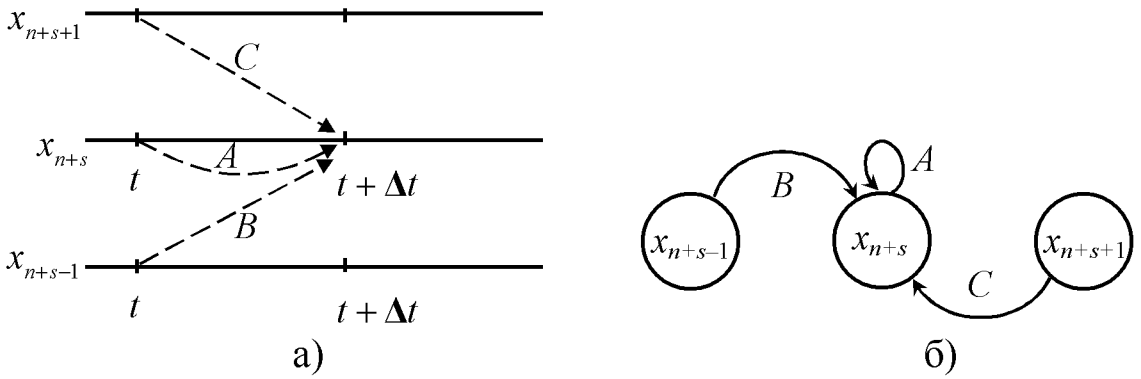


Рис. 3.7. Временная диаграмма (а) и участок диаграммы переходов (б) для состояния x_{n+s}

B – в момент t система была в состоянии x_{n+s-1}, а за время Δt перешла в состояние x_{n+s} (поступила одна заявка);

C – в момент t система была в состоянии x_{n+s+1}, а за время Δt перешла в состояние x_{n+s} (для этого либо один из каналов должен освободиться, и тогда одна из $s+1$ стоящих в очереди заявок займет его, либо одна из стоящих в очереди заявок должна уйти в связи с окончанием срока ожидания).

Найдем $P(A)$, вычислив сначала вероятность того, что за время Δt не поступит ни одной заявки, не освободится ни один из n каналов и ни одна из s стоящих в очереди заявок не уйдет:

$$P_{x_{n+s}\to x_{n+s}}(\Delta t)=P_0^{(\lambda)}\left[P_0^{(\mu)}\right]^n\left[P_0^{(v)}\right]^s=e^{-\lambda\Delta t}\left(e^{-\mu\Delta t}\right)^n\left(e^{-v\Delta t}\right)^s=e^{-(\lambda+n\mu+sv)\Delta t}.$$

Учитывая, что при $\Delta t\to 0$ выполняется $e^{-(\lambda+n\mu+sv)\Delta t}\approx 1-(\lambda+n\mu+sv)\Delta t$, получим

$$P(A)\approx p_{n+s}(t)P_{x_{n+s}\to x_{n+s}}(\Delta t)=p_{n+s}(t)\left[1-(\lambda+n\mu+sv)\Delta t\right].$$

Аналогично, с учетом того, что $(\Delta t)^2<<\Delta t$, имеем

$$P(B)=p_{n+s-1}(t)P_1^{(\lambda)}(\Delta t)=p_{n+s-1}(t)\lambda\Delta t\,e^{-\lambda\Delta t}\approx p_{n+s-1}(t)\lambda\Delta t(1-\lambda\Delta t)\approx p_{n+s-1}(t)\lambda\Delta t,$$

$$P(C)=p_{n+s+1}(t)\left[nP_1^{(\mu)}+(s+1)P_1^{(v)}\right]=p_{n+s+1}(t)\left[n\mu\Delta t\,e^{-\mu\Delta t}+(s+1)v\Delta t\,e^{-v\Delta t}\right]\approx$$

$$\approx p_{n+s+1}(t)\left[n\mu\Delta t(1-\mu\Delta t)+(s+1)v\Delta t(1-v\Delta t)\right]\approx p_{n+s+1}(t)\left[n\mu+(s+1)v\right]\Delta t,$$

где $nP_1^{(\mu)}$ – вероятность освобождения хотя бы одного из n каналов;
$(s+1)P_1^{(v)}$ – вероятность ухода из очереди одной из $s+1$ заявок.

Следовательно,

$$p_{n+s}(t+\Delta t)\approx p_{n+s}(t)\left[1-\lambda\Delta t-n\mu\Delta t-sv\Delta t\right]+p_{n+s-1}(t)\lambda\Delta t+p_{n+s+1}(t)\left[n\mu+(s+1)v\right]\Delta t,$$

откуда

$$\frac{d\,p_{n+s}(t)}{dt}=(-\lambda+n\mu+sv)\,p_{n+s}(t)+\lambda p_{n+s-1}(t)+\left[n\mu+(s+1)v\right]p_{n+s+1}(t). \quad (3.17)$$

Таким образом, мы получили для вероятностей состояний систему бесконечного числа ДУ:

$$\begin{cases} \dfrac{d\,p_0(t)}{d\,t} = -\lambda p_0(t) + \mu p_1(t), \\ \dfrac{d\,p_1(t)}{d\,t} = \lambda p_0(t) - (\lambda+\mu)\,p_1(t) + 2\mu p_2(t), \\ \cdots\cdots\cdots\cdots\cdots \\ \dfrac{d\,p_k(t)}{d\,t} = \lambda p_{k-1}(t) - (\lambda + k\mu)\,p_k(t) + (k+1)\mu p_{k+1}(t),\ 1 \le k \le n-1, \\ \cdots\cdots\cdots\cdots\cdots \\ \dfrac{d\,p_n(t)}{d\,t} = \lambda p_{n-1}(t) - [\lambda + n\mu]\,p_n(t) + (n\mu + v)\,p_{n+1}(t), \\ \cdots\cdots\cdots\cdots\cdots \\ \dfrac{d\,p_{n+s}(t)}{d\,t} = \lambda p_{n+s-1}(t) - [\lambda + n\mu + sv]\,p_{n+s}(t) + [n\mu + (s+1)v]\,p_{n+s+1}(t). \end{cases} \tag{3.18}$$

На практике вероятности $p_{n+s}(t)$ при возрастании s становятся пренебрежимо малыми, и соответствующие уравнения могут быть отброшены.

Получим формулы Эрланга для системы с ожиданием в установившемся режиме обслуживания (при $t \to \infty$). Из уравнений (3.18), полагая все $p_k(k = 0, 1, \ldots, n, \ldots, n+s, \ldots)$ постоянными, а все производные равными нулю, получим систему алгебраических уравнений:

$$\begin{cases} -\lambda p_0 + \mu p_1 = 0, \\ \lambda p_0 - (\lambda+\mu)\,p_1 + 2\mu p_2 = 0, \\ \cdots\cdots\cdots\cdots\cdots \\ \lambda p_{k-1} - (\lambda + k\mu)\,p_k + (k+1)\mu p_{k+1} = 0,\ 1 \le k \le n-1, \\ \cdots\cdots\cdots\cdots\cdots \\ \lambda p_{n-1} - [\lambda + n\mu]\,p_n + (n\mu + v)\,p_{n+1} = 0, \\ \cdots\cdots\cdots\cdots\cdots \\ \lambda p_{n+s-1} - [\lambda + n\mu + sv]\,p_{n+s} + (n\mu + (s+1)v)\,p_{n+s+1} = 0, \\ \cdots\cdots\cdots\cdots\cdots \end{cases} \tag{3.19}$$

К ним необходимо добавить условие нормировки $\sum_{k=0}^{\infty} p_k = 1$. Решив систему (3.19) для любого $k \le n$, получим: $p_k = \dfrac{\lambda^k}{k!\mu^k} p_0$. Вероятность того, что СМО будет полностью занята и в буфере будет очередь из s заявок

$$p_{n+s} = \frac{\lambda^{n+s} p_0}{n!\,\mu^n \prod_{m=1}^{s} (n\mu + mv)},\quad k > n,\ s \ge 1\ (k = n+s).$$

Для $k \le n$ и $s \ge 1$ величину p_0 можно выразить через основные параметры СМО:

$$p_0 = \frac{1}{\sum_{i=0}^{n} \frac{\lambda^i}{i!\mu^i} + \sum_{s=1}^{\infty} \frac{\lambda^{n+s}}{n!\,\mu^n \prod_{m=1}^{s}(n\mu + mv)}}.$$

Преобразуем выражения для p_k, p_{n+s} и p_0, введя в них приведенные плотности:

$$\left.\begin{aligned} \lambda/\mu = \lambda m_{t_{об}} = \rho, \\ v/\mu = v m_{t_{об}} = \beta, \end{aligned}\right\}$$

Параметр β выражает количество уходов заявок, стоящих в очереди, приходящееся на среднее время обслуживания одной заявки.

В итоге получим

$$p_k = \frac{\rho^k}{k!} p_0 \quad (0 < k \le n), \quad p_{n+s} = \frac{\frac{\rho^{n+s}}{n!} p_0}{\prod_{m=1}^{s}(n + m\beta)} \quad (s \ge 1), \quad p_0 = \frac{1}{\sum_{i=0}^{n} \frac{\rho^i}{i!} + \frac{\rho^n}{n!} \sum_{s=1}^{\infty} \frac{\rho^s}{\prod_{m=1}^{s}(n + m\beta)}}.$$

Подставляя p_0 в формулы для p_k и p_{n+s}, получим выражения для вероятностей состояний системы:

$$p_k = \frac{\rho^k / k!}{\sum_{i=0}^{n} \frac{\rho^i}{i!} + \frac{\rho^n}{n!} \sum_{s=1}^{\infty} \frac{\rho^s}{\prod_{m=1}^{s}(n + m\beta)}} \quad (0 \le k \le n), \quad p_{n+s} = \frac{\frac{\rho^n}{n!} \frac{\rho^s}{\prod_{m=1}^{s}(n + m\beta)}}{\sum_{i=0}^{n} \frac{\rho^i}{i!} + \frac{\rho^n}{n!} \sum_{s=1}^{\infty} \frac{\rho^s}{\prod_{m=1}^{s}(n + m\beta)}} \quad (s \ge 1).$$

Определим вероятность P_H того, что заявка покинет систему необслуженной из следующих соображений: в установившемся режиме вероятность $P_H = \bar{N}_{\text{ух}} / \bar{N}_{\text{пост}}$ – это отношение среднего числа заявок $\bar{N}_{\text{ух}}$, *уходящих* из очереди в единицу времени, к среднему числу заявок $\bar{N}_{\text{пост}}$, *поступающих* в единицу времени.

Найдем среднее число заявок, уходящих из очереди в единицу времени. Для этого сначала вычислим математическое ожидание m_s числа заявок, *находящихся* в очереди (среднее число заявок в очереди n-канальной СМО)

$$m_s = M\{s\} = \sum_{s=1}^{\infty} s p_{n+s} = \frac{\dfrac{\rho^n}{n!}\displaystyle\sum_{s=1}^{\infty}\frac{s\rho^s}{\displaystyle\prod_{m=1}^{s}(n+m\beta)}}{\displaystyle\sum_{i=0}^{n}\frac{\rho^i}{i!}+\frac{\rho^n}{n!}\sum_{s=1}^{\infty}\frac{\rho^s}{\displaystyle\prod_{m=1}^{s}(n+m\beta)}}\ . \qquad (3.20)$$

Чтобы получить P_H, нужно m_s умножить на коэффициент $\frac{v}{\lambda}=\frac{v/\mu}{\lambda/\mu}=\frac{\beta}{\rho}$. В результате получим: $P_H=\frac{\beta}{\rho}m_s$. Пропускная способность системы $q=1-P_H$ характеризуется вероятностью того, что заявка, попавшая в систему, будет обслужена.

Очевидно, что пропускная способность системы с ожиданием, при тех же λ и μ, будет всегда выше, чем пропускная способность системы с отказами: в случае наличия ожидания необслуженными уходят не все заявки, заставшие n каналов занятыми, а только некоторые. Пропускная способность увеличивается при увеличении среднего времени ожидания $m_{t_{ож}}=1/v$.

При $\beta\to\infty$ система с ожиданием превращается в систему с отказами (заявка мгновенно уходит из очереди), формулы для p_k превратятся в формулы Эрланга В (3.9).

В чистой системе с ожиданием ($\beta\to 0$) заявки вообще не уходят из очереди, и поэтому $P_H=0$ и каждая заявка рано или поздно дождется обслуживания. Однако в такой системе не всегда имеется предельный стационарный режим при $t\to\infty$, такой режим существует только при $\rho<n$, т. е. когда среднее число заявок, приходящееся на время обслуживания одной заявки, не выходит за пределы возможностей n-канальной системы. При $\rho\ge n$ число заявок в очереди будет неограниченно возрастать со временем, что видно из (3.20) при $\beta=0$.

Найдем предельные вероятности $p_k\ (0\le k\le n)$ для чистой системы с ожиданием при $\rho<n$ $(\beta\to 0)$:

$$p_0=\frac{1}{\displaystyle\sum_{i=0}^{n}\frac{\rho^i}{i!}+\frac{\rho^n}{n!}\sum_{s=1}^{\infty}\frac{\rho^s}{n^s}}=\frac{1}{\displaystyle\sum_{i=0}^{n}\frac{\rho^i}{i!}+\frac{\rho^{n+1}}{n!(n-\rho)}},$$

просуммировав прогрессию (это возможно только при $\rho<n$).

Из формул для p_k и p_{n+s} получаем при $k=n$ **формулу Эрланга С**

$$p_k = \frac{\rho^k/k!}{\sum_{i=0}^{n}\frac{\rho^i}{i!}+\frac{\rho^{n+1}}{n!(n-\rho)}}\ (0\le k\le n),\quad p_{n+s} = \frac{\frac{\rho^{n+s}}{n!n^s}}{\sum_{i=0}^{n}\frac{\rho^i}{i!}+\frac{\rho^{n+1}}{n!(n-\rho)}}\ ,\ k=n+s\ (s\ge 0).$$

Среднее число заявок в очереди при $\beta \to 0$

$$m_s = \frac{\rho^{n+1}}{nn!\left(1-\frac{\rho}{n}\right)^2\left[\sum_{i=0}^{n}\frac{\rho^i}{i!}+\frac{\rho^{n+1}}{n!(n-\rho)}\right]}.$$

Если входной поток является простейшим, то формулы Эрланга остаются справедливыми при любом законе распределения времени обслуживания.

Пример. На вход трехканальной ($n=3$) СМО с неограниченным временем ожидания поступает простейший поток заявок с плотностью $\lambda=4$ заявки/ч. Среднее время обслуживания одной заявки $m_{t_{об}}=30$ мин. Определить, существует ли установившийся режим обслуживания; если да, то найти вероятности p_0, p_1, p_2, p_3, вероятность наличия очереди и среднюю длину очереди m_s.

Решение. Имеем $\mu = 1/m_{t_{об}} = 2$; $\rho = \lambda/\mu = 2$. Так как $\rho < n$, установившийся режим существует. Вероятности состояний $p_0 = 1/9$; $p_1 = 2/9$; $p_2 = 2/9$; $p_3 = 8/54 \approx 0.148$. Вероятность величия очереди $P_{оч} = 1-(p_0+p_1+p_2+p_3) = 0.297$. Средняя длина очереди $m_s \approx 0.89$ заявки.

3.4. Контрольные вопросы

1. Дайте определение чистой системы с отказами.
2. Дайте определение чистой системы с ожиданием.
3. Нарисуйте и поясните диаграмму интенсивностей переходов для одноканальной СМО с отказами.
4. Нарисуйте и поясните диаграмму интенсивностей переходов для одноканальной СМО с очередью (объем буфера равен 2).
5. Что позволяют определить формулы Эрланга?
6. В чем отличие формул Эрланга от уравнений Эрланга?
7. Запишите формулу для вероятности простоя (все каналы свободны) для одноканальной СМО с отказами.

8. Запишите формулу для вероятности простоя (все каналы свободны) для двухканальной СМО с отказами.
9. Каковы начальные условия для уравнений Эрланга?
10. Что позволяет определить формула Эрланга B?
11. Запишите выражение для распределения Эрланга?
12. Запишите формулу для вероятности простоя (все каналы свободны) для одноканальной СМО с очередью.
13. Запишите формулу для вероятности простоя (все каналы свободны) для двухканальной СМО с очередью.
14. Как влияет размер буфера на вероятность отказа в СМО?
15. Как вычисляется средняя длина очереди в одноканальной СМО?
16. Как вычисляется средняя длина очереди в двухканальной СМО?
17. Что позволяет определить формула Эрланга C?
18. Каково условие предельного стационарного режима в чистой системе с ожиданием? Что происходит, если это условие не выполняется?
19. Поясните термин «интенсивность обслуженной нагрузки».
20. Поясните термин «интенсивность потенциальной нагрузки».
21. Что представляет собой относительная пропускная способность СМО?
22. Какие ограничения могут наложены на параметры СМО, связанные с ожиданием заявок в очереди?
23. Какие дисциплины очереди вы знаете? Поясните их сущность.
24. В какую СМО превращается система с ожиданием при $\beta \to \infty$?
25. В какую СМО превращается система с ожиданием при $\beta \to 0$?
26. При каком условии в чистой системе с ожиданием существует предельный стационарный режим?
27. Что происходит в чистой системе с ожиданием при выполнении условия $\rho \geq n$?

3.5. Задачи

1. Вычислите вероятность отказа для одноканальной СМО без очереди при нагрузке $\rho = 0.7$.
2. Чему равна нагрузка ρ в одноканальной СМО без очереди, если вероятность отказа равна $P_{\text{отк}} = 0.02$?
3. Вычислите вероятность отказа для одноканальной СМО без очереди, если интенсивность потока заявок $\lambda = 3$ выз/мин, а характеристика обслуживания μ=4 выз/мин ?

4. Чему равна нагрузка ρ в СМО, если интенсивность потока заявок $\lambda = 4$ выз/мин, а средняя длительность обслуживания $m_{t_{об}} = 0.2$ мин?
5. На вход системы с ожиданием, бесконечной длиной очереди и 3 каналами поступает простейший поток интенсивности $\lambda = 3$ выз/мин. Характеристика обслуживания каждого канала μ =4 выз/мин. Вычислить вероятность того, что в системе будут заняты все каналы.
6. На вход двухканальной СМО с неограниченным временем ожидания поступает простейший поток заявок с плотностью $\lambda =$ 3 выз/час. Среднее время обслуживания одной заявки $m_{t_{об}} = 0.2$ часа. Определить, существует ли установившийся режим обслуживания; если да, то найти среднюю длину очереди m_s.

4. ОБСЛУЖИВАНИЕ ПРИМИТИВНОГО ПОТОКА ЗАЯВОК

4.1. Распределения Эрланга, Энгсета, Бернулли и Пуассона

Для того чтобы получить характеристики обслуживания примитивного потока необходимо вначале рассмотреть вероятности потерь p_k при пуассоновском потоке заявок с условным параметром λ_i [12, с.80].

Рассмотрим частный случай уравнения (1.6), когда $\mu_i = i\mu$ и

$$p_k = \frac{\lambda_0\lambda_1\lambda_2 \ldots \lambda_{k-1}}{\mu^k k!} p_0 = \left[\prod_{i=0}^{k-1} \frac{\rho_i}{k!}\right] p_0$$

и с учетом условия нормировки (1.5) имеем

$$p_0 = \frac{1}{1 + \sum_{k=1}^{m} \prod_{i=0}^{k-1} \frac{\rho_i}{k!}}. \qquad (4.1)$$

Вероятность занятости k каналов

$$p_k = \left[\prod_{i=0}^{k-1} \frac{\rho_i}{k!}\right] \Big/ \sum_{k=0}^{m} \prod_{i=0}^{k-1} \frac{\rho_i}{k!}, \quad k = 0,1, \ldots, m. \qquad (4.2)$$

Выражение (4.2) представляет собой вероятность занятости k выходов полнодоступной неблокируемой системы в произвольный момент времени t при обслуживании с явными потерями пуассоновского потока заявок с условным параметром λ_i. Полученный результат (4.2) является частным случаем процесса «размножения-гибели» (п. 1.3.2).

Рассмотрим ряд логических трактовок вероятности p_k в формуле (4.2) [12]:

1. По ансамблю коммутационных систем (КС). Пусть имеется n ($n \to \infty$) одинаковых независимо функционирующих полнодоступных КС с m выходами и каждая система обслуживает с явными потерями пуассоновский поток заявок с условным параметром λ_i. Тогда вероятность p_k представляет собой долю систем, в которых в произвольный момент времени t занято точно k выходов, т. е.

$$p_k = \lim_{n\to\infty} n_k / n,$$

где n_k – число систем с k занятыми выходами в момент t.

2. По ансамблю моментов времени. Если фиксируются состояния КС в n_t ($n_t \to \infty$) произвольных моментов времени, то p_k есть доля моментов, в которых система находится в состоянии k, т. е. $p_k = \lim\limits_{n_t \to \infty} n_{tk} / n_t$, где n_{tk} – число моментов, когда зафиксировано состояние системы k.

3. По времени. Вероятность p_k представляет собой долю времени (на бесконечном интервале), в течение которого в рассматриваемой КС занятно ровно k выходов.

Рассмотрим основные частные случаи распределения (4.2).

1. Модель Mi/G/m, т. е. обслуживаемый поток заявок примитивный с параметром $\lambda_k = \alpha(N-k)$. Тогда

$$\prod_{i=0}^{k-1} \frac{\lambda_i}{k!} = \frac{\alpha N \cdot \alpha(N-1) \cdot \alpha(N-2) \cdot \ldots \cdot \alpha(N+1-k)}{k!} = C_N^k \alpha^k, \qquad C_N^k = \frac{N!}{k!(N-k)!}.$$

Подстановка этого выражения в (4.2) дает **распределение Энгсета** (усеченное распределение Бернулли):

$$p_k = \frac{C_N^k \alpha^k}{\sum\limits_{j=0}^{m} C_N^j \alpha^j}. \qquad (4.3)$$

2. Модель M/G/m, т. е. обслуживаемый поток заявок простейший с параметрами λ и μ. Из (4.2) непосредственно следует **первое распределение Эрланга** (усеченное распределение Пуассона):

$$p_k = \frac{\rho^k / k!}{\sum\limits_{j=0}^{m} \rho^j / j!}. \qquad (4.4)$$

Распределение Эрланга получается также из распределения Энгсета, если устремить $N \to \infty$, а $\alpha \to 0$, но так, что $N\alpha = \rho$.

3. Модель Mi/G/N, т. е. обслуживаемый поток заявок примитивный с параметром $\lambda_k = \alpha(N-k)$, способ обслуживания без потерь (число выходов в системе $m = N$). При этом выражение (4.3) с учетом бинома Ньютона

$$\sum_{j=0}^{N} C_N^j a^j b^{N-j} = (a+b)^N$$

принимает вид

$$p_k = \frac{C_N^k \alpha^k}{\sum_{j=0}^{N} C_N^j \alpha^j} = \frac{C_N^k \alpha^k}{(1+\alpha)^N}.$$

Обозначив $a = \alpha/(1+\alpha)$, можно получить **распределение Бернулли**:

$$p_k = C_N^k a^k (1-a)^{N-k}. \qquad (4.5)$$

При числе выходов $m = N$ за каждым источником как бы закрепляется определенный выход, поэтому занятие любого выхода происходит независимо от других. Если исследуется состояние СМО в произвольный момент времени, то каждое занятие очередного выхода можно рассматривать как очередное успешное испытание из общего числа N независимых испытаний. Этим объясняется, что в данном случае распределение вероятностей числа занятых выходов совпадает с распределением Бернулли для независимых испытаний. Величина a определяет вероятность занятия определенного выхода.

4. Модель M/G/∞, т. е. обслуживаемый поток заявок простейший с параметрами λ и μ, обслуживание – без потерь ($m \to \infty$). Из распределения (4.4) с учетом разложения экспоненциальной функции в ряд Маклорена $e^{\rho} = \sum_{j=0}^{\infty} \rho^j / j!$ получаем **распределение Пуассона**:

$$p_k = \frac{\rho^k / k!}{\sum_{j=0}^{\infty} \rho^j / j!} = \left(\rho^k / k!\right) e^{-\rho}. \qquad (4.6)$$

Распределение Пуассона (4.6) можно также получить из распределения Бернулли (4.5) при $N \to \infty$, а $a \to 0$, но так, что $Na = \rho$.

Таким образом, наиболее общим из рассматриваемых распределений является распределение Энгсета. Из него следует, с одной стороны, первое распределение Эрланга, а с другой – распределение Бернулли. Из последних двух разными способами можно получить распределение Пуассона. Во всех четырех распределениях параметры λ, α, a, определяющие нагрузку, выражены в заявках в единицу времени.

4.2. Распределение Энгсета

Вероятности состояний трехлинейной системы P_i, $i = 0,\ldots,3$, рассчитанные в соответствии с распределением Энгсета (4.3) в зависимо-

сти от интенсивности обслуженной Y (0…3 Эрл), потенциальной A (0…10 Эрл) и поступающей λ (0…∞ Эрл) нагрузок для N =10 источников, показаны на рис. 4.1, где также представлены зависимости вероятностей трех видов потерь от нагрузки. Сравнение со случаем первого распределения Эрланга [12, с.83] показывает, что в данном случае кривая, характеризующая зависимость $P_{\text{в}} = f(Y)$, проходит несколько ниже. Это объясняется последействием потока: уменьшением числа поступающих вызовов с ростом числа занятых линий.

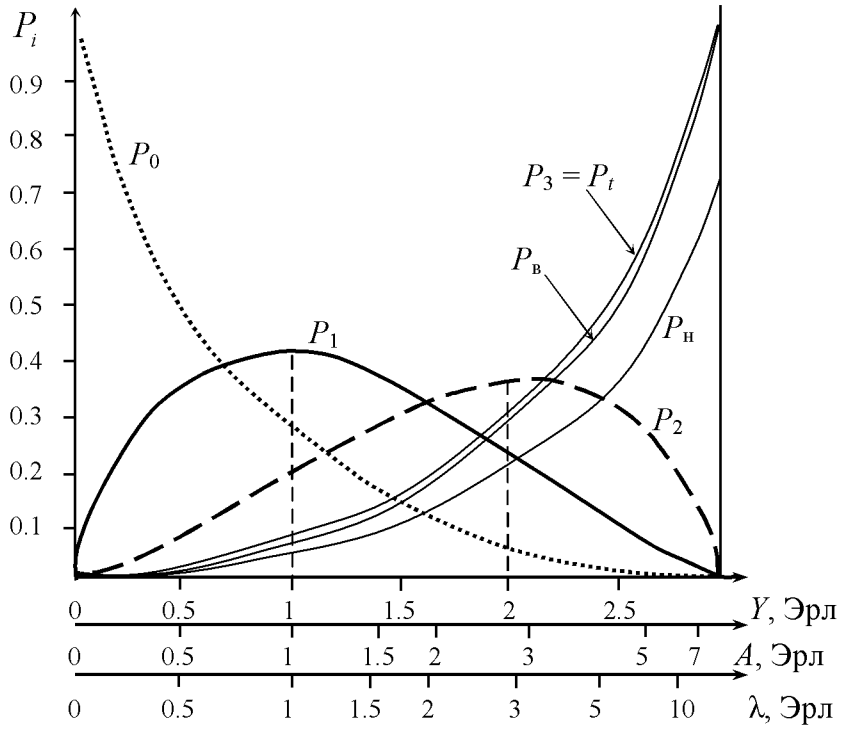


Рис. 4.1. Зависимость вероятностей состояний P_i (распределение Энгсета) и вероятностей потерь от интенсивности нагрузки для трехлинейной системы при числе источников N =10

Наибольшего значения вероятность P_i достигает при $Y = i$. Пересечение кривых P_i и P_{i-1} в соответствии с рекуррентным соотношением (1.6) происходит в точке $\lambda_{i-1} = \alpha(N+1-i) = i$. Поскольку на рис. 4.1 приведена усредненная по всем состояниям интенсивность поступающей нагрузки $\lambda = \sum_{i=0}^{3} \lambda_i P_i$, то точки пересечения кривых P_i и P_{i-1} не

выделены. Вероятность P_i достигает максимума при $i=\left[\lambda_{i-1}\right]$, причем если λ_{i-1} целое число, то имеют место две одинаковые наибольшие вероятности P_i и P_{i-1}.

Рассмотрим основные характеристики качества обслуживания.

1. Вероятность потерь по времени по аналогии с (3.10) определяется по формуле

$$P_t = P_n = \frac{C_N^n \alpha^n}{\sum_{j=0}^{n} C_N^j \alpha^j} = \xi(N,n,\alpha). \qquad (4.7)$$

Выражение (4.7) называют **формулой Энгсета**. Для ее вычисления можно построить алгоритмы, аналогичные алгоритмам, предложенным для функции $E_n(\rho)$.

2. Вероятность потери вызова:

$$P_{\text{в}} = \frac{R}{\lambda} = \frac{\lambda_n P_n}{\sum_{j=0}^{n} \lambda_j P_j} = \frac{C_{N-1}^n \alpha^n}{\sum_{j=0}^{n} C_{N-1}^j \alpha^j}. \qquad (4.8)$$

Сравнение формул (4.7) и (4.8) показывает, что всегда имеет место неравенство $P_t > P_{\text{в}}$. Равенство наступает только в предельном случае при $N \to \infty$, когда примитивный поток переходит в простейший. Формула (4.8) может быть получена непосредственно из (4.7) путем логических рассуждений. Вызов, поступивший от конкретного свободного источника, будет потерян, если в этот момент заняты все n выходов. Эта занятость должна быть обеспечена остальными $N-1$ источниками. Поскольку поступление вызова от свободного источника случайно (промежуток свободности распределен по показательному закону), то имеет место $P_{\text{в}}(N,n,\alpha) = P_t(N-1,n,\alpha)$.

3. Интенсивность обслуженной нагрузки

$$Y = \sum_{i=1}^{n} iP_i = \sum_{i=0}^{n-1} \lambda_i P_i = \frac{\alpha N \sum_{i=0}^{n-1} C_{N-1}^i \alpha^i}{\sum_{j=0}^{n} C_N^j \alpha^j} = \frac{\alpha N(1-P_{\text{в}})}{1+\alpha(1-P_{\text{в}})}. \qquad (4.9)$$

С ростом числа источников N пропускная способность системы снижается, приближаясь в пределе (при $N \to \infty$) к пропускной способности системы, обслуживающей простейший поток.

4. Интенсивность поступающей нагрузки (математическое ожидание параметра потока вызовов):

$$\lambda = \sum_{i=0}^{n} \lambda_i P_i = \sum_{i=0}^{n} \alpha\left(N - i\right) P_i = \alpha\left(N - Y\right) = \frac{\alpha N}{1 + \alpha\left(1 - P_{\text{в}}\right)}. \qquad (4.10)$$

5. Интенсивность потенциальной нагрузки

$$A = \sum_{i=1}^{N} i\, P_i = \sum_{i=0}^{N-1} \lambda_i P_i = \sum_{i=0}^{N-1} \alpha\left(N - i\right) C_N^i a^i \left(1 - a\right)^{N-i} = Na, \qquad (4.11)$$

где вероятность P_i определяется в соответствии с распределением Бернулли (3.12). Таким образом, величина a определяет интенсивность потенциальной нагрузки от одного источника.

