ФЕДЕРАЛЬНОЕ АГЕНТСТВО СВЯЗИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ТЕЛЕКОММУНИКАЦИЙ
им. проф. М. А. БОНЧ-БРУЕВИЧА»
(СПбГУТ)

Ю. Ф. Кожанов

ТЕОРИЯ ТЕЛЕТРАФИКА

УЧЕБНОЕ ПОСОБИЕ

САНКТ-ПЕТЕРБУРГ
2020

УДК 621.395(075.8)
ББК 32.88я73
К 58

Рецензенты:
доктор технических наук, профессор кафедры информационных систем и технологий
ФГБОУ ВПО «Санкт-Петербургский государственный экономический университет»
М. О. Колбанёв,

доктор технических наук, профессор кафедры
информационных управляющих систем СПбГУТ
Н. Н. Мошак

Утверждено редакционно-издательским советом СПбГУТ
в качестве учебного пособия

Кожанов, Ю. Ф.
К 58 Теория телетрафика : учебное пособие / Ю. Ф. Кожанов ; СПбГУТ. – СПб., 2020. – 203 с.

ISBN 978-5-89160-193-2

Излагаются основные результаты теории телетрафика, необходимые для проектирования телекоммуникационных систем с коммутацией каналов и пакетов. Особое внимание уделено современным системам с ожиданием, использующим механизмы приоритетного обслуживания в IP-сети для обеспечения требуемого качества обслуживания в соответствии с международными стандартами (МСЭ-Т, RFC). Приводится методика оценки функционирования подсистем на основе ключевых показателей эффективности. Рассмотрены методы обнаружения компрометации персональных данных на примере стандарта GSM.

Приведены программы для расчета наиболее важных характеристик сетевых элементов.

Предназначено для студентов, обучающихся по направлению 11.03.02 «Инфокоммуникационные технологии и системы связи».

УДК 621.395(075.8)
ББК 32.88я73

ISBN 978-5-89160-193-2
© Кожанов Ю. Ф., 2020

© Федеральное государственное бюджетное образовательное учреждение высшего образования «Санкт-Петербургский государственный университет телекоммуникаций им. проф. М. А. Бонч-Бруевича», 2020

СОДЕРЖАНИЕ

Введение 5

Глоссарий 6

1. ПОНЯТИЕ О КАЧЕСТВЕ ОБСЛУЖИВАНИЯ 9
   1.1. Параметры качества обслуживания 9
   1.2. Дозирование и выравнивание трафика 13
   1.3. Предотвращение перегрузки 14
   1.4. Обслуживание очередей 16

2. ПОТОКИ ВЫЗОВОВ, ВРЕМЯ ОБСЛУЖИВАНИЯ 22

3. ПРИНЦИПЫ ПОСТРОЕНИЯ СИСТЕМ КОММУТАЦИИ 25
   3.1. Понятие о нагрузке 25
   3.2. Типы систем коммутации 29
   3.3. Простейшие коммутационные устройства 36
   3.4. Стационарный процесс рождения и гибели в полнодоступном пучке 37

4. СИСТЕМА С ЯВНЫМИ ПОТЕРЯМИ (Blocked Calls Cleared, BCC) 42
   4.1. Модель Эрланга (M/M/V/0). Дискретная формула Эрланга 42
   4.2. Модель Эрланга (M/M/V/0). Интегральная формула Эрланга 43
   4.3. Система M/M/V/0/ /PPQ 46
   4.4. Модель Эрланга (M/M/V/0). Поток Пальма 48
   4.5. Модель Энгсета (M/M/V/0/N). Дискретная формула Энгсета 53
   4.6. Модель Энгсета (M/M/V/0/N). Интегральная формула Энгсета 55

5. СИСТЕМА С ОЖИДАНИЕМ БЕЗ ПОТЕРЬ (Blocked Calls Queued, BCQ) 62
   5.1. Система M/M/V. Стационарные вероятности 62
   5.2. Система M/M/V. Функция распределения 64
   5.3. Система M/M/1 66
   5.4. Система M/D/V 66
   5.5. Система M/D/1 68
   5.6. Однолинейная многофазная система обслуживания 69

6. СИСТЕМА С ОЖИДАНИЕМ И ПОТЕРЯМИ 75
   6.1. Система M/M/V/K. Стационарные вероятности 75
   6.2. Система M/M/1/K 77

7. СИСТЕМА С ОЖИДАНИЕМ И ПРИОРИТЕТАМИ 80
   7.1. Схема прохождения вызова 80
   7.2. Система с абсолютными приоритетами M/M/1/ / /NPQ 81
   7.3. Система с относительными приоритетами M/M/1/ / /CBWFQ 86
   7.4. Система с относительными приоритетами M/M/1/ / /WFQ 90

8. СИСТЕМА С ПОВТОРНЫМИ ВЫЗОВАМИ (Blocked Calls Repeated, BCR) 97

9. СИСТЕМА С КОЛЛИЗИЕЙ (Collisions Calls Repeated, CCR) 102

10. ОБЕСПЕЧЕНИЕ КАЧЕСТВА ОБСЛУЖИВАНИЯ В ИНТЕРНЕТЕ 106
   10.1. Общие принципы формирования пакетов 107
   10.2. Формирование речевых пакетов. Рекомендации G.711, G.729, G.723, G.726 108

11. СРАВНЕНИЕ СЕТЕЙ КОММУТАЦИИ КАНАЛОВ И ПАКЕТОВ 113
11.1. Сеть с интегрированным обслуживанием 114
11.2. Сеть с дифференцированным обслуживанием 118

12. РАСЧЕТ ОТДЕЛЬНЫХ ХАРАКТЕРИСТИК МАРШРУТИЗАТОРА 126
12.1. Расчет и использование метрик с учетом нагрузки на интерфейсе 127
12.2. Расчет метрики параллельных интерфейсов 131
12.3. Расчет буфера памяти маршрутизатора для UDP-соединений 133

13. РАСПРЕДЕЛЕНИЕ КАНАЛОВ СВЯЗИ НА «ПОСЛЕДНЕЙ МИЛЕ СЕРВЕРА» 140

14. ОЦЕНКА НАДЕЖНОСТИ СЕТЕВЫХ ЭЛЕМЕНТОВ 147

15. ПЕРЕХОДНЫЕ ВЕРОЯТНОСТИ СОСТОЯНИЙ В ПОЛНОДОСТУПНОМ ПУЧКЕ 151
15.1. Процесс рождения 151
15.2. Процесс рождения и гибели 154
15.3. Процесс гибели 156

16. КЛАСС ОБСЛУЖИВАНИЯ СЕТЕВЫХ ЭЛЕМЕНТОВ 160
16.1. Понятие о классе обслуживания 160
16.2. Виды связи 161
16.3. Параметры и формулы для комплексного контроля KPI 162
16.4. Параметры и формулы для контроля KPI пучка каналов 166
16.5. Параметры и формулы для контроля KPI управляющего устройства 169
16.6. Параметры и формулы для контроля KPI сигнальных каналов 169

17. ОБНАРУЖЕНИЕ КОМПРОМЕТАЦИИ ПЕРСОНАЛЬНЫХ ДАННЫХ 173
17.1. Использование чужих персональных данных 176
17.2. Неизвестные идентификационные данные 177
17.3. Одновременная регистрация в разных сетях 178

ЗАКЛЮЧЕНИЕ 179

СПИСОК ЛИТЕРАТУРЫ 181

Приложение 1. Программа вычисления формулы Пуассона 182
Приложение 2. Программа вычисления формулы Эрланга (система M/M/V/0) 183
Приложение 3. Программа классификации суммарного потока вызовов 184
Приложение 4. Программа вычисления формулы Энгсета (система M/M/V/0/N) 185
Приложение 5. Программа вычисления функции распределения системы M/M/V 186
Приложение 6. Программа вычисления функции распределения системы M/D/V 187
Приложение 7. Программа расчета системы M/M/1/ / /PPQ 188
Приложение 8. Программа расчета системы M/M/1/ / /CBWFQ 189
Приложение 9. Программа расчета системы M/M/1/ / /WFQ 191
Приложение 10. Программа расчета системы с повторными вызовами 193
Приложение 11. Программа расчета системы с интегрированным обслуживанием 194
Приложение 12. Программа расчета системы с дифференцированным обслуживанием 196
Приложение 13. Программа расчета коэффициента готовности 198
Приложение 14. Преобразование Лапласа – Стилтьеса 199

ВВЕДЕНИЕ

На существующем этапе большинство сетей связи исторически имеют смешанную структуру цифровой сети интегрального обслуживания (ISDN), сетей связи мобильных абонентов (GSM/UMTS/LTE) и сети следующего поколения (NGN), основанной на сети Интернет.

Каждый переход к следующей сети сопровождался совершенствованием технологии и усложнением функционирования, что неизбежно приводило к появлению новых принципов коммутации и, соответственно, к разработке новых, более сложных расчетов для проектирования сетей. Любую сложную конструкцию невозможно описать простыми математическими формулами без потери точности расчета. Именно поэтому для расчета и проектирования современных систем связи приходится использовать достаточно сложный математический аппарат, включающий теорию вероятностей, интегральное и дифференциальное исчисления, теорию бесконечно малых чисел, операционное исчисление и др. Достаточно сложные математические конструкции, используемые в книге, адекватны сложности реальных объектов.

Исходя из этого, в данной книге рассматриваются современные методы расчета систем коммутации каналов (системы с явными потерями и повторными вызовами) и систем пакетной коммутации (системы с ожиданием и системы с коллизией). Наличие многочисленных примеров расчета способствует лучшему пониманию материала.

Для удобства пользования в приложении на языке программирования C++ приведены программы расчета наиболее популярных систем массового обслуживания.

Основное содержание пособия базируется на материале, ранее изданной автором монографии «Качество обслуживания в сетях связи», в которую внесены некоторые исправления и дополнения.

Книга предназначена для студентов старших курсов по специальностям информационные системы и технологии, Сети связи и системы коммутации. Очевидно, она окажется полезной для научных работников, аспирантов, специалистов по телекоммуникации и обслуживающего персонала телекоммуникационных систем.

Электронную версию пособия можно найти в библиотеке СПбГУТ на сайте www.sut.ru.

Глоссарий

APQ (Absolute Priority Queueing) – механизм обслуживания очередей с абсолютным приоритетом, при котором вначале обслуживаются все пакеты из очереди high, затем – из очередей medium, normal и в конце – из очереди low. Имеется две опции ***NPQ*** и ***PPQ***.

ATM (Asynchronous Transfer Mode) – режим асинхронной передачи. Один из способов транспортировки информации в сети коммутации пакетов по ***VC***. Используются пакеты фиксированной длины 53 байта, из которых 5 байт занимает адресная часть, 48 байт – полезная нагрузка.

BCC (Blocked Calls Cleared) – система коммутации, в которой вызов, поступивший в момент занятости всех обслуживающих приборов, теряется.

BCQ (Blocked Calls Queued) – система коммутации, в которой вызов, поступивший в момент занятости всех обслуживающих приборов, становится в очередь на обслуживание.

BCR (Blocked Calls Repeated) – система коммутации, в которой вызов, поступивший в момент занятости всех обслуживающих приборов, теряется и сразу же поступает заново.

BHCA (Busy Hour Call Attempts) – час наибольшей нагрузки. Промежуток времени длиной 1 час, в течение которого средняя поступающая нагрузка от пользователей принимает максимальное значение.

BW (Band Width) – полоса передачи. Часть от скорости интерфейса, выделенная для определенных приложений.

CBWFQ (Class-Based Weighted Fair Queuing) – механизм обслуживания очередей с относительными приоритетами, при котором поступивший пакет в соответствии со своим приоритетом помещается в одну из выходных очередей. За один период прохода всех очередей из каждой очереди считывается различное число пакетов.

CCR (Collisions Calls Repeated) – система коммутации, в которой вызов, поступивший в момент занятости всех обслуживающих приборов, теряется и в течение случайного промежутка времени поступает заново. Вызов принимается к обслуживанию, если он единственный.

CCS (Common Channel Signalling) – общий канал сигнализации. Способ передачи сигналов управления и взаимодействия, в котором передача сигнальной информации производится пакетами на скорости 64 кбит/с.

DSCP (Differentiated Service Code Point) – код дифференцированной услуги. Способ маркирования пакетов для приоритетного обслуживания. Имеет поле длиной 1 байт, старшая часть определяет приоритет обслуживания, младшая – приоритет отбрасывания пакета при перегрузке сети.

Ethernet – самая распространенная технология коммутации пакетов в ***LAN***.

GoS (Grade of Service) – класс обслуживания. Методика контроля качества обслуживания сетевого элемента обслуживающим персоналом на основании измерения статистических данных параметров нагрузки.

IP (Internet Protocol) – межсетевой протокол. Обеспечивает доставку пакетов в IP-сети в соответствии с адресом назначения.

ISDN (Integrated Services Digital Network) – цифровая сеть интегрального обслуживания (ЦСИО). Обеспечивает доставку информации любого вида в сети с коммутацией каналов. Коммутируемой единицей является цифровой B-канал со скоростью передачи 64 кбит/с. Для широкополосных услуг используется несколько B-каналов.

ISP (Internet Service Provider) – провайдер интернет-услуг. Организация или юридическое лицо, решающие организационные, финансовые и технические вопросы по предоставлению доступа пользователей в Интернет.

ITU-T (International Telecommunication Union) – сектор стандартизации телекоммуникаций международного союза электросвязи.

KPI (Key Performance Indicator) – ключевой показатель эффективности. Показатели качества обслуживания подсистем сетевого элемента на основании GoS.

LAN (Local Area Network) – локальная сеть. Обычно ***Ethernet***.

MOS (Mean Opinion Scores) – шкала средней экспертной оценки для оценки передачи качества голосового сигнала.

MTU (Maximum Transmission Unit) – максимальный размер пакета, который может быть передан по определенному интерфейсу без фрагментации.

NGN (Next Generation Network) – сеть следующего поколения. Общее название сетей пакетной коммутации для доставки мультимедийной информации с заданным качеством обслуживания (***QoS***).

NPQ (Non Preemptive Queueing) – опция ***APQ***, при которой поступивший приоритетный вызов не прерывает обслуживание вызова, находящегося в обслуживании.

OSPF (Open Shortest Path First) – открытый протокол маршрутизации выбора первого кратчайшего пути. Обеспечивает сбор и рассылку известных сетей другим маршрутизаторам.

PPQ (Preemptive Priority Queueing) – опция ***APQ***, при которой поступивший приоритетный вызов прерывает обслуживание наименее приоритетного вызова.

PSTN (Public Switched Telephone Network) – телефонная сеть общего пользования. Коммутируемой единицей является канал тональной частоты 0,3–3,4 кГц.

QoS (Quality of Service) – качество сетевого обслуживания. Совокупность требований к сети связи для обеспечения доставки информации пользователя «из конца в конец» с заданными характеристиками.

RED (Random Early Detection) – алгоритм произвольного раннего обнаружения перегрузки. Предусматривает превентивное уничтожение поступающих пакетов с вероятностью, зависящей от текущей длины очереди.

RSVP (Resource Reservation Protocol) – протокол резервирования ресурсов. Резервирует сетевые ресурсы для гарантированного обеспечения ***QoS***.

SLA (Service Level Agreement) – соглашение об уровне обслуживания между пользователем и ***ISP***.

SNMP (Simple Network Management Protocol) – простой протокол управления сетью.

TCP (Transmission Control Protocol) – протокол управления передачей с гарантией доставки дейтаграмм.

TOS (Type of Service) – восьмибитовое поле в заголовке IP-пакета, определяющее приоритетность обслуживания пакета.

VC (Virtual Cirquit) – виртуальный канал. Создается между пользователями непосредственно перед обменом информацией и разрушается после окончания сеанса связи.

UDP (User Datagram Protocol) – протокол передачи дейтаграмм без гарантии доставки.

WFQ (Weighted Fair Queuing) – механизм обслуживания очередей с относительными приоритетами, при котором пакеты одного потока помещаются в индивидуальную выходную очередь. За один период прохода всех очередей считывается пакет с наименьшим местом ожидания.

1. ПОНЯТИЕ О КАЧЕСТВЕ ОБСЛУЖИВАНИЯ

1.1. Параметры качества обслуживания

При передаче информации по сети связи возникают следующие проблемы [4].

• Постоянные задержки прохождения сигналов от передатчика до приемника. Задержки в любых сетях связи существуют всегда. Абоненты не ощущают задержек и общаются в реальном масштабе времени, если суммарная задержка «из конца в конец» не превышает 150 мс [7]. Влияние задержки на качество воспроизведения речи приведено в табл. 1.1.

Таблица 1.1

Влияние сетевой задержки на качество обслуживания

Значение задержки «из конца в конец», мс	Оценка качества
0–150	Очень хорошее
150–300	Хорошее
300–400	Удовлетворительное
Свыше 400	Не удовлетворительное

Задержки, во-первых, возникают при формировании исходной информации. В цифровых сетях интегрального обслуживания с коммутацией каналов (ЦСИО) очередной фрагмент речи через каждые 125 мкс преобразуется в 8 бит (байт) и сразу же передается в заранее установленный канал связи, поэтому значение задержки ничтожно. В системе с коммутацией пакетов несколько байтов собираются в один пакет, иногда подвергаются анализу для устранения избыточности (сжимаются до меньшего значения), обрамляются заголовком и хвостовиком и только потом передаются. В результате такой сложной процедуры время формирования пакета *tcod* принимает значение от 10 до 37,5 мс. Некоторое время *tdec* (до 7,5 мс) требуется на приеме для обратных операций.

Во-вторых, в системе с коммутацией пакетов имеются задержки переприема (serialization), связанные с тем, что для определения маршрута следования поступившего пакета маршрутизатору необходимо прочитать адресную часть пакета, а для этого необходимо принять весь пакет или, по крайней мере, его заголовок. Следует отметить, что при наличии очереди задержка переприема определяется полосой передачи *bw*, занимающей часть скорости передачи по каналу связи *VEL*. Поэтому маршрутизатору для приема пакета длиной *LEN* байт от приложения, которому выделена полоса передачи *bw* кбит/с, может потребоваться время $8 \cdot LEN/bw$ мс. Например, для речевого пакета с использованием Рекомендации G.711

(*LEN* = 238 байт, *bw* = 95.2 кбит/с) время приема на одном маршрутизаторе может составить 20 мс. Поэтому для снижения задержки переприема коммутация в современных маршрутизаторах осуществляется сразу же после приема заголовка пакета.

В-третьих, при перегрузке системы с коммутацией пакетов маршрутизаторы устанавливают поступающие пакеты в выходную очередь. При большой интенсивности трафика создается дополнительная задержка *te* из-за прохождения очереди на передачу.

В-четвертых, задержки возникают из-за конечного времени распространения сигнала (примерно 6 мкс/км), которое имеет небольшое значение на коротких линиях, за исключением спутниковых каналов связи, где задержка может достигать до 400 мс.

- Потеря информации.

В процессе прохождения информации по сети связи она может быть:

– искажена из-за влияния помех;

– доставлена не по адресу из-за сбоя в программном обеспечении или в канале;

– отвергнута из-за поздней доставки.

Потеря информации приводит к различным последствиям в зависимости от способа преобразования аналогового сигнала, причем наиболее катастрофические последствия наступают в системе с коммутацией для пакетов, подвергшихся сжатию. Поэтому потеря пакетов свыше 5 % недопустима из-за клиппирования (пропадания) фрагментов речи.