6. Вероятность потерь по нагрузке может быть найдена с учетом формул (4.9) и (4.11):

$$P_{\text{н}} = \frac{A_{\text{пот}}}{A} = \frac{P_{\text{в}}}{1 + \alpha\left(1 - P_{\text{в}}\right)} = \frac{C_{N-1}^n \alpha^n}{\sum_{j=0}^{n} C_N^j \alpha^j} = \frac{N - n}{N} P_t. \qquad (4.12)$$

Из (4.12) следует, что при конечном значении α всегда $P_{\text{н}} < P_{\text{в}}$. Равенство имеет место только в предельном случае, когда $\alpha \to 0$ и примитивный поток переходит в простейший. С учетом соотношения $P_{\text{в}} \le P_t$ имеет место общее неравенство $P_{\text{н}} \le P_{\text{в}} \le P_t$ (рис. 4.1). Неравенство $P_{\text{н}} \le P_{\text{в}}$ обусловлено неравенством $A \le \lambda$. Рассмотрим разность

$$\lambda - A = \frac{\alpha N}{1 + \alpha\left(1 - P_{\text{в}}\right)} - \frac{\alpha N}{1 + \alpha} = \alpha A P_{\text{н}}. \qquad (4.13)$$

Различие между интенсивностями потенциальной и поступающей нагрузок обусловлено следующей особенностью рассматриваемой модели. При получении отказа в соединении источник сразу становится свободным и наравне с другими свободными источниками получает возможность посылать новые вызовы. Параметр потока и, соответственно, интенсивность поступающей нагрузки возрастают на величину $\alpha A P_{\text{н}}$. При этом произведение $AP_{\text{н}}$ определяет интенсивность потерянной нагрузки или среднее число источников, ставших свободными после получения отказа в соединении, а α является интенсивностью потока от одного свободного источника. Если предположить, что источник при получении отказа блокируется на время обслуживания, то увеличения λ не произойдет, и будет иметь место равенство $\lambda = A$. В этом случае вероятности числа занятых источников (обслуживаемых и блокируемых) подчиняются распределению Бер-

нулли, а распределение вероятностей числа занятых линий отличается от распределения Энгсета.

Таким образом, при обслуживании примитивного потока взаимодействие источника с системой обслуживания характеризуется четырьмя параметрами нагрузки, физический смысл которых необходимо чётко различать, поскольку их численные значения неодинаковы. Неправильное применение их при расчётах может привести к заметным погрешностям, особенно в области средних и больших потерь. Рассмотрим эти параметры:

1) интенсивность потока заявок от свободного источника в единицу времени $\alpha(0\ldots\infty)$ [Эрл];

2) средняя интенсивность потока заявок от источника в единицу времени, или интенсивность поступающей от источника нагрузки $\upsilon(0\ldots\infty)$ [Эрл]. Из (4.10) имеем

$$\upsilon = \frac{\lambda}{N} = \frac{\alpha}{1+\alpha(1-P_{\text{в}})}; \tag{4.14}$$

3) интенсивность потенциальной от источника $a(0\ldots1)$ [Эрл]

$$a = \frac{\alpha}{1+\alpha}; \tag{4.15}$$

4) интенсивность обслуженной нагрузки, отнесённая к одному источнику $y(0\ldots n/N)$ [Эрл]. Из (4.9) имеем

$$y = \frac{Y}{N} = \frac{\alpha(1-P_{\text{в}})}{1+\alpha(1-P_{\text{в}})}. \tag{4.16}$$

Из сравнения (4.14) – (4.16) следует

$$y \le a \le \upsilon \le \alpha. \tag{4.17}$$

Для удобства расчетов, кроме приведенных выше соотношений, можно использовать также другие:

$$\upsilon = \alpha(1-y); \tag{4.18}$$

$$Y = \upsilon(1-P_{\text{в}}) = a(1-P_{\text{н}}); \tag{4.19}$$

$$\alpha P_{\text{н}} = \upsilon P_{\text{в}}; \tag{4.20}$$

$$\alpha = \frac{Y}{(1-Y)(1-P_{\text{в}})} = \frac{\upsilon}{1-\upsilon(1-P_{\text{в}})}; \tag{4.21}$$

$$(1-n/N)P_t = (1-Y)P_{\text{в}} = P_{\text{н}}. \tag{4.22}$$

Таким образом, если при заданных значениях N, n и одного из перечисленных выше параметров нагрузки известна одна из характе-

ристик качества обслуживания, например $P_{в}$, то остальные характеристики легко определяются. Используя соотношение $P_t(N,n,a)=P_{в}(N+1,n,a)$ по таблицам [12, 31] можно определить вероятность потерь по времени, а также интенсивности поступающей нагрузки λ в зависимости от числа источников N, числа выходов n и нормы потерь $P_{в}$.

4.3. Распределение вероятностей занятия фиксированных выходов

Полученные выше распределения вероятностей P_i характеризуют i любых занятых выходов в системе. На практике, однако, часто возникает необходимость в определении вероятности H_i занятия i определенных, фиксированных выходов. При этом несущественно, заняты только отмеченные $i\,(i=0\ldots n)$ выходов или одновременно с ними заняты еще некоторые $j\,(j=0\ldots n-i)$ выходов (рис. 4.2). Значение вероятности H_i зависит от способа занятия свободных выходов. При последовательном занятии выражение для H_i получается громоздким и поэтому здесь не приводится. При случайном равновероятном занятии вывод упрощается, поскольку вероятности занятия любой комбинации с одним и тем же числом занятых выходов одинаковы и легко вычисляются.

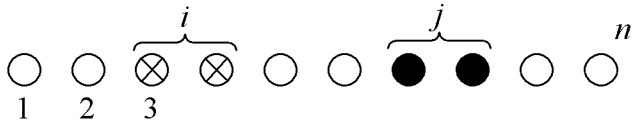


Рис. 4.2. К расчету распределения вероятностей занятия i фиксированных выходов

Если в системе с вероятностью P_{i+j} занято точно $i+j$ выходов, то вероятность занятия одной конкретной комбинации выходов равна отношению P_{i+j}/C_n^{i+j}. Отмеченные i выходов могут быть заняты совместно с любыми j выходами в C_{n-1}^{j} комбинациях. С учетом того, что j – любое число в пределах $j=0\ldots n-i$ можно получить вероятность H_i занятия фиксированных выходов:

$$H_i = \sum_{j=0}^{n-i} C_{n-i}^{j} P_{i+j} \Big/ C_n^{i+j} . \qquad (4.23)$$

Для основных распределений вероятностей P_i найдем значения соответствующих вероятностей H_i:

– распределение Эрланга

$$H_i = \sum_{j=0}^{n-i} \left[C_{n-i}^{j} \Big/ C_n^{i+j} \right] \left[\frac{\rho^{i+j}}{(i+j)! \sum_{k=0}^{n} \rho^k / k!} \right] = E_n(\rho) \Big/ E_{n-i}(\rho), \qquad (4.24)$$

где $E_n(\rho)$ – символическая запись формулы Эрланга;

– распределение Энгсета

$$H_i = \sum_{j=0}^{n-i} \left[C_{n-i}^{j} \Big/ C_n^{i+j} \right] \left[\frac{C_N^{i+j} \alpha^{i+j}}{\sum_{k=0}^{n} C_N^k \alpha^k} \right] = \xi(N,n,\alpha) \Big/ \xi(N-i,n-i,\alpha), \qquad (4.25)$$

где $\xi(N,n,\alpha)$ – символическая запись формулы Энгсета;

– распределение Бернулли

$$H_i = \sum_{j=0}^{N-i} \left[C_{n-i}^{j} \Big/ C_N^{i+j} \right] a^{i+j} (1-a)^{N-i-j} = a^i ; \qquad (4.26)$$

– для распределения Пуассона $H_i \to 0$, поскольку в этом случае $a \to 0$.

4.4. Обслуживание примитивного потока заявок. Модель Энгсета

Предположим, что СМО, рассмотренная в п.2.5, не имеет входного буфера (чистая система с отказами). Интенсивность входного (примитивного) потока линейно убывает с числом занятых серверов: $\lambda_k = \alpha(N-k)$. Максимальная нагрузка, поступающая на один сервер $\rho = \alpha/\mu$. Рассмотрим диаграмму интенсивностей переходов (рис. 4.3) и получим вероятность занятости *k* серверов в стационарном режиме. Имеем СМО типа M/M/m.

Интенсивности переходов:

$$\lambda_k = \begin{cases} (N-k)\alpha, & 0 \le k \le m-1, \\ 0, & k > m, \end{cases} \qquad \mu_k = \begin{cases} k\mu, & 0 < k \le m, \\ m\mu, & k \ge m. \end{cases}$$

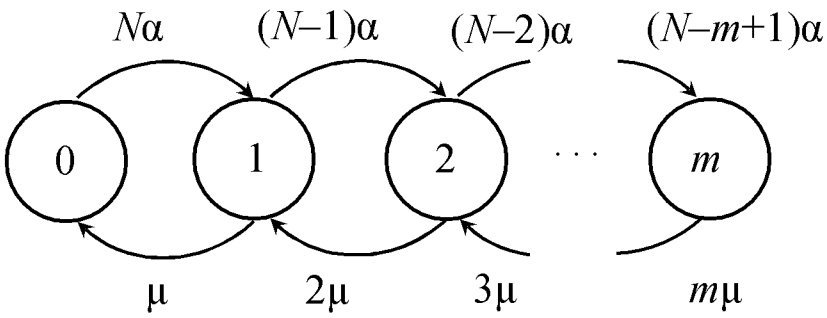


Рис. 4.3. Диаграмма интенсивностей переходов для модели Энгсета

Вероятности состояний могут быть определены из (1.6):

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} = p_0 \prod_{i=0}^{k-1} \frac{\alpha(N-i)}{(i+1)\mu} = p_0 \left(\frac{\alpha}{\mu}\right)^k \frac{N\cdot(N-1)\cdot\ldots\cdot(N-k+1)}{1\cdot 2\cdot 3\cdot\ldots\cdot k} = C_N^k p_0 \left(\frac{\alpha}{\mu}\right)^k,$$

$$p_{k+1} = \frac{\lambda_k}{\mu_{k+1}} p_k, \quad C_N^k = \frac{N!}{(N-k)!\,k!}, \quad 0 \le k \le m-1.$$

В итоге получим **распределение Энгсета**

$$p_k = \frac{C_N^k \rho^k}{\sum_{i=0}^{m} C_N^i \rho^i}, \quad k = 0,1,\ldots,m.$$

Вероятность занятости всех серверов (СМО заблокирована) $p_m = \dfrac{C_N^m \rho^m}{\sum_{i=0}^{m} C_N^i \rho^i}$. Данная формула является обобщением формулы Эрланга В (3.9).

Модель Энгсета часто применяется при небольшом числе ($N \le 300$) источников заявок, т. к. уменьшение интенсивности входного потока за счет исключения источника, получившего обслуживание, оказывается существенным. В пределе при $N \to \infty$, $\alpha \to 0$, так что $\lambda_0 = \alpha N = \text{const}$, $C_N^k \to 1/k!$ и формула Энгсета переходит в формулу Эрланга В (3.9)

$$p_k = \frac{\rho^k / k!}{\sum_{i=0}^{m} \rho^i / i!}.$$

На практике применима также **модель Молина**, или модель потерянных вызовов (Lost Calls Held - LCH) [14]. Она описывает блокировку телефонных вызовов, где блокированные заявки сохраняются в течение определенного времени задержки, но не обслуживаются.

Данная модель подобна модели, описываемой формулой Эрланга С. Вероятность блокировки для N линий, создающих интенсивность α_M для модели LCH:

$$P_B = e^{-\alpha_M} \sum_{k=N+1}^{\infty} \frac{\alpha_M^k}{k!}.$$

4.5. Контрольные вопросы

1. Что представляет собой вероятность p_k в качестве доли времени?
2. При каких условиях распределение Эрланга может быть получено из распределения Энгсета?
3. Что характеризует параметр α в распределении Энгсета?
4. При каких условиях распределение Пуассона можно получить из распределения Бернулли?
5. Чем принципиально отличаются вероятность потерь по времени и вероятность потери вызова?
6. Какие СМО описывает распределение Энгсета?
7. В каких случаях распределение Энгсета совпадает с распределением Эрланга?
8. Чем обусловлено различие между интенсивностями потенциальной и поступающей нагрузок?
9. При каком количестве источников заявок применима модель Энгсета?
10. Какой параметр СМО описывает модель Молина? Каковы при этом дополнительные условия?

5. АНАЛИЗ РАЗЛИЧНЫХ МОДЕЛЕЙ СМО

5.1. Базовая модель СМО и классификация Кендалла

Базовые составляющие математической модели СМО – это параметры входного потока заявок, характеристики очереди и размер буфера, а также параметры серверов СМО, которые имеют стандартные условные обозначения в соответствии с **классификацией Кендалла.**

Пуассоновский поток обозначается буквой **M** как типичный случай **Марковских процессов**. Серверы СМО описываются законом распределения временного интервала обслуживания одной заявки. Сервер с экспоненциальной ПРВ времени обслуживания обозначается также буквой **M** (марковский характер потока освобождений сервера).

Условные обозначения компонентов СМО:

M – марковская модель (экспоненциальное распределение интервала времени);

D – детерминированный (постоянный) интервал времени – регулярный поток;

E^k – распределение Эрланга k-го порядка;

U – равномерное (Uniform) распределение интервала времени;

G – произвольное (General) распределение интервала времени;

fBM – фрактальное броуновское движение (ФБД);

fGN – фрактальный гауссовский шум (ФГШ).

Рассмотрим примеры условных обозначений СМО.

– условное обозначение СМО M/M/1 означает, что на СМО с одним сервером поступает пуассоновский поток и время обслуживания в сервере имеет экспоненциальный ЗР. После трех основных символов через двоеточие могут быть дополнительные символы: a/b/c : d/e/f;

– M/M/m : Loss означает, что СМО обрабатывает пуассоновский поток заявок с помощью m серверов с полными потерями (чистая СМО с отказами). Если СМО без буфера, то обозначение Loss может быть опущено;

– M/D/3 : k – система с пуассоновским потоком на входе, детерминированным временем обслуживания на сервере, тремя серверами и буфером, вмещающим k заявок;

– G/G/2 – СМО без буфера (очереди) с произвольными ЗР интервалов времени между входными заявками и интервалов времени обслуживания заявок, содержащая два сервера;

– fBM/E^2/1 : ∞ – система с потоком типа фрактального броуновского движения на входе, временем обслуживания на сервере в виде потока Эрланга 2-го порядка, одним сервером и неограниченной очередью (заявки могут неограниченно накапливаться в буфере).

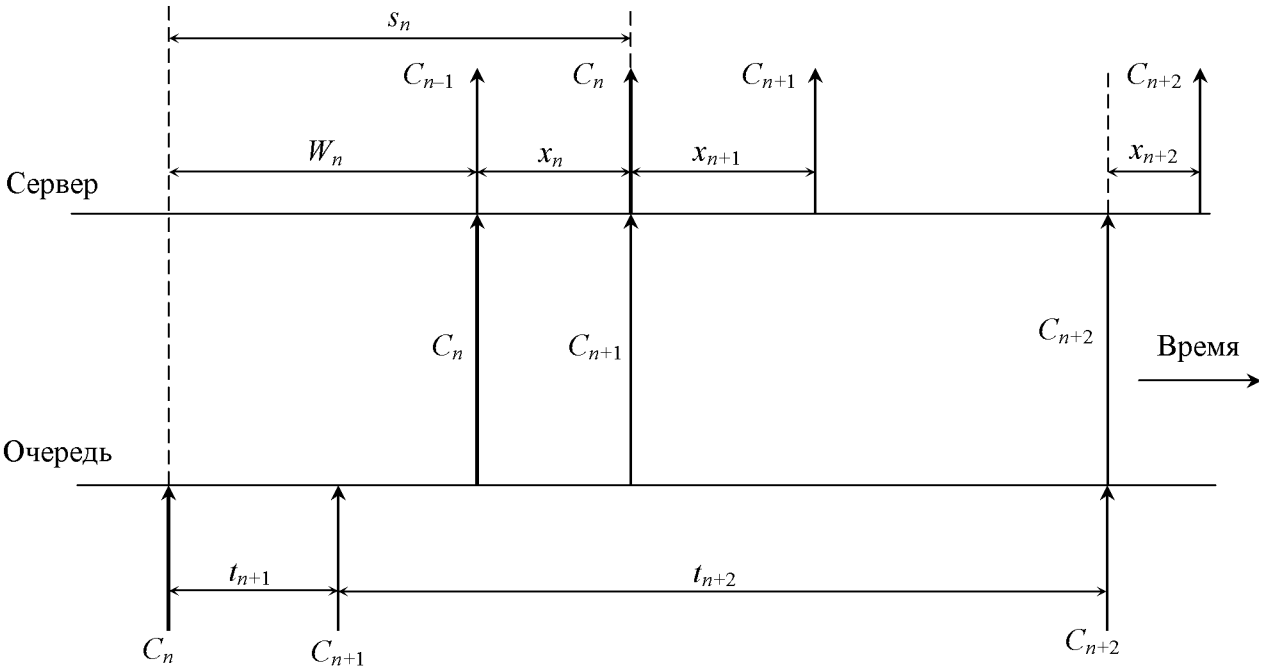


Рис. 5.1. Временная диаграмма работы СМО [10]

Последовательность поступления заявок, помещение их в буфер с очередью и их обработку серверами может быть описана с помощью временной диаграммы (рис. 5.1):

C_n – n-я входная заявка; x_n – время обслуживания заявки C_n;

t_n – интервал времени между приходом предыдущей заявки и C_n;

$s_n = W_n + x_n$ – общее время пребывания заявки в системе.

Определим соотношение между средним числом заявок в системе $\overline{N}$, интенсивностью входного потока λ и средним временем пребывания заявки в системе $\overline{T}$.

Число заявок, находящихся в системе в момент t:

$$N(t) = \alpha(t) - \delta(t), \qquad (5.1)$$

где $\alpha(t)$ – число заявок, поступающих в промежутке времени $(0, t)$; $\delta(t)$ – число исходящих из системы (обслуженных) заявок на интервале $(0, t)$ (рис. 5.2).

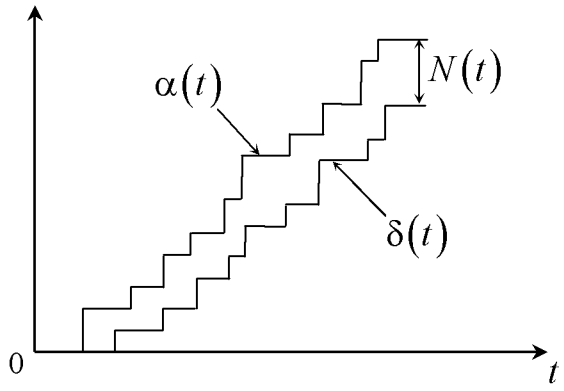


Рис. 5.2. Временные зависимости числа заявок в СМО

Площадь между кривыми $\alpha(t)$ и $\delta(t)$ от 0 до t дает общее время, проведенное всеми заявками в системе за время t:

$$\gamma(t) = \int_0^t N(t)\, dt\,.$$

Средняя интенсивность входного потока за время t равна $\overline{\lambda}(t) = \alpha(t)/t$, а время, проведенное одной заявкой в системе и усредненное по всем заявкам: $\overline{T}_t = \gamma(t)/\alpha(t)$. Среднее число заявок в системе в промежутке $(0, t)$: $\overline{N}_t = \gamma(t)/t$. Из последних трех уравнений следует, что $\overline{N}_t = \overline{\lambda}_t \overline{T}_t$, где $\overline{\lambda}_t = \overline{\lambda}(t)$.

Если в СМО существует стационарный режим, то при $t \to \infty$ имеют место равенства $\lambda = \lim\limits_{t\to\infty} \overline{\lambda}_t$, $\overline{T} = \lim\limits_{t\to\infty} \overline{T}_t$, откуда

$$\overline{N} = \lambda \overline{T}\,. \qquad (5.2)$$

Последнее соотношение называется **формулой Литтла** и означает, что среднее число заявок в системе равно произведению интенсивности поступления заявок в систему и среднего времени пребывания в системе независимо от ЗР входного потока и времени обслуживания. Формула Литтла справедлива и для СМО в виде очереди из заявок в буфере: средняя длина очереди равна произведению интенсивности входного потока заявок и среднего времени ожидания в очереди

$$\overline{N}_q = \lambda \overline{W}\,. \qquad (5.3)$$

Если рассматривать СМО только как серверы, то формула Литтла имеет вид: $\overline{N}_s = \lambda \overline{x}$, где $\overline{N}_s$ – среднее число заявок в серверах, а $\overline{x}$ – среднее время обработки в сервере. В итоге общее среднее время пребывания заявки в системе

$$\overline{T} = \overline{x} + \overline{W}\,. \qquad (5.4)$$

Коэффициент использования – это отношение интенсивности входного потока к пропускной способности $m/\bar{x}$ системы из m серверов (среднее значение долей занятости серверов):

$$\rho = \lambda \bar{x}/m, \quad 0 \le \rho < 1. \tag{5.5}$$

Если в СМО существует стационарный режим и известна вероятность p_0 того, что сервер окажется свободным, то $\rho = 1 - p_0$.

В ряде случаев в СМО заявки могут выбираться из входного потока (или очереди) и передаваться серверам на обработку в специфическом порядке. В СМО с **полнодоступными серверами** любой из освободившихся серверов может подбирать заявку из очереди на обработку. В СМО с **неполнодоступным включением серверов** существует некоторый механизм регулирования назначения серверов для тех или иных заявок [12]. Для этого во входном потоке заявок выделяются **классы заявок**, которые различаются механизмом распределения серверов – заявка определенного класса обрабатывается только серверами, закрепленными за данным классом.

5.2. СМО с приоритетным обслуживанием

В **СМО без приоритетов** заявки образуют естественную очередь и поступают на обслуживание строго в порядке очереди, т. е. действует правило «первый на входе – первый на выходе» (First Input First Output, FIFO). Имеются и другие правила, например, «последний на входе – первый на выходе» (Last Input First Output, LIFO). В общем случае может быть назначен механизм приоритетного обслуживания в соответствии с **дисциплинами обслуживания,** которые имеют условное обозначение в виде дополнительного символа в классификации Кендалла. Простейшая из дисциплин «первым пришел – первым обслужен» обозначается FCFS (First Came First Served). Она обычно не указывается, т. е. используется по умолчанию. Обратная дисциплина LCFS (Last Came First Served) (**стек**) должна быть указана: например: M/M/3 : k/LCFS. Существует также еще много различных дисциплин обслуживания: SPT/SJE – первоочередное обслуживание заявок с кратчайшим временем обслуживания; SRPT – первоочередное обслуживание заявок с кратчайшим временем дообслуживания; SEPT – первоочередное обслуживание заявок с кратчайшим средним временем обслуживания и др.

В системе с **абсолютным приоритетом** заявка в процессе обслуживания может быть удалена с сервера и возвращена в очередь при поступлении заявки с более высоким приоритетом. В системе **относительным приоритетом** обслуживание любой заявки, находящейся на сервере, не может быть прервано.

Предположим, что заявки принадлежат одному из P различных приоритетных классов $p = 1, 2, \ldots, P$. Каждая заявка в системе в момент времени t имеет приоритетную функцию $q_p(t)$ – чем больше ее значение, тем выше приоритет заявки (при выборе заявки на обслуживание – выбор делается в пользу заявки с наибольшим значением приоритетной функции). В простейшем случае в качестве приоритетной функции выбирается просто значение p – номер класса заявки (рис. 5.3).

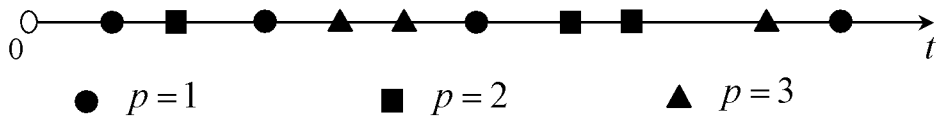


Рис. 5.3. Заявки различных приоритетных классов

Предположим, что заявки из приоритетного класса p образуют поток с интенсивностью λ_p заявок в секунду. Время обслуживания каждой заявки из класса p имеет ПРВ $b_p(x)$ и среднее значение

$$\overline{x}_p = \int_0^\infty x b_p(x)\, dx .$$

Введем следующие определения:

$$\lambda = \sum_{p=1}^{P} \lambda_p , \qquad \overline{x} = \sum_{p=1}^{P} \frac{\lambda_p \overline{x}_p}{\lambda} , \qquad \rho_p = \lambda_p \overline{x}_p , \qquad \rho = \lambda \overline{x} = \sum_{p=1}^{P} \rho_p .$$

Здесь ρ интерпретируется как доля времени, в течение которого сервер занят $(\rho < 1)$, а каждый из парциальных коэффициентов ρ_p – как доля времени, в течение которого сервер занят обслуживанием заявок из приоритетного класса с номером p.

Среднее время пребывания в системе заявок класса p:

$$\overline{T}_p = \overline{W}_p + \overline{x}_p ,$$

где $\overline{W}_p$ – среднее значение времени ожидания в очереди заявок из приоритетного класса p.

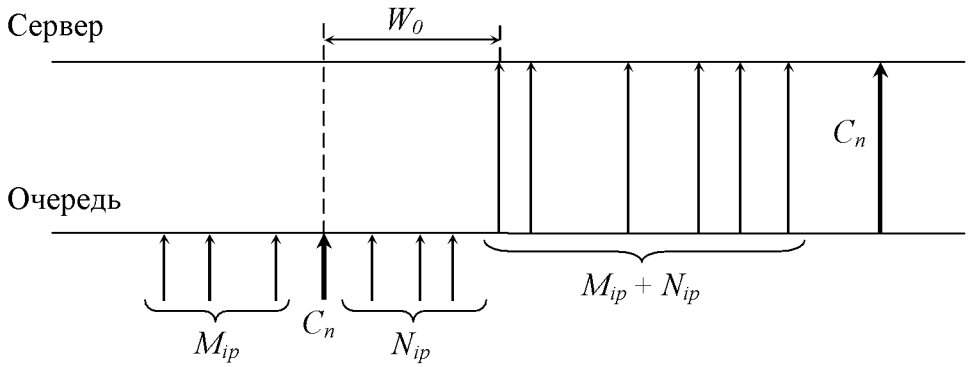


Рис. 5.4. Диаграмма функционирования СМО с меченой заявкой

Рассмотрим работу СМО с относительным приоритетом с момента поступления некоторой заявки из приоритетного класса p, которую далее будем именовать «меченой» (рис. 5.4). Первая составляющая времени ожидания для меченой заявки равна остаточному времени обслуживания W_0 другой заявки, которую она застает на сервере. Вторая составляющая времени ожидания для меченой заявки определяется тем, что перед меченой заявкой обслуживаются N_{ip} других заявок из класса i, которые застала в очереди меченая заявка (из класса p) и которые обслуживаются перед ней. Среднее значение данной величины

$$\overline{T}_N = \sum_{i=1}^{P} \overline{x}_i \overline{N}_{ip}.$$

Третья составляющая задержки связана с M_{ip} заявками, поступившими после того, как пришла меченая заявка, однако получившими обслуживание раньше нее. Среднее значение этой составляющей задержки

$$\overline{T}_M = \sum_{i=1}^{P} \overline{x}_i \overline{M}_{ip}.$$

В итоге среднее время ожидания в очереди для меченой заявки:

$$\overline{W}_p = \overline{W}_0 + \overline{T}_N + \overline{T}_M = \overline{W}_0 + \sum_{i=1}^{P} \overline{x}_i \left(\overline{N}_{ip} + \overline{M}_{ip} \right), \quad p = 1, \ldots, P,$$

где $\overline{W}_0$ – средняя задержка меченой заявки, связанная с наличием другой заявки на обслуживании.

Независимо от дисциплины обслуживания число заявок N_{ip} и M_{ip} в системе не может быть произвольным – существует некоторый набор соотношений, называемых **законами сохранения** и связывающих между собой задержки для каждой заявки из приоритетного класса.

Законы сохранения подразумевают, что незавершенная работа в любой СМО в течение любого интервала времени занятости не зависит от порядка обслуживания, если система является **консервативной** (заявки не исчезают внутри системы, и сервер не простаивает при непустой очереди).

Остаточное время обслуживания (на рис. 5.5 W_{i-1}) – это время, которое осталось затратить на обслуживание заявки, находящейся в сервере к моменту прихода следующей заявки.

Незавершенная работа R – это время, которое должно пройти до полного ухода всех заявок из системы с момента прекращения поступлений заявок c_{i+1} на ее вход. Незавершенная работа в момент прихода последней заявки c_{i+1} равна сумме времени ожидания заявки в очереди системы и времени дообслуживания предыдущей заявки $R_i = x_i + W_{i-1}$.

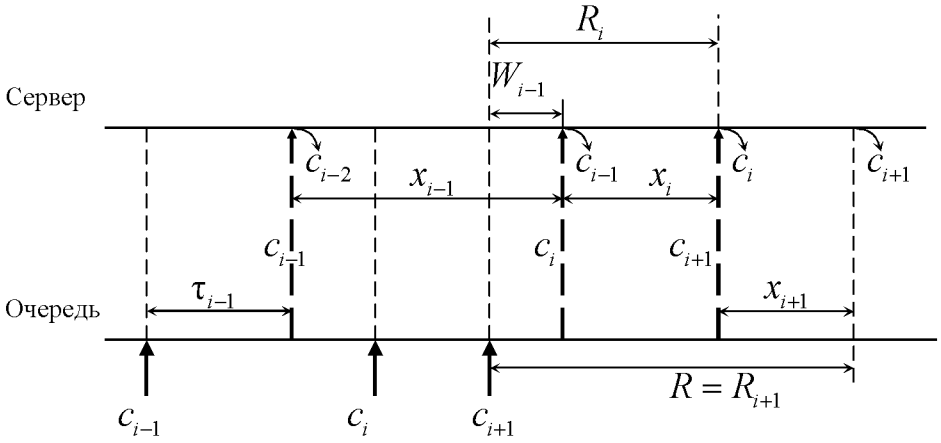


Рис. 5.5. Диаграмма, поясняющая незавершенную работу

В СМО любого типа для любой дисциплины обслуживания выполняется равенство (в правой части – разность средней незавершенной работы и остаточного времени обслуживания):

$$\sum_{p=1}^{P} \rho_p W_p = \overline{R} - \overline{W}_0, \qquad (5.6)$$

где $\overline{W}_0$ – среднее остаточное время обслуживания.

Взвешенная сумма времени ожидания никогда не изменяется, независимо от того, насколько сложна или искусно подобрана дисциплина обслуживания – если удается сократить задержку для одних заявок, то она немедленно возрастет для других.

Пример. На вход СМО поступают пакеты трех приоритетных классов $p=1,2,3$ с соответствующими плотностями: $\lambda_1=6$ пакетов/с, $\lambda_2=4$ пакета/с, $\lambda_3=3$ пакета/с (пуассоновские потоки). Среднее время обработки на сервере для пакетов различных приоритетных классов составляет: $\overline{x}_1=0.01$ с, $\overline{x}_2=0.02$ с, $\overline{x}_3=0.03$ с. Найти среднее время обработки пакетов в системе, а также среднюю нагрузку в системе.

Решение.

$$\lambda=\sum_{p=1}^{3}\lambda_p=6+4+3=13 \text{ пакетов/с},$$

$$\overline{x}=\sum_{p=1}^{3}\frac{\lambda_p\overline{x}_p}{\lambda}=\frac{6\cdot 0.01+4\cdot 0.02+3\cdot 0.03}{13}=\frac{0.06+0.08+0.09}{13}=\frac{0.23}{13}=0.0767 \text{ с},$$

$$\rho_p=\lambda_p\overline{x}_p,\quad \rho=\lambda\overline{x}=\sum_{p=1}^{3}\rho_p=13\cdot 0.767=0.23.$$

5.3. Анализ СМО типа M/M/m : ∞

5.3.1. Система M/M/1 : ∞

Вначале рассмотрим систему M/M/1, содержащую буфер, способный хранить очередь бесконечной длины (рис. 5.6). Так как входной поток ординарный, то в каждый момент времени к очереди может добавиться (или уйти из очереди на сервер) только одна заявка, т. е. имеет место процесс класса «гибели-размножения» (рис. 5.7).

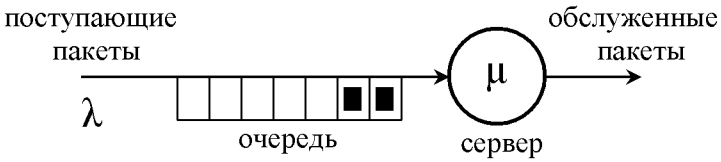


Рис. 5.6. Схема системы M/M/1

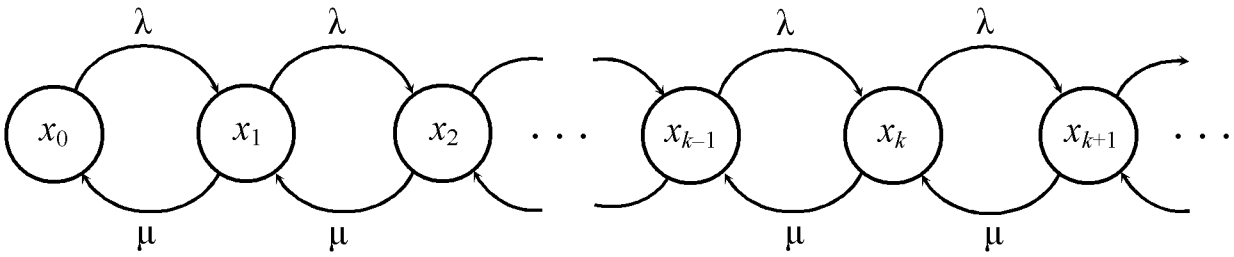


Рис. 5.7. Диаграмма интенсивностей переходов для рассматриваемой СМО

Так как поток стационарный, то интенсивности переходов будут иметь постоянные значения $\lambda_k = \lambda,\ k = 0, 1, 2, \ldots,\ \mu_k = \mu,\ k = 1, 2, 3, \ldots$.

Вероятность того, что в стационарном состоянии в системе будет находиться k заявок, может быть найдена по обобщенной формуле (1.6):

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

Начальное значение вероятности с учетом сходимости ряда в знаменателе:

$$p_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \left(\frac{\lambda}{\mu} \right)^k} = \frac{1}{1 + \frac{\lambda/\mu}{1 - \lambda/\mu}} = 1 - \frac{\lambda}{\mu} = 1 - \rho.$$

Вероятность длины очереди из k заявок имеет **геометрический** ЗР

$$p_k = (1 - \rho)\rho^k, \quad k = 0, 1, 2, \ldots. \qquad (5.7)$$

На рис. 5.8 приведен график вероятностей того, что в системе находится k заявок в установившемся режиме.

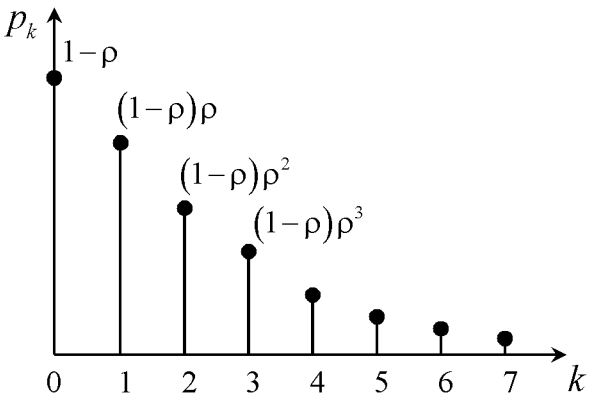
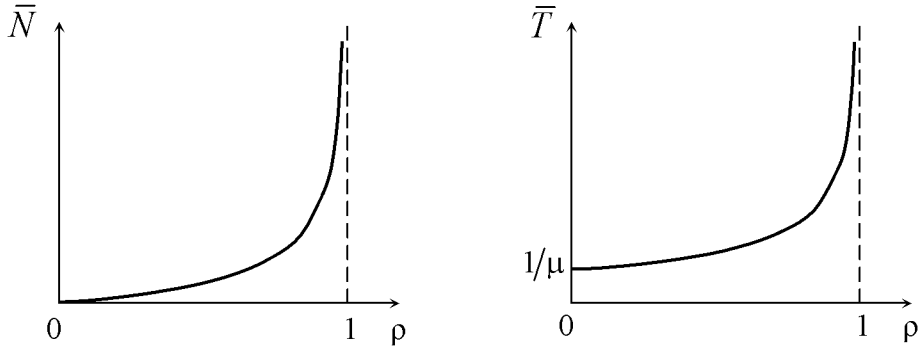


Рис. 5.8. Стационарное распределение вероятностей состояний в системе M/M/1

Найдем **среднее число заявок в системе**, вычислив математическое ожидание дискретной СВ k:

$$\bar{N} = \sum_{k=0}^{\infty} k p_k = (1 - \rho) \sum_{k=0}^{\infty} k \rho^k = \frac{\rho}{1 - \rho}. \qquad (5.8)$$

График среднего числа заявок в системе в зависимости от значения коэффициента использования (нагрузки) показан на рис. 5.9, а.



а) б)

Рис. 5.9. Зависимости среднего числа заявок (а) и среднего времени пребывания в системе (б) для СМО M/M/1

Дисперсия числа заявок в системе: $\sigma_N^2 = \sum_{k=0}^{\infty} \left(k - \overline{N}\right)^2 p_k = \dfrac{\rho}{(1-\rho)^2}$.

Среднее время пребывания заявки в системе определяется по формуле Литтла:

$$\overline{T} = \frac{\overline{N}}{\lambda} = \left(\frac{\rho}{1-\rho}\right)\frac{1}{\lambda} = \frac{1/\mu}{1-\rho}. \qquad (5.9)$$

На рис. 5.9, б приведен график зависимости среднего времени пребывания заявки в системе от значения коэффициента использования (нагрузки), который часто называют **нагрузочной кривой** СМО.

Из графиков на рис. 5.9 легко видеть, что при приближении коэффициента использования ρ к единице число заявок в системе и время пребывания неограниченно возрастают.

Пример. На вход односерверной СМО с буфером бесконечной длины поступает пуассоновский поток с плотностью λ=15 пакетов/с. Среднее время обработки пакетов в сервере составляет 0.06 с. Найти среднее время пребывания пакета в системе и среднее число пакетов в системе.

Решение. Нагрузка составляет $\rho = \lambda\overline{x} = 15 \cdot 0.06 = 0.9$,

$$T = \frac{\overline{N}}{\lambda} = \left(\frac{\rho}{1-\rho}\right)\frac{1}{\lambda} = \frac{\overline{x}}{1-\rho} = \frac{0.06}{0.1} = 0.6 \text{ с},$$

$\overline{N} = \lambda T = 15 \cdot 0.6 = 9$ пакетов.

Вероятность того, что в системе будет находиться не менее чем *m* заявок, а также вероятность, что в системе менее *m* заявок имеют вид:

$$\mathrm{P}\{k \ge m\} = \sum_{k=m}^{\infty} p_k = \sum_{k=m}^{\infty} (1-\rho)\rho^k = \rho^m, \quad \mathrm{P}\{k < m\} = 1 - \rho^m.$$

Рассмотрим теперь характеристики очереди. Поскольку среднее число заявок в системе складывается из средних чисел заявок в очереди и на сервере

$$\bar{N} = \bar{N}_q + \bar{N}_s, \tag{5.10}$$

то отсюда следует, что

$$\bar{N}_q = \bar{N} - \rho = \frac{\rho^2}{1-\rho}. \tag{5.11}$$

Время ожидания в очереди в среднем по всем заявкам можно найти по формуле Литтла:

$$\bar{W} = \frac{\bar{N}_q}{\lambda} = \frac{\rho}{\mu(1-\rho)}. \tag{5.12}$$

Нормированное время ожидания (пребывания в системе) показывает, во сколько раз оно возрастает по сравнению со средним временем обслуживания собственно на сервере по причине наличия очереди. Полное нормированное время пребывания в системе M/M/1:

$$\bar{T}_{\text{норм}} = \frac{\bar{T}}{\bar{x}} = \mu\bar{T} = \frac{1}{1-\rho}.$$

Пример. Терминальный концентратор состоит из 6 входных линий со скоростями 56 Кбит/с и единственной выходной линии со скоростью 128 Кбит/с (система M/M/1 : ∞). Средний размер пакета 450 байт, а скорость поступления в каждой входной линии равна 5 пакетов/с. Какова средняя задержка пакета и каково среднее число пакетов, хранящихся в концентраторе?

Решение.

Скорость обработки пакетов $\mu = \dfrac{128\,000}{(450/8)} = 35.56$ пакета/с.

Общая интенсивность $\lambda = 6\,\text{линий} \times 5\,\text{пакетов/с} = 30$ пакетов/с.

Нагрузка в концентраторе $\rho = \lambda/\mu = 0.844$.

Вероятность i-го состояния в системе M/M/1 $p_i = \rho^i p_0 = \rho^i(1-\rho)$.

Среднее число пакетов в системе

$$\bar{N} = \sum_{i=0}^{\infty} i\, p_i = \sum_{i=0}^{\infty} i\rho^i(1-\rho) = \frac{\rho}{1-\rho} = 5.4 \text{ пакетов.}$$

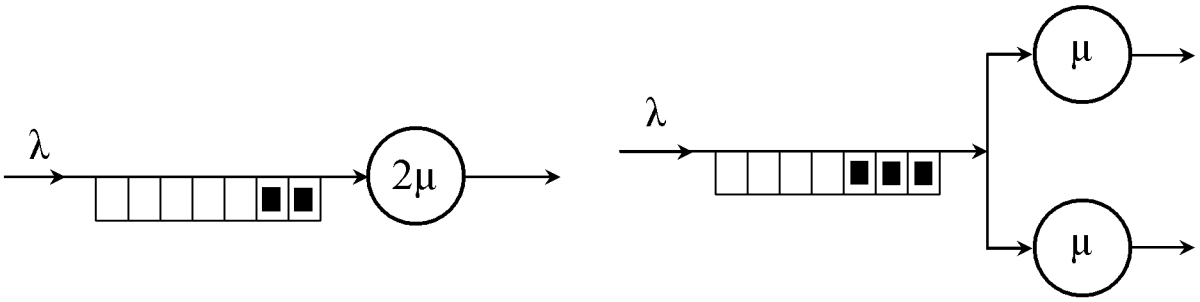
Средняя задержка T вычисляется исходя из формулы Литтла

$$\bar{N} = \lambda T,$$

$$T = \bar{N}/\lambda = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu - \lambda} = \frac{1}{35.56 - 30} = 179.8 \text{ мс. [9]}$$

5.3.2. СМО с двумя серверами (M/M/2 : ∞)

Вначале рассмотрим СМО с двумя серверами, любой из которых доступен для поступающих на вход заявок (полнодоступная СМО). При этом интерес представляет сравнение эффективности двух СМО: односерверной (M/M/1) со средним временем обслуживания $\bar{T}_{об} = 1/2\mu$ (рис. 5.10, а) и двухсерверной (M/M/2) с $\bar{T}_{об} = 1/\mu$ (рис. 5.10, б), т. е. необходимо выяснить, что эффективнее – удвоение скорости обработки или распараллеливание обработки заявок?



а) б)

Рис. 5.10. СМО с двумя серверами (а) и с одним сервером вдвое большей производительности (б)

СМО типа M/M/2 может быть представлена как процесс «размножения-гибели» (рис. 1.5) с параметрами:

$$\lambda_k = \lambda,\ \mu_1 = \mu,\ \mu_2 = 2\mu = 2\mu_1.$$

Наличие множителя 2 в формуле $\mu_2 = 2\mu$ определяется удвоением вероятности освобождения сервера $P_{2осв} = P_{1осв} + P_{1осв}$, т. к. для двух одновременно занятых серверов вероятность освобождения хотя бы одного из них вдвое больше, чем в случае занятия одного сервера $P_{1осв}$.

По формуле (1.6) распределение вероятностей в стационарном режиме (занятость k каналов):

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} = p_0 \left(\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{2\mu}\right)^{k-1} = \frac{p_0 \rho^k}{2^{k-1}} = 2 p_0 \rho_2^k, \qquad \rho_2 = \frac{\lambda}{2\mu},$$

и далее, с учетом условия нормировки (1.3) получим вероятность простоя (1.5)

$$p_0 = \frac{1}{1+\sum\limits_{k=1}^{\infty}\prod\limits_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}}.$$

Величину в знаменателе можно разложить в ряд и просуммировать его:

$$1+\sum_{k=1}^{\infty}\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}} = \left|\begin{matrix}\lambda_i=\lambda,\ \ \mu_1=\mu\\ \mu_2=\ldots=\mu_k=2\mu\end{matrix}\right| = 1+\frac{\lambda}{\mu}+\frac{\lambda}{\mu}\cdot\frac{\lambda}{2\mu}+\frac{\lambda}{\mu}\left(\frac{\lambda}{2\mu}\right)^2+\ldots+\frac{\lambda}{\mu}\left(\frac{\lambda}{2\mu}\right)^k+\ldots =$$

$$=1+\sum_{k=0}^{\infty}\frac{\lambda}{\mu}\left(\frac{\lambda}{2\mu}\right)^k = 1+\frac{\lambda}{\mu}\sum_{k=0}^{\infty}\left(\frac{\lambda}{2\mu}\right)^k = 1+\frac{\lambda}{\mu}\cdot\frac{1}{1-\lambda/2\mu} = \frac{1+\lambda/2\mu}{1-\lambda/2\mu},$$

откуда получаем:

$$p_0 = \frac{1-\lambda/2\mu}{1+\lambda/2\mu} = \frac{1-\rho_2}{1+\rho_2}, \qquad p_k = \frac{2(1-\rho_2)}{1+\rho_2}\rho_2^k.$$

Найдем среднее число заявок в двухсерверной СМО (M/M/2)

$$\bar{N}_2 = \sum_{k=0}^{\infty}kp_k = \frac{2(1-\rho_2)}{1+\rho_2}\sum_{k=0}^{\infty}k\rho_2^k = \frac{2(1-\rho_2)}{1+\rho_2}\cdot\frac{\rho_2}{(1-\rho_2)^2} = \frac{2\rho_2}{1-\rho_2^2}, \qquad (5.13)$$

где $\sum\limits_{k=1}^{\infty}k\rho_2^k = \frac{\rho_2}{(1-\rho_2)^2}$ [10].

Среднее время пребывания заявки в СМО M/M/2 (по формуле Литтла):

$$\bar{T}_2 = \frac{\bar{N}_2}{\lambda} = \frac{1}{\mu(1-\rho_2^2)}. \qquad (5.14)$$

Из (5.2) и (5.8) имеем среднее число заявок и среднее время пребывания заявки в односерверной СМО (M/M/1) с удвоенной производительностью 2μ:

$$\bar{N}_1' = \frac{\rho_2}{1-\rho_2}, \quad \bar{T}_1' = \frac{\bar{N}_1'}{\lambda} = \frac{1}{2\mu(1-\rho_2)}. \qquad (5.15)$$

Сравнение среднего числа заявок

$$\bar{N}_1' = \frac{\rho_2}{1-\rho_2} \le \bar{N}_2 = \frac{2\rho_2}{1-\rho_2^2} \le \bar{N}_1 = \frac{\rho}{1-\rho} = \frac{2\rho_2}{1-2\rho_2}$$

и среднего времени пребывания

$$\mu\overline{T}_1' = \frac{1}{2(1-\rho_2)} \le \mu\overline{T}_2 = \frac{1}{1-\rho_2^2} \le \mu\overline{T}_1 = \frac{1}{1-\rho} = \frac{1}{1-2\rho_2}.$$

M/M/1 (2μ) M/M/2 (μ) M/M/1 (μ)

Таблица 5.1

Сводная таблица численных значений $\mu\overline{T}_i$

ρ_2	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\mu\overline{T}_1'$ (2μ)	0.5	0.555	0.625	0.714	0.833	1	1.25	1.667	2.5	5
$\mu\overline{T}_2$ (μ)	1	1.01	1.042	1.099	1.19	1.333	1.562	1.961	2.78	5.263
$\mu\overline{T}_1$ (μ)	1	1.25	1.667	2.5	5	–	–	–	–	–

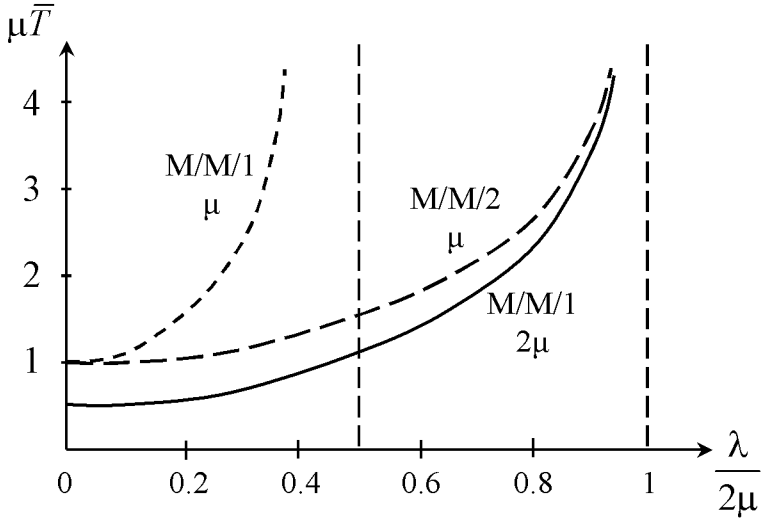


Рис. 5.11. Сравнение нагрузочных кривых для различных СМО

Из рис. 5.11 видно, что увеличение вдвое производительности сервера оказывается более эффективным, чем введение параллельного сервера той же производительности.

5.3.3. СМО с несколькими серверами (M/M/m : ∞)

Рассмотрим общий случай СМО с m серверами M/M/m. На рис. 5.12 представлена диаграмма интенсивностей переходов, на которой интенсивности переходов определяются следующим образом:

$$\lambda_i = \lambda,\ i = 0, 1, 2, \ldots; \qquad \mu_i = \min\{i\mu, m\mu\} = \begin{cases} i\mu, & 0 \le i \le m \\ m\mu, & i \ge m \end{cases}. \qquad (5.16)$$

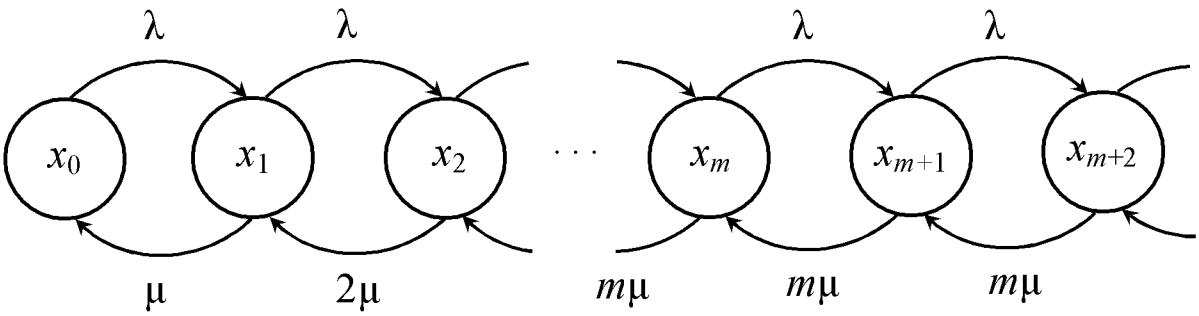


Рис. 5.12. Диаграмма интенсивностей переходов для m-серверной СМО

Состояние x_{m+k} означает занятость всех m серверов и наличие k заявок в очереди.

Для процессов «гибели-размножения» (п. 1.3.2):

$$p_k = \begin{cases} p_0 \dfrac{\rho^k}{k!}, & k \le m, \\ p_0 \dfrac{\rho^k}{m!\, m^{k-m}}, & k \ge m. \end{cases} \qquad (5.17)$$

Нагрузка в системе M/M/m:

$$\rho_m = \frac{\lambda}{m\mu} = \frac{\rho}{m} < 1. \qquad (5.18)$$

где $\rho = \lambda/\mu$ – общая входная нагрузка, ρ_m – удельная нагрузка на один сервер.

Докажем формулу (5.17). Исходя из (1.6) имеем следующие представления вероятностей

$$p_0 = \left[\sum_{i=0}^{\infty} \frac{\lambda_0 \lambda_1 \ldots \lambda_{i-1}}{\mu_1 \mu_2 \ldots \mu_i} \right]^{-1}, \qquad p_k = \frac{\lambda_0 \lambda_1 \ldots \lambda_{k-1}}{\mu_1 \mu_2 \ldots \mu_k} p_0$$

и с учетом (5.16), имеем при $k \le m$

$$p_k = \frac{\lambda^k}{\mu \cdot 2\mu \cdot 3\mu \cdot \ldots \cdot k\mu} p_0 = \frac{\lambda^k}{k!\, \mu^k} p_0 = \frac{\rho^k}{k!} p_0,$$

а при $k > m$ получаем

$$p_k = \frac{\lambda^k}{\underbrace{\mu \cdot 2\mu \cdot 3\mu \cdot \ldots \cdot m\mu}_{m} \cdot \underbrace{m\mu \cdot m\mu \cdot \ldots \cdot m\mu}_{k-m}} p_0 = \frac{\lambda^m \lambda^{k-m}}{m!\, \mu^m m^{i-m} \mu^{i-m}} p_0 = \frac{\rho^m}{m!} \left(\frac{\rho}{m} \right)^{k-m} p_0.$$

На основе (5.16) и (5.17) получим выражение для вероятности простоя одного сервера

$$p_0 = \left[\sum_{i=0}^{m} \frac{\lambda_0 \lambda_1 \dots \lambda_{i-1}}{\mu_1 \mu_2 \dots \mu_i} + \frac{\lambda_0 \lambda_1 \dots \lambda_{m-1}}{\mu_1 \mu_2 \dots \mu_m} \sum_{i=m+1}^{\infty} \frac{\lambda_m \lambda_{m+1} \dots \lambda_{i-1}}{\mu_{m+1} \mu_{m+2} \dots \mu_i} \right]^{-1} = \left| \begin{matrix} \lambda_i = \lambda \\ \mu_i = \mu \end{matrix} \right| =$$

$$= \left[\sum_{i=0}^{m} \frac{\rho^i}{i!} + \frac{\rho^m}{m!} \sum_{i=m+1}^{\infty} \left(\frac{\rho}{m} \right)^{i-m} \right]^{-1} = \left| \begin{matrix} k = i - m \\ \lambda/\mu < 1 \end{matrix} \right| = \left[\sum_{i=0}^{m} \frac{\rho^i}{i!} + \frac{\rho^m}{m!} \sum_{k=1}^{\infty} \left(\frac{\rho}{m} \right)^{k} \right]^{-1} =$$

$$= \left[\sum_{i=0}^{m} \frac{\rho^i}{i!} + \frac{\rho^m}{m!} \cdot \frac{\rho}{m-\rho} \right]^{-1} = \left[\sum_{i=0}^{m-1} \frac{\rho^i}{i!} + \frac{\rho^m}{m!} + \frac{\rho^m}{m!} \cdot \frac{\rho}{m-\rho} \right]^{-1} = \left[\sum_{i=0}^{m-1} \frac{\rho^i}{i!} + \frac{\rho^m}{m!} \left[1 + \frac{\rho}{m-\rho} \right] \right]^{-1} =$$

$$= \left[\sum_{i=0}^{m-1} \frac{\rho^i}{i!} + \frac{\rho^m}{m!} \left[\frac{m}{m-\rho} \right] \right]^{-1} = \left[\sum_{i=0}^{m-1} \frac{\rho^i}{i!} + \frac{\rho^m}{m!} \left[\frac{1}{1-\rho_m} \right] \right]^{-1}.$$

Определим теперь вероятность того, что поступающая заявка окажется в очереди, т. е. не застанет ни одного свободного сервера

$$P_t = p_{k \geq m} = \sum_{i=m}^{\infty} p_i = p_0 \frac{\rho^m}{m!} \sum_{i=m}^{\infty} \left(\frac{\rho}{m} \right)^{i-m} = \left| i - m = j \right| = p_0 \frac{\rho^m}{m!} \sum_{j=0}^{\infty} \left(\frac{\rho}{m} \right)^{j} = p_0 \frac{\rho^m}{m!} \cdot \frac{1}{1-\rho/m}.$$

Полученная формула называется **формулой Эрланга С,** или **второй формулой Эрланга**:

$$P_t = C(m,\rho) = \frac{\rho^m}{m!(1-\rho_m)} p_0 = \frac{E_m(\rho)}{1-(\rho/m)\left[1-E_m(\rho)\right]} = D_m(\rho). \qquad (5.19)$$

Из (5.19) следует, что $D_m(\rho) > E_m(\rho)$, т. е. при одинаковых значениях интенсивности нагрузки и числе линий вероятность ожидания в системе с ожиданием выше, чем вероятность потери заявки в системе с явными потерями (отказами). Это объясняется различием в дисциплинах обслуживания сравниваемых систем. В системе с явными потерями заявка, поступившая в момент занятости всех линий, теряется и никакого воздействия на систему в последующем не оказывает, а в системе с ожиданием заявка ставится в очередь. При освобождении линии в системе с явными потерями она предоставляется поступающей заявке, а в системе с ожиданием при наличии очереди – ожидающей заявке. Вновь поступающей заявке приходится становиться в очередь. Так как заявка не теряется, а лишь задерживается при обслуживании, то вероятность ожидания P_t называют **условными потерями**.

С учетом формул (5.17) и (5.3) среднее число заявок в очереди можно выразить следующим образом:

$$\bar{N}_q = \sum_{s=0}^{\infty} s p_s = p_0 \left[\frac{\rho^m}{m!} \sum_{s=0}^{\infty} s \left(\frac{\rho}{m} \right)^{s} \right] = \frac{\rho_m}{1-\rho_m} C(m,\rho), \qquad (5.20)$$

где s – текущее число заявок в очереди, а также использована вспомогательная формула $\sum_{s=0}^{\infty} sq^s = \frac{q}{(1-q)^2}$, $q<1$.

Вероятность того, что в СМО находятся k заявок:

$$p_k = \frac{\rho^m}{m!}\left(\frac{\rho}{m}\right)^{k-m} \cdot p_0 = \rho_m^{k-m}\left(1-\rho_m\right) C\left(m,\rho\right). \qquad (5.21)$$

Среднее число занятых серверов, с учетом формул (3.10) и (3.12):

$$\begin{aligned} \overline{B} &= \sum_{k=0}^{m} k p_k = \sum_{k=0}^{m} k \frac{\rho^k}{k!} p_0 = \sum_{k=0}^{m} k \frac{\rho^k}{k!} \Bigg/ \sum_{k=0}^{m} \frac{\rho^k}{k!} = \\ &= \rho \sum_{k=0}^{m-1} \frac{\rho^k}{k!} \Bigg/ \sum_{k=0}^{m} \frac{\rho^k}{k!} = \rho\left[1 - E_m(\rho)\right], \end{aligned} \qquad (5.22)$$

где $E_m(\rho)$ – первая формула Эрланга [31, с.59].

Среднее число свободных серверов: $\overline{I} = m - \overline{B}$.

5.3.4. Функция распределения времени ожидания в СМО M/M/m : ∞

Вероятность ожидания $P_t = D_m(\rho)$ характеризует долю заявок, обслуживаемых после некоторого времени ожидания [12]. Для абонента, заявка которого поставлена в очередь, кроме этой характеристики важно знать функцию распределения времени ожидания начала обслуживания. Обозначим через $W(t) = P(T_{ож} > t)$ вероятность того, что для поступившей в произвольный момент заявки длительность ожидания начала обслуживания превысит t; через $W_i(t) = P_i(T_{ож} > t)$ – условную вероятность того же неравенства в предположении, что заявка поступила в момент, когда система находилась в состоянии i. Так как ожидание возможно лишь в состояниях системы $i = m, m+1, m+2, \ldots$, то по формуле полной вероятности получим

$$W(t) = \sum_{i=m}^{\infty} P_i W_i(t), \qquad (5.23)$$

где P_i – вероятность того, что система находится в состоянии i.

Вероятность $W_i(t)$ можно рассматривать как вероятность того, что за длительность t после появления заявки произойдет не более чем $i-m$ освобождений, так как эта заявка является $(i-m+1)$-й в очереди.