- Вариация задержки (джиттер).

Дополнительные задержки возникают при прохождении систем коммутации. В цифровой системе с коммутацией каналов (ISDN) задержка возникает в каждом узле коммутации из-за перестановки входного временного интервала на более позднее значение на выходе узла и имеет постоянное значение (не более 125 мкс на ступень искания) на все время установления соединения. Суммарная постоянная задержка («из конца в конец») даже при большом количестве транзитных узлов легко укладывается в требуемую норму *tsum* = 150 мс.

В системе коммутации пакетов коммутацию выполняют маршрутизаторы, которые коммутируют поступающие пакеты в порядке их поступления и с учетом приоритета. Случайный процесс поступления пакетов от разных источников приводит к тому, что даже регулярный порядок поступающих на вход пакетов от одного источника на выходе маршрутизатора нарушается. Это приводит к тому, что в отличие от системы с коммутацией каналов в системе с коммутацией пакетов возникают дополнительные задержки переменной величины, которые компенсируются дополнительным буфером на приеме, емкость которого зависит от джиттера.

Для оценки передачи качества восприятия голосового сигнала Рекомендация ITU-T P.800 [10] использует шкалу средней экспертной оценки (Mean Opinion Scores, MOS), приведенную в табл. 1.2.

Таблица 1.2

Шкала MOS

MOS-оценка	Описание качества
5	Междугородное качество, соответствующее качеству передачи по медным физическим проводам
4	Качество, близкое к междугородному качеству с почти незаметными искажениями
3	Ощутимые искажения, приводящие к повторению части сказанного
2	Искажения весьма раздражают
1	Невозможно вести разговор

Для компенсации переменной сетевой задержки приемники пакетов хоста всегда оснащаются сглаживающим буфером, аналогично эластичной памяти в узлах ЦСИО. Поступающие пакеты записываются в момент их приема в очередную ячейку памяти буфера, а считывание пакетов происходит с некоторой постоянной задержкой *thj* (со смещением времени воспроизведения), определяемой емкостью буфера *hbuf*.

В системе с коммутацией каналов решение об обслуживании принимается сетью связи: если в момент поступления заявки ресурсов сети недостаточно, то заявка на соединение отвергается. В противном случае заявка принимается и всегда обеспечивается гарантированное качество обслуживания в соответствии с Рекомендацией ITU-T Q.543 [8], которая нормирует скорость установления соединения и долю отказов в обслуживании на каждом проходном сетевом элементе.

В системе с коммутацией пакетов заявка на обслуживание сетью принимается всегда. Опционально сеть связи может только уведомить пользователя о гарантии качества обслуживания, решение же о продолжении обслуживания принимается пользователем. Для понимания возникновения дефицита ресурсов сети рассмотрим типовую схему обслуживания услуг сетевым элементом сети коммутации пакетов – маршрутизатором (рис. 1).

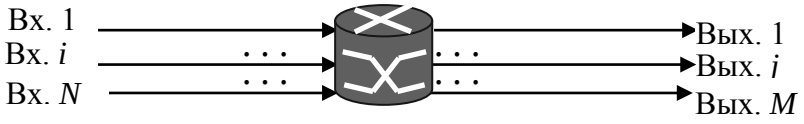


Рис. 1. Схема обслуживания вызовов маршрутизатором

Сеть с коммутацией пакетов предназначена для обслуживания различных типов услуг (видео, речь, данные). Для реализации каждой услуги

необходимо выделить определенный сетевой ресурс – производительность управляющих устройств, полосу передачи канала связи и т. д. Каждая услуга характеризуется своей полосой передачи *bw* и длиной пакета *LEN*. Услуга порождает поток пакетов (вызовов). На каждый из *N* входов (интерфейсов) сетевого маршрутизатора поступает множество пакетов, принадлежащих разным услугам. Каждый пакет может потребовать соединения с одним или несколькими из *M* выходов (интерфейсов) маршрутизатора. Если в момент поступления пакета имеется свободный требуемый выход, то пакет обслуживается немедленно, в противном случае он ставится в очередь. Дефицит канальных ресурсов возникает в моменты времени, когда несколько практически одновременно поступивших пакетов требуют соединения с одним и тем же выходом маршрутизатора, имеющего конечную скорость передачи *VEL*.

Традиционно в сети Интернет широко использовался способ распределения ресурсов «с наибольшими усилиями» (Best Effort), заключающийся в принципе «первый пришел – первый обслужен» (First In – First Out, FIFO). Такой алгоритм обслуживания хорошо работает, если приложения терпимы к задержкам пакетов из-за образования очередей или в случае, если сетевые ресурсы намного превышают потребности поступающих вызовов. Если имеются высокопроизводительное управляющее устройство и сверхскоростные каналы связи, то аппаратная задержка и время занятия канала связи для передачи пакета ничтожно и очереди в маршрутизаторе не возникают. Такая ситуация показана на рис. 2, а, когда на один выход (Вых.) поступают пакеты от трех входов (Вх. 1, Вх. 2, Вх. 3). Однако, запросы реально поступающих вызовов иногда значительно превышают возможности системы. На рис. 2, б показана та же ситуация при использовании каналов связи с малой скоростью.

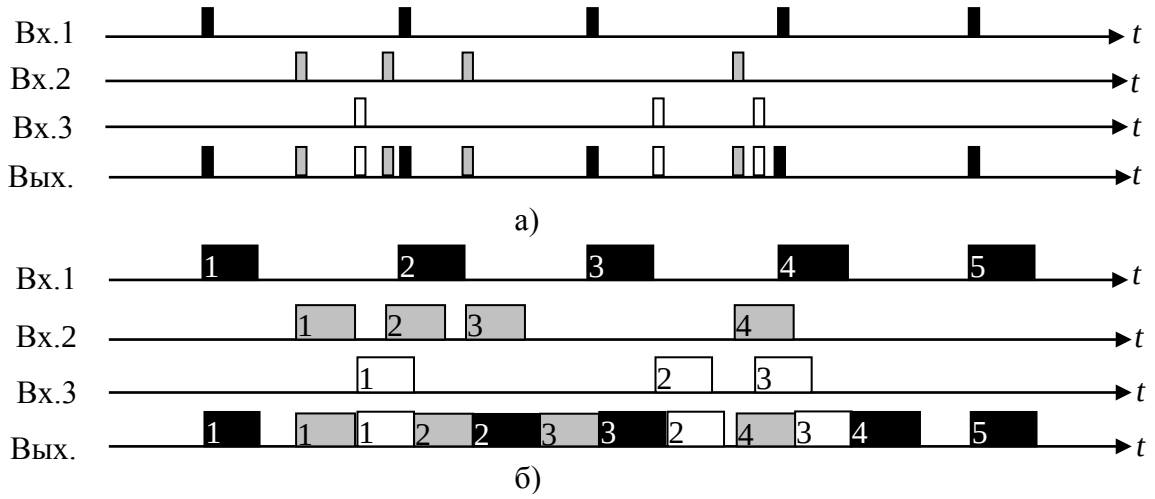


Рис. 2. Диаграмма обслуживания пакетов алгоритмом FIFO
на высокоскоростных (а) и низкоскоростных (б) каналах

Видно, что использование низкоскоростных каналов приводит к значительному джиттеру (нарушению периодичности следования) всех потоков, что совершенно неприемлемо для таких приложений, как передача речи. Следовательно, нужно создать такой порядок обслуживания потоков, чтобы приложения, чувствительные к задержкам, имели большую пропускную способность за счет снижения пропускной способности других приложений, терпимых к задержкам. В сети с коммутацией пакетов качество обслуживания (Quality of Service, QoS) достигается введением специальных механизмов: механизма дозирования и выравнивания трафика; механизма предотвращения перегрузки сети; механизма обслуживания очередей на основе приоритетов.

1.2. Дозирование и выравнивание трафика

Для устойчивой работы системы с коммутацией пакетов в каждом маршрутизаторе используются механизмы дозирования и выравнивания трафика (Traffic Shaping). Под дозированием понимается ограничение интенсивности входящего и/или исходящего потока пакетов из узла, а под выравниванием – их равномерное распределение во времени. Эти же механизмы используются в каждом пограничном маршрутизаторе для проверки соответствия интенсивности поступающих пакетов заданному профилю (согласованному значению). Дозатор и выравниватель трафика используют одну и ту же схему – «корзину маркеров» (token bucket). Схема «корзина маркеров» функционирует следующим образом (рис. 3).

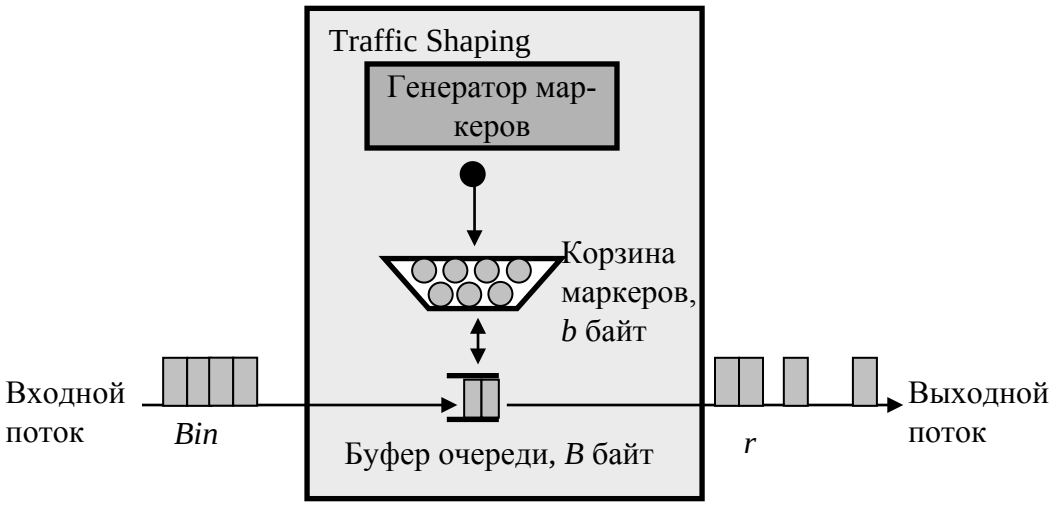


Рис. 3. Схема корзины маркеров

Каждому интерфейсу ставится в соответствие три параметра: согласованная (средняя) скорость передачи (Committed Information Rate – *r*, байт/с),

размер буфера корзины маркеров *b* (байт), размер буфера очереди *B* (байт). Их конфигурирование в маршрутизаторе задается параметрами *r, b, B*. Генератор маркеров периодически с интервалом времени $T0 = b/r$ пополняет корзину (счетчик) *b* маркерами (байтами). На вход дозатора/выравнивателя в буфер очереди емкостью *B* байт поступают пакеты со скоростью *Bin* байт в секунду. Если число маркеров в корзине равно или превышает длину первого в очереди пакета, то такой пакет передается, и соответствующее число маркеров изымается из корзины.

При недостатке маркеров в корзине пакет ожидает в буфере очереди, пока в корзине не накопится достаточное количество маркеров.

При заполнении корзины приток новых маркеров от генератора прекращается (на рис. 3 – черный маркер), при наличии в буфере очереди *B* байт вновь поступающие пакеты отбрасываются. Механизм функционирует нормально, если скорость передачи интерфейса превышает согласованную скорость передачи, т. е. при $VEL > r$.

1.3. Предотвращение перегрузки

Хосты заинтересованы передавать информацию с максимальной скоростью, которая, однако, зависит от пропускной способности сети. Поскольку к началу сеанса связи хосты ничего о ней не знают, то они пытаются определить пропускную способность сети самостоятельно, используя алгоритм медленного старта (slow start). В начале сеанса связи (рис. 4) в соответствии с алгоритмом медленного старта устанавливается размер окна, равный одному сегменту (win = 1 · MSS байт), а пороговое значение размера окна ssthresh = winr – принятому от получателя. При нормальном прохождении сегмента (получении подтверждения) передатчик каждый раз увеличивает размер окна вдвое, поэтому к моменту времени $t = 3$ текущее значение окна передатчика будет равно win = 8 · MSS байт.

Интенсивная передача пакетов приводит к увеличению очереди в проходных маршрутизаторах и снижению пропускной способности сети (пунктирная линия), что приводит к отбрасыванию девятого и последующих сегментов. Такой алгоритм отбрасывания последних пакетов при переполнении очереди получил название отбрасывания хвоста (tail drop).

Не получив в течение тайм-аута (Retransmit Timer Timeout, RTT) подтверждения о приеме девятого сегмента, TCP-источник осуществляет следующие действия:

– сбрасывает значение окна до одного сегмента (win = 1 · MSS байт);

– переустанавливает пороговое значение размера окна до величины, равной половине текущего значения (ssthresh = 8 · MSS байт/2 = 4·MSS байт),

хотя значение окна, принятое от получателя winr, может иметь гораздо большее значение;

– возобновляет передачу неподтвержденных сегментов, следуя алгоритму медленного старта и удваивая значение окна, пока оно не достигнет порогового значения (ssthresh = 4 · MSS байт). При достижении порогового значения размер окна начинает увеличиваться линейно с приращением MSS байт.

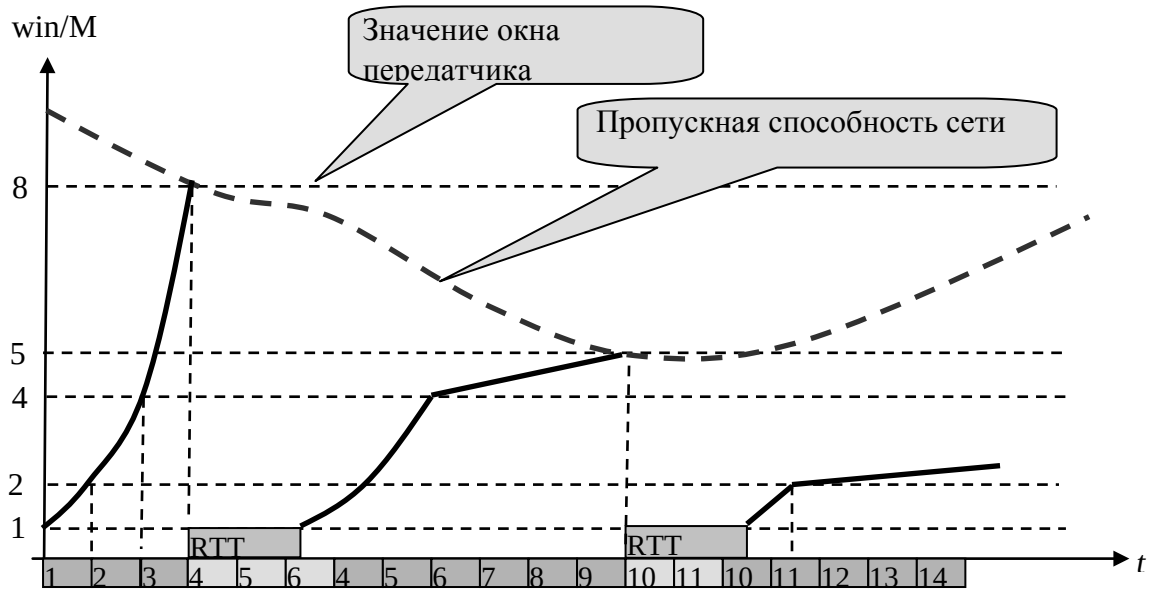


Рис. 4. Действие алгоритма медленного старта

В случае повторной перегрузки описанная процедура повторяется, но с новым значением ssthresh.

Подобная процедура ограничения трафика устраняет перегрузку в момент фактического переполнения очереди и хорошо работает в сети с небольшим количеством пользователей. Реально же в сети обычно находится большое количество TCP-соединений. Получив уведомление о перегрузке (по истечении RTT), множество TCP-источников практически одновременно перезапустят алгоритм медленного старта – сбросят значение размера окна до одного сегмента и начнут экспоненциально его увеличивать, что неизбежно скоро приведет к новой перегрузке. Резкое периодическое снижение интенсивности трафика и его быстрый рост приводят к пилообразному изменению размера очереди, получившему название глобальной синхронизации (global synchronization).

Алгоритм отбрасывания хвоста приводит к неравномерной загрузке сети и резкому снижению ее пропускной способности. С целью устранения эффекта глобальной синхронизации маршрутизаторы применяют другой алгоритм управления длиной очереди – алгоритм произвольного раннего

обнаружения перегрузки (Random Early Detection, RED). В соответствии с ним вместо ожидания фактического переполнения очереди маршрутизатор превентивно (заранее) начинает отбрасывать вновь поступающие пакеты с ненулевой вероятностью, зависящей от текущей длины очереди. Вероятностный характер отбрасывания пакетов приводит к отбрасыванию всего лишь нескольких произвольно выбранных пакетов, что является сигналом соответствующим TCP-источникам о необходимости перейти к перезапуску алгоритма медленного старта. Тем самым устраняется эффект глобальной синхронизации.

Отметим, что алгоритм RED не применим к протоколу UDP, который не чувствителен к потере пакетов.

Имеется несколько модификаций алгоритма RED, связанных с селекцией пакетов по значению поля IP-приоритета, по принадлежности к определенному потоку (Weighted Random Early Detection, WRED) и т. д.

1.4. Обслуживание очередей

Механизм обслуживания очередей в системе с коммутацией пакетов состоит из двух устройств – классификатора пакетов и планировщика очередей.

Классификатор пакетов (Packet Classifier) выполняет две функции: распознавание пакета и его маркировку (присвоение пакету определенного приоритета или класса). Он устанавливается в каждом пограничном маршрутизаторе. Функция распознавания зависит от пяти параметров – IP-адреса источника, IP-адреса назначения, порта источника, порта назначения, поля протокола IP-пакета (ICMP, TCP, UDP, …). В зависимости от принятого механизма обеспечения качества обслуживания пакет маркируется путем следующих действий:

– установки поля IP-приоритета (биты 0–2 в поле TOS заголовка IP-пакета);

– установки кода дифференцированной услуги (Differentiated Services Code Point, DSCP) взамен поля TOS в заголовке пакета;

– внесения внутренней записи о принадлежности пакета к QoS-группе.

Планировщик очередей (Packet Scheduler) сортирует маркированные пакеты, перемещая их в одну из соответствующих очередей интерфейса. Для обслуживания очередей он использует один из нижеперечисленных механизмов: взвешенного справедливого обслуживания потоков (Weighted Fair Queuing, WFQ), взвешенного справедливого обслуживания классов (Class-Based Weighted Fair Queuing, CBWFQ), с абсолютными приоритетами (Absolute Priority Queueing, APQ).

При использовании ***механизма WFQ*** планировщик очередей, используя маркировку классификатора пакетов, каждому пакету присваивает вес, значение которого используется для определения места в индивидуальной очереди на обслуживание. При использовании поля TOS вес пакета (с последующим округлением до наиближайшего целого) определяется по формуле

$$W = \frac{4096}{IP + 1},$$

где W – вес пакета; IP – значение поля IP-приоритета (3 бита) в поле TOS (8 бит). Место в очереди на обслуживание вновь поступившего пакета вычисляется как произведение веса на длину пакета с учетом наличия в очереди ранее поступивших пакетов того же потока, т. е.

$$Q_x = \begin{cases} Q_{x-1} + W \cdot LEN, & \text{при наличии в очереди пакета того же потока,} \\ Q_t + W \cdot LEN, & \text{при отсутствии в очереди пакета того же потока,} \end{cases}$$

где Q_x – место ожидания вновь поступившего пакета; Q_{x-1} – место в очереди ранее поступившего пакета того же потока; LEN – длина пакета; Q_t – текущее место обслуживания. Для подсчета места ожидания используется циклический счетчик с максимальным значением несколько Гигабайтов.

Пакеты одного потока помещаются в индивидуальную выходную очередь. За один период прохода всех очередей считывается пакет с наименьшим местом ожидания. Упрощенно механизм WFQ можно представить как одну очередь, в которой место ожидания тем меньше, чем выше приоритет и короче длина пакета.

Рассмотрим прохождение очередей при поступлении пакетов разных приоритетов (рис. 5).

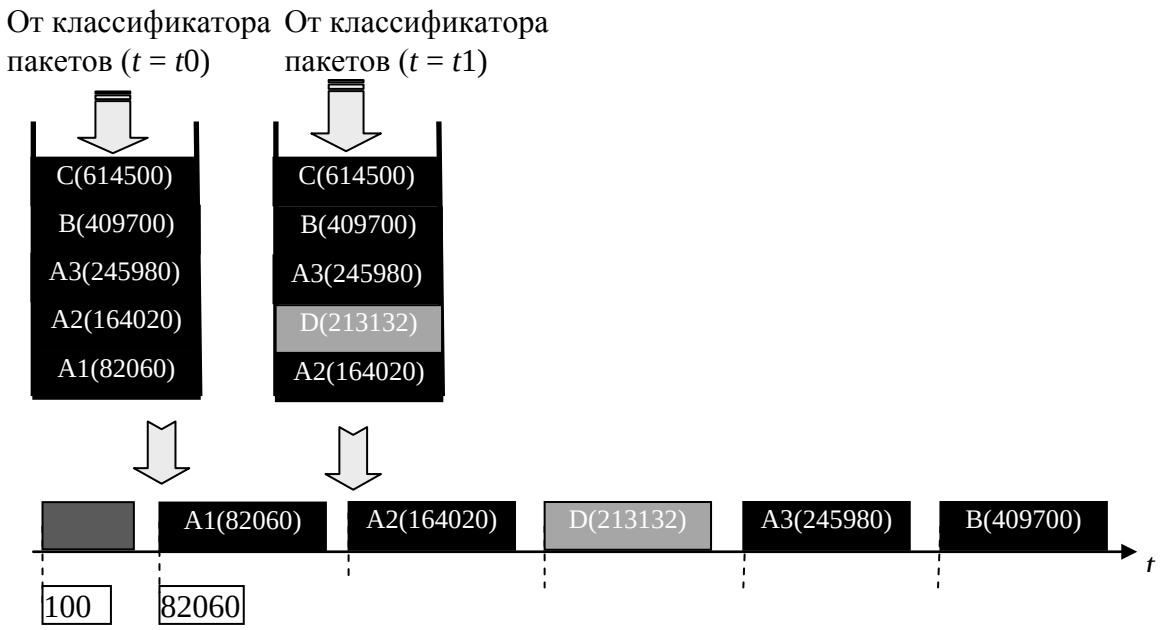


Рис. 5. Обслуживание пакетов планировщиком очередей

Предположим, что в момент времени $t0$ от классификатора пакетов в планировщик очередей поступают пакеты от трех потоков A, B и C. Текущее место обслуживания равно 100. Речевые пакеты потока A имеют приоритет 5 (вес = 683) и длину 120 байт, пакет данных потока B – приоритет 0 (вес = 4096) и длину 100 байт, пакет данных потока C – приоритет 0 (вес = 4096) и длину 150 байт. Места в очереди планировщик распределит следующим образом: Q(A1) = 100 + (683 · 120) = 82060, Q(A2) = 82060 + + (683 · 120) = 164020, Q(A3) = 164020 + (683 · 120) = 245980, Q(B) = 100 + + (4096 · 100) = 409700, Q(C) = 100 + (4096 · 150) = 614500.

В момент времени $t1$, когда в обслуживании уже находится пакет A1, от классификатора поступает новый пакет D с приоритетом 3 (вес = 1024) и длиной 128 байт. Поскольку счетчик текущего места обслуживания к этому моменту уже установлен на значение 82060, то пакет D будет поставлен на Q = 82060 + (1024 · 128) = 213132 место ожидания и будет обслужен раньше прежде поступивших пакетов A3, B и C.

Таким образом, чем меньше вес и длина пакета, тем быстрее пакет будет обслужен.

При использовании ***приоритетов на основании QoS-группы*** создается отдельная очередь, а вес пакета присваивается программой на основании какого-либо признака, не входящего в поле TOS или DSCP. Задаются общая зарезервированная полоса BW и максимальное значение полосы bwm, которая может быть выделена отдельному приложению. Обычно $bwm \le BW \le 0{,}75 \cdot VEL$. Вес пакета определяется по формуле

$$W = \frac{4 \cdot bwm}{bw},$$

где bwm – максимальное значение из всех полос, выделенных приложениям; bw – выделенная полоса передачи для конкретного приложения. Такое закрепление приоритетов используется в системе с интегрированным обслуживанием (протокол RSVP).

При использовании механизма CBWFQ все пакеты, имеющие общий признак (IP-приоритет, поле DSCP и т. д.), относятся к одному классу. Каждому классу в явном виде выделяется определенная полоса, выраженная в абсолютных или относительных единицах измерения.

В системе с дифференцированным обслуживанием поле DSCP делится на две части, старшая часть (3 бита) определяет класс приоритета обслуживания, а младшая (3 бита) – приоритет обслуживания пакетов внутри класса (механизм обслуживания с абсолютным приоритетом APQ). Поле DSCP обладает обратной совместимостью с полем TOS, т. е. каждому значению класса приоритета поля DSCP соответствует значение TOS.

Приоритет пакета определяется на основании кода DSCP и обслуживается с использованием механизма CBWFQ.

Пакеты одного класса одинаково маркируются *классификатором пакетов (Packet Classifier)*, а *планировщик очередей (Packet Scheduler)* помещает их в одну очередь. Более приоритетные пакеты помещаются в очередь с большей полосой передачи. В дальнейшем планировщик циклически обходит все очереди и считывает из каждой очереди число пакетов в соответствии с выделенной для нее полосой передачи.

Если *bwi* – абсолютная выделенная полоса для *i*-го класса, а *MTU* – максимальная длина пакета, то за один цикл из очереди *i*-го класса будет передано (считано) следующее количество пакетов объемом $qi = \frac{bwi}{\min\{bwi\}} \cdot MTU$ байт, где min*{bwi}* – минимальное значение из выделенных полос. Пусть, например, для классов A, B, C выделены полосы по 5, 60 и 35 кбит/с, соответственно, а максимальная длина пакета на интерфейсе – 620 байт. Тогда за один цикл из очереди класса A будет считано (5/5) · 620 = 620 байт, из очереди класса B – (60/5) · 620 = 7440 байт, а из очереди класса C – (35/5) · 620 = 4340 байт.

Если *bwi%* – доля выделенной полосы для *i*-го класса, а *MTU* – максимальная длина пакета, то за один цикл из очереди *i*-го класса будет передано пакетов объемом $qi = \frac{bwi\%}{\min\{bwi\%\}} \cdot MTU$ байт, где min*{bwi%}* – минимальное относительное значение из выделенных полос. Пусть, например, для классов A, B, C выделены полосы по 5, 60 и 35 % кбит/с, соответственно, а максимальная длина пакета на интерфейсе – 620 байт. Тогда за один цикл из очереди класса A будет считано (5 %/5 %) · 620 = 620 байт, из очереди класса B – (60 %/5 %) · 620 = 7440 байт, а из очереди класса C – (35 %/5 %) · 620 = 4340 байт.

Таким образом, чем выше приоритет, тем больше выделенная полоса передачи и, следовательно, тем быстрее пакет будет обслужен.

Механизм APQ реализует обслуживание с абсолютным приоритетом. Сетевой администратор определяет приоритетность обслуживания пакета созданием четырех классов: high, medium, normal, low. *Классификатор* маркирует пакеты, основываясь либо на основе значения IP-приоритета (поле TOS = 5 – high, …, TOS = 0 – low), либо на основе IP-адреса источника, либо на основе значения длины пакета.

Планировщик очередей (Packet Scheduler) помещает маркированные пакеты в соответствующую очередь и всегда начинает обслуживание очереди с класса high. Только после того, как все пакеты этого класса будут

обслужены, начнется обслуживание класса medium. Соответственно, передача класса low окажется возможной только после того, как будут обслужены все более приоритетные классы. Вполне возможно, что в моменты перегрузки сети практически вся скорость передачи интерфейса будет занята высокоприоритетными пакетами, а передача пакетов класса low будет происходить с большими задержками.

Таким образом, чем выше приоритет, тем выше класс обслуживания и, следовательно, тем быстрее пакет будет обслужен.

ЗАДАЧА 1. Дайте определение качества обслуживания.

Ответ. Совокупность требований к сети связи для обеспечения доставки информации пользователя «из конца в конец» с заданными характеристиками.

ЗАДАЧА 2. Всегда ли для установленного соединения в системе с коммутацией каналов обеспечивается нормированная доставка информации?

Ответ. Да.

ЗАДАЧА 3. Всегда ли для установленного соединения в системе с коммутацией пакетов обеспечивается нормированная доставка информации?

Ответ. Нет.

ЗАДАЧА 4. Дайте определение джиттера сети.

Ответ. Джиттер – это стандартное отклонение задержки от среднего значения при доставке пакета «из конца в конец».

ЗАДАЧА 5. Какова цель процедуры медленного старта?

Ответ. Выяснение хостом максимальной пропускной способности сети.

ЗАДАЧА 6. Поясните функции классификатора пакетов и планировщика очередей.

Ответ. Классификатор пакетов назначает приоритет обслуживания, а планировщик очередей реализует приоритетное обслуживание.

ЗАДАЧА 7. На занятый выходной интерфейс поступает пакет с полем TOS = 4 длиной 512 байт. Очередь на обслуживание отсутствует. На какое место ожидания будет поставлен пакет при использовании механизма WFQ?

Ответ. Поступивший пакет будет иметь вес $W = 819$ и будет поставлен на $Qx = 0 + 819 \cdot 512 = 419328$ место ожидания.

ЗАДАЧА 8. На свободный выходной интерфейс поступает пакет с полем TOS = 4 длиной 512 байт. Очередь на обслуживание отсутствует. На какое место ожидания будет поставлен пакет при использовании механизма WFQ?

Ответ. Пакет будет обслужен немедленно.

ЗАДАЧА 9. Почему более короткие пакеты в маршрутизаторе имеют преимущество в обслуживании?

Решение.

Пусть имеется два пакета. Первый пакет требует для обслуживания 4 с, а второй – 20 с. Предположим, что первым обслуживается длинный пакет. Тогда первый пакет будет обслужен за 20 с, а второй – за (20 + 4) с. Среднее время обслуживания равно 22 с. Если же первым обслуживается короткий пакет, то первый пакет будет обслужен за 4 с, а второй – за (4 + 20) с. Среднее время обслуживания стало равно 14 с.

Следовательно, для снижения среднего времени нахождения вызова в системе короткие пакеты должны обслуживаться первыми.

2. ПОТОКИ ВЫЗОВОВ, ВРЕМЯ ОБСЛУЖИВАНИЯ

Детерминированный поток вызовов – последовательность вызовов, в которой вызовы поступают в определенные, строго фиксированные неслучайные моменты времени.

Детерминированный поток вызовов может задаваться:

– в явном виде моментами поступления вызовов t_n, (n = 1, 2, ...);

– в виде рекуррентной последовательности $t_n = f(t_{n-1})$;

– в виде последовательности промежутков между вызовами z_n.

Случайный поток вызовов – последовательность вызовов, в которой моменты поступления вызовов и промежутки между вызовами являются случайными величинами.

Случайная величина описывается функцией распределения $F(t) = P(x < t)$, определяющей вероятность того, что случайная величина x расположена в промежутке $0 \le x < t$.

Однородный поток вызовов характеризуется только законом поступления вызовов.

В ***неоднородном*** потоке вызовов каждый вызов характеризуется двумя и более характеристиками, например, моментом поступления вызова и его направлением.

Рекуррентным потоком вызовов называется поток, у которого промежутки времени между соседними вызовами независимы друг от друга и распределены по одинаковому закону.

Поток вызовов является ***стационарным***, если число вызовов, поступивших за промежуток времени Dt, зависит только от длительности этого промежутка времени Dt и не зависит от нахождения этого промежутка на оси времени t.

Из теории вероятностей известно соотношение

$$P(A \cap B) = P(B) \cdot P(A / B),$$

откуда

$$P(A / B) = \frac{P(A \cap B)}{P(B)}.$$

Если произошло событие, подчиняющееся экспоненциальному распределению, уже наступило и длится время t_0, то вероятность его дальнейшего течения за время t

$$P\{\ge t \ge t_0\} = \frac{P\{\ge t\} \cdot P\{\ge t_0\}}{P\{\ge t_0\}} = P\{\ge t\} = e^{-\lambda t}$$

не зависит от t_0, т. е. от прошлого. Поток вызовов является ***потоком без последействия***, если вероятность поступления вызовов за некоторый промежуток времени t не зависит от процесса поступления вызовов до этого промежутка.

Поток вызовов является ***ординарным***, если вероятность поступления двух и более вызовов за бесконечно малый промежуток времени много меньше, чем вероятность поступления одного вызова.

В теории телетрафика часто используется простейший поток вызовов. ***Простейшим*** потоком вызовов называется ординарный стационарный поток без последействия. Простейший поток вызовов задается вероятностью поступления k вызовов за время t (формулой Пуассона)

$$\pi_k(t) = \frac{(\lambda \cdot t)^k}{k!} e^{-\lambda t},$$

где параметр λ – интенсивность поступающего потока вызовов, численно равен среднему количеству вызовов, поступающих за единицу времени. Вероятность $\pi_k(t)$ имеет максимум при $\lambda \cdot t = k$.

Функция распределения промежутков времени между вызовами – вероятность поступления хотя бы одного вызова за время $t > 0$

$$F(t) = P(x < t) = 1 - \pi_0(t) = 1 - e^{-\lambda t}.$$

Математическое ожидание или, то же самое, среднее время промежутка между вызовами

$$M[t] = \int_0^\infty t d\left(1 - e^{-\lambda t}\right) = -\int_0^\infty t de^{-\lambda t} = \frac{1}{\lambda},$$

дисперсия и среднеквадратическое отклонение

$$D[t] = \int_0^\infty \left(t - \frac{1}{\lambda}\right)^2 d\left(1 - e^{-\lambda t}\right) = \frac{1}{\lambda^2}, \quad \delta = \sqrt{D[t]} = \lambda^{-1}.$$

Симметричным потоком называется поток с простым последействием, параметр которого в любой момент времени зависит только от числа обслуживаемых в этот момент времени вызовов.

Примитивным называется такой симметричный поток, параметр которого прямо пропорционален числу свободных в данный момент источников $\lambda_i = (n - i)\alpha$, где n – общее число источников вызовов, α – параметр потока источника в свободном состоянии.

При экспоненциальном времени обслуживания функция распределения времени занятия системы коммутации имеет вид

$$H(t)=\begin{cases}1-e^{-\mu t}, & \text{если } \lambda,\mu,t>0,\\ 1, & \text{если } \lambda=0,\end{cases}$$

где параметр μ – интенсивность обслуживания вызова, численно равен среднему количеству обслуживаемых вызовов за единицу времени; λ – параметр потока поступающих вызовов. Математическое ожидание (оно же – среднее время обслуживания), дисперсия и среднеквадратическое отклонение

$$M[t]=\int_0^\infty t d\left(1-e^{-\mu t}\right)=\frac{1}{\mu}=t_s,\quad D[t]=\mu^{-2},\quad \delta_s=\sqrt{D[t]}=\mu^{-1}=t_s.$$

При постоянном времени обслуживания функция распределения времени занятия системы коммутации имеет вид

$$H(t)=\begin{cases}0, & \text{если } t<t_s \text{ или } \lambda=0,\\ 1, & \text{если } t\ge t_s, \text{ и } \lambda>0,\end{cases}$$

где t_s – среднее время обслуживания. Математическое ожидание, дисперсия и среднеквадратическое отклонение

$$M[t]=\int_0^\infty t d(H(t))=\frac{1}{\mu}=t_s,\qquad D[t]=0,\qquad \delta_s=0.$$

ЗАДАЧА 1. Доказать ординарность пуассоновского потока вызовов.

Решение.

Очевидно, что при фиксированном значении λ и $t \to 0$ произведение $\lambda t << 1$. Поэтому

$$\lim_{t\to 0}\pi_1(t)=\frac{(\lambda\cdot t)^1}{1!}e^{-\lambda t}>>\pi_2(t)=\frac{(\lambda\cdot t)^2}{2!}e^{-\lambda t}>>\pi_3(t)=\frac{(\lambda\cdot t)^3}{3!}e^{-\lambda t}>>\ldots$$

что и доказывает ординарность потока.

ЗАДАЧА 2. Определить вероятность поступления пяти вызовов за 10 мин при поступлении простейшего потока вызовов с параметром λ = 180 ч$^{-1}$.

Решение:

$$p_5\left(\tfrac{1}{6}\right)=\frac{\left(180\cdot \tfrac{1}{6}\right)^5}{5!}e^{-30}=\frac{30^5}{5!}e^{-30}=1{,}89\cdot 10^{-8}.$$

3. ПРИНЦИПЫ ПОСТРОЕНИЯ СИСТЕМ КОММУТАЦИИ

3.1. Понятие о нагрузке

Нагрузка есть суммарное время обслуживания вызовов за фиксированное время t. Единица измерения нагрузки – часо-занятие.

Нагрузка обладает аддитивным свойством:

$$Y(t, t + \tau) = Y(t) + Y(t + \tau).$$

Интенсивность нагрузки – количество вызовов, обслуженных в течение часа. Единица измерения – часо-занятие/час =***Эрланг*** (Эрл). Интенсивность нагрузки A численно равна числу вызовов λ, поступивших за время занятия ts, т. е. $A = \lambda \cdot ts$. В теории телетрафика для краткости интенсивность нагрузки называют нагрузкой.

В общем случае система коммутации (рис. 6) состоит из коммутационного поля (КП) и управляющего устройства (УУ).

Вызовы, поступающие в систему, характеризуются:

- параметром поступающего потока вызовов λ_0, численно равным числу вызовов, поступающих в единицу времени;
- средней длительностью занятия t_s, зависящей от скорости прохождения вызова через КП.