По условию задачи длительность обслуживания одной заявки без учета времени ожидания распределена по показательному закону

$$F(t)=P(t_з<t)=1-e^{-\mu t}.$$

Функция распределения промежутков между моментами освобождения линий пучка при условии занятости в пучке всех m линий есть $F_{ос}(t)=1-e^{-m\mu t}$. Отсюда, а также из формулы (5.16) следует, что при $i>m$ поток освобождений является простейшим с параметром $m\mu$. Для простейшего потока вероятность освобождения точно j линий за время t определяется по формуле Пуассона:

$$P_j=\frac{(m\mu t)^j}{j!}e^{-m\mu t}.$$

Вероятность освобождения не более $i-m$ линий за время t

$$W_i(t)=\sum_{j=0}^{i-m}P_j=\sum_{j=0}^{i-m}\frac{(m\mu t)^j}{j!}e^{-m\mu t}. \qquad (5.24)$$

Подставляя (5.24) и (5.17) в (5.23), получаем:

$$W(t)=\sum_{i=m}^{\infty}P_i\sum_{j=0}^{i-m}\frac{(m\mu t)^j}{j!}e^{-m\mu t}=\frac{\rho^m}{m!}p_0\,e^{-m\mu t}\sum_{j=m}^{\infty}\left(\frac{\rho}{m}\right)^{i-m}\sum_{j=0}^{i-m}\frac{(m\mu t)^j}{j!}=$$

$$=\frac{\rho^m}{m!}p_0\,e^{-m\mu t}\sum_{j=0}^{\infty}\frac{(m\mu t)^j}{j!}\sum_{j=m+j}^{\infty}\left(\frac{\rho}{m}\right)^{i-m}=\frac{\rho^m}{m!}p_0\,e^{-m\mu t}\sum_{j=0}^{\infty}\frac{(m\mu t)^j}{j!}\left(\frac{\rho}{m}\right)^j\sum_{j=m+j}^{\infty}\left(\frac{\rho}{m}\right)^{i-(m+j)}=$$

$$=\frac{(\rho^m/m!)p_0e^{-m\mu t}}{1-(\rho/m)}\sum_{r=0}^{\infty}\frac{(\lambda t)^r}{r!}=\frac{(\rho^m/m!)p_0}{1-(\rho/m)}e^{-(m\mu-\lambda)t}=D_m(\rho)e^{-\mu(m-\rho)t}.$$

Для того чтобы найти ПРВ для полного времени пребывания в системе, найдем сначала ПРВ времени ожидания в очереди для системы M/M/1. Поскольку вероятность отсутствия заявок в системе равна $p_0=1-\rho$, а для нескольких заявок время ожидания складывается из суммы независимых экспоненциально распределенных СВ с одинаковым средним равным $\overline{x}=1/\mu$, то их суммарная ПРВ описывается гамма-распределением [14]:

$$W(x)=\frac{\mu^n x^{n-1}e^{-\mu x}}{\Gamma(n)}.$$

ПРВ для времени ожидания в очереди [10, с.223]:

$$W_p(t)=\mu e^{-\mu t}(1-\rho)\rho\sum_{n=1}^{\infty}\frac{(\mu\rho t)^{n-1}}{(n-1)!}=\mu\rho(1-\rho)e^{-\mu(1-\rho)t}.$$

ПРВ для полного времени пребывания в системе [10]:

$$W(t)=\mu(1-\rho)e^{-\mu(1-\rho)t}.$$

5.4. СМО с ограниченным буфером (M/M/1 : N–1)

Рассмотрим СМО M/M/1, в которой присутствуют N заявок ($N-1$ – в буфере и одна на обслуживании). При этом любая заявка, поступившая на вход СМО получает отказ (блокировка). Такая СМО может быть описана процессом «гибели-размножения» с параметрами:

$$\lambda_k = \begin{cases} \lambda, & k < N, \\ 0, & k \geq N, \end{cases} \qquad \mu_k = \mu, \quad k = 1, 2, \ldots, N.$$

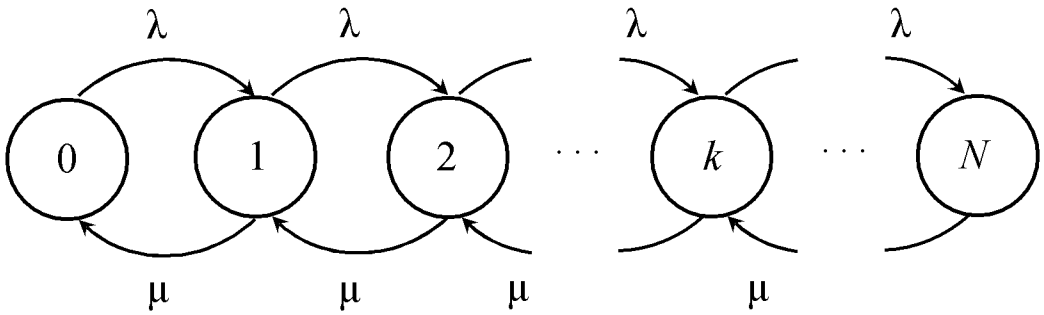


Рис. 5.13. Диаграмма интенсивностей переходов для СМО M/M/1 : N–1

Исходя из (1.6) имеем распределение вероятностей в стационарном режиме:

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda}{\mu}, \quad k \leq N, \qquad p_k = \begin{cases} p_0 (\lambda/\mu)^k = p_0 \rho^k, & k \leq N, \\ 0, & k > N. \end{cases}$$

Условие нормировки $$\sum_{k=1}^{N} p_k = 1. \qquad (5.25)$$

Вероятность простоя для СМО M/M/1 : N–1 определяется с учетом (5.25):

$$p_0 = \frac{1}{1 + \sum_{k=1}^{N} \rho^k} = \left[1 + \frac{\rho\left(1 - \rho^N\right)}{1 - \rho} \right]^{-1} = \frac{1 - \rho}{1 - \rho^{N+1}}.$$

Здесь использовано разложение:

$$\sum_{k=1}^{N} x^k = x\left(1 + x + x^2 + \ldots + x^{N-1}\right) = \frac{x\left(1 - x^N\right)}{1 - x}.$$

Окончательная формула для стационарных вероятностей:

$$p_k = \begin{cases} \dfrac{1-\rho}{1-\rho^{N+1}}\rho^k, & 0 \le k \le N, \\ 0, & k > N. \end{cases} \qquad (5.26)$$

При этом вероятность блокировки будет равна

$$P_B = p_N = \frac{(1-\rho)\rho^N}{1-\rho^{N+1}}. \qquad (5.27)$$

Рассмотрим пример расчета характеристик QoS для такой СМО, а именно вероятности блокировки (потери заявки) равной вероятности переполнения буфера. При коэффициенте использования ρ=0.5 и N=18 имеем $P_B \approx 1.9 \cdot 10^{-6}$, а при N=19, $P_B \approx 0.95 \cdot 10^{-6}$, т. е. для достижения $P_B \le 10^{-6}$ буфер объема $N-1$ должен вмещать не менее 18 заявок.

Среднее число заявок в системе:

$$\bar{N} = \sum_{k=0}^{N} k p_k = \frac{(1-\rho)\rho}{1-\rho^{N+1}} \sum_{k=0}^{N} k\rho^{k-1} = \frac{\rho}{1-\rho} - \frac{(N+1)\rho^{N+1}}{1-\rho^{N+1}}, \qquad (5.28)$$

где $\sum_{k=1}^{N} k\rho^{k-1} = \frac{\partial}{\partial\rho}\left(\sum_{k=1}^{N}\rho^k\right) = \frac{\partial}{\partial\rho}\left(\frac{\rho(1-\rho^N)}{1-\rho}\right) = \frac{1-(N+1)\rho^N + N\rho^{N+1}}{(1-\rho)^2}$.

Среднее время пребывания заявки в системе $\bar{T} = \bar{N}/\lambda$.

Определим производительность СМО как число заявок, фактически обслуженных системой в единицу времени $\gamma = \lambda(1-P_B)$. Реальная производительность на выходе СМО выражается как $\gamma = \mu(1-p_0)$, где p_0 – вероятность простоя (в системе нет ни одной заявки). Для СМО с бесконечным буфером ($N = \infty$), подставляя значение $p_0 = 1-\rho$, получим $\gamma = \mu\left[1-(1-\rho)\right] = \lambda$, откуда $P_B = 0$.

Для СМО с конечным буфером (при $\rho^N << 1$):

$$\gamma = \mu(1-p_0) = \mu\left[1 - \frac{1-\rho}{1-\rho^{N+1}}\right], \text{ откуда } P_B = \frac{(1-\rho)\rho^N}{1-\rho^{N+1}} \approx (1-\rho)\rho^N.$$

В качестве примера рассмотрим концентратор сети с коммутацией пакетов. Средняя длина поступающего пакета 1200 бит; если максимальная скорость в линии υ_{line}=2400 бит/с, то средняя пропускная способность составит μ=2 пакета/с. Если входной поток λ=1 пакет/с, то ρ=0.5 и при размере буфера N=9 пакетов по 1200 бит по формуле (5.27) имеем $P_B \approx 0.001$. Чтобы получить $P_B \le 10^{-6}$, необходим буфер N=19 пакетов по 1200 бит, т. е. всего около 2850 байт.

5.5. СМО с произвольным распределением времени обслуживания (M/G/1)

Подобные системы по классификации Кендалла обозначаются как M/G/m. Рассмотрим простейший случай M/G/1 с входным пуассоновским потоком интенсивности λ и временем обслуживания $x>0$ в сервере, имеющим ПРВ $b(x)$. При этом средняя длительность обслуживания равна $\overline{x}=m_{t_{об}}=\int_0^\infty x\,b(x)\,dx$.

Определим коэффициент использования сервера, т. е. долю времени, затрачиваемого им на обработку заявок, как $\rho=\lambda\overline{x}=\lambda/\mu$. Число заявок, находящихся в системе $N(t)$ не является Марковским процессом. В среднем каждая заявка, поступившая в СМО, будет ожидать обслуживания в течение времени обслуживания всех находящихся в очереди (пришедших ранее) заявок $\overline{N}_q\,\overline{x}$, а также времени освобождения $\overline{R}$ сервером последней из заявок (средней незавершенной работы). В итоге общее время ожидания

$$\overline{W}=\overline{N}_q\,\overline{x}+\overline{R}=\overline{W}\lambda\,\overline{x}+\overline{R}=\frac{\overline{R}}{1-\rho} \qquad (5.29)$$

с учетом формулы Литтла для числа заявок в очереди $\overline{N}_q=\lambda\overline{W}$.

Незавершенная работа (НР) в СМО с течением времени линейно убывает, пока сервер обрабатывает заявку и скачком увеличивается до величины x_i с каждым новым поступлением заявки на сервер (рис. 5.14).

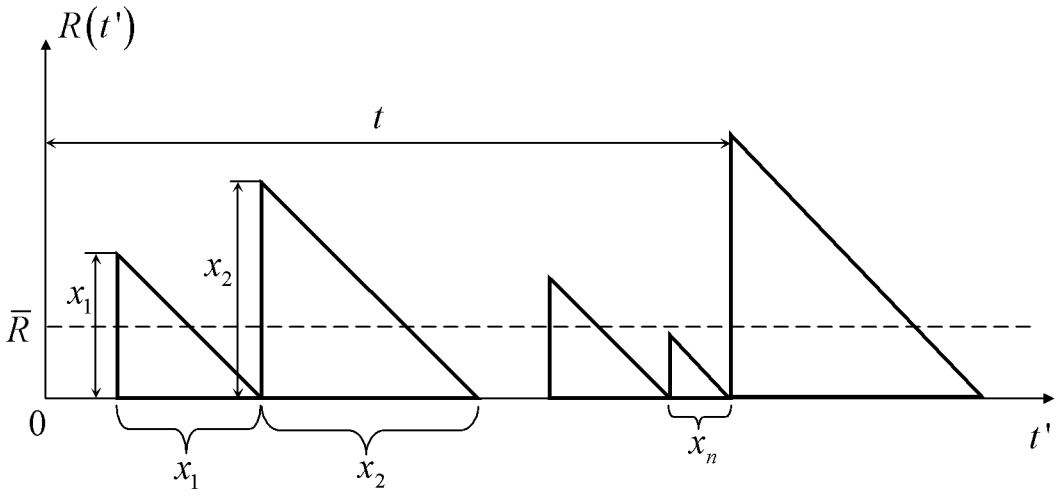


Рис. 5.14. График изменения незавершенной работы

Все участки имеют вид равнобедренных прямоугольных треугольников и $\bar{R}$ представляет собой их суммарную площадь [3]:

$$R=\frac{1}{t}\int_0^t R(t')dt'=\frac{1}{t}\sum_{i=1}^{n(t)}\frac{1}{2}x_i^2=\frac{1}{2}\cdot\frac{n(t)}{t}\cdot\frac{\sum_{i=1}^{n(t)}x_i^2}{n(t)},$$

$$\bar{R}=\lim_{t\to\infty}\frac{1}{t}\int_0^t R(t')dt'=\frac{1}{2}\lim_{t\to\infty}\left(\frac{n(t)}{t}\right)\cdot\lim_{t\to\infty}\left(\frac{1}{n(t)}\sum_{i=1}^{n(t)}x_i^2\right)\approx\frac{\lambda\overline{x^2}}{2},$$

$$\lambda=\lim_{t\to\infty}\left(\frac{n(t)}{t}\right),\quad \overline{x^2}=\lim_{t\to\infty}\left(\frac{1}{n(t)}\sum_{i=1}^{n(t)}x_i^2\right),$$

где $R(t')$ – остаточное время обслуживания, т. е. время, необходимое для завершения обслуживания заявки в текущий момент времени t'; $n(t)$ – количество заявок, для которых обслуживание уже завершилось на отрезке $[0,t]$.

Среднее время ожидания обслуживания в СМО типа M/G/1 определяется по **формуле Поллачека-Хинчина**:

$$\bar{W}=\frac{\bar{R}}{1-\rho}=\frac{\lambda\overline{x^2}}{2(1-\rho)}. \qquad (5.30)$$

Получим выражение для среднего времени пребывания заявки в СМО типа M/G/1 $\bar{T}=\bar{x}+\bar{W}$. Введем коэффициент вариации (формы) $C_v^2=(\sigma/\bar{x})^2$, где $\sigma^2=\overline{x^2}-(\bar{x})^2$. Отсюда можно получить:

$$\bar{W}=\frac{\lambda\left[\sigma^2+\bar{x}^2\right]}{2(1-\rho)}=\frac{\lambda\left[\left(\sigma^2/\bar{x}^2\right)+1\right]\bar{x}^2}{2(1-\rho)}=\frac{1+C_v^2}{2}\cdot\frac{\rho}{1-\rho}\cdot\bar{x}, \qquad (5.31)$$

$$\bar{T}=\bar{x}+\frac{\lambda\overline{x^2}}{2(1-\rho)}=\left[1+\frac{1+C_v^2}{2}\cdot\frac{\rho}{1-\rho}\right]\bar{x}. \qquad (5.32)$$

По формуле Литтла вычислим среднее число заявок в очереди:

$$\bar{N}_q=\lambda\bar{W}=\frac{\lambda^2\overline{x^2}}{2(1-\rho)}=\frac{1+C_v^2}{2}\cdot\frac{\rho^2}{1-\rho}$$

Среднее число заявок в системе:

$$\bar{N}=\lambda\bar{T}=\lambda\bar{x}+\frac{\lambda^2\overline{x^2}}{2(1-\rho)}=\rho+\frac{1+C_v^2}{2}\cdot\frac{\rho^2}{1-\rho}. \qquad (5.33)$$

Проверим формулу (5.30) для СМО типа M/M/1:

$$b(x)=\mu e^{-\mu x},\ x>0;\qquad \overline{x^2}=\int_0^\infty x^2 b(x)dx=\mu\int_0^\infty x^2 e^{-\mu x}dx=\frac{2}{\mu^2}.$$

Подставляя $\overline{x^2}$ в формулу (5.30), имеем:

$$\overline{N} = \rho + \frac{\rho^2}{1-\rho} = \frac{\rho}{1-\rho}; \quad \overline{T} = \left(1 + \frac{\rho}{1-\rho}\right)\overline{x} = \frac{\overline{x}}{1-\rho},$$

где $C_v^2 = 1$, т. е. $\sigma = \overline{x}$.

Рассмотрим СМО с детерминированным временем обслуживания (M/D/1), примером которой может являться система обработки пакетов постоянной длины. Для таких систем $\sigma = 0$, т.е. $\overline{x^2} = \overline{x}^2$ и, соответственно, $C_v^2 = 0$.

Соответствующие характеристики СМО типа M/D/1:

$$\overline{N} = \rho + \frac{\rho^2}{2(1-\rho)}; \quad \overline{T} = \left[1 + \frac{\rho}{2(1-\rho)}\right]\overline{x}.$$

Обычно пакеты имеют заголовок постоянной длины (чем заголовок длиннее – тем целесообразнее применять формулу (5.30)) и экспоненциальное распределение $b(x) = \mu e^{-\mu x}$, $x > 0$ не всегда удовлетворительно описывает время обслуживания пакетов.

Пример. На вход СМО типа M/D/1 : ∞ поступает пуассоновский поток с плотностью λ=20 пакетов/с. Длительность обслуживания в сервере имеет постоянное значение 0.04 с. Найти:

1) среднее время пребывания заявки в системе;

2) среднее число заявок в системе.

Решение. Нагрузка в системе $\rho = \lambda \overline{x} = 20 \cdot 0.04 = 0.8$. Среднее время пребывания заявки в системе

$$T = \left[1 + \frac{\rho}{2(1-\rho)}\right]\overline{x} = \left[1 + \frac{0.8}{2(1-0.8)}\right]0.04 = 0.12 \text{ с.}$$

Среднее число заявок в системе

$$\overline{N} = \rho + \frac{\rho^2}{2(1-\rho)} = 0.8 + \frac{0.8^2}{2(1-0.8)} = 2.4.$$

5.6. СМО с самоподобным входным потоком и детерминированным временем обслуживания (fBM/D/1)

Традиционные модели трафика, основанные на кратковременно зависимых процессах (в частности, пуассоновском процессе), часто не в состоянии описать характер реального трафика. В этом случае ста-

новится необходимым исследовать СМО с долговременно зависимыми процессами на входе.

Рассмотрим СМО, на вход которой поступает поток, который описывается ФБД с интенсивностью λ [27]. Число поступивших пакетов за время t определяется величиной

$$N(t) = \sup_{s \le t} \{A(t) - A(s) - \mu(t-s)\}, \tag{5.34}$$

причем

$$A(t) = \lambda t + \sqrt{\lambda a}\, Z(t), \tag{5.35}$$

где a – масштабный коэффициент; $Z(t)$ – нормированное ФБД с параметром Хёрста H.

Многие свойства ФБД отличаются от тех, которые имеют место у большинства стохастических процессов, обычно выступающих в качестве моделей трафика. Следовательно, большинство стандартных методов теории очередей становятся непригодными. Роли трех параметров модели трафика (5.35) могут быть разграничены таким образом: H и a характеризуют качество трафика, а средняя интенсивность λ характеризует его объём.

Несколько интересных свойств фрактальной броуновской модели накопления $N(t)$ могут быть прослежены только из предположения самоподобия, если рассматривать $N(t)$ в качестве стохастического процесса в различных масштабах времени.

Типичным требованием телекоммуникационных приложений является то, что вероятность превышения величиной нагрузки в системе некоторого уровня n (размера очереди в бесконечной модели накопления) должна быть равна «параметру качества обслуживания» ε. Поэтому соотношение

$$\varepsilon = P\{N > n\} \tag{5.36}$$

должно выполняться при максимальной допустимой нагрузке.

Вероятность того, что количество пакетов N в системе превысит заданную величину n, определяется выражением

$$P\{N(t) > n\} \ge \Phi\left(\frac{1}{\sqrt{a\lambda}}\left(\frac{1-\lambda}{H}\right)^{H}\left(\frac{n}{1-H}\right)^{1-H}\right), \tag{5.37}$$

где $\Phi(y) = P(Z(1) > y) \approx \dfrac{1}{\sqrt{2\pi}(1+y)} e^{-y^2/2}$ – дополнительная функция распределения стандартного гауссова распределения и ее аппроксимация. Выражение (5.37) характеризует вероятность блокировки па-

кета при ограниченном объеме буфера n пакетов. Если задать некоторую величину $p_B = P(N > n) = \varepsilon$ вероятности блокировки пакета в системе, то можно получить формулу для расчета объема буфера [27]:

$$\frac{1-\rho}{\rho^{1/(2H)}}\mu^{\left(H-\frac{1}{2}\right)/H} n^{(1-H)/H} = a^{1/(2H)}\Phi^{-1}(\varepsilon) = \text{const}, \qquad (5.38)$$

где $\rho = \lambda/\mu$ – входная нагрузка. Если при фиксированном значении μ выразить величину n (средняя длина очереди или необходимый объем буфера), то можно получить

$$\overline{n} = \text{const} \cdot \frac{\rho^{1/(2-2H)}}{(1-\rho)^{H/(1-H)}}. \qquad (5.39)$$

Из формулы (5.39) видно, что когда H велико (находится в диапазоне 0.8…0.9), существенное увеличение коэффициента ρ требует гораздо большей емкости буфера (рис. 5.15).

Параметр Хёрста имеет большое значение для оценки длины очереди, показывая, насколько неадекватны традиционные модели трафика, когда реальный трафик является самоподобным. В ряде работ [26-30] показано, что для пакетного трафика без установления соединения коэффициент использования ρ не может быть существенно улучшен за счет увеличения размера буфера (рис. 5.15).

Полагая H=0.5 (броуновское движение) в (5.38), получим СМО, близкую к M/M/1, для которой

$$\overline{n} = \text{const}\frac{\rho}{1-\rho}. \qquad (5.40)$$

При аппроксимации интенсивного трафика для СМО M/D/1 из рис. 5.15 видно, что снижение относительной свободной пропускной способности $1-\rho$ наполовину приводит к удвоению размера очереди. Интенсивность обслуживания μ также воздействует на коэффициент использования ρ. Фиксируя n, можно выразить μ из (5.38) и получить

$$\mu = \mu(\rho) = \text{const} \cdot \frac{\rho^{1/(2H-1)}}{(1-\rho)^{H/(H-0.5)}}. \qquad (5.41)$$

Важное практическое следствие формулы (5.41) заключается в том, что каналы передачи с более высокой пропускной способностью могут быть задействованы с бо́льшим коэффициентом использования ρ без увеличения размера буфера. Это объясняется повышенной эффективностью при мультиплексировании.

Используя формулу Литтла, можно определить среднее время пребывания пакета в системе типа fBM/D/1 [14]

$$\mu T = \text{const} \cdot \frac{\rho^{(H-0.5)/(1-H)}}{(1-\rho)^{H/(1-H)}}. \qquad (5.42)$$

Аналогичные выражения для системы M/M/1 $\mu T = \dfrac{1}{1-\rho}$

и для системы M/D/1 $\mu T = \dfrac{1}{1-\rho}\left(1-\dfrac{\rho}{2}\right)$.

Как видно из рис. 5.15, наибольшее время пребывания заявки (время обработки) в системе имеет место для системы с самоподобным трафиком.

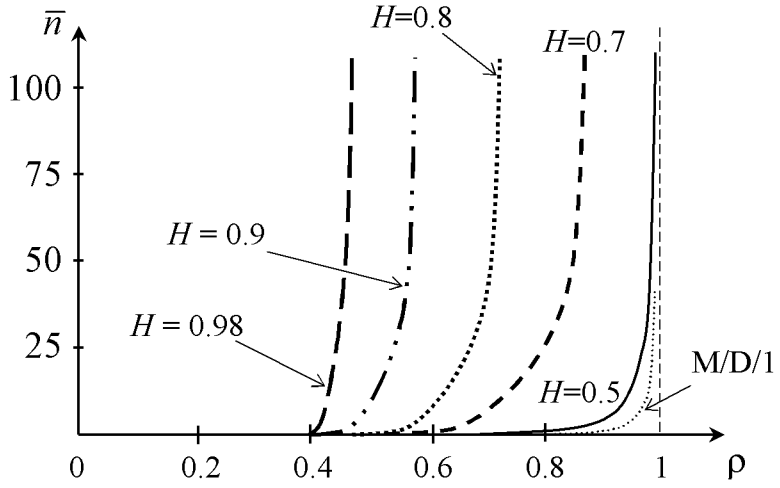


Рис. 5.15. Средняя длина очереди как функция от нагрузки

Неожиданным результатом для исследователей стал тот факт, что время обработки превысило даже оценку, полученную в предположении экспоненциального закона распределения интервала между поступающими заявками и времени их обработки, т. е. системы типа M/M/1. Ранее считалось, что система M/M/1 дает самую завышенную оценку для времени пребывания, по сравнению со всеми другими способами оценивания. Таким образом, системы с самоподобным входным потоком приводят к наибольшему ухудшению характеристик QoS по сравнению с любыми другими типами входного потока.

Несмотря на то, что представленные результаты справедливы для частного случая, когда самоподобный процесс представляет собой ФБД, они позволяют получить представление о влияния самоподобия и долговременной зависимости на характеристики СМО.

5.7. СМО с самоподобным характером времени обслуживания

Измерения реального телекоммуникационного трафика показали, что для некоторых цифровых сетей связи, например Ethernet, Telnet, Opnet и др., адекватными являются модели, в которых интервалы времени между поступающими заявками могут быть описаны с помощью экспоненциального распределения, а время обслуживания с долговременной зависимостью – с помощью ПРВ, например, Парето или Вейбулла [26-30]. При этом зависимости задержки и вероятности потери от размера буфера, получаемые теоретически и основанные на марковских моделях трафика, сильно отличаются от тех зависимостей, которые наблюдаются на практике. С увеличением емкости буфера вероятность потери снижается значительно медленнее, чем в случае экспоненциальной зависимости (для марковских моделей). Средняя задержка пакета всегда увеличивается с ростом емкости буфера, отличаясь от марковских моделей, для которых средняя задержка не превышает фиксированного предела независимо от размера буфера.

Рассмотрим однолинейную СМО типа M/G/1 с буфером неограниченного объема и одним сервером, на входе которой имеем простейший поток со средней интенсивностью λ. По формуле Поллачека-Хинчина (5.30) для среднего числа пакетов в системе можно записать [27]

$$\overline{n} = \rho + \frac{1+C_v^2}{2}\cdot\frac{\rho^2}{1-\rho} = \left(\frac{\rho}{1-\rho}\right)\left[1-\frac{\rho}{2}\left(1-\mu^2\sigma^2\right)\right], \qquad (5.43)$$

где σ^2 – дисперсия длины пакета на входе.

Из (5.43) следует, что для модели M/M/1, когда $\sigma^2 = 1/\mu^2$, справедливо соотношение (5.40), а для модели M/D/1, когда $\sigma^2 = 0$, справедливо соотношение

$$\overline{n} = \left(\frac{\rho}{1-\rho}\right)\left(1-\frac{\rho}{2}\right). \qquad (5.44)$$

Определим среднее время нахождения пакета в системе и среднее число пакетов в системе для случая, когда время обработки пакетов подчиняется ПРВ Парето (модель M/fBM/1). На основе ПРВ Парето (п. 6.2.2) можно определить среднее время обработки одного пакета на сервере:

$$\overline{x} = \int_k^\infty \frac{x\alpha k^\alpha}{x^{\alpha+1}}dx = \frac{\alpha k}{\alpha-1},$$

где $\alpha > 0$, $k > 0$ – параметры ПРВ Парето.

Входную нагрузку определим как

$$\rho = \lambda \bar{x} = \frac{\lambda \alpha k}{\alpha - 1}.$$

Определим среднее время ожидания пакета в буфере с помощью (5.30)

$$\bar{W} = \frac{\lambda \overline{x^2}}{2(1-\rho)},$$

где $\overline{x^2}$ – второй момент для времени обслуживания, который определяется как

$$\overline{x^2} = \int_0^\infty x^2 w(x) dx.$$

Для дальнейших расчетов данную величину удобно выразить через дисперсию и математическое ожидание времени обслуживания $\overline{x^2} = \sigma^2 + \bar{x}^2$.

В случае распределения Парето дисперсия и средний квадрат

$$\sigma^2 = \frac{\alpha k^2}{(\alpha-1)^2(\alpha-2)}, \qquad \overline{x^2} = \frac{\alpha k^2}{(\alpha-2)}.$$

Среднее время ожидания пакета в буфере системы:

$$\bar{W} = \frac{\lambda}{2(1-\rho)} \frac{\alpha k^2}{(\alpha-2)}.$$

Используя формулу Литтла, найдем среднее число пакетов в буфере:

$$\bar{N}_q = \lambda W.$$

Аналогично можно определить среднее время пребывания пакета в системе, а также среднее число пакетов. Кроме того, подставляя вместо величин σ^2 и $\bar{x}$ значения для других ПРВ, например, Вейбулла, можно подобным образом выполнять анализ цифровых систем для других ПРВ.

5.8. Контрольные вопросы

1. Запишите и поясните формулу Литтла.
2. Расшифруйте обозначение M/M/3.
3. Расшифруйте обозначение E^k/G/2 : 5/FCFS.
4. Расшифруйте обозначение fGN/D/3 : ∞.
5. Расшифруйте обозначение G/fBM/3 : 7.

6. Поясните особенности СМО с неполнодоступным включением серверов?
7. В соответствии с каким признаком обычно подразделяются различные классы заявок?
8. Как вычисляется средняя длина очереди в одноканальной СМО?
9. Как вычисляется средняя длина очереди в двухканальной СМО?
10. Какому ЗР подчиняется вероятность длины очереди в СМО типа M/M/1?
11. Как вычисляется среднее число заявок в СМО типа M/M/1?
12. Как изменяется число заявок в системе и время пребывания при приближении коэффициента использования ρ к единице?
13. Какие дисциплины обслуживания Вы знаете?
14. Как функционирует СМО с приоритетами? Какие типы приоритета вы знаете?
15. В чем особенность СМО с абсолютным приоритетом?
16. Из каких трех составляющих складывается среднее время ожидания в очереди для меченой заявки?
17. Дайте понятие о незавершенной работе в СМО.
18. Как выглядит закон сохранения для консервативной СМО?
19. Что представляет собой остаточное время обслуживания?
20. Какой параметр в СМО может быть описан с помощью геометрического закона распределения?
21. Запишите и поясните формулу Поллачека-Хинчина.
22. Для каких СМО (по классификации Кендалла) может быть использована формула Поллачека-Хинчина?
23. Каков характер убывания незавершенной работы в СМО с течением времени, пока сервер обрабатывает заявку?
24. Каким образом изменяется незавершенная работа в СМО при поступлении новой заявки на сервер?
25. Что является более эффективным: увеличение вдвое производительности сервера в СМО M/M/1 или введение параллельного сервера той же производительности (СМО M/M/2)? Обоснуйте этот вывод формулой.
26. Входные потоки какого типа приводят к наибольшему ухудшению характеристик QoS по сравнению с любыми другими типами входного потока в СМО?
27. Какие параметры модели трафика в СМО типа fBM/D/1 характеризуют качество трафика?
28. В чем заключается особенность самоподобного трафика при H =0.5?

29. Какой параметр в СМО резко возрастает при больших значениях H и малой нагрузке ρ ?
30. Каким образом изменяется средняя задержка пакета с ростом емкости буфера в СМО с самоподобным характером времени обслуживания?
31. Каким законам распределения обычно подчиняется время обработки пакетов в СМО с самоподобным характером времени обслуживания?

5.9. Задачи

1. Вычислите дисперсию числа заявок в СМО типа M/M/1 с нагрузкой $\rho=0.5$.
2. Вычислите вероятность того, что в СМО типа M/M/1 будет находиться не менее 3 заявок.
3. Определите время ожидания в очереди в среднем по всем заявкам в СМО типа M/M/1 при нагрузке $\rho=0.8$ и $\mu=2$.
4. Определите полное нормированное время пребывания заявки в системе M/M/1 при нагрузке $\rho=0.4$.
5. Найдите среднее время пребывания в системе в СМО типа M/M/1 при нагрузке $\rho=0.5$ и $\mu=2$.
6. Определите среднее время пребывания заявки в СМО типа M/M/2 при нагрузке $\rho=0.6$ и $\mu=3$.
7. Определите среднее число заявок в СМО типа M/M/1 : 3 при нагрузке $\rho=0.6$.
8. Определите вероятность отказа в СМО типа M/M/1 : 4 при нагрузке $\rho=0.8$.
9. Определите среднее число заявок в СМО типа M/G/1 при нагрузке $\rho=0.5$ и $C_v^2=0.5$.
10. Определите среднее число заявок в СМО типа M/D/1 при нагрузке $\rho=0.9$.