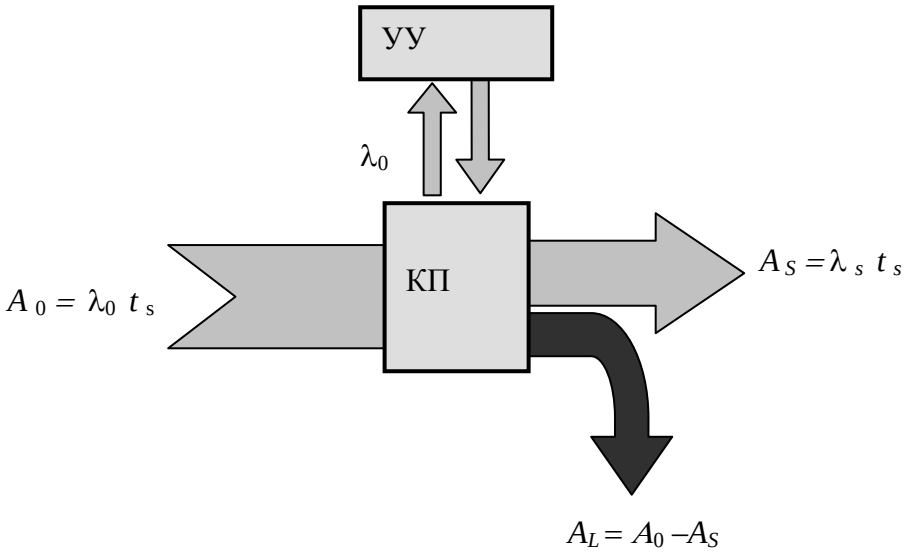


Рис. 6. Система коммутации

Каждый поступающий вызов содержит информацию пользователя и служебную информацию. Служебная информация содержит адрес пункта назначения для доставки информации пользователя и используется для

маршрутизации вызова в сети связи. Она обрабатывается УУ, темп ее поступления совпадает с параметром поступающего потока вызовов λ_0. УУ располагает всеми данными о ресурсах системы и, если необходимые ресурсы для установления соединения имеются, то вызов принимается к обслуживанию. Время обработки вызова управляющим устройством образует аппаратную задержку, которая не зависит от скорости прохождения вызова через КП. Принятые к обслуживанию вызовы обрабатываются (коммутируются) в КП и образуют обслуженную нагрузку $A_s = \lambda_s t_s$.

В отдельные моменты времени число одновременно поступивших вызовов может превышать ресурсы системы, что приводит к блокировке некоторых вызовов. Блокировка – событие, состоящее в невозможности обслуживания вызова в момент его поступления из-за дефицита ресурсов, например, линий связи или скорости передачи. Блокированные вызовы, образующие блокированную нагрузку A_L, обслуживаются различно в зависимости от типа системы коммутации. Они могут сразу же удаляться из системы (система с явными потерями), ставиться в очередь на обслуживание (система с ожиданием) или обслуживаться иным образом.

Как и поток вызовов, нагрузка подвержена резким колебаниям в течение суток, по временам года и т. д. Например, наблюдения за величиной нагрузки показывают, что интенсивность нагрузки значительно изменяется в течение суток. Оператор сети заинтересован в максимизации отношения «прибыль/затраты», что равнозначно отношению «обслуженная нагрузка/объем (цена) коммутационного оборудования». Однако составляющие этого отношения имеют одинаковые тенденции к увеличению. Поэтому для определения оптимального объема коммутационного оборудования необходимо правильно выбрать диапазон времени измерения статистических данных нагрузки.

На рис. 7 показано типовое распределение нагрузки A в течение суток. Жирная пунктирная линия условно определяет объем коммутационного оборудования V, которое способно обслужить поступающую нагрузку A немедленно. Затемненная область нагрузки, расположенная выше этой линии, является блокированной, которая в зависимости от системы обслуживания либо удаляется из системы (система с явными потерями), либо, например, будет поставлена в очередь на обслуживание (система с ожиданием).

В первом случае пользователь немедленно получает уведомление о невозможности обслуживания (акустический сигнал «Занято»).

Во втором случае пользователь получает обслуживание с ухудшенным качеством, которое при пакетной коммутации ощущает в виде задержек фрагментов информации (пакетов данных или речи).

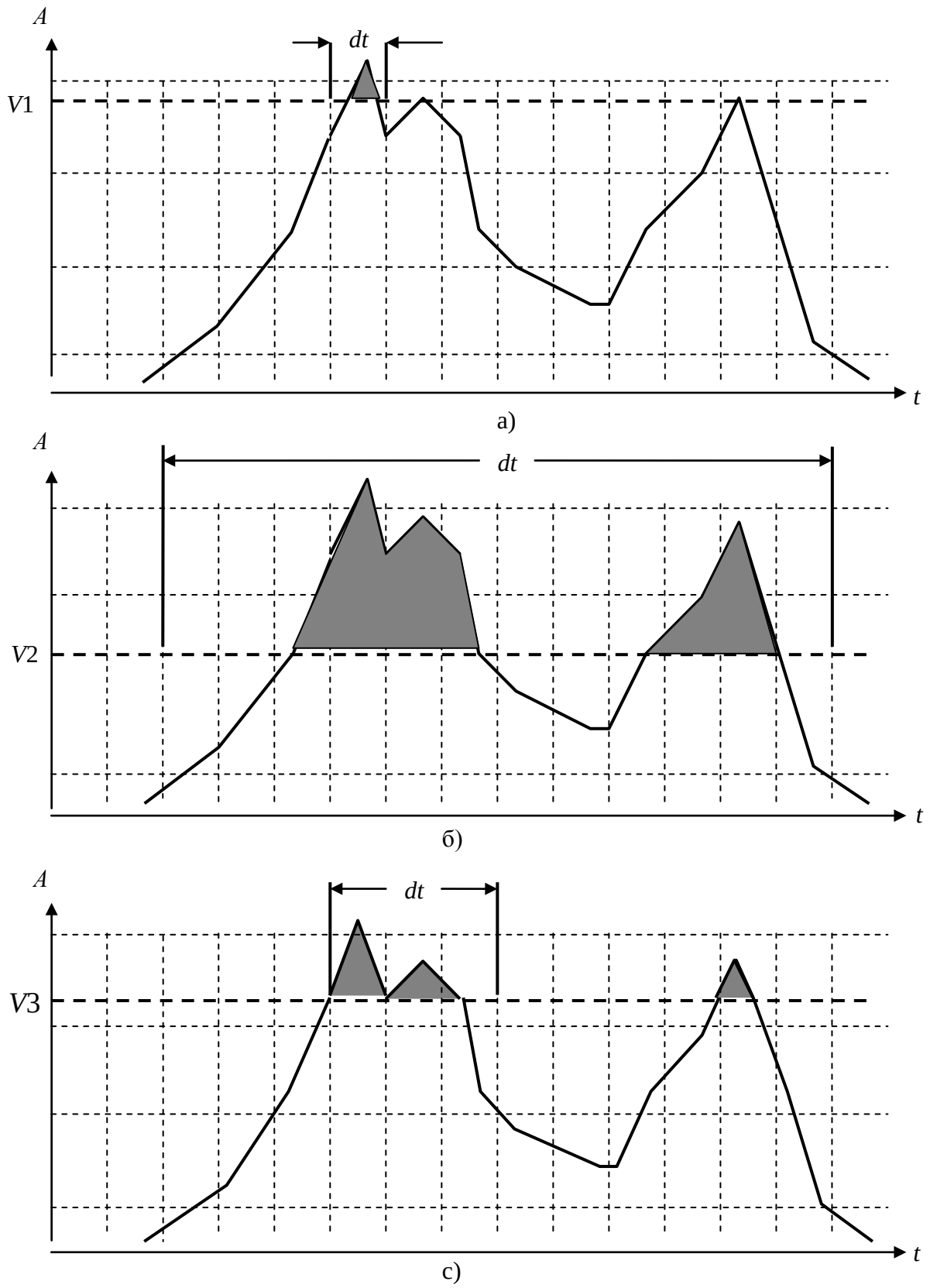


Рис. 7. Сравнение различных критериев выбора диапазонов измерения максимального среднего значения нагрузки:
а) малого; б) большого; с) оптимального

Рис. 7, а представляет случай, когда объем оборудования $V1$ выбирается в соответствии с максимальным средним значением измеренной нагрузки среди коротких диапазонов времени dt. Высокое качество обслуживания пользователей достигается за счет очень большого объема коммутационного оборудования $V1$, которое в течение суток используется крайне неэффективно за исключением короткого промежутка времени dt, что крайне невыгодно операторам сетей связи.

Рис. 7, б представляет случай, когда объем оборудования $V2$ выбирается в соответствии с максимальным средним значением измеренной нагрузки из длинного диапазона времени dt. Низкое качество обслуживания пользователей образуется за счет небольшого объема коммутационного оборудования $V2$, которое в течение промежутка времени dt используется достаточно интенсивно.

Рис. 7, с представляет компромиссный случай, когда объем оборудования $V3$ выбирается в соответствии с максимальным средним значением измеренной нагрузки среди небольших (оптимальных) диапазонов времени dt. При этом получается приемлемое качество обслуживания для пользователей при приемлемых затратах на коммутационное оборудование для операторов сети связи.

Считается, что в течение еще многих лет телефонная нагрузка в мультимедийных сетях будет существенна. В ходе продолжительных наблюдений за статическими данными телефонной нагрузки был сделан вывод о том, что наилучшие (оптимальные) статистические данные дают результаты измерения нагрузки за интервалы времени dt = 1 час. Промежуток времени в 1 час, где среднее значение нагрузки имеет максимальное значение, носит название ***часа наибольшей нагрузки*** (ЧНН, BHCA – Busy Hour Call Attempts). Наибольшая телефонная нагрузка приходится на периоды между 11–12 и 19.30–20.30 часами, в эти временные промежутки обслуживается 10–12 % суточной нагрузки. При расчете оборудования следует ориентироваться на статистические данные этого часа, когда средняя нагрузка имеет наибольшее значение. Мотив такого подхода прост: если уж в течение ЧНН обеспечивается требуемое качество обслуживания, то в любой другой промежуток времени (час) высокое качество обслуживания будет обеспечиваться тем более.

Согласно теории вероятностей средняя нагрузка, обслуженная за время t, равна ее математическому ожиданию

$$M(t) = \int_0^t M_i(\tau)d\tau.$$

В стационарном режиме среднее число занятых линий $M_i(\tau) = V$ и не зависит от времени. Поэтому $M(t) = Vt$. Принимая t = 1 час, получаем,

что средняя обслуженная нагрузка за единицу времени (час) численно равна среднему числу занятых линий за то же время.

Статистические данные параметров телефонной нагрузки таковы.

Среднее число вызовов с одного телефонного аппарата λ (в час): квартирного – 2; уличного (монетного автомата) – 15; учрежденческого – 5. Средняя длительность разговора *ts* (в секундах): квартирного – 200, уличного – 120, учрежденческого – 150.

Масштаб параметров нагрузки в системах коммутации каналов и пакетов существенно различается.

В системе коммутации каналов время разговора фактически является временем удержания канала связи, в течение которого параметр потока вызовов равен нулю. Если пользователь осуществляет 1 вызов в час со средней длительностью разговора *ts* = 3 мин, то параметр потока вызовов в ЧНН $\lambda = 1$ ч$^{-1}$.

В системе коммутации пакетов поток носит взрывной характер. Например, при использовании кодека G.711 со временем формирования пакета *tcod* = 20 мс во время одного разговора *ts* = 3 мин параметр потока вызовов $\lambda = 1/tcod = 50$ с$^{-1}$, но параметр сетевого потока в ЧНН всего лишь $\lambda = 1 \cdot ts/tcod = 1 \cdot 3 \cdot 60/(0{,}02 \cdot 3600)$ с$^{-1}$ = 2,5 с$^{-1}$.

3.2. Типы систем коммутации

В зависимости от типа системы коммутации, блокированные вызовы обслуживаются различно. Они могут сразу же удаляться из системы (система с явными потерями), ставиться в очередь на обслуживание (система с ожиданием) или обслуживаться иным образом.

В системе с явными потерями (Blocked Calls Cleared, BCC) блокированные вызовы надолго покидают систему. Параметр поступающей нагрузки A_0 определяется средним значением в час наибольшей нагрузки. Поэтому при фиксированных ресурсах сети V (например, числе каналов связи) и числе пользователей $N > V$, даже если $V > A_0$, в некоторые промежутки времени не все поступающие вызовы принимаются к обслуживанию и возникают потери. Типовая зависимость обслуженной нагрузки A_S от поступающей нагрузки A_0 показана на рис. 8.

При увеличении поступающей нагрузки A_0 экспоненциально увеличивается и обслуженная нагрузка A_S, однако ее значение не может превысить числа каналов связи V. Разность между поступающей и обслуженной нагрузкой образует потерянную нагрузку $A_L = A_0 - A_S$. Потери будут отсутствовать только при наличии числа каналов, равному числу абонентов, т. е. при $N = V$, при этом обслуженная нагрузка будет равна поступающей.

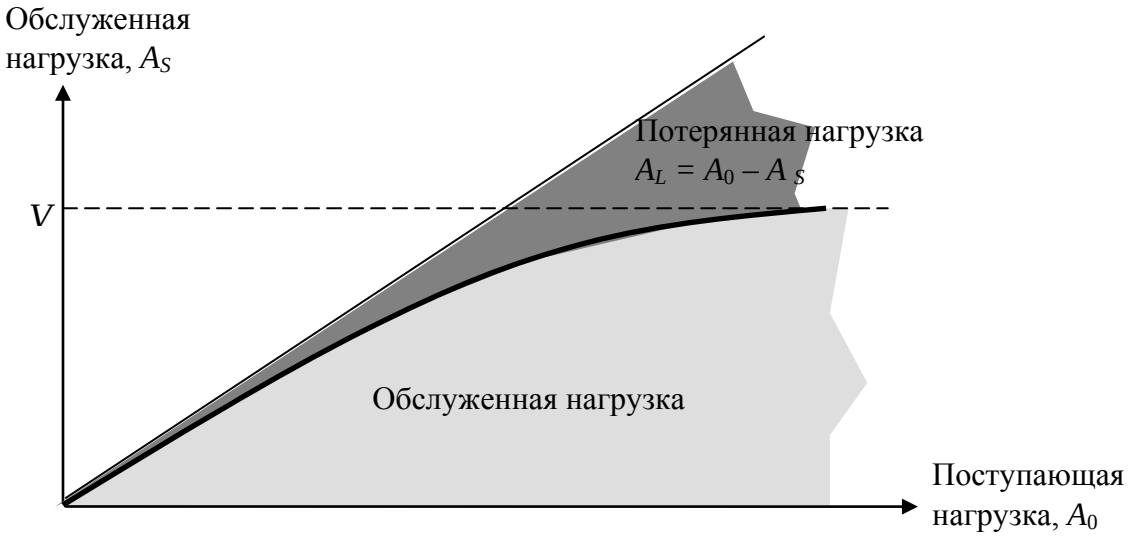


Рис. 8. Пропускная способность системы с явными потерями

В реальных системах с явными потерями всегда $N > V$. Следствием соотношения $N > V$ является потеря некоторых вызовов, поэтому параметр обслуженного потока вызовов λ_s всегда меньше параметра поступающего потока λ_0. Поступающая, обслуженная и потерянная нагрузки связаны соотношением $A_S = A_0(1 - P) = A_0 - A_L$,, где P – вероятность потерь. Возможный сценарий обслуживания вызовов системой с явными потерями в стационарном режиме при A_0 = 3 Эрл, V = 4, As = 2 Эрл показан на рис. 9 (серым цветом помечены обслуженные вызовы, черным цветом – потерянные, среднее время занятия – 5 мин).

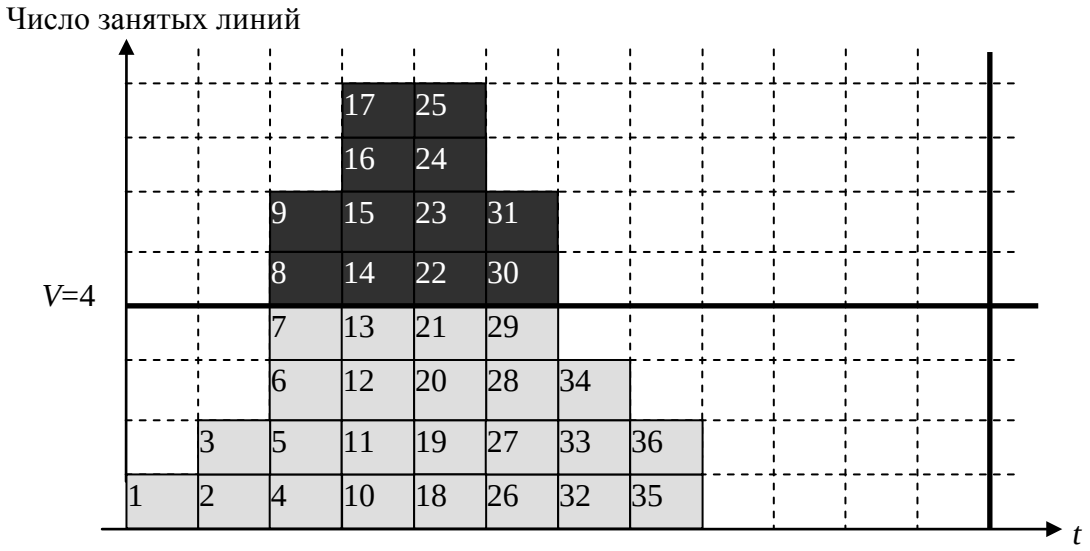


Рис. 9. Диаграмма работы системы с явными потерями
при A_0 = 3 Эрл; V = 4; As = 2 Эрл

Пропускная способность систем с явными потерями оценивается нагрузкой, обслуженной немедленно с заданным качеством (потерями). На рис. 10 показано, что при $A_0 < A^*$ большую пропускную способность имеет система 2 (имеет меньше потери), а при $A_0 \geq A^*$ – система 1.

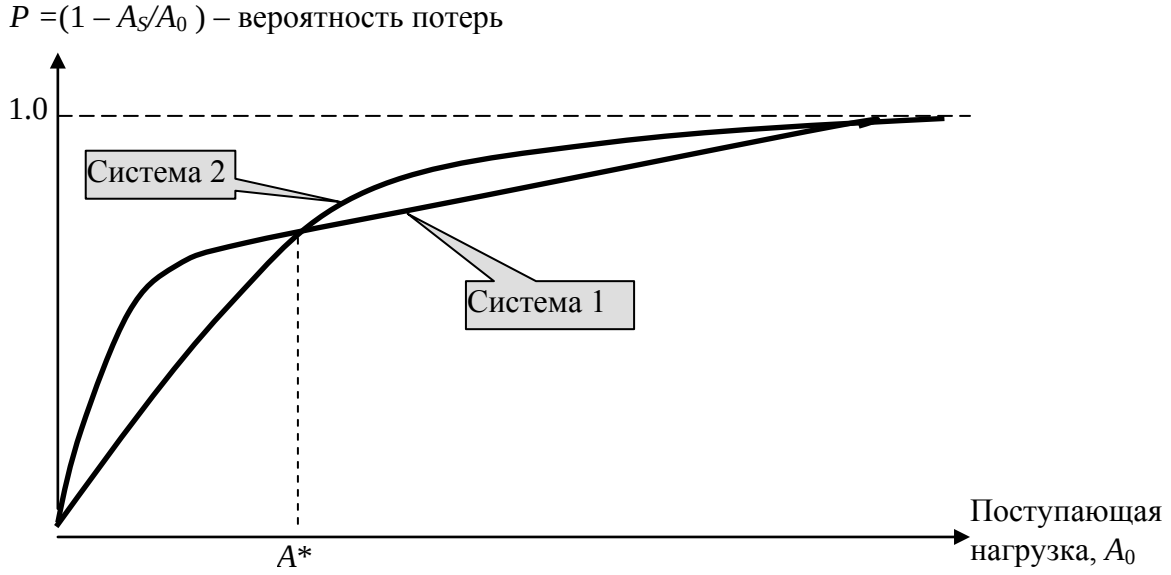


Рис. 10. Сравнение пропускной способности систем с явными потерями

В системе с явными потерями различают потери по времени, вызовам и нагрузке. Потери по времени численно равны доле времени *dt,* в течение которого появляются блокировки, к общему времени наблюдения *T* (обычно *T* = 1 ч)*:* $p_t = dt/T$. Потери по вызовам численно равны доле потерянных вызовов $c(dt)$ за время *dt* к числу поступивших вызовов за общее время наблюдения *T:* $p_c = c(dt)/c(T)$. Потери по нагрузке численно равны отношению потерянной нагрузки к поступившей: $p_y = y(dt)/y(T)$.

Общие потери в системе с явными потерями образуют две составляющие: потери из-за внутренних блокировок и потери в пучке каналов. Потери из-за внутренних блокировок возникают из-за недостаточности ресурсов коммутационного оборудования, а потери в пучке – из-за недостаточности ресурсов линейного оборудования (каналов связи). Внутренней блокировкой в многозвенных КП называется ситуация, когда соединение некоторых свободных входов-выходов (интерфейсов) невозможно из-за неблагоприятного сочетания свободных выходов и свободных промежуточных линий в КП.

Детали образования внутренних блокировок в системе с явными потерями подробно рассмотрены в [2, 3]. Необходимый минимальный объем ресурсов коммутационного оборудования для устранения внутренних

блокировок определил Ч. Клоз (C. A. Clos), который показал, что их устранение достигается удвоением числа промежуточных линий в КП по сравнению с числом интерфейсов.

В системе с ожиданием (Blocked Calls Queued, BCQ) блокированные вызовы помещаются в бесконечную выходную очередь и обслуживаются в порядке поступления. Следствием блокировки является замедление обслуживания вызовов. В стационарном режиме, если $A_0 < V$, параметр обслуженного потока вызовов λ_S всегда равен параметру поступающего потока λ_0.

Возможный сценарий обслуживания вызовов системой с ожиданием в стационарном режиме показан на рис. 11 (серый цвет – вновь поступившие вызовы, темно-серый – вызовы, обслуженные из очереди).

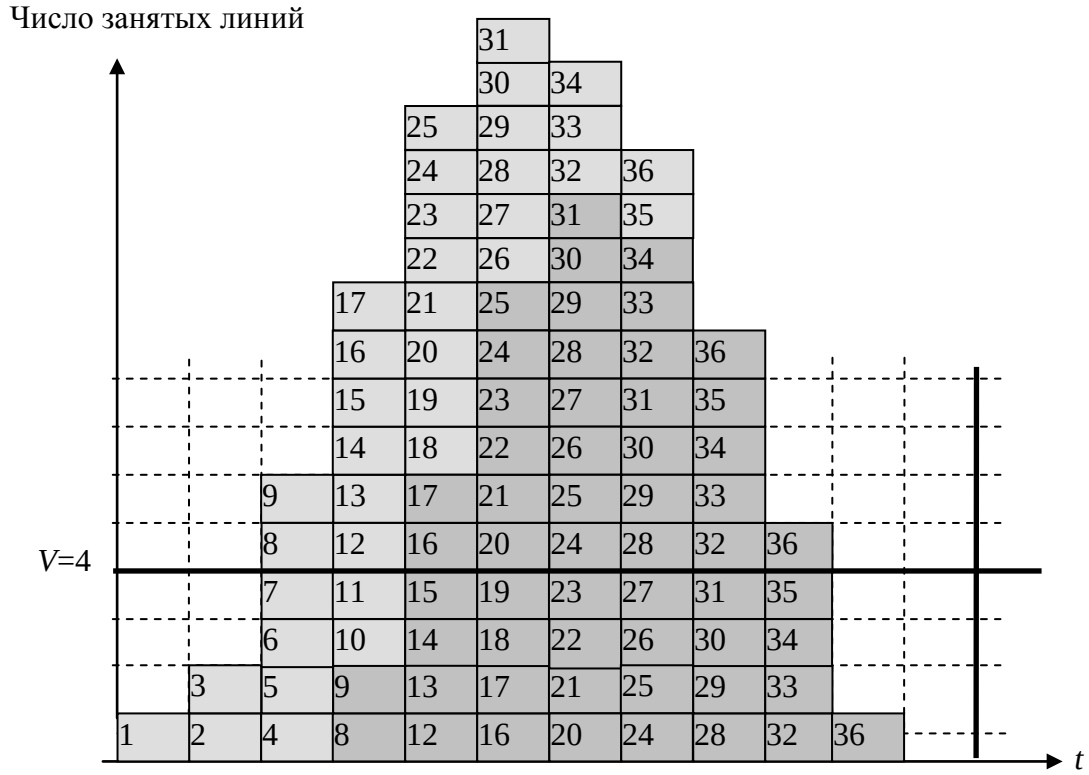


Рис. 11. Диаграмма работы системы с ожиданием при A_0 = 3 Эрл; V = 4; As = 3 Эрл

Пропускная способность систем с ожиданием оценивается нагрузкой, доставленной сетью «из конца в конец» с заданной вероятностью за нормируемое время t_e. В качестве оценки используется функция распределения $F(t_e)$, обычно нормируемая на уровне 0,95. Последнее означает, что не менее 95 % поступивших от источника пакетов должны быть доставлены получателю за время, не превышающее t_e. Для доставки информации в реальном масштабе времени $F(t_e)$ нормируется на уровне 0,999.

На рис. 12 показано, что при одинаковой нагрузке система 1 (сеть 1) удовлетворяет заданной пропускной способности (более 95 % вызовов обслуживаются за время, не превышающее t_e), а система 2 – нет. Хотя среднее время доставки пакета системой 1 может быть и больше, чем среднее время доставки системой 2.

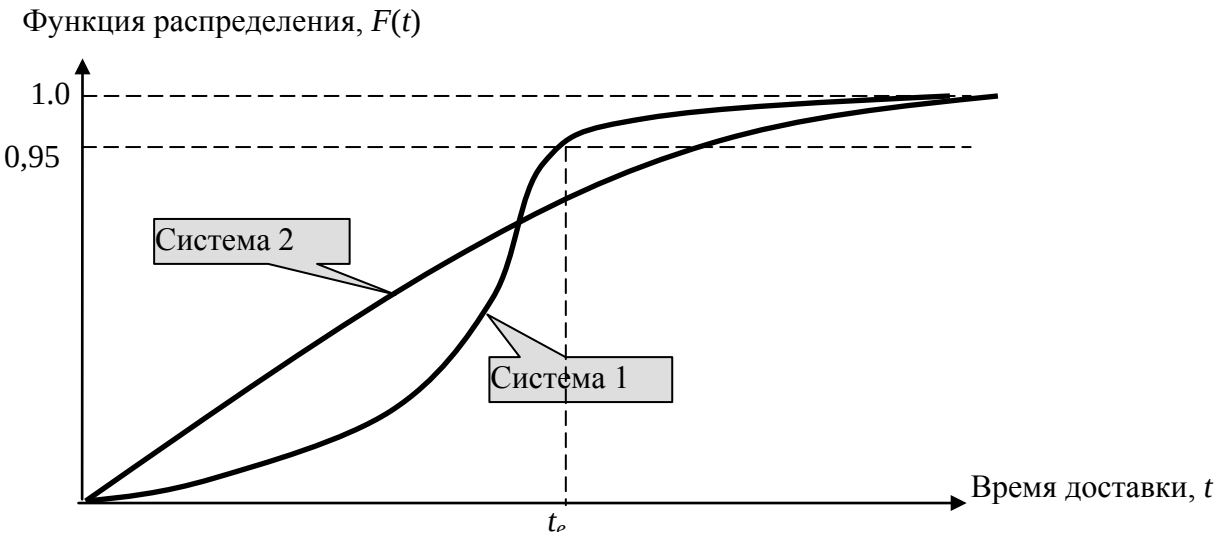


Рис. 12. Сравнение пропускной способности систем с ожиданием

Общую задержку обслуживания пакета в системе с ожиданием образуют две составляющие: задержка из-за внутренних блокировок и задержка в пучках каналов. Задержка из-за внутренних блокировок возникает из-за недостаточности ресурсов коммутационного оборудования, а задержка в пучке – из-за недостаточности ресурсов линейного оборудования (каналов связи).

Задержка из-за внутренней блокировки в КП возникает в случае, когда суммарная интенсивность (скорость) поступающих пакетов превышает интенсивность (скорость) обслуживания пакетов внутри коммутатора. Интенсивность обслуживания пакетов внутри коммутатора определяется производительностью процессора обработки пакетов и скоростью внутренней шины коммутатора. Внутренняя блокировка приводит к накоплению пакетов во входном буфере интерфейса; при отсутствии внутренних блокировок во входном буфере не может быть более одного пакета.

Задержка в пучке возникает в случае, когда суммарная интенсивность (скорость) поступающих пакетов на выходной интерфейс превышает интенсивность (скорость) обслуживания пакетов этим интерфейсом. Это возникает в случае, когда несколько практически одновременно поступивших пакетов требуют соединения на один и тот же интерфейс. Для устранения задержек в пучке скорость передачи любого выходного интерфейса должна

быть равна суммарной скорости приема всех входных интерфейсов, что крайне невыгодно по технико-экономическим соображениям. Поэтому наличие очереди в выходном буфере является нормальной ситуацией. Задержка в пучке приводит к накоплению пакетов в выходном буфере интерфейса, при отсутствии задержек в пучке в выходном буфере не может быть более одного пакета.

В системе с повторными вызовами (Blocked Calls Repeated, BCR) вызов, не принятый к обслуживанию в момент занятости V линий пучка, удаляется из системы и практически сразу вновь поступает на вход системы. Следствием блокировки является только увеличение загрузки управляющего устройства, так как повторный вызов обрабатывается как обычный. Обслуживающие приборы занимаются на очень короткое время, близкое к нулю. В стационарном режиме параметр обслуженного потока вызовов λ_s всегда равен параметру поступающего потока λ_0. Процесс обслуживания вызовов системой с повторными вызовами показан на рис. 13 (серым цветом изображены поступившие (первичные) вызовы, темно-серым – повторные).

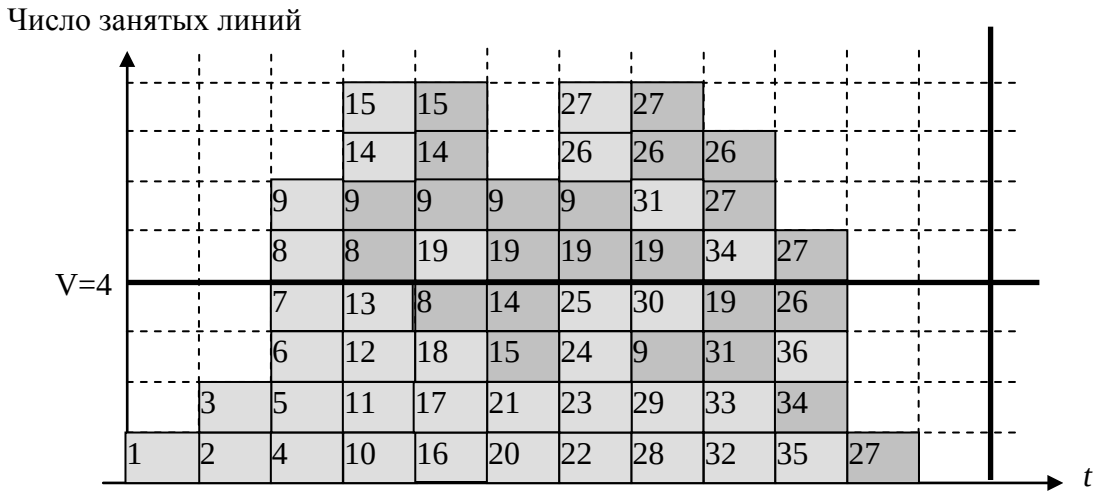


Рис. 13. Диаграмма работы системы с повторными вызовами при A_0 = 3 Эрл; V = 4; As = 3 Эрл

В системе с коллизией (Collisions Calls Repeated, CCR) вызов, не принятый к обслуживанию в момент занятости V линий пучка, ставится на ожидание в виртуальной входной очереди. Если вызов в очереди единственный, то при освобождении любой линии он обслуживается, в противном случае все вызовы из очереди удаляются и через случайный интервал времени вновь поступают на вход системы. Следствием блокировки при небольшой нагрузке является увеличение времени обслуживания, а при превышении некоторого предела поступающей нагрузки $A_{0\max}$ – полное

самопроизвольное прекращение обслуживания вызовов системой из-за переполнения очереди. Процесс обслуживания вызовов системой с коллизией в стационарном режиме показан на рис. 14 (серый цвет – первичные вызовы, темно-серый – с коллизией).

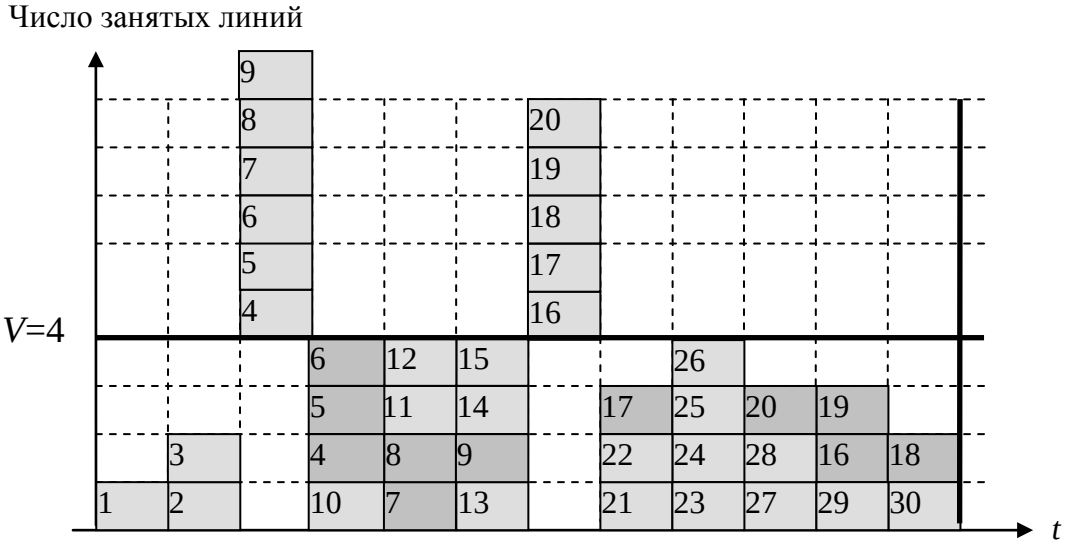


Рис. 14. Диаграмма работы системы с коллизией
при A_0 = 2,5 Эрл; V = 4; As = 2,5 Эрл

Зависимость обслуженной нагрузки от поступающей нагрузки для рассмотренных систем приведена на рис. 15. Видно, что для всех систем существует свой предел обслуженной нагрузки A_s, который достигается при различных значениях поступающей нагрузки A_0.

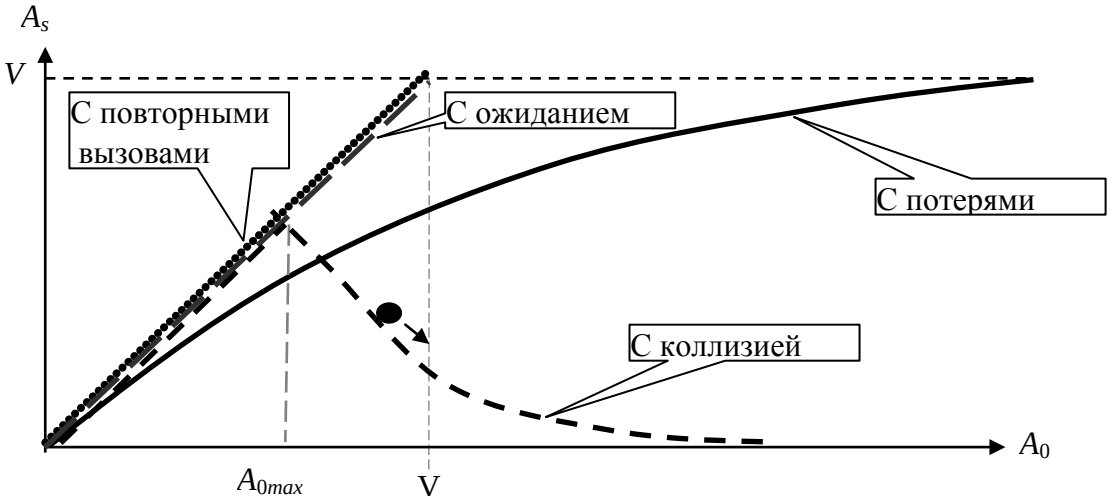


Рис. 15. Характеристики различных систем коммутации

3.3. Простейшие коммутационные устройства

К простейшим коммутационным устройствам относят коммутационный элемент, соединитель и коммутатор.

Коммутационный элемент – двухполюсник вида 1*x*1 (рис. 16, а) с одной точкой коммутации между входом (входным интерфейсом) и выходом (выходным интерфейсом), имеет два возможных состояния: замкнуто или разомкнуто.

Соединитель – многополюсник вида *nx*1 (рис. 16, б) с *n* точками коммутации, полученный путем объединения выходов *n* коммутационных элементов, в котором любой из *n* входов может соединиться с выходом. Соединитель хорошо отображает функционирование концентратора в локальной сети Ethernet.

Коммутатор – многополюсник вида *nxm* (рис. 16, в) с *nm* точками коммутации, полученный путем объединения одноименных *n* входов *m* соединителей, в котором любой из *n* входов может соединиться с любым из *m* выходов. Коммутатор хорошо отображает функционирование IP-модуля. Входы коммутатора отображают входные интерфейсы, по которым в IP-модуль поступают пакеты на обслуживание, а выходы коммутатора – выходные интерфейсы, по которым пакеты покидают IP-модуль в соответствии с адресом пункта назначения.

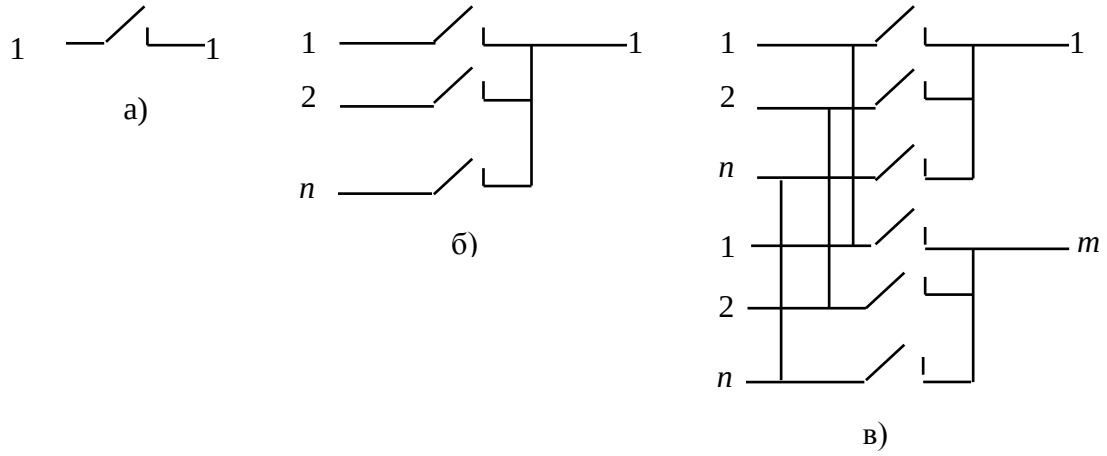


Рис. 16. Простейшие коммутационные устройства:
а) коммутационный элемент; б) соединитель; в) коммутатор

Коммутационное поле (КП) – многополюсник вида *NxM*, в котором любой поступивший на вход вызов может соединиться с любым из *M* выходов в соответствии с заданным режимом установления соединения. Коммутационные поля, в которых соединение входа-выхода происходит через один коммутационный элемент (точку коммутации) называются ***однозвенными***. Если КП обладает таким свойством, что между любым вхо-

дом и свободным выходом ***всегда*** имеется свободный соединительный путь, то оно называется ***полнодоступным***. Коммутатор – однозвенное полнодоступное КП.