6. АНАЛИЗ И МОДЕЛИРОВАНИЕ ТРАФИКА В СЕТЯХ

6.1. Задачи анализа и моделирования ТС

Задача анализа ТС заключается в измерении и анализе реальных характеристик QoS ТС и сравнении их с проектными с целью определения причин ухудшения QoS. Для этого часто используют физические модели ТС (или типового фрагмента сети) в виде сетеобразующего оборудования, соединенного между собой имитаторами каналов реальной сети. Более эффективными являются математические (имитационные) модели (ММ) ТС в виде компьютерных программ (GPSS, OpNet) – такие модели удобны для внесения всевозможных изменений и наглядного представления (визуализации внутренних процессов) и их можно строить задолго до построения самой реальной системы [8, 14, 22].

6.2. Моделирование трафика в сетях

6.2.1. Моделирование пуассоновского потока

Для получения СВ, описывающей интервалы времени между соседними заявками, поступающими на вход ТС и имеющей показательный (экспоненциальный) закон распределения [4, 6]

$$w(y)=\lambda e^{-\lambda y},\ F(y)=1-e^{-\lambda y},\ y\ge 0,$$

решая уравнение $F(y)=x$, т. е. $1-e^{-\lambda y}=x$, находим обратную функцию $y=-\frac{1}{\lambda}\ln x$. Таким образом, показательную СВ y можно сформировать из СВ x, равномерно распределенной в интервале [0, 1], с помощью функционального преобразования $y=-\frac{1}{\lambda}\ln x$.

СВ с показательным законом распределения можно также получить путем преобразования системы двух независимых нормальных случайных чисел x_1 и x_2 с параметрами $(0,\sigma^2)$ в виде

$$y=x_1^2+x_2^2.$$

При этом параметр λ показательного распределения связан с параметром σ исходного нормального распределения соотношением $\lambda=0.5\sigma^2$.

6.2.2. Моделирование самоподобных случайных процессов

Реальный трафик в цифровых сетях связи (Internet, Ethernet, Telnet и др.) лучше всего описывается самоподобными случайными процессами (пульсирующие источники). Ниже представлены модели самоподобного трафика, выраженные в виде ЗР соответствующих параметров трафика.

Рассмотрим логнормальное распределение [28], которое формируется из нормального распределения

$$Z = \ln X,$$

где Z – нормально распределенная СВ с нулевым средним; X – СВ распределенная по логнормальному закону с ПРВ

$$w(x) = \frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\ln x - m)^2}{2\sigma^2}}, \text{ при } x > 0,$$

где σ – среднеквадратическое отклонение СВ Z; $m = \ln a$ – математическое ожидание. Эти параметры можно определить на основе экспериментальных данных:

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i; \ \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\ln(x_i) - \hat{m}\right)^2}.$$

На основе известных значений m и σ нормально распределенной СВ Z можно определить математическое ожидание и дисперсию для логнормально распределенной СВ X:

$$m_x = e^{m+\frac{\sigma^2}{2}} = a e^{\sigma^2/2}; \qquad \sigma_x = e^m \sqrt{e^{\sigma^2}\left(e^{\sigma^2} - 1\right)} = a\sqrt{e^{\sigma^2}\left(e^{\sigma^2} - 1\right)}.$$

Генерировать логнормальные СВ X можно путем следующего функционального преобразования:

$$x_i = e^{u_i},$$

где u_i – нормально распределенная СВ с нулевым средним и единичной дисперсией.

Для моделирования размеров передаваемых файлов, типовых сообщений, Web-страниц часто используется распределение Парето с ПРВ

$$w(x) = \frac{\alpha k^\alpha}{x^{\alpha+1}}; \ \alpha > 0; \ x \geq k,$$

где α – параметр формы; k – параметр, определяющий нижнюю границу для случайной величины.

ЭВМ позволяет генерировать равномерно распределенные СВ y в диапазоне от 0 до 1. Для того чтобы смоделировать СВ x с ПРВ Па-

рето необходимо найти функциональное преобразование $x = f(y)$ из выражения

$$w\left(x = f\left(y\right)\right) = w\left(y\right)\left|\frac{dy}{dx}\right|,$$

откуда

$$y = \int w\left(x\right)dx = \int \frac{\alpha k^{\alpha}}{x^{\alpha+1}}dx = 1 - \left(\frac{k}{x}\right)^{\alpha},\ \text{при}\ x > k,\ \alpha > 0.$$

Далее легко можно получить следующее выражение для моделирования СВ, имеющей ПРВ Парето:

$$x = \frac{k}{\sqrt[\alpha]{1-y}},$$

где y – случайное число с равномерной ПРВ, генерируемое на ЭВМ в диапазоне от 0 до 1. Параметр α связан с показателем Хёрста выражением

$$\alpha = 3 - 2H.$$

Параметр α обычно вычисляется на основе метода максимального правдоподобия по известным результатам измерения интенсивности реального трафика $X = \{x_1, x_2, \ldots, x_n\}$:

$$\hat{\alpha} = \frac{n-1}{\sum_{i=1}^{n} \ln\left(x_i\right) - n\log\left(\hat{k}\right)},$$

где $\hat{k} = \min_i x_i$.

Рассмотрим распределение Вейбулла с ПРВ

$$w\left(x\right) = \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}},$$

где α и β – некоторые параметры, влияющие на форму ПРВ [28]. Для моделирования СВ x с ПРВ Вейбулла используется функциональное преобразование вида

$$x = \sqrt[\beta]{\frac{-\ln\ y}{\alpha}},$$

где y – случайное число с равномерной ПРВ, генерируемое на ЭВМ в диапазоне от 0 до 1.

На основе вычисленных значений трафика $X = \{x_1, x_2, \ldots, x_n\}$ можно путем моделирования определить среднюю длину очереди в буфере при передаче данных по каналу связи с заданной пропускной способностью v (например, кбит/с). Алгоритм моделирования заключается в следующем. Если значение передаваемого трафика (па-

кетов или байт в единицу времени) $x_i > v$, то в буфере передачи запоминается объем данных равный $s_i = x_i - v$, в противном случае $s_i = 0$. В следующий момент времени $i+1$ объем передаваемых данных составит величину $x_{i+1} + s_i$. Если данное значение больше пропускной способности канала передачи v, то в буфере запоминается объем данных $s_{i+1} = x_{i+1} + s_i - v$. И так далее, до тех пор, пока все данные не будут переданы по каналу связи. Таким образом, получаем набор значений $\{s_i\}$, $i = 1,2,...,N$, которые будут характеризовать объем данных, находящихся в буфере в моменты времени t_i. Следовательно, средняя длина очереди в буфере:

$$\overline{N}_q = \frac{1}{n}\sum_{i=1}^{n} s_i .$$

Если размер буфера резко возрастает при уменьшении v, то канал с такой пропускной способностью «не справляется» с требуемым объемом передаваемых данных, и, следовательно, значение v нужно увеличивать.

Зная средний объем использования буфера, по формуле Литтла можно определить среднее время нахождения единицы данных (байт, пакет и т. д.) в буфере:

$$\overline{T} = \overline{N}_q / \overline{\lambda},$$

где $\overline{\lambda}$ – средняя интенсивность входного потока, которая определяется как

$$\overline{\lambda} = \int_0^{\infty} x\, w(x)\, dx,$$

где $w(x)$ – используемая ПРВ при моделировании входного потока.

Рассмотрим алгоритм моделирования **фрактального броуновского движения** (ФБД) на основе **RMD-метода** [25].

Шаг 1. Формируются два отсчета $x(0)$ и $x(1)$, причем $x(0) = 0$ (по условию), а $x(1) = \xi(1)$ – нормально распределенная СВ с нулевым МО и дисперсией σ_0^2 (рис. 5.8).

Шаг 2. На интервале от 0 до 1 берется центральный отсчет $x(1/2)$, который определяется как

$$x(1/2) = 1/2 \cdot (x(0) + x(1)) + d_1,$$

где d_1 – гауссовская СВ с нулевым средним и дисперсией, выбранной так, чтобы дисперсия сформированного отсчета $x(1/2)$ была равна

$$\sigma_0^2\left|\Delta t\right|^{2H} = \sigma_0^2(t_2 - t_1)^{2H} = \sigma_0^2(1/2)^{2H},$$

где $\Delta t = 1/2$ – интервал между отсчетами на данном шаге алгоритма.

Дисперсия линейной аппроксимации $1/2 \cdot (x(0) + x(1))$ равна $\sigma_0^2/4$, тогда можно записать

$$\sigma_1^2 = \left(\frac{1}{2}\right)^{2H} \sigma_0^2 - \frac{1}{4}\sigma_0^2 = \left(\frac{1}{2}\right)^{2H} \left(1 - 2^{2H-2}\right)\sigma_0^2.$$

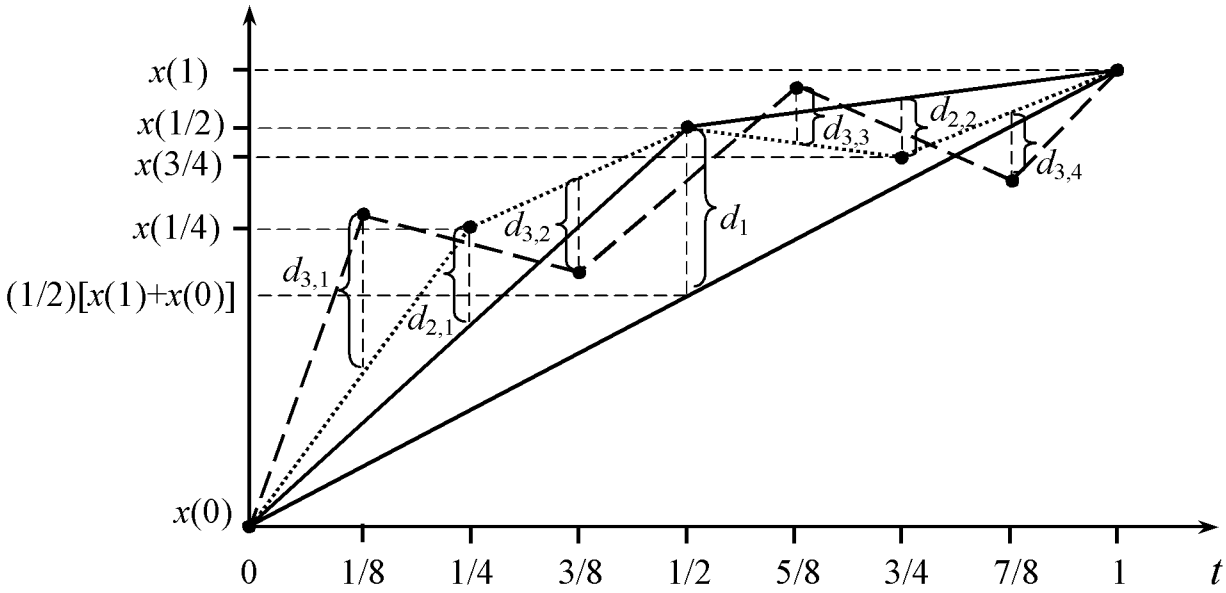


Рис. 6.1. Шаги генерации ФБД

Шаг 3. Рассматриваются поочередно два интервала (0; 1/2) и (1/2; 1), в которых выделяются центральные отсчеты $x(1/4)$ и $x(3/4)$. Значения этих отсчетов формируются аналогично величине $x(1/2)$:

$$x(1/4) = \frac{1}{2}\left(x(0) + x(1/2)\right) + d_{2,1};\ x(3/4) = \frac{1}{2}\left(x(1/2) + x(1)\right) + d_{2,2},$$

где $d_{2,1}$ и $d_{2,2}$ – гауссовские СВ с нулевым средним и дисперсией выбранной так, чтобы дисперсии сформированных отсчетов $x(1/4)$ и $x(3/4)$ удовлетворяли условию

$$\sigma_0^2\left|\Delta t\right|^{2H} = \sigma_0^2(t_2 - t_1)^{2H} = \sigma_0^2(1/4)^{2H},$$

где $\Delta t = 1/4$ – интервал между отсчетами на данном шаге алгоритма.

Для отсчета $x(1/4)$ дисперсия линейной аппроксимации $1/2 \cdot (x(0) + x(1/2))$ равна $\sigma_0^2(1/2)^{2H}/4$, тогда можно записать

$$\sigma_2^2 = \left(\frac{1}{4}\right)^{2H} \sigma_0^2 - \left(\frac{1}{2}\right)^{2H} \frac{\sigma_0^2}{4} = \left(\frac{1}{2^2}\right)^{2H} \left(1 - 2^{2H-2}\right)\sigma_0^2.$$

Аналогичное значение дисперсии сохраняется и для величины $d_{2,2}$.

Шаги 4…n. На данных шагах повторяются действия, описанные на втором и третьем шагах. Причем для значений дисперсий случайных добавок d_{ij} имеем следующую формулу:

$$\sigma_{ij}^2 = \left(\frac{1}{2^i}\right)^{2H} \left(1 - 2^{2H-2}\right)\sigma_0^2, \quad j = 1, 2, \ldots$$

6.3. Сети Петри как эффективная модель СМО

Сети Петри позволяют детально описывать и исследовать дискретные динамические системы и разработаны специально для моделирования систем с взаимодействующими параллельными компонентами [14]. **Формальная теория сетей Петри** занимается разработкой основных средств, методов и понятий, необходимых для применения сетей Петри. **Прикладная теория сетей Петри** связана главным образом с применением сетей Петри к моделированию вычислительных систем и их анализу. Различные расширения сетей Петри направлены на учет временных, вероятностных характеристик, использование данных, построение иерархических моделей и т. д. Одно из основных достоинств сетей Петри заключается в том, что они могут быть представлены как в графической форме (наглядность), так и в аналитической.

Впервые сети Петри предложил немецкий математик К.А. Петри в своей докторской диссертации «Kommunikation mit Automaten» («Связь автоматов»), где он сформулировал основные понятия теории связи асинхронных компонентов вычислительной системы и подробно рассмотрел описание причинных связей между событиями.

При графической интерпретации сеть Петри является графом особого вида, состоящим из вершин двух типов – **позиций (position)** и **переходов (transition)**, соединенных ориентированными дугами, причем каждая дуга может связывать лишь разнотипные вершины (позицию с переходом или переход с позицией). Вершины-позиции обозначаются кружками, вершины-переходы – прямоугольниками (или черточками) (рис. 6.2, а).

Переходы соответствуют событиям, присущим исследуемой системе, а позиции – условиям их возникновения. Переход (событие) характеризуется определенным числом входных и выходных позиций, соответствующих предусловию и постусловию данного события. Совокупность переходов, позиций и дуг позволяет описать статическую систему.

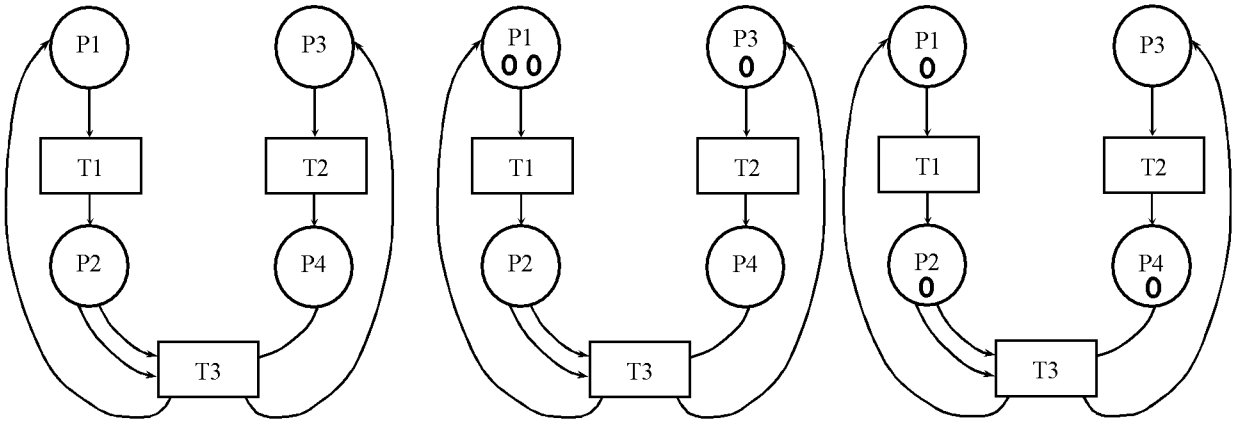


Рис. 6.2. Графическая интерпретация сети Петри

Для описания динамики вводится **маркер (token)**, или метка позиции, которая соответствует выполнению того или иного условия (обозначается точкой внутри позиции). Расположение маркеров в позициях называется **разметкой сети**. Переход считается **активным**, если в каждой его входной позиции есть хотя бы один маркер, что равносильно выполнению всех необходимых условий для наступления события. Наступление события в терминах сетей Петри представляется **срабатыванием перехода** (на рис. 6.2, б, в – маркированная сеть Петри), при этом маркеры из входных позиций изымаются и добавляются в каждую выходную позицию. **Текущее состояние** исследуемой системы определяется распределением маркеров по позициям сети, а динамика поведения системы отображается перемещением маркеров по позициям сети.

Рассмотрим наиболее важные расширения сетей Петри. **Приоритетные сети** – это сети, учитывающие приоритетные соотношения между переходами. В сетях данного типа при наличии двух и более активных переходов сработать может лишь переход, имеющий высший приоритет. **Структурированные сети** служат для моделирования иерархических систем, которые наряду с неделимыми компонентами содержат составные компоненты, сами представляющие собой системы. В **раскрашенных сетях** каждому переходу ставится в соответствие функция, определяющая маркирование выходных позиций в зависимости от цветов входных маркеров. Расширение простых сетей в цветные заключается в добавлении перечисленной ниже информации к элементам сети:

– маркеры вместо простого обозначения выполнения условия пре-

образуются в объект, который может содержать в себе один или более параметров, каждый из которых способен принимать дискретный набор значений. В соответствии с этим маркеры различаются по типам параметров (переменных). Чтобы отличать маркеры различных типов, их можно окрашивать в различные цвета (поэтому сети называют цветными);

– к местам добавляется информация о типах маркеров, которые могут находиться в данном месте;

–к переходам может быть добавлена информация с предикатом активизации перехода, в зависимости от переменных, содержащихся в маркерах;

– к начальной маркировке сети добавляется информация о значениях переменных, содержащихся в маркерах.

В дополнение к описанным выше существуют расширения сетей Петри, с помощью которых можно аналитически определить некоторые количественные характеристики исследуемых систем.

Временные и стохастические сети Петри позволяют определить аналитически количественные характеристики исследуемых систем. **Стохастические сети** – это сети Петри, в которые вводятся некоторые вероятностные атрибуты, например вероятности или плотности вероятностей срабатывания активных переходов. Это позволяет наиболее полно позволяют описать элементы СМО.

На рис. 6.3 сеть Петри схематично описывает простейшую СМО, состоящую из источника заявок, сервера и очереди. Маркер в позиции P1 соответствует готовности источника заявок к выдаче очередной заявки. Обратная связь перехода T1 с позицией P1 необходима для генерации последующих заявок в каждую единицу времени, таким образом формируется входной поток заявок.

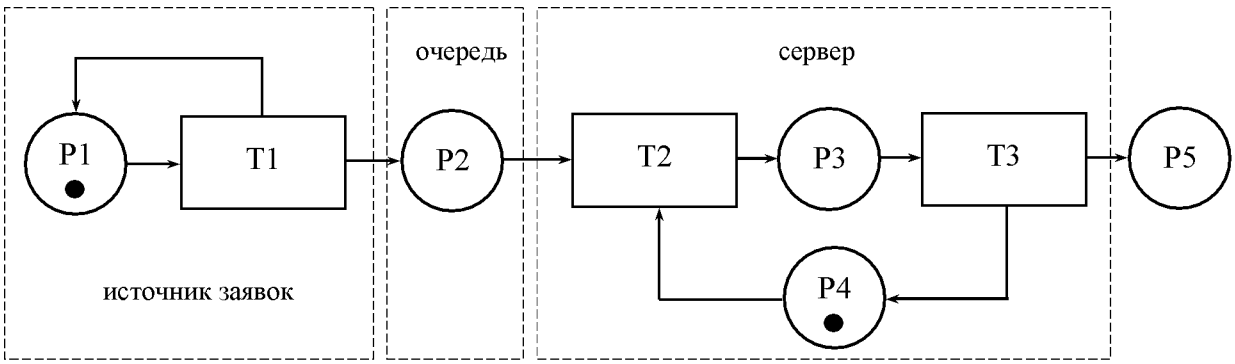


Рис. 6.3. Простейшая СМО в виде сети Петри

Позиция P2 моделирует очередь, которая в данном случае может быть бесконечна (т. к. на нее не наложены никакие ограничения), но может быть и всегда пустой (если сервер обладает бесконечной производительностью). Маркер в позиции P4 моделирует свободное состояние сервера (т. к. переход T2 может сработать и забрать очередную заявку из очереди только при наличии этого маркера). Соответственно отсутствие маркера в позиции P4 говорит о том, что в данный момент сервер занят.

Однако данная модель не представляет практической ценности, т. к. в приведенной модели не хватает информации о параметрах СМО. Поэтому данную сеть Петри необходимо преобразовать в стохастическую:

1) добавить информацию о характере входного трафика; для этого в переходе T1 следует записать математическую модель генерируемого потока заявок (с пуассоновским или другим ЗР);

2) добавить информацию о времени обслуживания сервера – она заносится в переход T3;

3) к позиции P2 можно добавить ограничение на длину очереди или на время пребывания в ней заявок.

В результате получаем модель, в которой переход T1 по заданному закону генерирует заявки, далее заявки поступают в очередь P2, и если сервер свободен, то обслуживаются и попадают в обслуженные заявки.

Бесплатные пакеты программ, реализующие принципы сетей Петри [14]:

– Visual Petri – разработка Института теоретической механики;

– ARP – низкоуровневые сети Петри: быстрое моделирование, пространство состояний, структурный анализ (MS-DOS);

– CPN Tools – высокоуровневые временные сети Петри: графический редактор; анимированное перемещение маркеров; импорт-экспорт данных (MS-Windows).

Имеется также коммерческий пакет Artifex (www.artis-software.com).

6.4. Измерение параметров трафика

6.4.1. Основные параметры трафика

Трафик может быть охарактеризован количественно **объемом**. В цифровых системах – число бит переданных в единицу времени, как мера занятости времени («пространства для движения»), в пределах которого доступен ресурс сети. **Объем трафика**, пропущенного ресурсом, – это суммарный интервал времени, в течение которого данный ресурс был занят за анализируемый период времени (**работа** ресурса за заданное время). Единицей работы можно считать **секундозанятие** (или часозанятие) ресурса. Рекомендации ITU [38] дают размерность объема трафика в **эрлангочасах** [14].

Если в каждый момент времени из интервала (t_1,t_2) число занятых обслуживанием трафика ресурсов (серверов) на заданном **пуле** (наборе) ресурсов равно $A(t)$, то **средняя интенсивность трафика** (нагрузки):

$$\overline{A}(t_1,t_2)=\frac{1}{t_2-t_1}\int_{t_1}^{t_2}A(t)dt\,.$$

Интенсивность трафика – среднее число ресурсов, занятых обслуживанием трафика на заданном интервале времени. Единица измерения интенсивности трафика – один **эрланг** (1 Эрл, 1 Е) – интенсивность трафика, которая требует полной занятости одного ресурса в единицу времени (ресурсом выполняется работа величиной в одно секундозанятие за одну секунду). **1 эрланг** = одна поступившая заявка за единицу времени с продолжительностью соединения, равной этой единице времени [8, 14].

Средняя интенсивность трафика вычисляется по формуле $A=\lambda T$, где λ – среднее число заявок в час, T – средняя продолжительность соединения. При λ=6 заявок/ч и длительности разговора T=2 мин, средняя нагрузка в линии 0.2 Эрл. Телефонная абонентская линия (единственный ресурс) обслуживает максимальную интенсивность трафика в один эрланг. За каждый час эта линия пропустит максимальный объем трафика в 1 эрлангочас, а за минуту – 1/60 эрлангочаса. В среднем один абонент создает нагрузку на местную АТС порядка 200 миллиэрлангов при усреднении за сутки, т. е. 1/5 времени в среднем в сутки абонентская линия занята, а оставшиеся 4/5 – простаивает.

Пример. Интенсивность трафика в линии зависит от времени $A(t)=e^{-at}$, где $a=0.02$. Вычислить среднюю интенсивность трафика в линии на промежутке времени от t_1=0 до t_2=60 секунд.

Решение.

$$\overline{A}(t_1,t_2)=\frac{1}{t_2-t_1}\int_{t_1}^{t_2}A(t)dt=\frac{1}{60}\int_{0}^{60}e^{-0.02t}dt=\frac{1}{1.2}\left(1-e^{-1.2}\right)=0.5823 \text{ Эрл.}$$

6.4.2. Распределение интенсивности нагрузки во времени

Рассмотрим СМО, имеющую несколько ресурсов (серверов), одновременно обслуживающих некоторый трафик (поток заявок). **Диаграмма Ганта** изображает процесс обслуживания заявок пулом серверов. При этом объем трафика вычисляется как суммарная длина всех отрезков диаграммы Ганта. Разность между объемами поступающего и пропущенного трафиков называется *объемом избыточного трафика* (overflow traffic). Приведенная на рис. 6.4 диаграмма отражает объем трафика за время в 10 секунд (в эрлангосекундах или секундозанятиях):

$$U(0,10) = 3\cdot2 + 2\cdot1 + 1\cdot1 + 2\cdot2 + 2\cdot1 = 15 \text{ [Es]}.$$

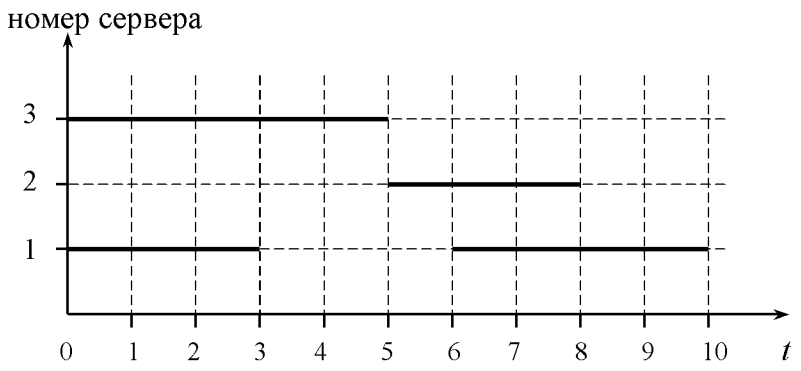


Рис. 6.4. Диаграмма Ганта для системы с тремя серверами

Каждому временному интервалу по оси абсцисс соответствует число серверов, занятых на этом интервале – **мгновенная интенсивность** $A(t)$. Если дискретность измерения составляет 1 секунду, то имеем последовательность:

$$A(t) = \{2, 2, 2, 1, 1, 1, 2, 2, 1, 1\}.$$

Средняя за период T=10 с интенсивность трафика, обслуженного системой из трех серверов:

$$\overline{A} = \frac{A(t)}{T} = 1{,}5[\text{E}].$$

В зависимости от выбора интервала усреднения оценка средней интенсивности будет различной. Рекомендации ITU E.500 [35] определяют для оценки интенсивности трафика в ТС интервал времени усреднения в 15 минут. По графику (рис. 6.5) можно определить часовой интервал с максимальной интенсивностью (сумма четырех соседних значений является наибольшей). Время, соответствующее этому интервалу, называют **часом наибольшей нагрузки** (ЧНН). Обычно интервал времени, соответствующий ЧНН, повторяется каждые сутки, например, с 11 до 12 часов ежедневно.

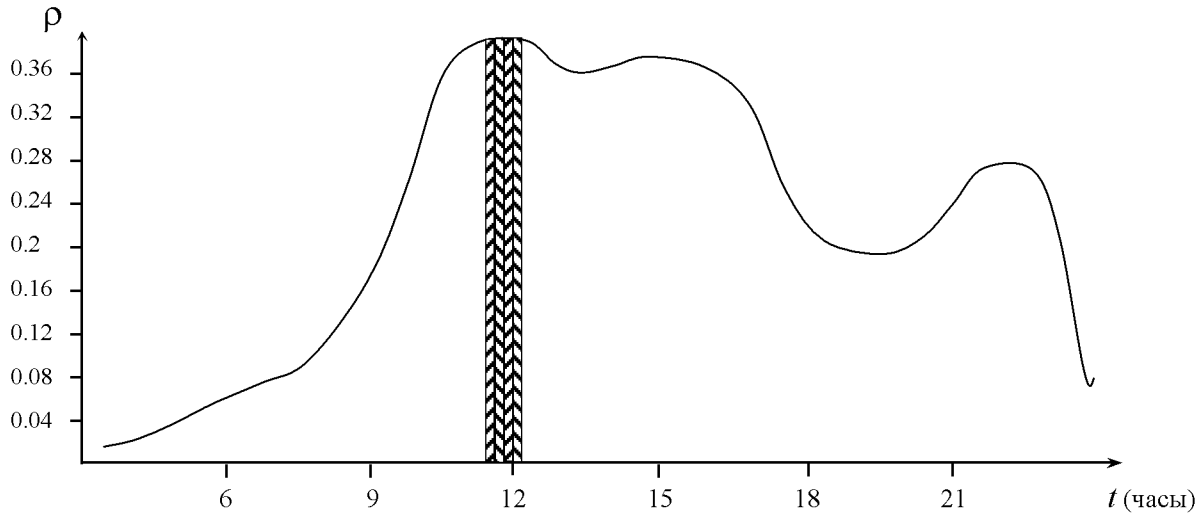


Рис. 6.5. Час наибольшей нагрузки (ЧНН) – 4 интервала по 15 минут

Более строго определение ЧНН производится следующим образом. Рекомендация ITU E.500 предписывает проанализировать данные об интенсивности за 12 месяцев, выбрать из них 30 наиболее загруженных дней, найти в эти дни наиболее загруженные часы и усреднить результаты измерения на этих интервалах (нормальная оценка интенсивности трафика в ЧНН, уровень А) [36]. Для получения более жесткой оценки проводится усреднение за 5 самых загруженных дней 30-дневного периода (повышенная оценка, уровень В).

Средняя интенсивность потока заявок (среднее число поступлений заявок в единицу времени, средняя частота поступлений):

$$\lambda = n/T,$$

где n – число обслуживаемых заявок, T – длительность интервала мониторинга.

Суммарная длина n отрезков случайной длины на диаграмме Ганта:

$$s = n\bar{x}$$

где $\bar{x}$ – средняя длительность занятия (обслуживания).

Средняя интенсивность трафика (нагрузка ρ) – отношение суммарного времени обслуживания к общему времени мониторинга:

$$\rho = \frac{s}{T} = \frac{n\bar{x}}{T} = \lambda\bar{x}\,.$$

Эта величина также называется **средней нагрузкой** и ассоциируется с **коэффициентом использования ресурса** (для одного сервера $\rho < 1$).

6.4.3. Измерение трафика в сетях

Трафик может существовать между любыми элементами сети (физическими или логическими устройствами, имеющими собственный сетевой адрес). Сообщения могут быть представлены пакетами символов различной длины и передаваемых друг за другом по каналу связи, соединяющем между собой **узлы сети** – устройства объединения, ответвления и перенаправления пакетов. Различают входящий и исходящий трафик для каждого сетевого элемента. Измерение трафика сводится к подсчету числа входящих или исходящих пакетов и измерению длины каждого из них [35].

Средняя за время T **интенсивность** трафика на данном ресурсе – отношение времени занятого пакетами к общему времени измерения:

$$A = \frac{\sum_{i=1}^{n} \frac{l_i}{R}}{T} = \frac{\sum_{i=1}^{n} l_i}{RT} \quad [\text{Эрл}],$$

где n – число пакетов, прошедших за время T [с]; l_i – длина i-го пакета [бит]; R – скорость передачи информации по каналу [бит/с].

В практике пакетных сетей интенсивность трафика оценивают величиной

$$I = \frac{\sum_{i=1}^{n} l_i}{T} = A \times R \ \text{ [бит/с или байт/с]}.$$

Интенсивность трафика измеряется не тем, насколько быстро перемещаются пакеты, а тем, насколько «плотно» они движутся. Если

сказать, что интенсивность трафика 300 пакетов/с со средней длиной пакета 1000 бит, то для сети Ethernet, где **пропускная способность** (максимальная скорость передачи) 10 Мбит/с, это одно (A = 0,003 Эрл), а для телефонного модема, работающего на максимальной скорости 32 000 бит/с, – совсем другое (A = 0,9375 Эрл). В последнем случае модем просто перегружен и имеют место недопустимые задержки.

6.4.4. Средства измерения параметров трафика

Чтобы измерить интенсивность трафика в сетях передачи пакетов необходимо регистрировать время появления каждого пакета и его длину. Это делают с помощью специальных приборов – **анализаторов пакетов.**

В настоящее время для исследователей доступны огромные архивы трафиковых данных из реально функционирующих сетей: CCSN/SS7, ISDN, Ethernet LAN, WAN, NSFNet (National Science Foundation Network), VBR (Variable Bit-Rate) [14].

Network sniffer (сетевой анализатор) – средство для проверки текущего состояния сетевого оборудования, перехвата сетевого трафика, а также сбора и анализа содержимого пакетов, включая расшифровку паролей.

Например, анализаторы протоколов серии RC-88/-100/-88WL/-100WL позволяют определять время доставки пакетов выбранного типа протокола, длину пакета, выделять ошибочные пакеты, систематизировать полученную информацию в виде графиков и таблиц. Измерения можно проводить и с помощью обычного компьютера со специальной программой – **монитором сети**. Большинство программ для мониторинга сети работает под управлением ОС UNIX. Программа **Network Monitor** компании Microsoft позволяет наблюдать сетевой трафик, проходящий через данный компьютер, а ее расширенная версия помогает захватывать весь сетевой трафик любого сегмента сети. Анализатор трафика может быть установлен на прокси-сервере (proxy) и улавливать, таким образом, всю информацию в выделенном Интернет-канале (2 Мбит/с).

Мониторинг пучка линий (**транк** (trunk) из 30–120 линий, обслуживающих трафик на некотором направлении), может быть проведен системой на базе многоканального преобразователя «напряжение-код» и компьютера. Преобразователь измеряет напряжение на каждой линии и в виде двоичного кода вводит эти данные в компьютер. Спе-

циальная программа измеряет напряжение на каждом канале с дискретностью около 5–10 секунд и после этого принимает решение о состоянии каждой линии (busy/idle), усредняет по интервалам в 15 минут и записывает результаты в log-файл.

Набор сетевых анализаторов Netboy австралийской компании NDG Software (http://www.ndgsoftware.com) содержит три отдельные сетевые утилиты: PacketBoy, EtherBoy и WebBoy, каждая из которых представляет собой спецпрограмму для сбора данных – пакетов, информации о функционировании сети Ethernet и протокола HTTP, соответственно.

Утилита PacketBoy является анализатором и декодером протоколов и предназначена для просмотра активности пользователей и анализа сетевых проблем. Это приложение захватывает и анализирует все сетевые пакеты, позволяет установить, настроить правила для включения захвата пакетов при определенных условиях, разрешает настроить фильтры для уже захваченных пакетов, а также производит декодирование протоколов TCP/IP (включая протоколы уровня приложений, такие как NFS, RPC, HTTP), протоколов Novell Netware (IPX) (включая NCP, SPX, SAP и RIP) и протоколов Appletalk.

Утилита EtherBoy обеспечивает мониторинг в реальном режиме времени сетевых потоков по различным протоколам. Она дает возможность просмотра всего трафика ЛВС (локальной вычислительной сети), а также обнаруживает все сетевые устройства в сети, включая нарушающие систему безопасности. Приложение формирует отчеты в различных вариантах, в том числе HTML, отображает в реальном времени статистику, трафик и проблемы с сетевыми соединениями на конкретных хостах.

Утилита WebBoy – почти точная копия утилиты EtherBoy с той лишь разницей, что она предназначена для мониторинга трафика в среде Web. В генерируемых утилитой WebBoy отчетах в основном содержатся сведения, касающиеся активности адресов URL в локальных или региональных сетях, но наряду с этим она может представить отчеты общего характера о трафике в Интернете, например, списки последних соединений, распределенные по главным машинам и протоколам.

Одной из свободно распространяемых программ является **CommView**, TamoSoft Inc. (http://www.tamos.com), которая предназначена для мониторинга сетевой активности путем сбора и анализа пакетов любой Ethernet-сети. С помощью этой программы можно ви-

деть список сетевых соединений, IP-статистику и исследовать отдельные пакеты (например, IP-пакеты декодируются вплоть до самого низкого уровня с полным анализом распространенных протоколов). Перехваченные пакеты могут быть сохранены в файл для последующего анализа, а также экспортированы в другие форматы.

6.5. Контрольные вопросы

1. Запишите функциональное преобразование, позволяющее получить СВ с показательным (экспоненциальным) ЗР из СВ с равномерным ЗР.
2. Запишите формулу, позволяющую сгенерировать СВ с показательным ЗР из системы двух независимых нормальных случайных чисел x_1 и x_2 с параметрами $(0,\sigma^2)$. Какова при этом зависимость между параметром λ показательного распределения и параметром σ исходного нормального распределения?
3. Каким образом связаны между собой СВ с логнормальным распределением и СВ с нормальным распределением?
4. Запишите функциональное преобразование, позволяющее получить СВ с распределение Парето из СВ с равномерным ЗР.
5. Запишите функциональное преобразование, позволяющее получить СВ с распределение Вейбулла из СВ с равномерным ЗР.
6. Поясните сущность RMD-метода моделирования фрактального броуновского движения.
7. Назовите вершины двух типов, составляющие сеть Петри при ее представлении в виде графа.
8. Какую роль играет маркер (метка позиции) в сети Петри?
9. Что происходит при срабатывании перехода в сети Петри?
10. В каких единицах измеряется работа ресурса за заданное время?
11. Поясните термин «интенсивность трафика».
12. Приведите пример системы с нагрузкой в 1 Эрл/час.
13. Какой интервал времени называют часом наибольшей нагрузки (ЧНН)?
14. В чем принципиальное отличие между интенсивностью трафика и пропускной способностью линии?
15. Каковы основные функции сетевого анализатора?

6.6. Задачи

1. СМО содержит три канала. Первый канал занят в течение 20% от общего времени работы СМО, второй канал занят 30% времени, третий канал занят 50% времени. Чему равна нагрузка в системе в Эрлангах?
2. СМО содержит два канала. Первый канал занят в течение 40% от общего времени работы СМО, второй канал занят 80% времени. Чему равна нагрузка в системе в Эрлангах?
3. Производительность источника равна 200 пакетов/с со средней длиной пакета 500 бит. Найти для сети Ethernet с пропускной способностью в 10 Мбит/с среднюю за время T интенсивность трафика на данном ресурсе.

7. МЕТОДЫ АНАЛИЗА СИСТЕМ И СЕТЕЙ ТЕЛЕКОММУНИКАЦИЙ

7.1. Анализ сетей массового обслуживания. Марковские сети без потерь

Однофазными называются СМО, в которых каждая заявка проходит только одну операцию обслуживания. В **многофазных** системах, которые обычно называют **сетями массового обслуживания** (Queuing Networks), заявка получает обслуживание более чем в одном сервере. В общем случае каждый узел такой сети может содержать СМО определенного типа, а заявки способны поступать в сеть в различных точках и, получив обслуживание в одном узле, переходить на обслуживание в другой для дальнейшей обработки (рис. 7.1). При исследовании сети необходимо задать ее топологическую структуру, т. к. она определяет возможные переходы заявок между узлами, а также описать маршруты отдельных заявок и вероятностные модели потоков заявок между узлами сети [3, 10].

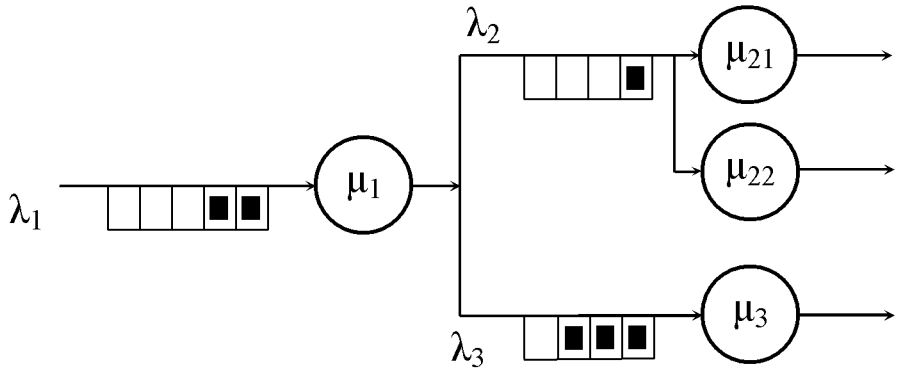


Рис. 7.1. Многофазная СМО

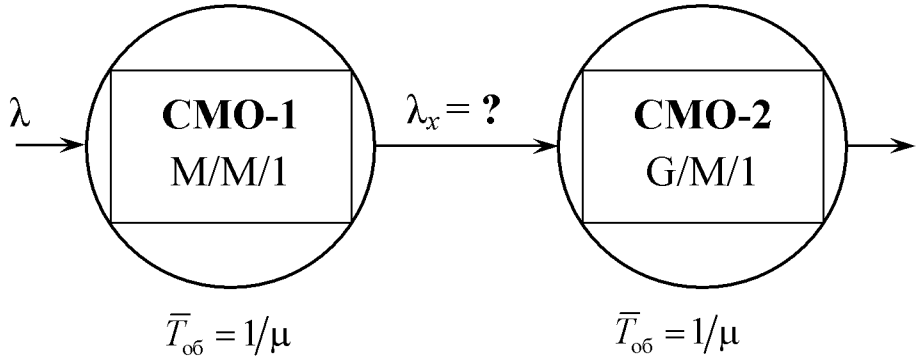


Рис. 7.2. Топологическая структура двухфазной СМО

Рассмотрим простейшую последовательную систему с двумя узлами (рис. 7.2) [10]. На вход первого узла (СМО-1 типа M/M/1 с показательным ЗР времени обслуживания $T_{об}$ и средним значением $\overline{T}_{об} = 1/\mu$) поступает пуассоновский поток плотности λ. Второй узел (СМО-2) состоит из единственного сервера с такими же характеристиками. Необходимо получить ЗР промежутков времени между соседними заявками на выходе СМО-1.

Пусть $d(t)$ – ПРВ промежутков времени $T_{об}$ между соседними заявками на выходе СМО-1. Выразим эту функцию для двух случаев:

1) когда при уходе (окончании обслуживания) заявки СМО-1 не пуста (наличие в буфере заявки) (рис. 7.3, а);

2) когда при уходе заявки из СМО-1 в буфере не было другой заявки, т. е. СМО-1 пуста (рис. 7.3, б).

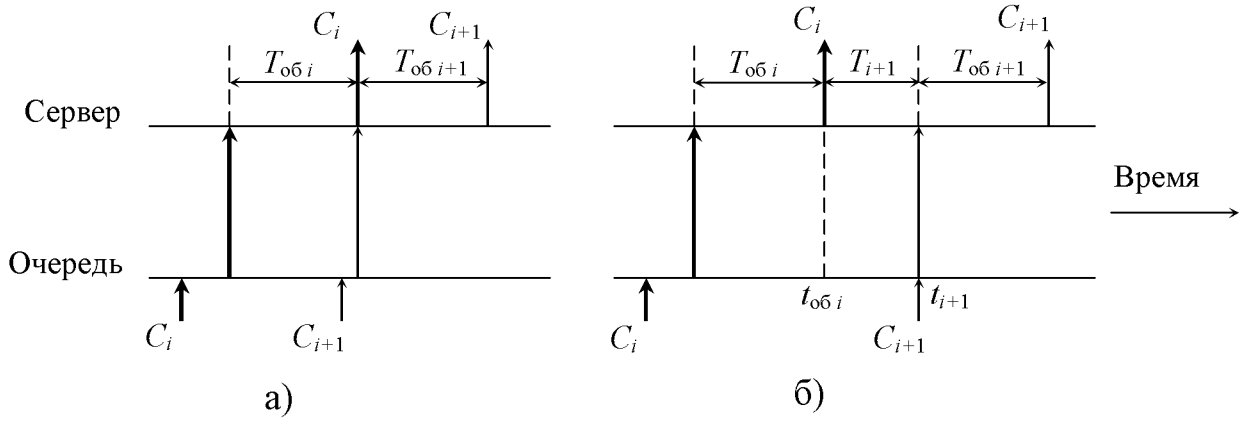


Рис. 7.3. Временные диаграммы процессов обслуживания заявок

В первом случае очередная $(i+1)$-я заявка C_{i+1} появится на выходе СМО-1 сразу через интервал времени обслуживания $T_{об\,i+1}$ (рис. 7.3). Вероятность этого события равна $1-p_0$ (вероятность, что СМО не пуста). Следовательно, ПРВ интервала времени между выходами заявки из СМО-1 будет совпадать с ПРВ длительности обслуживания $T_{об\,i}$ в СМО-1 $b(x)$.

Во втором случае очередная $(i+1)$-я заявка C_{i+1} покидает сервер через интервал, равный сумме интервала $T_{i+1} = t_{i+1} - t_{об\,i}$ (между моментами поступления на вход СМО-1 $(i+1)$-й заявки и окончания обслуживания i-й заявки; при $T_{i+1} > 0$ – буфер пустой) и времени обслужи-

вания $T_{\text{об}\,i+1}$. Вероятность этого события (простаивание системы) равна $p_0 = 1-\rho$. Поскольку эти два интервала T_{i+1} и $T_{\text{об}\,i+1}$ распределены независимо, то ПРВ их суммы равна свертке ПРВ суммируемых СВ. Для ее вычисления удобно воспользоваться преобразованием Лапласа (ПЛ) [10].

Обозначим ПЛ для ПРВ суммы двух промежутков времени: T_{i+1} и $T_{\text{об}\,i+1}$ как

$$D_0(s) = L\left\{w_D\left(T_{i+1} + T_{\text{об}\,i+1}\right)\right\} = L\left\{w_D\left(\Delta T_{i+1}\right)\right\},$$

где $w_D(\Delta T) = \int_{-\infty}^{\infty} w_A(T)\, w_B(\Delta T - T)\, dT$, $\Delta T_{i+1} = T_{i+1} + T_{\text{об}\,i+1}$.

ПЛ для ПРВ суммы равно произведению преобразований исходных ПРВ. Таким образом, ПЛ для ПРВ промежутка времени между выходными заявками для случая СМО-1 с пустым буфером будет

$$D_0(s) = A(s)B(s),$$

где $A(s)$ – ПЛ для ПРВ интервалов между событиями входного потока; $B(s)$ – ПРВ интервалов времени обслуживания в сервере.

Поскольку в системе M/M/1 ПРВ интервалов являются экспоненциальными как для входных событий, так и для времени обслуживания, имеем:

$$A(s) = \int_0^{\infty} \lambda e^{-\lambda t} e^{-st}\, dt = \frac{\lambda}{s+\lambda}, \qquad B(s) = \int_0^{\infty} \mu e^{-\mu t} e^{-st}\, dt = \frac{\mu}{s+\mu}.$$

Общее распределение интервала между выходными заявками можно записать как

$$D(s) = (1-p_0)B(s) + p_0 D_0(s) = \rho\frac{\mu}{s+\mu} + (1-\rho)\frac{\lambda}{s+\lambda}\frac{\mu}{s+\mu} = \frac{\lambda}{s+\lambda},$$

где $\rho = \lambda/\mu$ – нагрузка в узле.

Следовательно, ПРВ промежутков времени между заявками на выходе СМО-1 экспоненциальная с тем же самым параметром, т. е. СМО типа M/M/1 превращает пуассоновский поток на входе в пуассоновский поток на выходе с тем же самым параметром (**теорема Бёрке**) [3]. Данный факт имеет место для всех СМО типа M/M/m. На основании этой теоремы можно исследовать многофазные последовательные схемы.

Основная трудность анализа сетей из линий связи заключается в том, что интервалы между моментами поступления пакетов после прохождения пакетов через первую очередь начинают коррелировать с их длинами. Оказывается, что если каким-либо образом устранить эту корреляцию и использовать рандомизацию для разделения потока по различным маршрутам, то можно найти среднее число пакетов в системе, рассматривая каждую очередь в сети как систему M/M/1. Данный результат выражается теоремой Джексона [3, 10].

Рассмотрим сеть (рис. 7.4), содержащую N узлов, причем каждый i-й узел состоит из m_i серверов с показательным распределением $T_{\text{об}}$ с параметром μ_i. В каждый узел извне поступает пуассоновский поток заявок с интенсивностью γ_i. Покидая i-й узел, заявка с вероятностью r_{ij} поступает в j-й узел или выходит из сети с вероятностью $1-\sum_{j=1}^{N} r_{ij}$.

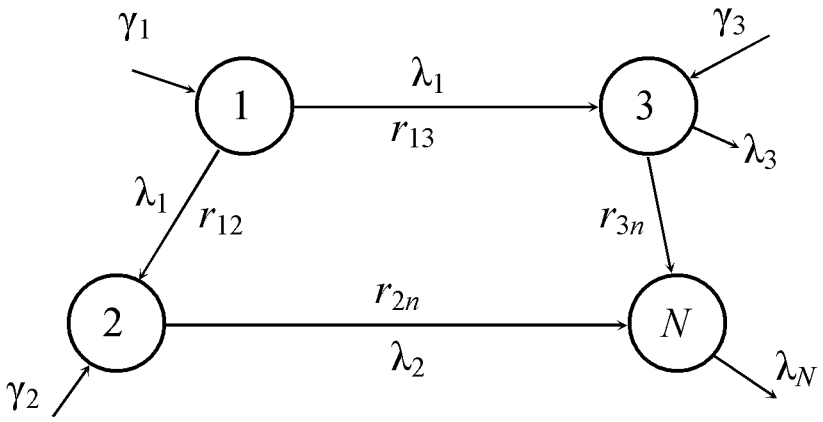


Рис. 7.4. Модель многоузловой сети

Обозначив λ_j суммарную интенсивность потока, поступающего в j-й узел, можно показать, что должно выполняться **условие баланса**

$$\lambda_j = \gamma_j + \sum_{i=1}^{N} \lambda_i r_{ij}, \quad i = 1,2,\ldots,N,$$

где γ_j – поток из внешней сети; λ_i – потоки из смежных узлов (типичный пример – узел 2).

Вероятность того, что заявка после обслуживания в i-м узле вообще покинет сеть, равна $1-\sum_{j=1}^{N} r_{ij}$. Выполнение **условия эргодичности** марковской модели каждого узла будет обеспечено, если $\lambda_i > m_i \mu_i$.

Джексоном было доказано, что каждый узел в сети ведет себя как независимая СМО типа M/M/m с входящим пуассоновским потоком λ_i. В общем случае полный входящий поток не является пуассоновским. Состояние сети с N узлами описывается вектором с компонентами в виде количества заявок в каждом из узлов сети $(k_1,k_2,\ldots,k_N)$ [10].

Теорема Джексона: стационарная вероятность состояния сети с N узлами раскладывается в произведение безусловных распределений

$$p(k_1,k_2,\ldots,k_N)=p_1(k_1)\,p_2(k_2)\ldots p_N(k_N),$$

которые представляют собой стационарные вероятности для системы M/M/m.

Исследование сети с помощью изложенной здесь методики приводит к громоздким системам уравнений для определения интенсивностей входных потоков. Для получения многих результатов вместо условий глобального равновесия для сети в целом можно применять условия локального равновесия для отдельных подсистем, что позволит значительно упростить задачу.

Определим систему уравнений локального равновесия как систему уравнений, в которых интенсивность потока из данного состояния сети за счет ухода заявок из узла i приравнивается к интенсивности потока в данное состояние сети за счет поступления заявок в узел i (рис. 7.5) [10].

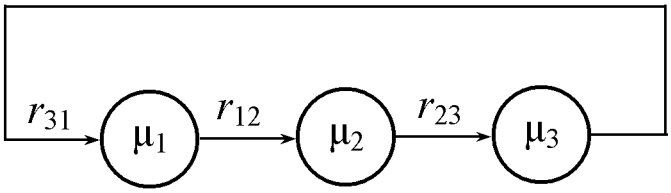


Рис. 7.5. Замкнутая сеть с тремя узлами

Пусть в замкнутой сети с тремя (N=3) узлами циркулируют две заявки (K=2). При этом для остальных индексов: $r_{12}=r_{23}=r_{31}=1$, $r_{ij}=0$. Состояние сети описывается тройками: (k_1,k_2,k_3), причем $k_1+k_2+k_3=2$. Всего в сети возможно $C_{N+K-1}^{N-1}=C_4^2=6$ различных состояний (рис. 7.6).

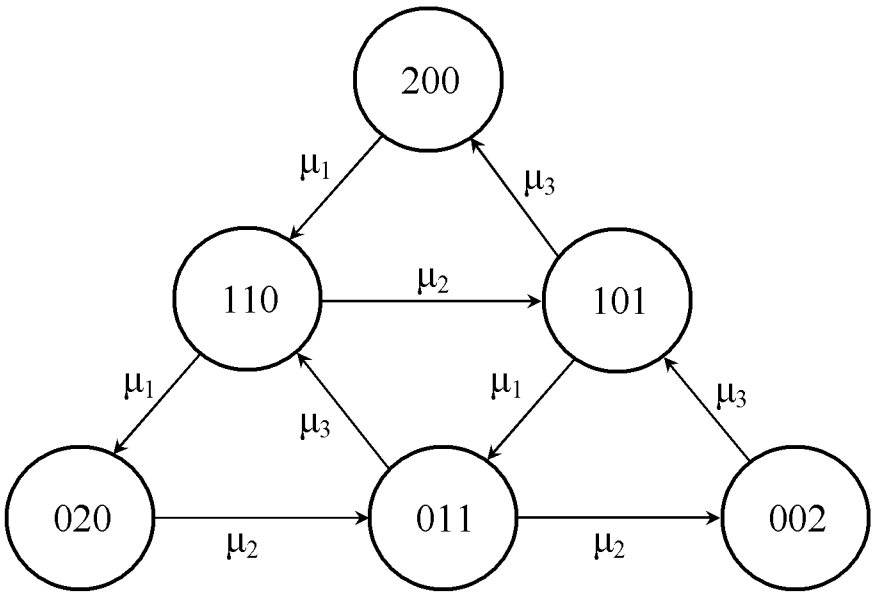


Рис. 7.6. Диаграмма интенсивностей переходов между состояниями

Вероятности переходов (между состояниями): $\mu_1 = r_{12}$, $\mu_2 = r_{23}$, $\mu_3 = r_{31}$.
Система уравнений глобального равновесия состоит из шести уравнений, одно из которых избыточно за счет условия нормировки (сумма вероятностей состояний p равна 1):

$$\mu_3\, p(2,0,0) = \mu_1\, p(1,1,0) \quad \text{(состояние 200)}$$

$$\mu_1\, p(0,2,0) = \mu_2\, p(0,1,1) \quad \text{(состояние 020)}$$

$$\mu_1\, p(1,1,0) + \mu_3\, p(1,1,0) = \mu_2\, p(1,0,1) + \mu_1\, p(0,2,0) \quad \text{(состояние 110)}$$

$$\mu_2\, p(0,1,1) + \mu_1\, p(0,1,1) = \mu_2\, p(0,0,2) + \mu_3\, p(1,1,0) \quad \text{(состояние 011)}$$

$$\mu_3\, p(1,0,1) + \mu_2\, p(1,0,1) = \mu_1\, p(0,1,1) + \mu_3\, p(2,0,0) \quad \text{(состояние 101)}$$

В каждом уравнении правая часть соответствует потоку, исходящему из данного состояния, а левая – потоку, входящему в это состояние. Из системы пяти уравнений можно получить уравнения локального равновесия:

$$\mu_1\, p(1,1,0) = \mu_2\, p(1,0,1) \quad \text{(локальный баланс для узла 1)}$$

$$\mu_3\, p(1,1,0) = \mu_1\, p(0,2,0) \quad \text{(локальный баланс для узла 2)}$$

Всего образуется 9 уравнений локального равновесия, из которых 4 избыточны. Решение дается следующими формулами:

$$p(1,0,1) = \frac{\mu_3}{\mu_2}\, p(2,0,0), \qquad p(1,1,0) = \frac{\mu_3}{\mu_1}\, p(2,0,0), \qquad p(0,1,1) = \frac{(\mu_2)^2}{\mu_1 \mu_2}\, p(2,0,0),$$

$$p(0,0,2) = \left(\frac{\mu_3}{\mu_2}\right)^2 p(2,0,0), \qquad p(0,2,0) = \left(\frac{\mu_3}{\mu_1}\right)^2 p(2,0,0),$$

$$p(2,0,0)=\left[1+\frac{\mu_3}{\mu_2}+\frac{\mu_3}{\mu_1}+\frac{(\mu_2)^2}{\mu_1\mu_2}+\left(\frac{\mu_3}{\mu_2}\right)^2+\left(\frac{\mu_3}{\mu_1}\right)^2\right]^{-1}.$$

Решение уравнений локального равновесия получить значительно проще. В любом случае поиск стационарных вероятностей сводится к решению больших систем линейных уравнений. Расчеты будут верны, если только сеть массового обслуживания удовлетворяет свойствам эргодической марковской цепи.

7.2. Сети с блокировками (потерями)

Выше были рассмотрены сети массового обслуживания в предположении, что каждый узел содержит бесконечный накопитель, и все заявки будут обслужены через некоторое время. Рассмотрим сети с узлами, в которых может быть СМО с блокировкой заявок – обычно это коммутационные схемы с конечными пучками соединительных линий. Другой моделью являются сети с множественным доступом к фиксированному числу каналов.

Рассмотрим в качестве примера (рис. 7.7) подключение абонента С через абонентскую линию с блокиратором в пункте В к АТС, которая, в свою очередь, имеет два канала связи с АТС в пункте А. Требуется определить вероятность блокировки звонка абоненту С из пункта А. Поставим в соответствие рассматриваемой сети так называемый **вероятностный граф** (граф Ли), с вершинами А, В и С и ребрами (звеньями) a, b, c, соответствующими потокам заявок. Каждому звену a, b, c присвоим соответствующую вероятность занятия w_a, w_b, w_c.

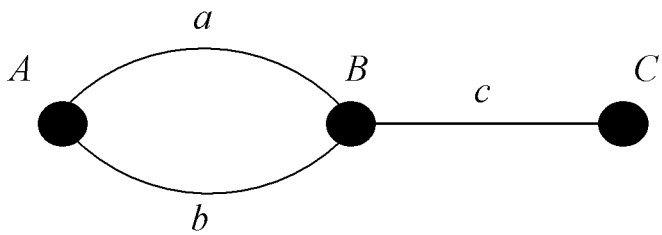


Рис. 7.7. Вероятностный граф

Метод Ли состоит в том, что вероятность блокировки пути между любыми вершинами графа может быть рассчитана как вероятность совместного занятия всех соединяющих эти вершины звеньев в пред-

положении, что вероятности занятия каждого из звеньев независимы. Вероятность совместного занятия может быть рассчитана с помощью известных теорем теории вероятностей для сложных событий.

Вероятности того, что звено свободно:

$$q_a = 1 - w_a, \quad q_b = 1 - w_b, \quad q_c = 1 - w_c.$$

Вероятность блокировки пути AB будет определяться как совместная вероятность занятости a и b: $w_a w_b$. Вероятность свободности этого пути: $1 - w_a w_b$.

Общая вероятность свободности пути AC:

$$(1 - w_a w_b) q_c = (1 - w_a w_b)(1 - w_c).$$

Вероятность блокировки пути AC: $p_B = 1 - (1 - w_a w_b)(1 - w_c)$.

Рассмотренный выше граф является параллельно-последовательным. Расчет вероятностей на таких графах производится по следующим правилам:

1) параллельное включение звеньев: полная вероятность занятости

$$w = w_1 w_2 \ldots w_i \ldots w_n;$$

2) последовательное включение звеньев: полная вероятность свободности

$$q = q_1 q_2 \ldots q_i \ldots q_n.$$

В ряде случаев граф сети не сводится к параллельно-последовательным схемам, например, мостиковый граф (рис. 7.8).