В режиме свободного искания любой поступивший на вход вызов соединяется с любым свободным выходом.

В режиме группового искания выходы КП разбиваются на h направлений, и за каждым направлением закрепляется V_j (j = 1, …h) выходов так, что $\Sigma V_j = M$. Выходы одного направления образуют ***пучок линий***. Поступивший на вход вызов соединяется с любым свободным выходом выбранного направления.

В режиме линейного искания поступивший вызов соединяется с одним фиксированным выходом.

3.4. Стационарный процесс рождения и гибели в полнодоступном пучке

Функционирование многих реальных сетевых элементов хорошо описывается Марковским процессом рождения и гибели, где под рождением понимается поступление вызова (пакета) и занятие им одной обслуживающей линии, а под гибелью – окончание его обслуживания и освобождение линии. Процесс предполагает наличие ограниченного пучка обслуживающих линий. Момент поступления вызова и время его обслуживания являются случайными величинами. При отсутствии свободных линий обслуживание вновь поступивших вызовов прекращается. Такой принцип обслуживания применяется в системе с коммутацией каналов [5].

Результаты имеют фундаментальное значение и широко используются в частных случаях, которые приведены в последующих разделах.

Вероятностный процесс называется *Марковским*, если будущее поведение процесса не зависит ни от каких сведений о прошлом. В *уравнении Колмогорова – Чепмана* это выражено в безусловных вероятностях перехода состояний:

$$p_{ji}(t+\tau) = \sum_{j=0}^{V} p_j(t)\, p_{ji}(\tau).$$

Пусть $p_x(t + \tau)$ – вероятность состояния $\{x\}$ пучка с x = 0,…V занятыми линиями в момент времени ($t + \tau$), $p_x(t)$ – та же вероятность в момент времени t. Согласно определению марковского процесса в момент времени $\tau \to 0$ может произойти не более одного элементарного события. Поэтому, если в момент времени ($t + \tau$) пучок находится в состоянии $\{x\}$, то это возможно при наступлении за время τ одного из трех возможных событий:

– в момент времени t пучок находился в состоянии $\{x-1\}$, а за время τ поступил один вызов;

– в момент времени t пучок находился в состоянии $\{x+1\}$, а за время τ произошло одно освобождение;

– в момент времени t пучок находился в состоянии $\{x\}$, а за время τ не поступил ни один вызов, и не произошло ни одного освобождения.

Графически вероятностные процессы часто изображают в виде *диаграммы переходов*, в которой соседние состояния соединяются линиями, отображающими интенсивности переходов между ними. Диаграмма переходов процесса рождения и гибели изображена на рис. 17.

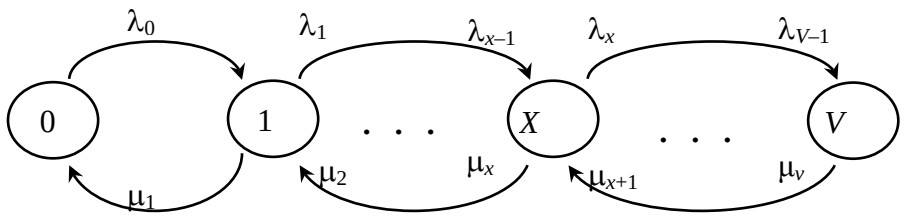


Рис. 17. Диаграмма переходов

Вероятность поступления в состоянии $\{x-1\}$ за время τ хотя бы одного вызова

$$p_{+1}(\tau)=1-e^{-\lambda_{x-1}\tau}=1-\left(1-\frac{(\lambda_{x-1}\tau)^1}{1!}+\frac{(\lambda_{x-1}\tau)^2}{2!}-\ldots\right)_{\tau\to 0}=\lambda_{x-1}\tau.$$

Вероятность освобождения в состоянии $\{x+1\}$ за время τ хотя бы одной линии пучка

$$p_{-1}(\tau)=1-e^{-\mu_{x+1}\tau}=1-\left(1-\frac{(\mu_{x+1}\tau)^1}{1!}+\frac{(\mu_{x+1}\tau)^2}{2!}-\ldots\right)_{\tau\to 0}=\mu_{x+1}\tau.$$

Вероятность нахождения пучка за время τ в том же состоянии $\{x\}$

$$p_0(\tau)=e^{-\lambda_x\tau}e^{-\mu_x\tau}=e^{-(\lambda_x+\mu_x)\tau}{}_{\tau\to 0}=1-\lambda_x\tau-\mu_x\tau.$$

Следовательно, из уравнения Колмогорова – Чепмана получаем

$$p_x(t+\tau)=p_{x-1}(t)p_{+1}(\tau)+p_{x+1}(t)p_{-1}(\tau)+p_x(t)p_0(\tau)=$$
$$=\lambda_{x-1}\tau\ p_{x-1}(t)+\mu_{x+1}\tau\ p_{x+1}(t)+p_x(t)(1-\lambda_x\tau-\mu_x\tau),\qquad x=0,\ldots V,$$

или

$$\frac{p_x(t+\tau)-p_x(t)}{\tau}=-\lambda_{x-1}p_{x-1}(t)-\left(\lambda_x+\mu_x\right)p_x(t)+\mu_{x+1}p_{x+1}(t),\quad x=0,\ \ldots,V.$$

Перейдя к пределу и учитывая то, что $\lambda_{-1} = \mu_0 = \lambda_V = \mu_{V+1} = 0$, получаем систему дифференциальных уравнений

$$\frac{dp_0(t)}{dt} = -\lambda_0 p_0(t) + \mu_1 p_1(t), \qquad x = 0,$$

$$\frac{dp_x(t)}{dt} = \lambda_{x-1} p_{x-1}(t) - \left(\lambda_x + \mu_x\right) p_x(t) + \mu_{x+1} p_{x+1}(t), \qquad x = 1, \ldots V-1,$$

$$\frac{dp_V(t)}{dt} = \lambda_{V-1} p_{V-1}(t) - \mu_V p_V(t), \qquad x = V.$$

В стационарном режиме вероятности состояний не зависят от времени наблюдения, т. е.

$$\frac{dp_0(t)}{dt} = \frac{dp_1(t)}{dt} = \ldots = \frac{dp_V(t)}{dt} = 0,$$

поэтому $p_x(t) \to [x]$, $x = 0,\ldots V$, где $[x]$ – стационарная вероятность занятия точно x линий пучка.

Следовательно,

$$\mu_1[1] - \lambda_0[0] = \mu_2[2] - \lambda_1[1] = \ldots = \mu_V[V] - \lambda_{V-1}[V-1] = 0$$

или, иначе:

$$[1] = \frac{\lambda_0}{\mu_1}[0],$$

$$[2] = \frac{\lambda_1}{\mu_2}[1] = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2}[0],$$

$$\cdot \quad \cdot \quad \cdot$$

$$[x] = \frac{\lambda_{x-1}}{\mu_x}[x-1] = \prod_{i=0}^{x-1} \frac{\lambda_i}{\mu_{i+1}}[0],$$

$$\cdot \quad \cdot \quad \cdot$$

$$[V] = \frac{\lambda_{V-1}}{\mu_V}[V-1] = \prod_{i=0}^{V-1} \frac{\lambda_i}{\mu_{i+1}}[0].$$

Из условия нормировки $\sum_{x=0}^{V} [x] = 1$, находим

$$[0] = \left(\sum_{x=0}^{V} \prod_{j=0}^{x-1} \frac{\lambda_j}{\mu_{j+1}} \right)^{-1},$$

а затем

$$[x] = \prod_{j=0}^{x-1} \frac{\lambda_j}{\mu_{j+1}}[0] = \frac{\prod_{i=0}^{x-1} \frac{\lambda_i}{\mu_{i+1}}}{\sum_{x=0}^{V} \prod_{j=0}^{x-1} \frac{\lambda_j}{\mu_{j+1}}}, \quad x = 0,\ldots V. \qquad (3.1)$$

Математическое описание функционирования реальных объектов (систем) принято называть *моделями*. Наиболее распространенным на практике моделям часто присваивают имена их исследователей.

В 1953 году ***Кендаллом*** (***D. G. Kendall***) была предложена классификация моделей (***Kendall's notation***), основанная всего на трех обязательных символах. Для описания сложных процессов функционирования современных инфокоммуникационных систем эта классификация была дополнена еще тремя символами.

В общем виде модели представляют в виде сокращенной записи A/S/c/K/N/D из шести символов. Здесь первый символ A обозначает вид поступающего потока вызовов (M – пуассоновский поток вызовов, D – с фиксированными промежутками между вызовами), S – вид времени занятия (M – экспоненциальное, D – постоянное, G – произвольное), c – число обслуживающих приборов, K – число мест ожидания, N – число источников нагрузки, D – дисциплина обслуживания очереди. Если последние три значения не указаны, то по умолчанию они принимают значение $K = \infty$, $N = \infty$, D = *FIFO*.

Например, модель M/M/100/0 описывает систему, на которую поступает пуассоновский поток вызовов, время обслуживания имеет экспоненциальное распределение, число обслуживающих приборов равно 100, мест ожидания нет, число источников нагрузки не ограничено, дисциплина обслуживания очереди FIFO.

Модель M/D/V описывает систему, на которую поступает пуассоновский поток вызовов, время обслуживания имеет постоянное значение, число обслуживающих приборов равно V, число мест ожидания и источников нагрузки не ограничено, дисциплина обслуживания очереди FIFO.

Модель M/M/10/0/ /PPQ описывает систему, на которую поступает пуассоновский поток вызовов, время обслуживания имеет экспоненциальное распределение, число обслуживающих приборов равно 10, места ожидания отсутствуют, число источников нагрузки не ограничено, используется приоритетное обслуживание с прерыванием обслуживания неприоритетных вызовов.

Модель M/M/10/ / /WFQ описывает систему, на которую поступает пуассоновский потоков вызовов, время обслуживания имеет экспоненциальное распределение, число обслуживающих приборов равно 10, число мест ожидания в индивилуальной очереди и число источников нагрузки не ограничены, при освобождении линии обслуживается вызов любого потока с наименьшим местом ожидания.

ЗАДАЧА 1. На систему за 1 мин поступило 15 вызовов. Время занятия 30 с. Найти поступившую нагрузку.

Решение.

$$A = \lambda \cdot ts = \frac{15}{\text{мин}} \cdot 30\ \text{с} = \frac{450\ \text{с}}{60\ \text{с}} = 7{,}5\ \text{Эрл}.$$

ЗАДАЧА 2. Определить телефонные нагрузки, создаваемые телефонными аппаратами, используя статистические данные, приведенные в табл. 3.1.

Таблица 3.1

Статистические данные телефонной нагрузки

Вид телефонного аппарата	Параметр потока вызовов λ, ч	Средняя длительность разговора *ts*, с
Квартирный	2	200
Уличный	15	120
Учрежденческий	5	150

Решение.

Для квартирного телефонного аппарата $\lambda = 2\ \text{ч}^{-1} = 5{,}5 \cdot 10^{-4}\ \text{с}^{-1}$; $ts = 200$ с; $a = \lambda \cdot ts = 0{,}11$ Эрл.

Для уличного телефонного аппарата $\lambda = 15\ \text{ч}^{-1} = 4{,}1 \cdot 10^{-3}\ \text{с}^{-1}$; $ts = 120$ с; $a = \lambda \cdot ts = 0{,}5$ Эрл.

Для учрежденческого телефонного аппарата $\lambda = 5\ \text{ч}^{-1} = 1{,}4 \cdot 10^{-3}\ \text{с}^{-1}$; $ts = 150$ с; $a = \lambda \cdot ts = 0{,}21$ Эрл.

ЗАДАЧА 3. От квартирного аппарата поступает телефонная нагрузка, соответствующая приведенным выше статистическим данным. Телефон подключен через формирователь пакетов к ИКМ-тракту со скоростью *VEL* = 2048 кбит/с = 256 Кбайт/с. Формирователь пакетов реализован в соответствии с Рекомендацией G.711 (время формирования пакета *tcod* = 20 мс, длина пакета *LEN* = 238 байт). Определить параметр потока пакетов λ*pac*, среднее время обслуживания пакета *tpac*, пакетную нагрузку *apac* в ЧНН.

Решение.

$$\lambda pac = \lambda \frac{ts}{tcod} = \frac{a}{tcod} = 5{,}55\ \text{с}^{-1}, \qquad tpac = \frac{LEN}{VEL} = 0{,}00093\ \text{с},$$

$$apac = \lambda pac \cdot tpac = 0{,}0051\ \text{Эрл}.$$

4. СИСТЕМА С ЯВНЫМИ ПОТЕРЯМИ (Blocked Calls Cleared, BCC)

4.1. Модель Эрланга (M/M/V/0). Дискретная формула Эрланга

В системе с коммутацией каналов большой емкости, когда число источников нагрузки N (абонентов) велико, а нагрузка от одного источника a мала, поведение одного источника (наличие или отсутствие от него вызовов) мало влияет на суммарный поток вызовов $A = Na$. В этом случае суммарный поток вызовов является практически постоянной величиной и не зависит от состояния $\{x\}$, (x = 0, 1, ...V) системы. Такой поток называется простейшим. Для него $\lambda_0 = \lambda_1 = ... = \lambda_{V-1} = \lambda > 0$, $\mu_x = x\mu$.

Модель Эрланга для расчета вероятности потерь справедлива при следующих предположениях:

– вызовы, поступающие на вход системы, образуют пуассоновский поток постоянной интенсивности с параметром $\lambda > 0$;

– длительность занятия линии подчиняется экспоненциальному распределению с параметром $\mu > 0$;

– вызов, заставший хотя бы одну свободную линию, обслуживается немедленно, в противном случае теряется, не влияя на моменты поступления последующих вызовов;

– любой из V выходов пучка доступен, когда он свободен, для любого поступающего вызова;

– исходной для расчета является поступающая нагрузка;

– система находится в стационарном режиме.

Подставляя значения параметров $\lambda_x = \lambda$ и $\mu_x = x \cdot \mu$ в (3.1) вероятностей стационарного процесса рождения и гибели, получим

$$[x] = \frac{(\lambda/\mu)^x}{x!}\left[\sum_{i=0}^{V}\frac{(\lambda/\mu)^i}{i!}\right]^{-1} = \frac{A^x}{x!}[0], \quad x = 0, 1, \ldots V, \tag{4.1}$$

где $A = \lambda/\mu$ – поступающая нагрузка ***первого рода***.

Финальная вероятность

$$[V] = E_V(A) = \frac{A^V}{V!}\left[\sum_{i=0}^{V}\frac{A^i}{i!}\right]^{-1} = \frac{A^V}{V!}[0] \tag{4.2}$$

определяет потери по времени в полнодоступном пучке и носит название ***первой формулы Эрланга***.

В модели Эрланга потери по времени, вызовам, нагрузке совпадают, параметр потерянного потока – $\lambda E_V(A)$, потерянная нагрузка – $AE_V(A)$. Обслуженная нагрузка (число вызовов в системе)

$$A_s = \sum_{k=0}^{V} k[k] = A[0]\sum_{k=0}^{V-1}\frac{A^k}{k!} = A[0]\left(\sum_{k=0}^{V}\frac{A^k}{k!} - \frac{A^V}{V!}\right) = A\left(1 - E_V(A)\right).$$

Прямой расчет формулы Эрланга во многих практических случаях невозможен из-за переполнения разрядной сетки вычислительного устройства (при больших значениях A и V). Поэтому для ее расчета пользуются рекуррентным соотношением

$$E_V(A) = \frac{AE_{V-1}(A)}{V + AE_{V-1}(A)}, \tag{4.3}$$

последовательно вычисляя $E_1(A)$, $E_2(A)$, . . . $E_{V-1}(A)$, $E_V(A)$, при начальном значении $E_0(A0) = 1$.

Следует запомнить:

$$E_0(A) = 1, \quad E_V(0) = 0, \quad E_0(0) = 0,$$

$$E_V(A) = \left.\frac{\sum_{k=0}^{V} A^k/k! - \sum_{k=0}^{V-1} A^k/k!}{\sum_{k=0}^{V} A^k/k!}\right|_{A>>V} \approx 1 - \frac{A^{V-1}/(V-1)!}{A^V/V!} = \frac{A-V}{A}.$$

4.2. Модель Эрланга (M/M/V/0). Интегральная формула Эрланга

Пусть $\{x\}$ – состояние V-линейного пучка (наличие x установленных соединений), $[x]$ – стационарная вероятность состояния $\{x\}$, ($x = 0,\ldots V$). Тогда первый потерянный вызов, поступивший на (V+1)-ю фиктивную линию пучка при условии неосвобождения занятых линий, будет принадлежать потоку Эрланга (V–x)-го порядка, поскольку потери наступят только после поступления (V–x+1)-го вызова.

Рассмотрим функцию распределения промежутков времени между вызовами потока k-го порядка. Для этого обратимся к потоку Эрланга нулевого порядка (простейшему потоку вызовов) с интенсивностью λ, у которого функция распределения, плотность, изображение плотности и математическое ожидание промежутков между вызовами соответственно равны

$$F_0(t) = 1 - e^{-\lambda t};$$

$$f_0(t) = dF_0(t)\big/dt = \lambda e^{-\lambda t};$$

$$f_0(s) = L\left[\, f_0(t)\right] = \int_0^\infty \lambda e^{-\lambda t} e^{-zt} dt = \frac{\lambda}{\lambda + z};$$

$$M[t_0] = \int_0^\infty t f_0(t) dt = \frac{1}{\lambda},$$

где L – оператор Лапласа.

Поток Эрланга k-го порядка образуется из порождающего потока нулевого порядка, когда k вызовов потока нулевого порядка пропускаются, (k+1)-й учитывается, затем снова k вызовов пропускаются, следующий учитывается и т. д. Используя теорему о свертке, получим изображение потока Эрланга k-го порядка:

$$f_k(z) = L\left[\, f_k(t)\right] = \frac{\lambda^{k+1}}{(\lambda + z)^{k+1}},$$

из которого разложением Хэвисайда для рациональных алгебраических дробей находим оригинал:

$$f_k(t) = L^{-1}\left[\, f_k(z)\right] = \lambda \frac{(\lambda t)^k}{k!} e^{-\lambda t},$$

где L^{-1} – оператор обратного преобразования Лапласа.

Теперь легко находим функцию распределения промежутков между вызовами

$$F_k(t) = \int_0^\infty f_k(t) dt = 1 - \sum_{i=0}^{k} \frac{(\lambda t)^i}{i!} e^{-\lambda t},$$

математическое ожидание промежутков времени между ними

$$M_k[t] = \int_0^\infty t f_k(t) dt = \frac{k+1}{\lambda}$$

и параметр просеянного потока Эрланга k-го порядка

$$\lambda_k = 1 \big/ M_k[t] = \frac{\lambda}{k+1}.$$

Вернемся к нашему случаю. На (x+1)-ую линию пучка в состоянии $\{k\}$, k = 1, 2,…x поступают условно потерянные вызовы потока Эрланга (x–k)-го порядка, ибо k вызовов уже «потеряны» на предыдущих линиях пучка. Поэтому плотность потока условно потерянных вызовов, поступающих на (x+1)-ю линию пучка, находим по формуле полной вероятности:

$$f_x(t)=\sum_{k=0}^{V}[k]f_{x-k}(t)=\sum_{k=0}^{x}[k]f_{x-k}(t)+\sum_{k=x+1}^{V}[k]f_0(t)=$$

$$=\sum_{k=0}^{x}\frac{(\lambda/\mu)^k}{k!}[0]\cdot\lambda\frac{(\lambda t)^{x-k}}{(x-k)!}e^{-\lambda t}+\{1-F(x)\}\lambda e^{-\lambda t}=$$

$$=\lambda[x]e^{-\lambda t}\sum_{k=0}^{x}C_x^k(\mu t)^{x-k}+\{1-F(x)\}\lambda e^{-\lambda t}=\lambda[x](1+\mu t)^x e^{-\lambda t}+\{1-F(x)\}\lambda e^{-\lambda t},$$

где $F(x)=\sum_{k=0}^{x}[x]$ – функция распределения занятости x линий пучка.

Очевидно, $\int_0^\infty f_x(t)dt=1,$ откуда для любого значения x

$$\lambda[x]\int_0^\infty(1+\lambda t)^x e^{-\lambda t}dt+\{1-F(x)\}=1,$$

а функция распределения занятости x линий пучка

$$F(x)=\lambda[x]\int_0^\infty(1+\mu t)^x e^{-\lambda t}dt=[x]\frac{\int_0^\infty(A+z)^x e^{-z}dz}{A^x}=\frac{[x]I(x,A)}{A^x},$$

где $I(x;A)=\int_0^\infty(A+z)^x e^{-z}dz$ – интеграл первого рода.

В частности, при $x = V$, очевидно что $F(V) = 1$,

$$[V]=E_V(A)=\frac{A^V}{I(V;A)}=\frac{A^V}{\int_0^\infty(A+z)^V e^{-z}dz}. \qquad (4.4)$$

Выражение (4.4) – ***интегральная формула Эрланга***, позволяющая вычислять потери при любых действительных положительных A и V.

Из очевидного равенства

$$[V]=E_V(A)=\frac{A^V}{V!}\left[\sum_{i=0}^{V}\frac{A^i}{i!}\right]^{-1}=\frac{A^V}{\int_0^\infty(A+z)^V e^{-z}dz}$$

при $A = 0$ следует, что $V!=I(V;0)=\int_0^\infty z^V e^{-z}dz$ для любого действительного положительного V. Например, 5,6! = 344,69; $\sum_{i=0}^{5,6}\frac{3,2^i}{i!}=\frac{I(5,6;3,2)}{I(5,6;0)}=22,96$.

Для интегральной формулы Эрланга верно рекуррентное соотношение

$$E_V(A) = \frac{AE_{V-1}(A)}{V + AE_{V-1}(A)}$$

с начальным значением

$$E_\nu(A) = \frac{A\nu}{I(\nu; A)},$$

где $\nu = V - INT\,(V)$ – дробная часть емкости пучка.

Также, очевидно, что интегральное распределение Эрланга имеет следующий вид:

$$F(x) = \frac{\sum_{k=0}^{x} A^k / k!}{\sum_{k=0}^{V} A^k / k!} = \frac{I(x;A)}{I(x;0)} \frac{I(V;0)}{I(V;A)}.$$

Интегральная формула Эрланга применяется при нецелочисленном значении емкости пучка (см. задачу 7 в конце главы).

4.3. Система M/M/V/0/ /PPQ

Для системы M/M/V/0/ /PPQ с абсолютным приоритетом первого потока и прерыванием вызовов второго потока справедливы следующие предположения:

– на вход системы поступают два пуассоновских потока с интенсивностями $\lambda 1$ и $\lambda 2$, соответственно;

– длительность занятия подчиняется экспоненциальному распределению с параметром μ;

– если в момент поступления вызова первого потока все линии заняты вызовами первого потока, то вызов теряется; если нет, то поступивший вызов замещает занятую линию вторым потоком. Потерянные или замещенные вызовы покидают систему;

– любой из V выходов пучка доступен, когда он свободен, для любого поступающего вызова;

– исходной для расчета является поступающая нагрузка;

– система находится в стационарном режиме.

В системе M/M/V/0 без приоритетов распределение вероятностей для первого и второго потоков пропорциональны поступающим нагрузкам

$$A1 = \lambda 1/\mu, \quad A2 = \lambda 2/\mu, \quad q1 = \frac{A1}{A} = \frac{A1}{A1 + A2},$$

$$E_V(A1+A2)=\frac{(A1+A2)^V\ /V!}{\sum_{i=0}^{V}(A1+A2)^i\ /i!}.$$

Общая обслуженная нагрузка

$$As=(A1+A2)\cdot\left[1-E_V(A1+A2)\right],$$

обслуженная нагрузка первого потока

$$As1=As\cdot q1=A1\cdot\left[1-E_V(A1+A2)\right],$$

обслуженная нагрузка второго потока

$$As2=As\cdot(1-q1)=A2\cdot\left[1-E_V(A1+A2)\right].$$

В системе M/M/V/0/ /PPQ с абсолютным приоритетом первого потока вызовы второго потока не оказывают никакого влияния на вызовы первого потока, поэтому

$$E_V(A1+A2)=\frac{(A1+A2)^V\ /V!}{\sum_{i=0}^{V}(A1+A2)^i\ /i!},$$

$$E_V(A1)=\frac{A1^V\ /V!}{\sum_{i=0}^{V}A1^i\ /i!},$$

общая обслуженная нагрузка

$$As=(A1+A2)\cdot\left[1-E_V(A1+A2)\right],$$

обслуженная нагрузка первого потока

$$As1=A1\cdot\left[1-E_V(A1)\right],$$

обслуженная нагрузка второго потока

$$As2=As-As1.$$

В системе с абсолютным приоритетом произошло увеличение доли первого потока с $q1=\frac{A1}{A1+A2}$ до $q1p=\frac{A1\cdot\left(1-E_V(A1)\right)}{(A1+A2)\cdot\left[1-E_V(A1+A2)\right]}$. Соответственно доля второго потока снизилась на такую же величину.

4.4. Модель Эрланга (M/M/V/0). Поток Пальма

Промежутки времени между вызовами потерянного потока, кроме первого, распределены по одинаковому закону, поэтому потерянный поток вызовов является рекуррентным и называется ***потоком Пальма***.

Для потерянного на V-линейном полнодоступном пучке простейшего потока вызовов Пальм вывел уравнение

$$\varphi_x(t) = \varphi_{x-1}(t) - \int_0^t (1 - e^{-\mu\tau})\varphi_x(t-\tau)d\varphi_{x-1}(\tau), \quad x = 1,\ldots V, \tag{4.5}$$

где $\varphi_x(t)$ – вероятность того, что в промежутке времени (t_0, $t_0 + \tau$) ни один вызов на x-й линии пучка не будет потерян, при условии, что в данный момент t_0 на этой линии теряется вызов, μ – интенсивность обслуживания вызова.

Применим к (4.5) преобразование Лапласа:

$$\varphi_x(z) = \varphi_{x-1}(z) - \varphi_x(z)d\varphi_{x-1}(z) + \varphi_x(z)d\varphi_{x-1}(z+\mu) =$$
$$= \varphi_{x-1}(z) - z\varphi_x(z)\varphi_{x-1}(z) + (z+\mu)\varphi_x(z)\varphi_{x-1}(z+\varphi),$$

откуда

$$\varphi_x(z) = \frac{\varphi_{x-1}(z)}{1 + z\varphi_{x-1}(z) - (z+\mu)\varphi_{x-1}(z+\mu)}. \tag{4.6}$$

Введем новую функцию $F_{x-1}(t) = 1 - \varphi_{x-1}(t)$ – функцию распределения промежутков времени между вызовами на выходе (x–1)-й линии пучка. Для нее

$$\varphi_{x-1}(z) = \frac{1}{z} - F_{x-1}(z) = \frac{1 - zF_{x-1}(z)}{z}. \tag{4.7}$$

Подстановкой (4.7) в (4.6) получаем

$$\varphi_x(z) = \frac{1 - zF_{x-1}(z)}{z[1 - zF_{x-1}(z) + (z+\mu)F_{x-1}(z+\mu)]}$$

или

$$F_x(z) = \frac{1}{z} - \varphi_x(z) = \frac{(z+\mu)F_{x-1}(z+\mu)}{z\left[1 - zF_{x-1}(z) + (z+\mu)F_{x-1}(z+\mu)\right]} =$$
$$= \frac{f_{x-1}(z+\mu)}{z\left[1 - f_{x-1}(z) + f_{x-1}(z+\mu)\right]}, \tag{4.8}$$

где $f_{x-1}(z)$ – преобразование Лапласа от плотности потока, потерянного на (x–1) линиях пучка.

Последовательно вычисляя

$$f_i(z+j\mu)=\frac{f_{i-1}\left(z+(j+1)\mu\right)}{1-f_{i-1}(z+j\mu)+f_{i-1}\left(z+(j+1)\mu\right)},\quad i=1,\ldots x,\quad j=0,\ldots x,$$

при начальных

$$f_0(z+j\mu)=\frac{\lambda}{\lambda+z+j\mu},\qquad j=0,\ldots x,$$

итерационным методом из решения уравнения

$$1-f_{x-1}(z)+f_{x-1}(z+\mu)=0$$

находим корни γ_k, $k = 1,\ldots x + 1$.

Тогда вследствие того, что $f_x(0) = 1$, (4.8) можно представить в виде

$$F_x(z)=\frac{\prod\limits_{k=1}^{x+1}\gamma_k}{z\prod\limits_{k=1}^{x+1}(z+\gamma_k)},$$

которое имеет оригинал

$$F_x(t)=1-\sum_{k=1}^{x+1}\frac{\prod\limits_{k=1}^{x+1}\gamma_k}{\gamma_k\prod\limits_{\substack{i=1\\ j\neq k}}^{x+1}(\gamma_j+\gamma_k)}e^{\gamma_k t}. \tag{4.9}$$

Таким образом, в неявном виде получено аналитическое выражение (4.9) для расчета функции распределения промежутков времени между вызовами потерянного потока на x линиях пучка. Любое аналитическое выражение, полученное в неявном виде, обладает, по крайней мере, двумя недостатками.

Во-первых, оно лишено наглядности для анализа зависимости рассчитываемого параметра от исходных данных, поскольку последние в конечном аналитическом выражении отсутствуют.

Во-вторых, дальнейшее использование рассчитанного параметра в качестве составной части другого сложного аналитического выражения еще более усложняет решение и теряет наглядность. Эти причины заставляют искать приближенную оценку точного решения.

Для приближенной оценки функции распределения можно воспользоваться формулой в виде

$$F_x(t)=1-\exp\left(-\lambda t E_x(A)\right). \tag{4.10}$$

Для сравнения (4.9) и (4.10) в табл. 4.1 приведены результаты расчета функции распределения промежутков времени между вызовами потерянного потока при следующих исходных данных: $\lambda = 10$ ч$^{-1}$; $\mu = 10$ ч$^{-1}$; $x = 5$. Точное решение вычислялось с использованием (4.9)

$$F_5(t) = 1 - 1{,}0039e^{-0{,}02t} + 0{,}0069e^{-10{,}5t} - 0{,}0043e^{-22{,}5t} +$$
$$+ 0{,}0016e^{-37{,}8t} - 0{,}0003e^{-56{,}8t} + 0{,}00003e^{-82{,}9t}.$$

Таблица 4.1

Значения функций распределения, вычисленных по (4.9) и (4.10)

Значение t, час	Значение $F_x(t)$, вычисленное по формуле	
	(4.9)	(4.10)
0,50	0,0060	0,0152
1,00	0,0159	0,0302
2,00	0,0354	0,0595
5,00	0,0916	0,1421
10,00	0,1780	0,2641
50,00	0,6306	0,7842
100,00	0,8640	0,9534

Как видно из таблицы, приближение дает завышенные оценки функции распределения во всем диапазоне времени.

В процессе обслуживания вызовов сетью связи каждый узел коммутации каналов анализирует адресную часть поступившего сигнального сообщения и устанавливает соединение в соответствии с таблицей маршрутов. Если возникает дефицит ресурсов (каналов) на пути первого выбора, то на альтернативные пути поступает просеянный поток вызовов, не принятый к обслуживанию на пути первого выбора. На последующих узлах происходит объединение собственного потока вызовов с просеянным потоком вызовов.

Рассмотрим фрагмент сети связи, состоящей из трех узлов коммутации (A, B, C), изображенной на рис. 18, и проследим прохождение потоков при установлении соединений от узла A к узлу B.

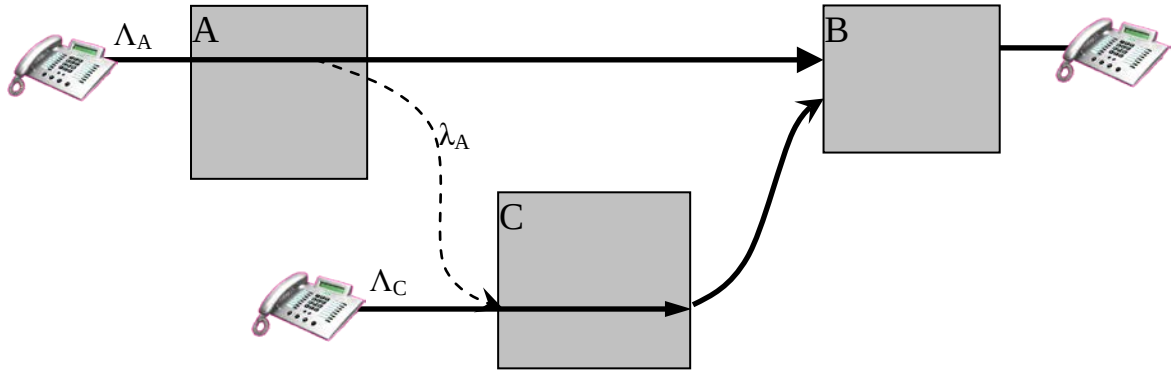


Рис. 18. Объединение и просеивание потоков в сети

Основной поток вызовов пойдет по пути первого выбора (на рис. 18 – жирная линия), при недостаточной емкости основного пучка – по пути второго выбора через узел С (на рис. 18 – пунктирная линия). Таким образом, в пучке соединительных линий от узла А к узлу В происходит просеивание потока собственных вызовов узла А с параметром Λ_A , а на входе узла С – объединение двух потоков: потока собственных вызовов с параметром Λ_C и просеянного потока абонентов узла А с параметром λ_A.

При расчете характеристик системы коммутации в предыдущих разделах было необходимо знать тип и параметры поступающего потока вызовов. Если поток собственных вызовов узла С можно вполне считать простейшим с параметром Λ_C, то степень влияния просеянного потока с параметром λ_A на поток Λ_C можно определить, только зная результат просеивания потока Λ_A узлом А. Тогда встают вопросы: как классифицировать суммарный поток вызовов с параметром (Λ_C + λ_A)? В частности, в каких случаях суммарный поток вызовов можно считать простейшим и уверенно пользоваться соответствующими формулами?

Ранее были рассмотрены характеристики двух просеянных потоков: в 3.3 – потока Эрланга и в 3.4 – потока Пальма. На рис. 19 для примера показаны: а) порождающий поток вызовов б) просеянный поток Эрланга четвертого порядка и в) потерянный поток Пальма четвертого порядка.

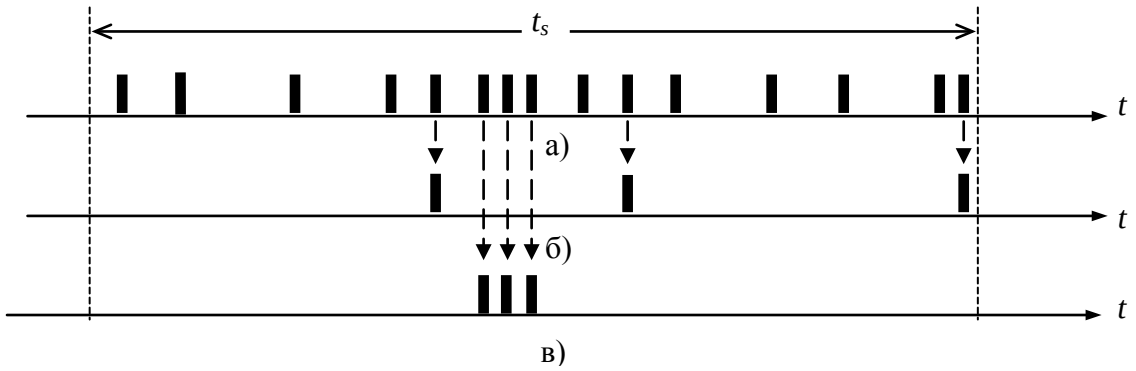


Рис. 19. Потоки вызовов:
а) порождающий поток вызовов; б) просеянный поток Эрланга четвертого порядка; в) потерянный поток Пальма четвертого порядка

Как видно из рисунка, при одинаковой интенсивности просеянных потоков на интервале среднего времени занятия t_s характер потоков существенно разнится, хотя они имеют один и тот же параметр потока.

Просеянный поток Эрланга *k*-го порядка описывается достаточно простым аналитическим выражением:

$$F_k(t)=\int_0^\infty f_k(t)dt=1-\sum_{i=0}^{k}\frac{(\lambda t)^i}{i!}e^{-\lambda t}=1-\frac{I(k;\lambda t)}{I(k;0)}e^{-\lambda t},$$

где $I(k;\lambda t)=\int_0^\infty(\lambda t+z)^k e^{-z}dz$ – интеграл первого рода, $\lambda=\Lambda/(k+1)$ – параметр просеянного потока, Λ – параметр порождающего потока вызовов.

Потерянный поток Пальма считается только численными методами и не имеет явного аналитического выражения, хотя его параметр после просеивания полнодоступной схемой известен – $\Lambda E_V(A)$. Поэтому для аналитического описания потерянного потока вызовов на полнодоступном пучке приходится производить подмену потока Пальма потоком Эрланга той же интенсивности.

Пусть Λ_i – параметр простейшего порождающего i-го потока вызовов, λ_i – параметр просеянного i-го потока вызовов. Тогда процедура классификации суммарного n-мерного потока вызовов сводится к следующему.