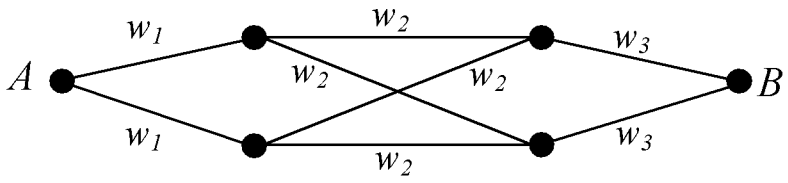
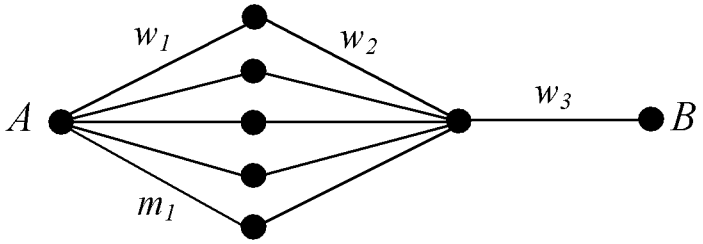


Рис. 7.8. Мостиковый граф

Для такого графа можно получить вероятность блокировки пути AB в виде

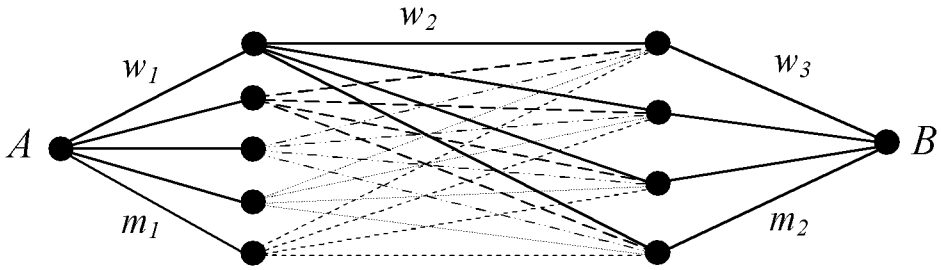
$$p_{B(AB)} = (1 - w_2)^2 \left[w_1 + w_2^2 (1 - w_1) \right]^2 + 2 w_3 (1 - w_3) \left[w_1 + w_2 (1 - w_1) \right]^2 + w_3^2.$$

Графы, типа приведенных выше, часто встречаются при анализе многозвенных коммутационных схем. Для этих графов можно получить явные выражения для вероятности блокировки пути AB (рис. 7.9, 7.10).



$$p_{B(AB)} = 1 - \left\{1 - \left[1 - (1 - w_1)(1 - w_2)\right]^{m_1}\right\}(1 - w_3)$$

Рис. 7.9. Пример параллельно-последовательного графа



$$p_{B(AB)} = \sum_{k=0}^{m_2} C_{m_2}^{k} w_3^{m_2-k} (1 - w_3)^{k} \left[w_1 + w_2^{k}(1 - w_1)\right]^{m_1}$$

Рис. 7.10. Пример не параллельно-последовательного графа

В том случае, если граф получается слишком сложным, можно пользоваться **методом оценочных графов**, заключающимся в следующем: строятся граф оценки сверху путем разделения вершин и отбрасывания ребер для упрощения расчета и граф оценки снизу путем объединения части вершин; рассчитываются вероятности блокировки для оценочных графов, которые станут служить соответственно верхней и нижней границами, между которыми и будет находиться значение вероятности блокировки для исходного графа.

7.3. Анализ и оптимизация коммутационных схем

7.3.1. Понятие коммутационной схемы

Простейшая коммутационная схема (КС) – это однофазная (однозвенная) схема (рис. 7.11), имеющая n входов и m выходов и называемая **коммутатором** [12, 23]. При $n > m$ коммутатор также выполняет функцию концентратора, а при $n < m$ – функцию расширителя.

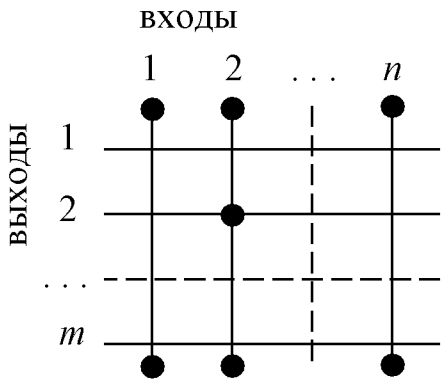


Рис. 7.11. Простейший коммутатор

Выходы иногда объединяются в группы, которые называются **направлениями**. Коммутатор является **неблокирующим**, если выполнено соотношение $n \le m$, и блокирующим в противном случае. Блокируемость коммутатора определяется невозможностью части входов получить доступ ни к одному из выходов. Блокируемость может быть общей, когда все выходы рассматриваются равнозначными, и в определенном направлении, когда недоступными оказываются все выходы данного направления.

Число точек поля коммутации КС – число это управляемых точек соединения, равное $C = n\,m$ («каждый с каждым»). Например, для АТС на 10 000 абонентов реализация в виде простого коммутатора привела бы к необходимости построения цепи с 10^8 электронных ключей.

Существуют две основные технологии коммутации – пространственная, основанная на реальных матрицах электронных ключей, и временная, использующая временное мультиплексирование входных потоков и последующее перекрестное демультиплексирование. Число точек коммутации ограничивается параметрами элементной базы и стоимостными факторами и требует снижения.

В связи с этим для построения систем коммутации со многими входами и выходами применяют многозвенные схемы, которые позволяют обеспечить управляемое соединение входов и выходов, используя меньшее, чем в простом коммутаторе, число точек коммутации. Многозвенные схемы, кроме коммутаторов, содержат фиксированные соединения между ними, называемые **промежуточными линиями** (ПЛ).

Рассмотрим двухзвенную схему, приведенную на рис. 7.12, у которой любой выход схемы доступен любому входу (полнодоступный

пучок выходов) [12]. Схема изображена в общем виде и имеет k коммутаторов в первом звене на n входов и m выходов каждый и m коммутаторов во втором звене на k входов и l выходов каждый. Выходы схемы разбиты на **группы** (**направления**). На рис. 7.12 показаны два направления: H_i, к которому отнесены по два выхода в каждом коммутаторе второго звена и имеющее, таким образом, $2m$ выходов, и H_j, имеющее m выходов (по одному выходу в каждом коммутаторе второго звена). В общем случае число выходов в каждом коммутаторе, отводимых для одного направления, может быть равно q, и тогда суммарное число выходов в направлении составит mq.

В двухзвенной КС для установления соединения входа с выходом требуются две точки коммутации и одна ПЛ. КС, содержащие два и более звеньев в соединительном пути, называют звеньевыми. В общем случае звеньевая схема – это схема, имеющая входы, выходы, коммутаторы и ПЛ. Если все ПЛ и выход, составляющие соединительный путь, свободны, то и этот путь свободен. Соединительный путь считается занятым, если хотя бы одна из ПЛ или выход заняты.

Любая звеньевая КС имеет конечное, хотя и довольно большое, число состояний. Каждое состояние отличается комбинацией занятых входов, выходов и ПЛ. Поэтому для звеньевых схем, представляющих практический интерес, система уравнений для вероятностей состояний во многих случаях не может быть решена, а в отдельных случаях не может быть даже выписана.

Исследование звеньевых схем сложно не только из-за их большого числа состояний. Дополнительные усложнения возникают также и из-за того, что между процессами, происходящими в разных направлениях выходов звеньевой схемы, существует взаимная зависимость. Из рис. 7.12 видно, что для установления соединения к выходам направлений H_i и H_j используются одни и те же ПЛ. Поэтому занятие ПЛ для подключения к выходам одного направления изменяет вероятность занятия выходов другого направления.

В общем случае условные вероятности блокировки зависят не только от числа занятых выходов, но и от структуры схемы, поступающей нагрузки и алгоритма установления соединения, что усложняет задачи исследования звеньевой схемы. В связи с этим инженерный расчет звеньевых схем основывается на априорных предположениях относительно способа математического описания результатов воздействия поступающего потока вызовов на отдельные звенья соединения. Обычно предполагается, что процессы, протекающие в раз-

личных звеньях схемы, независимы и могут быть описаны каким-нибудь простым законом распределения; кроме того, используются и другие упрощающие предположения. Это облегчает решение задачи, однако вносит отклонение от истинных характеристик, имеющих место в процессе функционирования схемы.

В большинстве случаев нельзя заранее указать, в какой степени то или иное упрощающее предположение искажает истинную величину отыскиваемого показателя (например, вероятности потерь), поэтому для определения степени погрешности приближенных методов можно воспользоваться сравнением с результатами компьютерного моделирования [31].

Современные методы расчета числа соединительных устройств в значительной степени базируются на результатах статистического моделирования на ЭВМ. Полученные результаты, как правило, аппроксимируются простыми функциональными зависимостями. Так как практически невозможно получить числовые данные для любых значений нагрузки и параметров структуры, которые могут встретиться при расчетах, то подобные методы предполагают интерполяцию и экстраполяцию в области, где числовые данные не получены.

7.3.2. Комбинаторный метод Якобеуса

Рассмотрим метод анализа КС, который применим при малом числе звеньев и дает весьма точные результаты – **комбинаторный метод Якобеуса** на примере двухзвенной КС с полнодоступным включением ПЛ [2, 12]. Количество выходов из каждого коммутатора второго звена в этой схеме для направления H_j равно единице ($q=1$). Пусть, например, к рассматриваемому моменту вызов поступил на один из входов (например, второй вход) первого коммутатора. Установление соединения через схему заключается в использовании одной из свободных ПЛ и одного из свободных выходов рассматриваемого направления H_j, взаимно доступных друг другу. Для обслуживания поступившего вызова могут быть использованы m ПЛ и m выходов требуемого направления, которые выделены на рис. 7.12 жирными линиями. Соединение может быть установлено, если имеется пара свободных и взаимно доступных звеньев. Блокировка (потеря вызова) наступит в трех случаях:

1) заняты все ПЛ, которые могут быть использованы для обслуживания;

2) заняты все выходы в требуемом направлении;

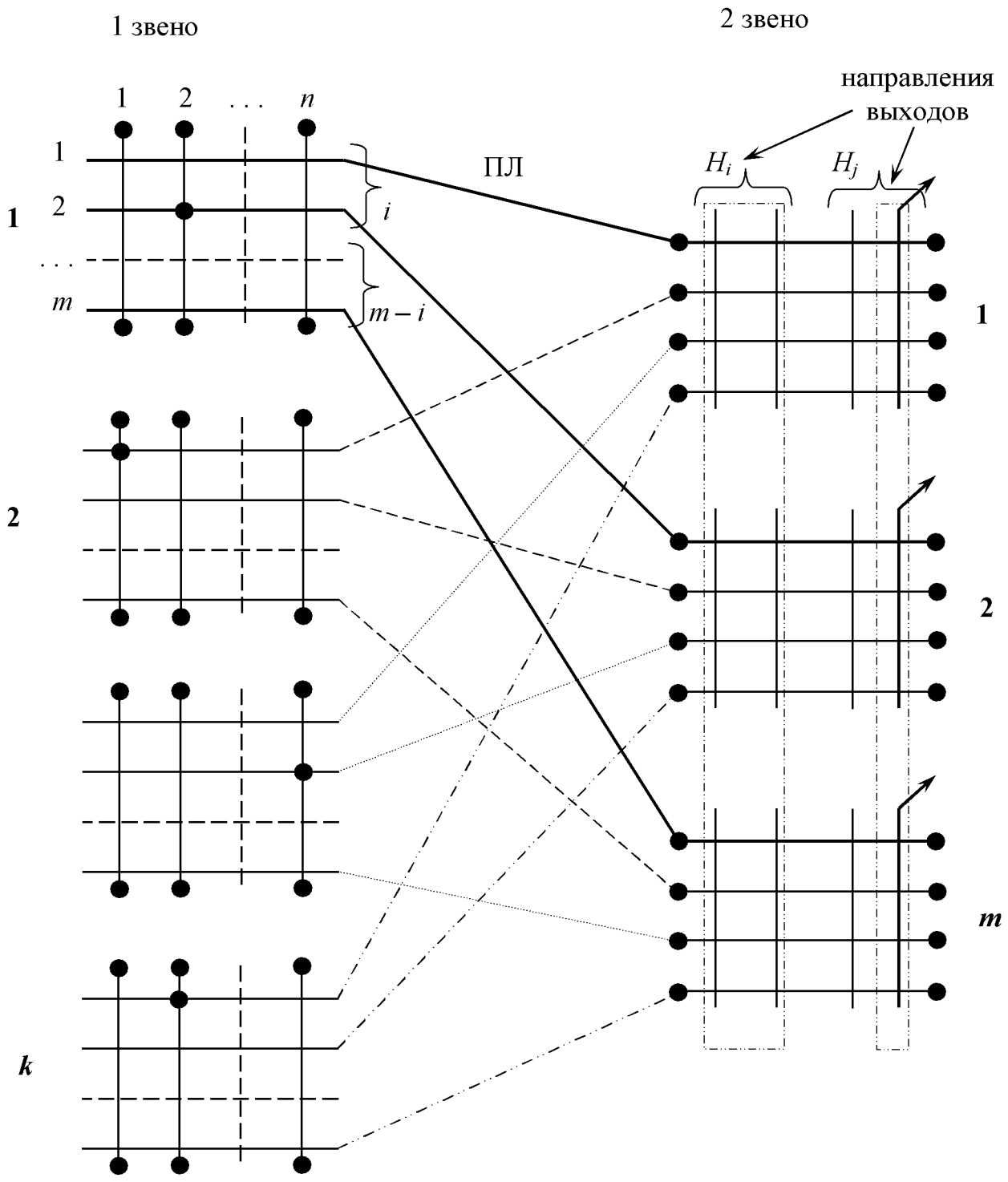


Рис. 7.12. Двухзвенная коммутационная схема

3) комбинация свободных ПЛ и свободных выходов требуемого направления не является взаимно доступной.

Если считать, что рассматриваемый вызов поступил на отмеченный вход первого коммутатора в момент, когда i промежуточных ли-

ний из m, подключенных к выходам данного коммутатора, заняты, то для подключения входа к одному из выходов требуемого направления могут быть использованы только оставшиеся $m-i$ ПЛ. Если же выходы требуемого направления, соответствующие этим $m-i$ линиям, заняты, то возникнут потери. Это утверждение справедливо для любого i, лежащего в пределах $0 \le i \le m$, и охватывает два случая занятости:

1) заняты все ПЛ, исходящие из первого звена ($i = m$);

2) все ПЛ, исходящие из первого звена, свободны, но заняты все выходы в данном направлении ($i = 0$) второго звена.

Если вероятность занятия любых i из m ПЛ, принадлежащих одному коммутатору первого звена, обозначить W_i, а вероятность занятия определенных $m-i$ выходов (соответствующих свободным ПЛ) обозначить H_{m-i}, то вероятность блокировки КС может быть вычислена по формуле [12, 23]:

$$p_B = \sum_{i=0}^{m} W_i H_{m-i} . \qquad (7.1)$$

Данная формула справедлива при выполнении следующих двух допущений:

1. Независимость событий, описываемых вероятностями W_i и H_{m-i}. Допущение является условным, так как ПЛ и выходы занимаются парами.

2. Случайное (равновероятное) занятие ПЛ и выходов. При этом все вероятности занятия i ПЛ считаются в (7.1) равными между собой вне зависимости от того, какие i из m линий заняты. При наличии определенного порядка занятия ПЛ это допущение несправедливо.

Метод Якобеуса предполагает, что события, определяемые вероятностями W_i и H_{m-i}, независимы и могут быть заданы распределениями Эрланга или Бернулли.

При распределении Эрланга вероятность занятия i любых линий в пучке из m серверов при интенсивности поступающей нагрузки ρ [Эрл/пучок] вычисляется по формуле

$$E_{i,m}(\rho) = B(m,\rho) = \frac{\rho^i / i!}{\sum_{j=0}^{m} \frac{\rho^j}{j!}} = W_i , \qquad (7.2)$$

а вероятность занятия $m-i$ фиксированных линий из m линий в пучке [12]

$$H_{m-i,m}(\rho) = H_{m-i} = E_m(\rho) / E_i(\rho) , \qquad (7.3)$$

где $E_m(\rho)$ – потери в полнодоступном пучке из m линий при интенсивности поступающей нагрузки ρ [Эрл/пучок], вычисленные по формуле Эрланга

$$E_m(\rho) = \frac{\rho^m / m!}{\sum_{j=0}^{m} \frac{\rho^j}{j!}},$$

а $E_i(\rho)$ – потери при той же интенсивности нагрузки в пучке из i линий

$$E_i(\rho) = \frac{\rho^i / i!}{\sum_{j=0}^{i} \frac{\rho^j}{j!}}.$$

Если использовать распределение Энгсета (4.3) для вероятностей занятия линий, то при равенстве числа линий и количества источников трафика $(N = m)$ для вероятностей занятия фиксированного числа линий получается распределение Бернулли (4.5):

$$p_k = \frac{C_m^k \alpha^k}{\sum_{j=0}^{m} C_m^j \alpha^j} = \frac{C_m^k \alpha^k}{(1+\alpha)^m} = C_m^k a^k (1-a)^{m-k},$$

где $a = \alpha/(1+\alpha)$ – средняя нагрузка, обслуженная одной линией в пучке (или вероятность занятости линии в системе без потерь), а величина $(1-a)$ – вероятности простоя этой линии.

При использовании распределения Бернулли вероятности занятия любых i линий из m устройств:

$$W_i = C_m^i \eta^i (1-\eta)^{m-i}, \qquad (7.4)$$

а вероятность занятия $m-i$ фиксированных линий при тех же условиях

$$H_{m-i} = \eta^{m-i}, \qquad (7.5)$$

где η – средняя нагрузка на одну линию в пучке.

Распределение Эрланга предполагает неограниченное число источников нагрузки, а (7.2) и (7.3) основываются на интенсивности поступающей нагрузки. Распределение Бернулли предполагает ограниченное число источников нагрузки, не превышающее число линий $(N \le m)$, а в (7.4) и (7.5) входит обслуженная нагрузка.

Естественно, что вероятность потерь при использовании различных распределений получится различной. Метод рекомендует принимать распределение Эрланга при определении вероятности занятия

тех линий, для которых число источников нагрузки больше числа линий ($N > m$). Использование распределения Бернулли считается целесообразным при числе источников нагрузки, примерно равном числу линий ($N \approx m$), для которых определяются вероятности занятия.

Расчетные формулы для определения вероятности потерь в двухзвенной схеме можно получить, если в общее выражение для потерь (7.1) подставить выражение для W_i и H_{m-i}.

7.3.3. Потери в двухзвенных схемах без сжатия и расширения

В схеме без сжатия и расширения коммутаторы первого звена имеют равное число входов и выходов n и для ПЛ можно принять распределение Бернулли, так как число источников нагрузки, которыми являются входы, равно числу ПЛ [12]. Для выходов двухзвенной схемы можно также принять распределение Бернулли, считая, что количество коммутаторов первого звена небольшое. При этом, относя W_i к ПЛ, а H_{m-i} к выходам коммутатора первого звена, будем иметь для W_i и H_{m-i} следующие выражения:

$W_i = C_m^i b^i (1-b)^{m-i}$, где C_m^i – число сочетаний из m по i; b – средняя интенсивность нагрузки, обслуженной одной ПЛ [Эрл];

$H_{m-i} = c^{m-i}$, где c – средняя интенсивность нагрузки, обслуженной одним выходом рассматриваемого направления [Эрл].

Подставляя значения W_i и H_{m-i} в (7.1), получаем вероятность блокировки (с учетом формулы бинома Ньютона)

$$p_B = \sum_{i=0}^{m} C_m^i b^i (1-b)^{m-i} c^{m-i} = (b + c - bc)^m . \qquad (7.6)$$

Если число коммутаторов k первого звена достаточно велико, то целесообразно для выходов данного направления принять распределение Эрланга. При этом, относя W_i к направлению, а H_{m-i} к ПЛ, получаем

$$W_i = \frac{\rho^i / i!}{\sum_{j=0}^{m} \frac{\rho^j}{j!}} \quad \text{и} \quad H_{m-i} = b^{m-i},$$

где ρ – интенсивность поступающей нагрузки на направление [Эрл]. Подставляя эти выражения в (7.1) и вынося несуммирующиеся множители за знак суммы, находим

$$p_B = \sum_{i=0}^{m} \frac{\rho^i / i!}{\sum_{j=0}^{m} \rho^j / j!} b^{m-i} = \frac{b^m}{\sum_{j=0}^{m} \rho^j / j!} \sum_{i=0}^{m} \frac{(\rho/b)^i}{i!},$$

откуда с учетом формулы $E_m(\rho/b) = \dfrac{(\rho/b)^m / m!}{\sum_{i=0}^{m} (\rho/b)^i / i!}$ имеем

$$p_B = \frac{E_m(\rho)}{E_m(\rho/b)}. \qquad (7.7)$$

Если для образования каждого направления в каждом коммутаторе второго звена отводится не один, а q выходов, то для случая, когда и занятие выходов коммутатора и занятие ПЛ можно описать распределением Бернулли, будем иметь

$$W_i = C_m^i b^i (1-b)^{m-i}; \quad H_{(m-i)q} = c^{(m-i)q}.$$

Подставляя эти выражения в (7.1) и учитывая формулу бинома Ньютона, получаем

$$p_B = \left(b + c^q - bc^q\right)^m. \qquad (7.8)$$

Если занятие выходов подчиняется распределению Эрланга, а занятие ПЛ распределению Бернулли, то в этом случае выражение для потерь при некоторых дополнительных ограничениях может быть преобразовано к виду

$$p_B = \frac{E_{mq}(\rho)}{E_{mq}(\rho/b)}. \qquad (7.9)$$

В соответствии с рассматриваемым методом формула (7.9) может применяться и для дробных значений q. Выражения (7.8) и (7.9) имеют более общий вид и включают в себя соответственно (7.6) и (7.7), которые можно получить из первых двух, полагая $q = 1$.

7.3.4. Потери в двухзвенных схемах при наличии сжатия и расширения

В схемах со **сжатием** (концентрацией) число входов n в коммутатор первого звена больше числа выходов m из этого коммутатора. В таких схемах потери возникают из-за наличия неудачных сочетаний занятых ПЛ и выходов, а также при поступлении на входы коммутатора первого звена более m вызовов.

Если при $q \geq 1$ и распределении Бернулли для ПЛ и выходов W_i отнести к ПЛ, а $H_{(m-i)q}$ – к выходам рассматриваемого направления, то можно записать

$$W_i = C_n^i a^i (1-a)^{n-i}, \quad H_{(m-i)q} = c^{(m-i)q},$$

где a – средняя интенсивность нагрузки, обслуженной одним входом коммутатора первого звена. Вероятность блокировки (потери) в данном случае будет

$$p_B = \sum_{i=0}^{m} W_i H_{(m-i)q} + \sum_{i=m+1}^{n} W_i = \sum_{i=0}^{m} C_n^i a^i (1-a)^{n-i} c^{(m-i)q} + \sum_{i=m+1}^{n} C_n^i a^i (1-a)^{n-i}.$$

В этом выражении первое слагаемое учитывает потери из-за неудачных сочетаний при занятиях ПЛ и выходов, а второе – потери за счет поступления более m вызовов в один коммутатор первого звена.

Если искание свободных выходов в схемах с $q > 1$ производить в два этапа, т. е. таким образом, чтобы в первую очередь занимались все выходы в $q-1$ столбцах (группах) выходов и только после этого занимались бы выходы последнего столбца (группы) q, то можно приближенно выразить потери для схем с концентрацией при $q \geq 1$:

$$p_B = b^m + (b + c - bc)^{mq}, \qquad (7.10)$$

где $b = (n/m)a$.

Для случая неупорядоченного занятия выходов в направлении достаточно точные результаты дает выражение (7.8).

Если для первого звена сохранить распределение Бернулли, а для второго звена принять распределение Эрланга, то для двухэтапного искания можно получить следующее приближенное выражение для потерь:

$$p_B = b^m + \frac{E_{mq}(\rho)}{E_{mq}(\rho/b)}. \qquad (7.11)$$

В схемах с **расширением** число входов n в каждый коммутатор первого звена меньше числа выходов m из этого коммутатора. В такой схеме число одновременных вызовов не превышает n, а следовательно, меньше m, поэтому потери возникают только за счет неудачных сочетаний занятых ПЛ и выходов. Если и для ПЛ и для выходов справедливо распределение Бернулли, то при $q \geq 1$ и W_i отнесенном к ПЛ, можно записать $W_i = C_n^i a^i (1-a)^{n-i}$, $H_{(m-i)q} = c^{(m-i)q}$. Подставим значения этих вероятностей в (6.1). Тогда

$$p_B = \sum_{i=0}^{n} C_n^i a^i (1-a)^{n-i} c^{(m-i)q} = \sum_{i=0}^{n} C_n^i a^i (1-a)^{n-i} \left(c^q\right)^{n-i} \left(c^q\right)^{m-n}.$$

Учитывая формулу бинома Ньютона, получаем окончательное выражение для потерь:

$$p_B = c^{q(m-n)} \left(a + c^q - ac^q\right)^n. \qquad (7.12)$$

Если, сохранив распределение Бернулли для ПЛ, принять распределение Эрланга для выходов, то в данном случае вероятность потерь

$$p_B = \frac{E_{mq}(\rho)}{E_{nq}(\rho/b)}. \qquad (7.13)$$

Рассмотренные выше схемы относятся к случаю односвязного двухзвенного включения, при котором один коммутатор первого звена соединен с коммутатором второго звена одной ПЛ. При наличии f ПЛ между парой коммутаторов первого и второго звеньев в соответствии с комбинаторным методом будут справедливыми все полученные выше формулы, если a заменить на a^f, а b заменить на b^f.

7.3.5. Многозвенные схемы

Рассмотрим трехзвенную КС (рис. 7.13), которая содержит N/n входных и N/n выходных коммутаторов, образующих соответственно первую и третью ступени коммутации. Вторая ступень коммутации состоит из k квадратных коммутаторов с N/n входами и выходами. С помощью ПЛ каждый выход коммутатора первой ступени соединяется с разными коммутаторами второй ступени, так что подключаемый вход соответствует месту коммутатора первой ступени.

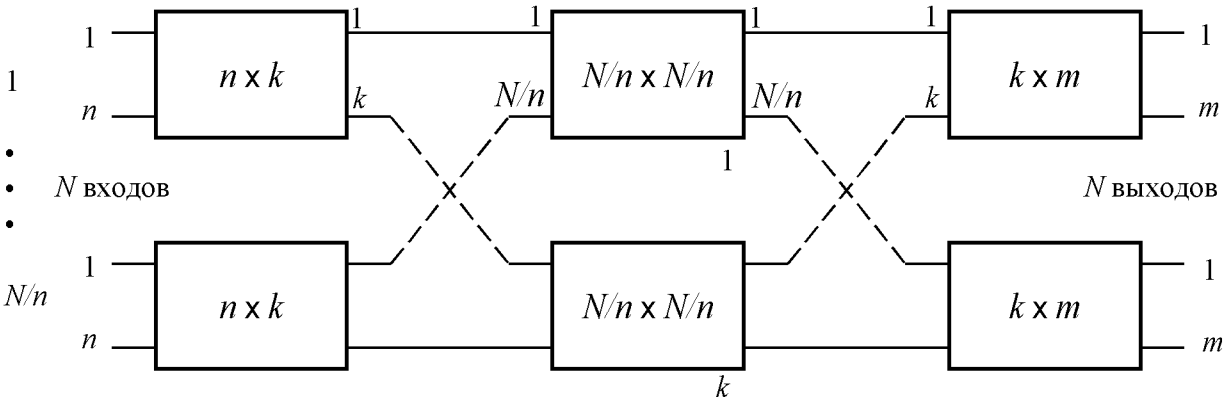


Рис. 7.13. Трехзвенная коммутационная схема

Число точек коммутации для этой схемы (при $m = n$):

$$C = 2\left(\frac{N}{n}\right)nk + k\left(\frac{N}{n}\right)^2 = 2Nk + k\left(\frac{N}{n}\right)^2,$$

где N – общее число входов КС. При надлежащем выборе параметров n и k сложность КС может быть существенно ниже, чем при однозвенном построении, когда количество точек коммутации в точности равно $N \times N$.

Обеспечивая выигрыш в количестве точек коммутации, многозвенная схема может, как было показано на примере двухзвенной структуры, привести к возникновению блокировок. Для трехзвенной схемы **условие неблокируемости** было получено Ч. Клосом [2]:

$$k \geq n + m - 1.$$

Это соотношение можно объяснить наихудшим случаем, когда в конкретном блоке 1-й ступени с n входами заняты $n-1$ выходов (исходящие звонки), а также в конкретном блоке 3-й ступени с m выходами заняты $m-1$ входов (входящие звонки) (см. рис. ниже), причем все эти занятые линии не совпадают. При этом общее число занятых блоков 2-й ступени будет $(n-1)+(m-1)$, и для установления n-го соединения необходим еще один коммутатор второй ступени. Отсюда

$$k = (n-1)+(m-1)+1 = n+m-1.$$

Число точек коммутации для неблокирующей трехзвенной схемы:

$$C_{\text{нб}} = 2N(2n-1) + (2n-1)\left(\frac{N}{n}\right)^2.$$

Чтобы минимизировать число точек коммутации, варьируя количество входов n, нужно продифференцировать $C_{\text{нб}}$ по n и, решая уравнение $\frac{dC_{\text{нб}}}{dn} = 0$, можно получить: $n_{opt} \approx \sqrt{N/2}$, $C_{opt} \approx 4\sqrt{2}N^{3/2}$. При N =100000 для однозвенной схемы число точек коммутации составило бы значение 10^{10}. Для трехзвенной неблокирующей схемы оно составит около $1.7 \cdot 10^8$.

Большей экономии в сложности схемы можно достигнуть, применяя многозвенные схемы с блокировкой, но на очень низком уровне. Для анализа таких схем применяют метод Якобеуса (п.7.3.2) и более простой метод вероятностных графов Ли (п.7.2). Рассмотрим пример такого анализа для приведенной выше трехзвенной схемы.

На рис. 7.14 (слева) представлен вероятностный граф, отражающий две группы каналов, которые должны соединяться друг с другом. Блокировка может возникнуть, если $k < 2n-1$. Рассмотрим систему, в которой не используется концентрация, т. е. $k > n$, и блокировка входных или выходных коммутаторов исключена [2].

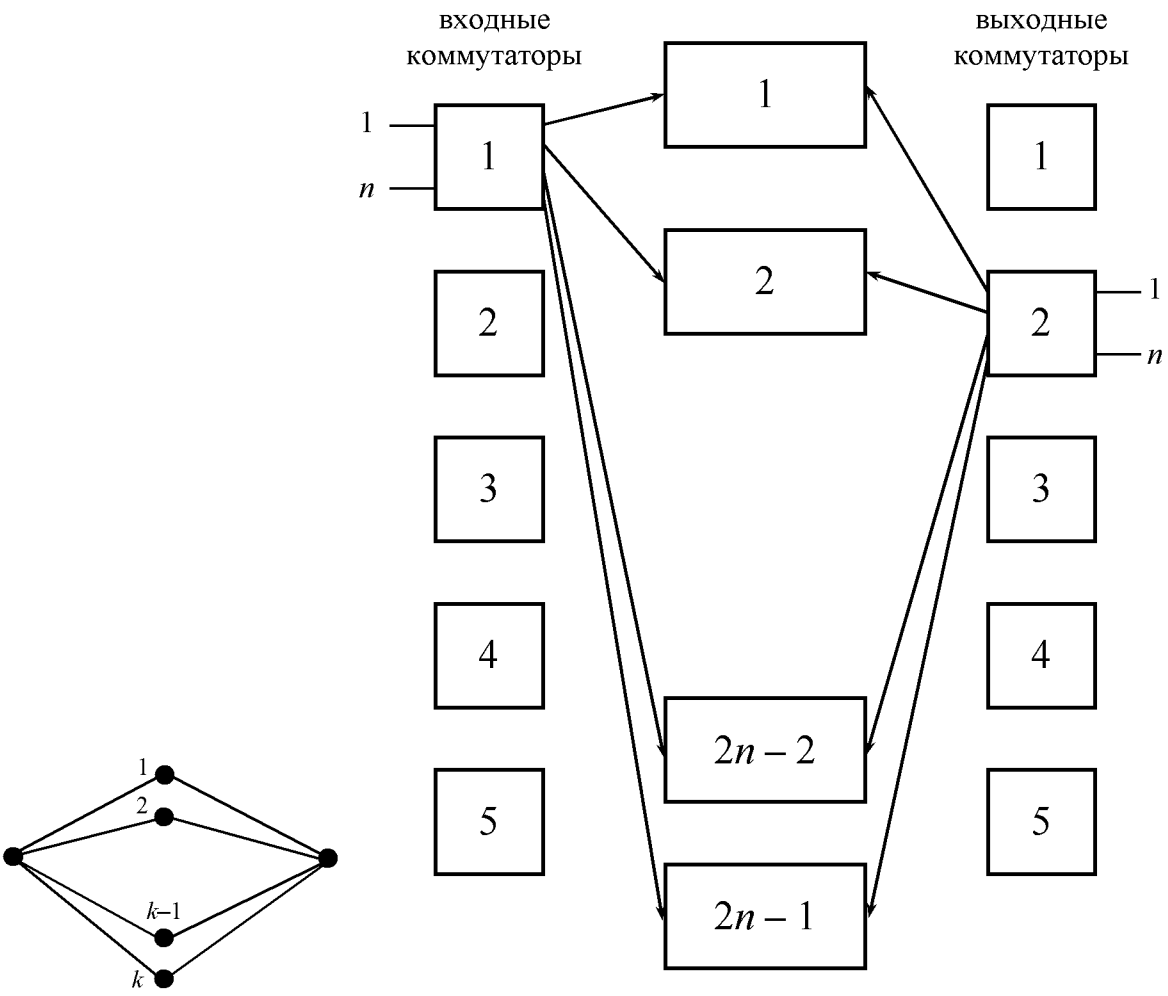


Рис. 7.14. Трехзвенная схема (слева – граф схемы)

Если положить вероятность занятия входного канала равной значению параметра биноминального распределения $a = \dfrac{\lambda}{\lambda + \mu}$ и предполагаемая нагрузка будет распределена равномерно между всеми коммутаторами промежуточного звена, то вероятность занятия ПЛ будет равна $w = an/k$. Для графа Ли с параллельно-последовательной структурой можно записать $p_B = \left[1-(1-w)^2\right]^k$. Нужно заметить, что эта вероятность не равна нулю даже при выполнении условия небло-

кируемости Клоса. Это говорит о погрешности формул, получаемых с помощью графов Ли. Ниже представлена таблица 7.1 для сравнения результатов расчета вероятности блокировки трехзвенной схемы методом графов Ли и метода Якобеуса.

Таблица 7.1

Сравнительный анализ вероятностей блокировки

n	k	p_B по Ли	p_B по Якобе-
64	64	0.002	0.002
64	68	0.0002	0.00016
64	72	$1.5\cdot10^{-5}$	$6.2\cdot10^{-6}$
64	76	$8\cdot10^{-7}$	$1.5\cdot10^{-7}$
120	128	10^{-7}	$3.5\cdot10^{-8}$
240	256	10^{-14}	$1.3\cdot10^{-15}$

Из анализа полученных результатов видно, что метод Ли дает несколько завышенную оценку, особенно заметную для больших размеров КС.

7.4. Управление доступом к среде

Рассмотрим методы управления доступом к среде (Media Access Control – MAC). Первой схемой является ALOHA (Н. Абрамсон, Гавайский ун-т, 1970 г.). Б. Меткалф разработал сеть Ethernet и положил в ее основу алгоритмы ALOHA [3, 14].

Объединение нескольких абонентов в сеть путем простого соединения их между собой общей средой передачи:

1) эфир – куда производят излучение радиопередатчики абонентов, а их приемники принимают из эфира все излучаемые сигналы;

2) общий коаксиальный кабель, соединяющий компьютеры;

3) хабы (hubs), объединяющие между собой сетевые карты ПК.

В любом случае используется принцип эфира «все слышат всех» (общий канал). Пусть передача сообщений в такой сети производится пакетами, включающими в себя адрес получателя, передаваемый абонентами некоторым случайным образом по мере необходимости. В силу случайности возникновения пакетов любой метод разделения требует избыточного ресурса. В протоколе ALOHA никакого резервирования ресурсов не производится. Здесь каждый абонент может

полностью использовать полосу пропускания общего канала. Расплатой за такой агрессивный характер использования является возможность конфликта, столкновения пакетов от различных абонентов. Анализ последствий такого конфликта и является одной из задач, решаемых с помощью теории телетрафика.

7.4.1. Анализ протокола «классическая ALOHA»

Протокол ALOHA будет работать хорошо, если время передач мало по сравнению с общим временем доступности канала. На рис. 7.15 показаны случаи бесконфликтного обмена пакетами и случай столкновения. Предположим, что все пользователи канала в среднем генерируют λ пакетов в секунду. Пакеты имеют фиксированную длину с продолжительностью передачи T секунд [3, 14].

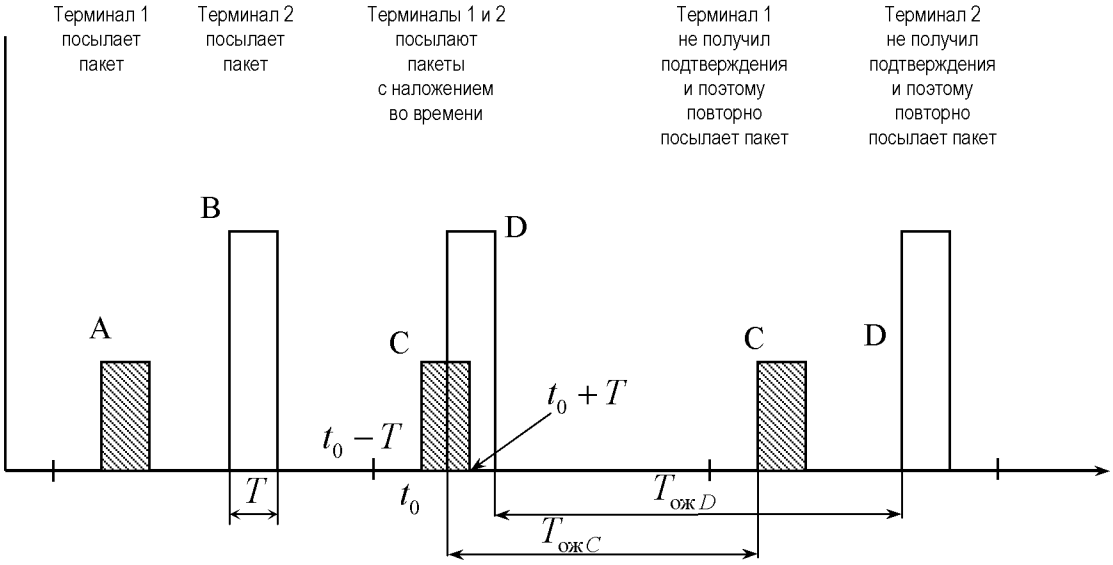


Рис. 7.15. Временная диаграмма для протокола «классическая ALOHA»

Все пользователи генерируют пакеты независимо и так, что вероятность передачи пакета в течение любого интервала времени описывается законом Пуассона. Следовательно, модель протокола ALOHA может быть представлена СМО типа M/D/1. Коэффициент использования канала $\rho = \lambda T$ характеризует отношение полезного времени для передачи пакетов к общему времени. Общее время расходуется на передачу не только пакетов, сгенерированных источниками сообщений, но и повторно передаваемых пакетов, возникающих вследствие конфлик-

тов. Фактическая интенсивность потока пакетов λ' в общем канале будет больше, чем интенсивность λ их генерирования источниками

$$\lambda' = \lambda\left(1 + R + R^2 + R^3 + \ldots\right),$$

где R – доля пакетов, передаваемых пользователем повторно из-за столкновений. Найдем вероятность столкновения и бесконфликтной передачи. Вероятность того, что за время t будет сгенерировано n пакетов, равна:

$$P_n(t) = \frac{(\lambda' t)^n e^{-\lambda' t}}{n!},$$

а вероятность, что за это время не будет сгенерировано ни одного пакета: $P_0(t) = e^{-\lambda' t}$.

Предположим, что в момент t_0 сгенерирован некоторый конкретный пакет. Конфликт произойдет, если какой-либо другой пакет будет сгенерирован в интервале $(t_0 - T, t_0 + T)$. Для того чтобы избежать конфликта, нужно, чтобы за время $2T$ не появилось ни одного пакета. Вероятность такого события равна $P_0(2T) = e^{-\lambda' 2T}$ – вероятность того, что пакет будет передан без конфликта. При этом вероятность повторной передачи пакета $R = 1 - e^{-\lambda' 2T}$.

Иногда конфликт происходит с повторно передаваемым пакетом, и его приходится передавать еще раз и т. д. Среднее число передач одного пакета составит

$$N = 1 + R + R^2 + R^3 + \ldots = \frac{1}{1 - R} = e^{\lambda' 2T}.$$

Соотношение между числом сгенерированных пользователем пакетов λ' и числом успешно переданных по каналу пакетов λ:

$$\lambda = \lambda'(1 - R) = \lambda' e^{-\lambda' 2T}.$$

Подставляя λ в $\rho = \lambda T$, получим: $\rho = \lambda' T e^{-\lambda' 2T}$.

Подставляя выражение для экспоненты через N, получим: $\rho = \frac{\ln N}{2N}$.

На рис. 7.16 приведен график функции ρ от N. Возникает цепная реакция, при которой повторно передаваемые пакеты сами порождают необходимость повторных передач, в результате при некоторых нагрузках поведение системы становится нестабильным.

Дифференцируя функцию $\rho(N)$ и приравнивая производную нулю, можно найти, что максимальное значение коэффициента использования канала

$$\rho_{max} = \frac{1}{2}e^{-1} \cong 0.184,$$

т. е. не превышает 18.4%. Например, если канал позволяет передавать пакеты между любыми двумя абонентами со скоростью 1 Мбит/с, то реально, при подключении к нему по протоколу ALOHA нескольких абонентов, скорость передачи в нем будет составлять не более 184 кбит/с.

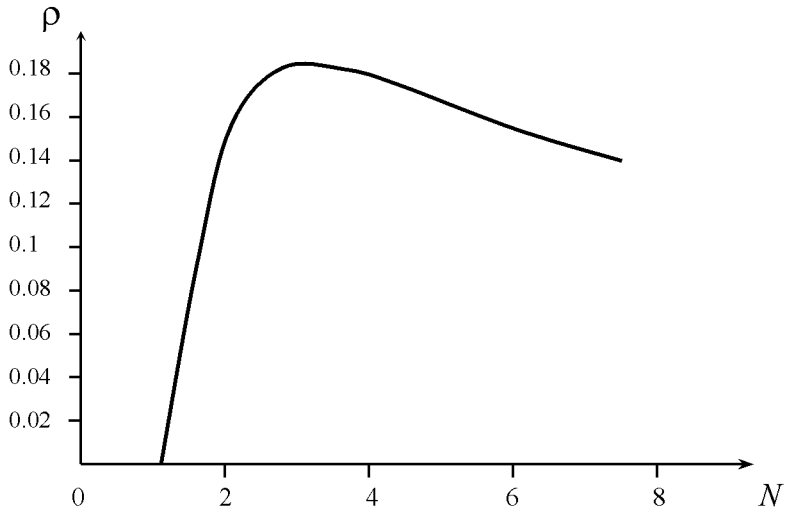


Рис. 7.16. График нагрузки в схеме «классическая ALOHA»

Очевидно, что для улучшения использования ресурса канала нужно, чтобы пользователи воздерживались от передачи, если нагрузка начинает приближаться к критической. К примеру, выдерживать паузу каждый раз, когда пакет пришлось передавать дважды. Это один из алгоритмов борьбы с перегрузкой в сети. Другим эффективным способом улучшения использования канала при случайном доступе является тактирование передач.

7.4.2. Анализ протокола «тактированная ALOHA»

Здесь имеется в виду не полная синхронизация всех передач, а привязка начала передачи пакета, случайно порождаемого пользователем, к временным меткам (тактам) [14]. На рис. 7.17 показана диаграмма передач пакетов в тактируемую систему. Как видно, временная ось разделяется теперь на временные интервалы, одинаковые для всех сетевых устройств, и любой пакет передается только в начале такого интервала. Снова предположим, что время передачи пакета равно T. Положим длительность слота равной этому времени.

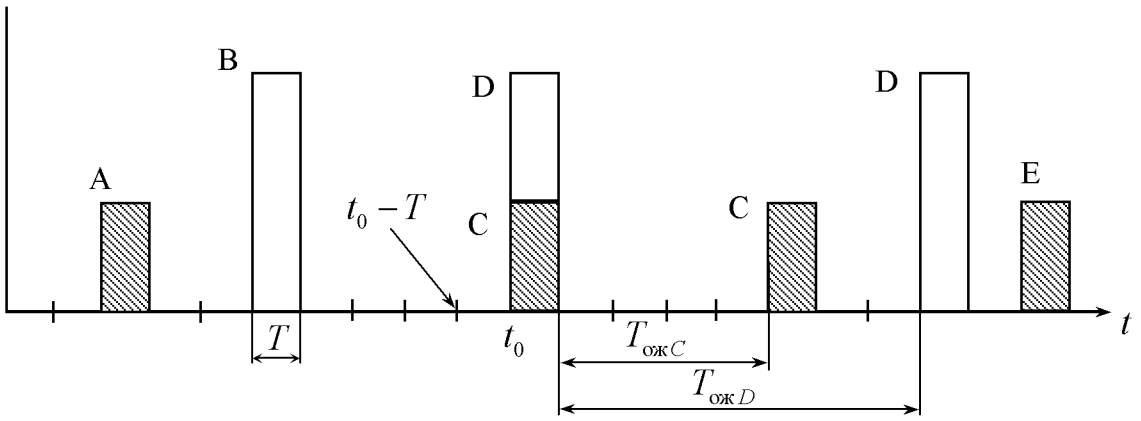


Рис. 7.17. Бесконфликтная передача (A, B, E) и столкновение пакетов (C, D)

Если передача пакета начата в момент t_0, то конфликт с другим пакетом произойдет, когда этот второй пакет будет сгенерирован в интервале $(t_0 - T, t_0)$, т. е. чтобы избежать конфликта, ни одного пакета не должно появиться на интервале длиной T. Вероятность такого события при пуассоновском характере входного потока равна $e^{-\lambda' T}$. Уравнение для коэффициента использования приобретает вид $\rho = \lambda' T e^{-\lambda' T}$.

Среднее число передач одного пакета в случае тактирования будет $N = e^{-\lambda' T}$. Соотношение для связи среднего числа попыток и коэффициента использования:

$$\rho = \frac{\ln N}{N}.$$

Максимальный коэффициент использования канала увеличился вдвое и достиг 36.8%. Недостатком здесь является необходимость иметь общую систему синхронизации для всех сетевых устройств. При передаче сигнала синхронизации по тому же каналу преимущества тактирования исчезают и проявляются лишь в случае использования встроенных высокоточных часов или отдельного канала синхронизации.

7.4.3. Предупреждение столкновений

Любой метод для предотвращения одновременной передачи пакетов несколькими устройствами управления может повысить эффективность канала. Один из методов связан с применением протокола резервирования временных слотов. Другой метод заключается в применении **системы предупреждения столкновений** (collision

avoidance) [14]. Терминалы разделяются по приоритетам, и терминалы с высоким приоритетом, прежде чем начать передачу в фиксированном слоте, посылают сигнал предупреждения. Все устройства управления передачей воспринимают его. Терминалы с низким приоритетом при этом вообще не выходят на передачу в соответствующем такте. Если же два терминала имеют одинаковый приоритет, то каждый из них должен решить, переносить ему время передачи или нет. Вероятность конфликтов при этом снижается, но они по-прежнему могут происходить, а некоторые временные слоты останутся неиспользованными.

Наиболее эффективным способом увеличения коэффициента использования ресурса при случайном доступе является метод, называемый **контролем несущей** (CSMA – Carrier Sense Multiple Access) [3]. В данном методе предполагается, что каждый терминал по наличию сигнала в канале обнаруживает, что другой терминал ведет передачу и сам при этом воздерживается от передачи, чтобы не вызвать конфликта. Как должен поступать терминал, в котором возникла необходимость передачи пакета при снятии сигнала «занято»? Он может ждать момента снятия и немедленно начинать передачу. Это вариант контроля несущей с так называемыми **настойчивыми терминалами**. Терминал захватывает канал сразу, как только он освободится. Однако при этом, может оказаться, что другой настойчивый терминал поступит так же, и конфликт произойдет сразу после освобождения канала. Альтернативой является протокол с **ненастойчивыми терминалами**. В этом случае, если канал оказывается занятым, терминал с помощью датчика случайных чисел откладывает попытку передачи на более поздний момент. Повысить эффективность использования канала могут оба варианта контроля несущей – как с настойчивыми, так и с ненастойчивыми терминалами. Задержки при этом увеличиваются, но незначительно, по сравнению с задержками из-за повторных запросов в перегруженном канале. Третий вариант контроля несущей позволяет получить такую же эффективность использования канала, как в варианте с ненастойчивыми терминалами, но при существенно меньших задержках. Его называют методом с ***p*-настойчивыми терминалами**. В нем при освобождении терминал с вероятностью p передает пакет немедленно, а с вероятностью $1-p$ откладывает передачу на случайный интервал времени. Можно подобрать вероятность немедленной передачи для получения максимальной производительности в зависимости от трафика в сети (рис. 7.18).

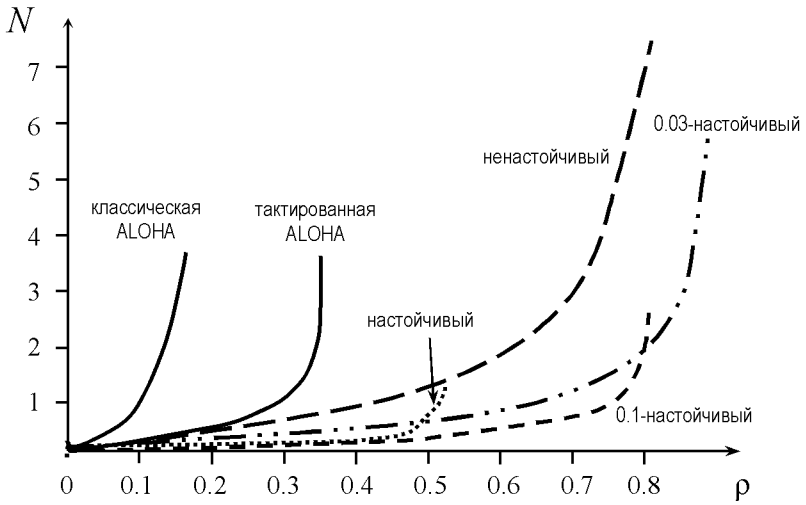


Рис. 7.18. Влияние контроля несущей на эффективность использования канала

Варианты с ненастойчивыми и p-настойчивыми терминалами позволяют получить коэффициент использования до 80% ресурса канала (результаты получены имитационным моделированием на GPSS). Эффективность контроля несущей зависит от того, насколько задержки распространения в канале малы по сравнению со временем передачи пакета. Это условие выполняется в низкоскоростных сетях или сетях с малой протяженностью трасс. В гигабитных сетях или сетях со спутниковыми геостационарными ретрансляторами добиться его выполнения не удается. Здесь приходится применять более сложные методы, например, передачу отдельного сигнала занятости перед передачей пакета [3].

7.5. Контрольные вопросы

1. Какие СМО называют однофазными?
2. Какие СМО называют многофазными?
3. Приведите пример сети массового обслуживания.
4. Что утверждается в теореме Бёрке?
5. Сформулируйте теорему Джексона.
6. В чем заключаются условия эргодичности марковской модели узла сети?
7. Что описывают уравнения локального равновесия в замкнутой сети?

8. Приведите пример параллельно-последовательного вероятностного графа.
9. Поясните понятие «направление» в коммутационной схеме.
10. Какой коммутатор называется неблокирующим?
11. При каком условии происходит блокировка в коммутаторе?
12. В чем преимущество многозвенных коммутационных схем перед однозвенными?
13. Какую функцию выполняют промежуточные линии в коммутационной схеме?
14. Назовите три случая, когда наступает блокировка в коммутационной схеме.
15. Каким должно быть число источников нагрузки в СМО, чтобы можно было бы считать адекватным применение распределения Эрланга?
16. Чем ограничено число источников нагрузки при использовании распределения Бернулли?
17. Что представляет собой коммутационная схема с расширением?
18. Что представляет собой коммутационная схема со сжатием?
19. Для чего в системах связи используется концентратор? Какова его основная функция?
20. Сформулируйте условие неблокируемости для трехзвенной коммутационной схемы.
21. В чем основной недостаток протокола «классическая ALOHA» по сравнению с протоколом «тактированная ALOHA»?
22. Поясните функционирование системы предупреждения столкновений.
23. В чем заключается механизм контроля несущей?
24. Чем отличается настойчивый терминал от ненастойчивого?
25. Поясните понятие «p-настойчивый терминал».

ЗАКЛЮЧЕНИЕ

В учебном пособии рассмотрены основные положения теории телетрафика, изложены методологические основы теории систем массового обслуживания, дано развернутое определение математических моделей различных потоков, проведена классификация и анализ основных характеристик различных систем массового обслуживания, представлены практические примеры анализа трафика в элементах сетей связи.

Анализируемые системы и случайные потоки в них представлены в виде марковских моделей. Подробно рассмотрены самоподобные (фрактальные) модели трафика. Значительное внимание уделяется выводу уравнений Эрланга, характеризующих параметры обслуживания простейшего потока заявок в системах с отказами и ожиданием.

Подробно описаны методы статистического моделирования потоков в телекоммуникационных системах. Дано краткое описание моделирования систем массового обслуживания посредством сетей Петри. Заключительная глава посвящена методам анализа сетей массового обслуживания и методам управления доступом к среде в системах радиодоступа.

За рамками изложения остались: анализ потоков с гиперэкспоненциальным и гиперэрланговским распределением [12]; системы обслуживания с повторными вызовами и анализ трафика в мультисервисных сетях [21, 22]; анализ СМО с обходными направлениями [12]; анализ трафика в системах с динамическим распределением ресурса [22].

СПИСОК СОКРАЩЕНИЙ

АКФ - автокорреляционная функция
ДУ – дифференциальное уравнение
ЗР – закон распределения
КС – коммутационная схема
КФ – корреляционная функция
МСЭ – Международный союз электросвязи
ПЛ – 1) преобразование Лапласа; 2) промежуточная линия
ПРВ – плотность распределения вероятностей
СВ – случайная величина
СМО – система массового обслуживания
СП – случайный процесс
ТС – телекоммуникационная система
ФБД – фрактальное броуновское движение
ФГШ – фрактальный гауссовский шум
ФРВ – функция распределения вероятностей
ЦПТ – центральная предельная теорема
ЧНН – час наибольшей нагрузки

ITU – International Telecommunication Union
QoS – Quality of Service

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Башарин, Г. П. Лекции по математической теории телетрафика : учебное пособие / Г. П. Башарин. – 2-е изд., испр. и доп. – М. : Изд-во РУДН, 2007. – 268 с.

2.* Беллами, Д. Цифровая телефония / Д. Беллами; пер. с англ., под ред. А. Н. Берлина, Ю. Н. Чернышева. – 3-е изд. – М. : Эко-Трендз, 2004. – 640 с.

3. Бертсекас, Д. Сети передачи данных / Д. Бертсекас, Р. Галлагер; пер. с англ., под ред. Б. С. Цыбакова. – М. : Мир, 1989. – 544 с.

4. Васильев, К. К. Математическое моделирование систем связи : учебное пособие / К. К. Васильев, М. Н. Служивый. – 2-е изд., перераб. и доп. – Ульяновск : УлГТУ, 2010. – 170 с.

5. Вентцель, Е. С. Теория вероятностей : учебник для втузов / Е. С. Вентцель. – 8-е изд., перераб. и доп. – М. : Физматлит, 1999. – 576 с.

6. Вентцель, Е. С. Прикладные задачи теории вероятностей / Е. С. Вентцель, Л. А. Овчаров. – М. : Радио и связь, 1983. – 416 с.

7. Ершов, В. А. Мультисервисные телекоммуникационные сети / В. А. Ершов, Н. А. Кузнецов. – М. : Изд-во МГТУ им. Н. Э. Баумана, 2003. – 432 с.

8.* Иверсен, В. Б. Разработка телетрафика и планирование сетей : учебное пособие / В. Б. Иверсен; пер. с англ., под ред. А. Н. Берлина. – М. : НОУ «Интуит»; Бином. Лаборатория знаний, 2011. – 526 с.

9. Ирвин, Дж. Передача данных в сетях : инженерный подход / Дж. Ирвин, Д. Харль; пер. с англ. – СПб. : БХВ-Петербург, 2003. – 448 с.

Символом * отмечена основная литература по дисциплине «Теория телетрафика»

10. Клейнрок, Л. Теория массового обслуживания / Л. Клейнрок; пер. с англ. И. И. Грушко, под ред. В. И. Неймана. – М. : Машиностроение, 1979. – 432 с.

11. Клейнрок, Л. Вычислительные системы с очередями / Л. Клейнрок; пер. с англ., под ред. Б. С. Цыбакова. – М. : Мир, 1979. – 600 с.

12.* Корнышев, Ю. Н. Теория телетрафика : учебное пособие / Ю. Н. Корнышев, А. П. Пшеничников, А. Д. Харкевич. – М. : Радио и связь, 1996. - 272 с.

13. Кроновер, Р. М. Фракталы и хаос в динамических системах : учебное пособие для вузов / Р. М. Кроновер; пер. с англ., под ред. Т. Э. Кренкеля. – 2-е изд., доп. – М. : Техносфера, 2006. – 488 с.

14.* Крылов, В. В. Теория телетрафика и ее приложения : учебное пособие / В. В. Крылов, С. С. Самохвалова. – СПб. : BHV-Санкт-Петербург, 2005. – 288 с.

15. Кузин, Л. Т. Основы кибернетики. В 2 т. Т. 2 : Основы кибернетических моделей : учебное пособие для вузов / Л. Т. Кузин. – М. : Энергия, 1979. – 584 с.

16. Лагутин, В. С. Сети связи : проблемы эффективности использования ресурсов цифровых линий / В. С. Лагутин. – М. : Радио и связь, 1999. – 229 с.

17. Лидский, Э. А. Задачи трафика в сетях связи : учебное пособие / Э. А. Лидский. – Екатеринбург : УГТУ-УПИ, 2006. – 202 с.

18. Платонов, Г. А. Поезда, пассажиры и… математика / Г. А. Платонов, М. А. Файнберг, М. С. Штильман. – М. : Транспорт, 1977. – 240 с.

19. Розенберг, В. Я. Что такое теория массового обслуживания / В. Я. Розенберг, А. И. Прохоров. – 2-е изд. – М. : Сов.радио, 1965. – 256 с.

20. Саати, Т. Л. Элементы теории массового обслуживания и ее приложения / Т. Л. Саати; пер. с англ. Е. Г. Коваленко, под ред. И. Н. Коваленко. – М. : Сов.радио, 1971. – 520 с.

21. Степанов, С. Н. Основы телетрафика мультисервисных сетей / С. Н. Степанов. – М. : Эко-Трендз, 2010. – 392 с.

22.* Степанов С.Н. Теория телетрафика : концепции, модели, приложения / С. Н. Степанов. – М. : Горячая линия – Телеком, 2015. – 868 с.

23. Теория телетрафика : основы расчета систем проводной связи / Х. Штермер, Э. Белендорф, Н. Бининда и др.; пер. с нем., под ред. Г. П. Башарина. – М. : Связь, 1971. - 320 с.

24. Тихонов, В. И. Марковские процессы / В. И. Тихонов, М. А. Миронов. – М. : Советское радио, 1977. – 488 с.

25. Украинцев, Ю. Д. Теория телетрафика для анализа современных телекоммуникационных сетей : учебное пособие / Ю. Д. Украинцев. – Ульяновск : УлГУ, 2012. – 148 с.

26. Шелухин, О. И. Самоподобие и фракталы. Телекоммуникационные приложения / О. И. Шелухин, А. В. Осин, С. М. Смольский; под ред. О. И. Шелухина. – М. : Физматлит, 2008. – 368 с.

27. Шелухин, О. И. Фрактальные процессы в телекоммуникациях : монография / О. И. Шелухин, А. М. Тенякшев, А. В. Осин. – М. : Радиотехника, 2003. – 480 с.

28. Шелухин, О. И. Моделирование информационных систем : учебное пособие / О. И. Шелухин, А. М. Тенякшев, А. В. Осин; под ред. О. И. Шелухина. – М. : Радиотехника, 2005. – 368 с.

29.* Шелухин, О. И. Моделирование информационных систем : учебное пособие для вузов / О. И. Шелухин. – 2-е изд., перераб. и доп. – М. : Горячая линия - Телеком, 2012. – 536 с.

30. Шелухин, О. И. Мультифракталы. Инфокоммуникационные приложения / О. И. Шелухин. – М. : Горячая линия - Телеком, 2011. – 576 с.

31. Шнепс, М. А. Системы распределения информации. Методы расчета : справочное пособие / М. А. Шнепс. – М. : Связь, 1979. – 344 с.

32. Эллдин, А. Основы теории телетрафика / А. Эллдин, Г. Линд; пер. с англ., под ред. А. Д. Харкевича. – М. : Связь, 1972. – 200 с.

33. Fiche, G. Communicating Systems & Networks: Traffic & Performance / G. Fiche, G. Hebuterne. – Kogan Page Limited, 2004. – 528 p.

Интернет-ресурсы

34. http://www.itu.int/rec/T-REC-E.600/en (Terms and definitions of traffic engineering) (дата обращения: 15.10.2015)

35. http://www.itu.int/rec/T-REC-E.500/en (Traffic intensity measurement principles) (дата обращения: 15.10.2015)

36. http://www.itu.int/rec/T-REC-E.800/en (Definitions of terms related to quality of service) (дата обращения: 15.10.2015)

37. http://www.itu.int/rec/T-REC-E.800SerSup1/en (Table of the Erlang formula) (дата обращения: 15.10.2015)

38. http://www.itu.int/rec/T-REC-E.800SerSup2/en (Curves showing the relation between the traffic offered and the number of circuits required) (дата обращения: 15.10.2015)

ОГЛАВЛЕНИЕ

Учебное издание

НАМЕСТНИКОВ Сергей Михайлович
СЛУЖИВЫЙ Максим Николаевич
УКРАИНЦЕВ Юрий Дмитриевич

ОСНОВЫ ТЕОРИИ ТЕЛЕТРАФИКА

Учебное пособие

Редактор Н.А. Евдокимова
ЛР №020640 от 22.10.97
Подписано в печать 17.06.2016.
Формат 60×84/16. Усл. печ. л. 7,21.
Тираж 100 экз. Заказ 1000

Ульяновский государственный технический университет
432027, г. Ульяновск, ул. Сев. Венец, 32.
ИПК «Венец» УлГТУ, 432027, г. Ульяновск, ул. Сев. Венец, 32.