1. Определяется эквивалентный параметр порождающего потока вызовов

$$\Lambda_s=\frac{\lambda_i}{\sum_{i=1}^{n}\lambda_i}\Lambda_1+\frac{\lambda_2}{\sum_{i=1}^{n}\lambda_i}\Lambda_2+\ldots+\frac{\lambda_n}{\sum_{i=1}^{n}\lambda_i}\Lambda_n=\frac{\sum_{i=1}^{n}\lambda_i\Lambda_i}{\sum_{i=1}^{n}\lambda_i}.$$

2. Определяется суммарный параметр просеянного потока вызовов

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

3. Вычисляется коэффициент просеивания эквивалентного порождающего потока вызовов

$$k=\Lambda_s/\lambda_s-1.$$

4. Рекомендуются следующие функции распределения:

$$F(t)=1-\exp\left(-\sum_{i=1}^{n}\lambda_s t\right),\qquad k<1,$$

$$F(t)=1-\frac{I\left(k;\Lambda_s t\right)}{I(k;0)}\exp(-\Lambda_s t),\quad k\geq 1.$$

4.5. Модель Энгсета (M/M/V/0/N). Дискретная формула Энгсета

В системе с коммутацией каналов малой емкости, когда число источников нагрузки N (абонентов) мало, параметр потока от одного источника сравним с суммарным потоком вызовов. В этом случае суммарный поток вызовов зависит от числа источников и от состояния $\{x\}$, (x = 0, 1, ...V) системы обслуживания. Такой поток называется примитивным потоком с простым последействием. Для него $\lambda_0 \neq \lambda_1 \neq ... \neq \lambda_{V-1} > 0$, $\mu_x = x\mu$.

Модель Энгсета для расчета вероятности потерь справедлива при следующих предположениях:

– параметр поступающего потока вызовов в момент занятости x выходов пропорционален числу свободных источников, т. е.

$$\lambda_x = (N - x)\alpha, \quad 0 \le x \le V,$$

где N – число источников вызовов (число входов в КП), $\alpha > 0$ – интенсивность поступления вызова от свободного источника;

– длительность занятия подчиняется экспоненциальному распределению с параметром $\mu > 0$;

– вызов, заставший хотя бы одну свободную линию, обслуживается немедленно; в противном случае теряется, не влияя на моменты поступления последующих вызовов;

– любой из V выходов пучка доступен, когда он свободен, для любого поступающего вызова;

– исходной для расчета является поступающая нагрузка;

– система находится в стационарном режиме.

Подставляя значения параметров λ_x и μ_x в формулу (3.1) вероятностей стационарного процесса рождения и гибели, получим

$$[x] = C_N^x A^x \left[\sum_{i=0}^{V} C_N^i A^i \right]^{-1}, \tag{4.11}$$

где $A = \alpha/\mu$ – пуассоновская нагрузка ***второго рода***.

Финальная вероятность

$$P_t = [V] = E_{N,V}(A) = C_N^V A^V \left[\sum_{i=0}^{V} C_N^i A^i \right]^{-1} = C_N^V A^V [0]$$

определяет потери по времени и носит название ***формулы Энгсета***.

Рассмотрим источник без потерь. Для него $N = V = 1$, $[1] = a$ – реальная удельная нагрузка, поэтому из (4.11) следует, что

$$[1] = \frac{A}{1+A} = a = \alpha_r t_s, \quad A = \frac{\alpha}{\mu} = \frac{a}{1-a},$$

где α_r – реальный параметр потока, поступающий от источника в системе без потерь, t_s – среднее время занятия.

При численных расчетах формулы (4.11) пуассоновскую нагрузку второго рода находят, используя реальную удельную нагрузку a по формуле $A = a/(1 - a)$.

Параметр потерянного потока вызовов

$$\lambda_\pi = \lambda_V[V] = (N-V)\alpha[V] = (N-V)\alpha E_{N,V}(A) = (N-V)\alpha P_t.$$

Потери по вызовам

$$P_c = \frac{[V]\lambda_V}{\sum_{x=0}^{V}[x]\lambda_x} = \frac{C_N^V A^V [0](N-V)\alpha}{\sum_{x=0}^{V} C_N^x A^x [0](N-x)\alpha} = \frac{C_{N-1}^V A^V}{\sum_{x=0}^{V} C_{N-1}^x A^x} = E_{N-1,V}(A).$$

Потери по нагрузке

$$P_y = \frac{Na - \sum_{x=0}^{V} x[x]}{Na} = \frac{C_{N-1}^V A^V}{\sum_{x=0}^{V} C_N^x A^x} = \frac{N-V}{N} E_{N,V}(A) = \left(1 - \frac{V}{N}\right) P_t.$$

Обслуженная нагрузка (число вызовов в системе)

$$A_s = \sum_{x=0}^{V} x[x] = \left(N - (N-V)P\right)_t a = Na\left(1 - P_y\right).$$

Прямой расчет формулы Энгсета во многих практических случаях невозможен из-за переполнения разрядной сетки вычислительного устройства (при больших значениях A и V). Поэтому для ее расчета пользуются рекуррентным соотношением

$$E_{N,V}(A) = \frac{A(N-V+1)E_{N,V-1}(A)}{A(N-V+1)E_{N,V-1}(A) + V},$$

последовательно вычисляя $E_{N,1}(A)$, $E_{N,2}(A)$, ... $E_{N,V}(A)$, при начальном значении $E_{N,0}(A) = 1$.

Следует запомнить что:

$$E_{N,0}(A) = 1, \quad E_{N,V}(0) = 0, \quad E_{N,0}(0) = 0,$$

$$P_{t\left|\begin{smallmatrix}N>V\\a\to 1\end{smallmatrix}\right.}=\frac{(N-V+1)A-V}{(N-V+1)A},\quad P_{c\left|\begin{smallmatrix}N>V\\a\to 1\end{smallmatrix}\right.}=\frac{(N-V)A-V}{(N-V)A},$$

$$P_{y\left|\begin{smallmatrix}N>V\\a\to 1\end{smallmatrix}\right.}=\frac{N-V}{N}\ \frac{(N-V+1)A-V}{(N-V+1)A}.$$

В инженерной практике при выполнении условия $N > 100$ для расчета потерь рекомендуется пользоваться моделью Эрланга, в противном случае следует использовать модель Энгсета.

4.6. Модель Энгсета (M/M/V/0/N). Интегральная формула Энгсета

Пусть $\{x\}$ – состояние V-линейного пучка (наличие x установленных соединений), $[x]$ – стационарная вероятность состояния $\{x\}$, $(0 \le x \le V)$. Тогда первый потерянный вызов, поступивший на $(V+1)$-ю фиктивную линию пучка при условии неосвобождения занятых линий, будет принадлежать потоку Энгсета $(V–x)$-го порядка, поскольку потери наступят только после поступления $(V–x+1)$-го вызова.

Изображение плотности условно потерянного потока Энгсета $(V–k)$-го порядка имеет вид

$$f_k(z)=L\left[f_k(t)\right]=\prod_{i=k}^{V}\frac{\lambda_i}{\lambda_i+z},$$

а сам оригинал

$$f_k(t)=L^{-1}\left[f_k(z)\right]=\sum_{i=k}^{V}e^{-\lambda_i t}\prod_{i=k}^{V}\lambda_i\Bigg/\prod_{j=k,\,j\neq k}^{V}\left(\lambda_j-\lambda_i\right).$$

Плотность потока условно потерянных вызовов, поступающих на $(V+1)$-ю фиктивную линию пучка, находим по формуле полной вероятности

$$f_V(t)=\sum_{x=0}^{V}[x]\,f_{V-x}(t)=\sum_{x=0}^{V}\prod_{j=0}^{x-1}\frac{\lambda_j}{\mu_{j+1}}[0]\sum_{i=x}^{V}e^{-\lambda_j t}\frac{\prod\limits_{j=k}^{V}\lambda_j}{\prod\limits_{j=k,\,j\neq i}^{V}\left(\lambda_j-\lambda_i\right)}=$$

$$=[V]\lambda_V\sum_{x=0}^{V}e^{-\lambda_x t}\frac{\prod\limits_{j=x+1}^{V}\mu_j}{\prod\limits_{j=x+1}^{V}\left(\lambda_j-\lambda_x\right)}\left(1+\sum_{j=1}^{x}\frac{\prod\limits_{i=j}^{x}\mu_j}{\prod\limits_{i=j-1}^{x-1}\left(\lambda_i-\lambda_x\right)}\right). \qquad (4.12)$$

Подставляя в (4.12) $\lambda_x = (N - x)\alpha$, $\mu_x = x\mu$, получим

$$f_V(t) = [V](N-V)\alpha e^{-N\alpha t}\sum_{x=0}^{V} C_V^x (-1)^{V-x}\left(1+\mu/\alpha\right)^x e^{x\alpha t}\left(\mu/\alpha\right)^{V-x} =$$

$$= [V](N-V)\alpha e^{-N\alpha t}\left(\left(1+\mu/\alpha\right)e^{\alpha t} - \mu/\alpha\right)^V.$$

Очевидно, что $\int\limits_0^\infty f_V(t)dt = 1$, откуда для любого значения V

$$[V] = E_{N,V}(A) = \frac{1}{\int\limits_0^\infty (N-V)\alpha\left((1+\mu/\alpha)e^{\alpha t} - \mu/\alpha\right)^V e^{-N\alpha t}dt} =$$

$$= \frac{(\alpha/\mu)^V}{\int\limits_0^\infty \left(1+\alpha/\mu - e^{-\alpha t}\right)^V e^{-(N-V)\alpha t} d\left((N-V)\alpha t\right)} = \qquad (4.13)$$

$$= \frac{A^V}{\int\limits_0^\infty \left(1 + A - e^{-\frac{z}{N-V}}\right)^V e^{-z}dz} = \frac{A^V}{I(N;V;A)},$$

где $I(N;V;A) = \int\limits_0^\infty \left(1 + A - e^{-\frac{z}{N-V}}\right)^V e^{-z}dz$ – интеграл второго рода.

Выражение (4.13) – ***интегральная формула Энгсета***, позволяющая производить расчет потерь при любых действительных положительных значениях *N, V, A*.

Из очевидного равенства

$$[V] = E_{N,V}(A) = C_N^V A^V \left[\sum_{i=0}^{V} C_N^i A^i\right]^{-1} = \frac{A^V}{\int\limits_0^\infty \left(1 + A - e^{-\frac{z}{N-V}}\right)^V e^{-z}dz}$$

при $A = 0$ следует, что $C_N^V = \dfrac{1}{\int\limits_0^\infty \left(1 - e^{-\frac{z}{N-V}}\right)^V e^{-z}dz}$ для любых действительных положительных V и N. Например, $C_{5,8}^{3,3} = 16{,}874$.

Для интегральной формулы Энгсета верно рекуррентное соотношение

$$E_{N,V}(A) = \frac{A(N-V+1)E_{N,V-1}(A)}{A(N-V+1)E_{N,V-1}(A)+V}$$

с начальным значением

$$E_{N,v}(A) = \frac{A^v}{I(N;v;A)},$$

где $v = V - INT\,(V)$ – дробная часть емкости пучка.

Интегральная формула Энгсета используется при расчете потерь в пучке малой емкости при нецелочисленных значениях N и V.

ЗАДАЧА 1. Время занятия подчинено экспоненциальному распределению с параметром b. Пучок находится в состоянии $\{x\}$. Определить вероятность того, что за время t: 1) освободятся все линии пучка, 2) не освободится ни одной линии, 3) освободится хотя бы одна линия.

Решение.

1) $(1-e^{-bt})^x$; 2) e^{-xbt}; 3) $1-e^{-xbt}$.

ЗАДАЧА 2. Для системы M/M/V/0 определить: а) какая максимальная нагрузка может поступать на двухлинейный пучок; б) какая максимальная нагрузка может быть обслужена двухлинейным пучком; в) какая должна быть емкость пучка, чтобы без потерь обслужить нагрузку 1 Эрланг.

Решение.

а) ∞; б) 2 Эрланг; в) ∞.

ЗАДАЧА 3. Вывести рекуррентную формулу Эрланга.

Решение.

$$\frac{1}{E_V(A)} = \frac{\dfrac{A^V}{V!} + \sum\limits_{i=0}^{V-1} A^i \big/ i!}{A^V \big/ V!} = 1 + \frac{V}{AE_{V-1}(A)} = \frac{V + AE_{V-1}(A)}{AE_{V-1}(A)}.$$

ЗАДАЧА 4. На систему M/M/3/0 поступает нагрузка первого рода интенсивностью A = 3 Эрл. Определить распределение стационарных вероятностей пучка.

Решение.

Используем (4.1):

$$[0] = [0];$$

$$[1] = [0]\frac{A}{1} = 3[0];$$

$$[2]=[1]\frac{A}{2}=4,5[0];$$

$$[3]=[2]\frac{A}{3}=4,5[0].$$

Из нормирующего условия [0] + 3[0] + 4,5[0] + 4,5[0] = 1, находим [0] = 2/26, а затем – искомые стационарные вероятности [1] = 6/26; [2] = 9/26; [3] = 9/26.

ЗАДАЧА 5. На систему M/M/3/0 поступает нагрузка первого рода интенсивностью A = 3 Эрл. Определить вероятность потерь, используя рекуррентную формулу Эрланга.

Решение.

Используем (4.3):

$$E_0(3)=1;$$

$$E_1(3)=\frac{3\cdot 1}{1+3}=0,75;$$

$$E_2(3)=\frac{3\cdot 0,75}{2+3\cdot 0,75}=\frac{9}{17};$$

$$E_3(3)=\frac{3\cdot 9/17}{3+3\cdot 9/17}=\frac{9}{26}.$$

ЗАДАЧА 6. В телефонной сети общего пользования провайдер имеет N = 20 000 пользователей. Каждый пользователь в среднем осуществляет сеанс связи в Интернет в дневное время (длительностью 12 часов) один раз в два дня. Для выхода в Интернет используется модемный пул, который закрепляет за каждым пользователем индивидуальный станционный модем на время сеанса связи. Среднее время пребывания в Интернете составляет 1 час. Определить необходимое число станционных модемов, чтобы вероятность отказа в сеансе связи из-за их недостатка не превосходила бы значения 0,1.

Решение.

Находим параметр потока вызовов от одного абонента λ = 1/2 дня = = 1/24 ч = 0,0416 ч$^{-1}$ и поступающую нагрузку

$$A=\lambda\ *N*ts=0,0416*20000*1=833,3\ \text{Эрл.}$$

Из первой формулы Эрланга

$$[V]=E_V(A)=\frac{A^V}{V!}\left[\sum_{i=0}^{V}\frac{A^i}{i!}\right]^{-1}$$

при $E_V(A)$ = 0,1; A = 833,3 Эрл находим необходимое число станционных модемов V = 760, что гораздо меньше числа пользователей N = 20 000.

ЗАДАЧА 7. На полнодоступный 1000-линейный пучок в ЧНН поступает нагрузка A = 900 Эрл. В течение 20 мин ЧНН пучок был исправен, в течение последующих 20 мин вышли из строя 200 линий пучка, а затем неисправность устранили. Определить обслуженную нагрузку.

Решение.

Средняя емкость пучка в ЧНН составляет V = 1000/3 + 800/3 + + 1000/3 = 933,33. Подставляя значения A = 900, V = 933,33 в интегральную формулу Эрл (4.4), находим потери $E_{933,33}(900)$ = 0,0081 и обслуженную нагрузку $As = A(1 - E_V(A))$ = 892,658 Эрл.

ЗАДАЧА 8. Станция мобильных абонентов стандарта GSM имеет V = 1392 радиоканала. Сколько абонентов одновременно может обслужить станция с вероятностью 0,99, если каждый абонент производит λ = 2 вызова в час с длительностью разговора ts = 5 мин?

Решение.

Нагрузка от одного абонента $a = \lambda \cdot ts$ = 0,166 Эрл.

Вероятность потерь 0,01.

Из рекуррентной формулы Эрланга $0{,}01 \geq E_{1392}(0{,}166 \cdot N)$ находим число одновременно обслуживаемых абонентов N = 8204.

ЗАДАЧА 9. Показать, что суммарный поток, полученный объединением простейших потоков, – простейший.

Решение.

В этом случае $\Lambda_i = \lambda_i$.

$$\Lambda_s = \frac{\sum_{i=1}^{n} \Lambda_i^2}{\sum_{i=1}^{n} \Lambda_i}, \qquad \lambda_s = \sum_{i=1}^{n} \Lambda_i, \qquad k = \frac{\sum_{i=1}^{n} \Lambda_i^2}{\left(\sum_{i=1}^{n} \Lambda_i\right)^2} - 1 < 0.$$

ЗАДАЧА 10. Показать, что суммарный поток, полученный объединением многих просеянных потоков, стремится к простейшему.

Решение.

$$\Lambda_s = \frac{\sum_{i=1}^{n} \lambda_i \Lambda_i}{\sum_{i=1}^{n} \lambda_i}, \qquad \lambda_S = \sum_{i=1}^{n} \lambda_i, \qquad k = \Lambda_s / \lambda_s - 1.$$

При $n \to \infty$ среднее значение порождающего потока Λ_s возрастает гораздо медленнее, чем λ_s, что приводит к постепенному снижению значения $k = \Lambda_s / \lambda_s - 1$, а это и означает, что по своим свойствам суммарный поток вызовов все более будет приближаться к простейшему.

В частности, при объединении n одинаково просеянных потоков с параметром λ каждый, среднее значение параметра порождающего потока остается постоянным $\Lambda_s = \Lambda$, суммарный параметр просеянного потока $\lambda_s = n\lambda$, $k = \Lambda/n\lambda - 1|_{n \to \infty} = -1$.

ЗАДАЧА 11. Классифицировать суммарный поток, полученный объединением простейшего потока с параметром $\Lambda_1 = 100$ ч$^{-1}$ и потока с параметром $\Lambda_2 = 200$ ч$^{-1}$, потерянном на шестилинейном пучке. Время занятия – 3 мин.

Решение.

$\mu = 20$ ч$^{-1}$, A2 = $\Lambda_2 / \mu = 10$ Эрл,

$\Lambda_1 = 100$ ч$^{-1}$, $\lambda_1 = 100$ ч$^{-1}$, $k_1 = \Lambda_1/\lambda_1 - 1 = 0$,

$\Lambda_2 = 200$ ч$^{-1}$, $\lambda_2 = 200\, E_6(10) = 96{,}9$ ч$^{-1}$, $k_2 = \Lambda_2/\lambda_2 - 1 \approx 1$,

$\Lambda_s \approx 150$ ч$^{-1}$, $\lambda_s \approx 200$ ч$^{-1}$, $k = \Lambda_s/\lambda_s - 1 < 0$.

Суммарный поток – простейший с функцией распределения

$$F(t) = 1 - \exp(-200\, t).$$

ЗАДАЧА 12. Классифицировать суммарный поток, полученный объединением простейшего потока с параметром $\Lambda_1 = 100$ ч$^{-1}$ и потока с параметром $\Lambda_2 = 200$ ч$^{-1}$, потерянном на 17–линейном пучке. Время занятия – 3 мин.

Решение.

$\mu = 20$ ч$^{-1}$, A2 = $\Lambda_2 / \mu = 10$ Эрл,

$\Lambda_1 = 100$ ч$^{-1}$, $\lambda_1 = 100$ ч$^{-1}$, $k_1 = 0$,

$\Lambda_2 = 200$ ч$^{-1}$, $\lambda_2 = 200\, E_{17}(10) = 2{,}6$ ч$^{-1}$, $k_2 = 76$,

$\Lambda_s = 102{,}5$ ч$^{-1}$, $\lambda_s = 102{,}6$ ч$^{-1}$, $k < 0$.

Суммарный поток – простейший с функцией распределения

$$F(t) = 1 - \exp(-102{,}6\, t).$$

ЗАДАЧА 13. Классифицировать суммарный поток, полученный объединением потока с параметром $\Lambda_1 = 100$ ч$^{-1}$, потерянном на шестилинейном пучке и потока с параметром $\Lambda_2 = 200$ ч$^{-1}$, потерянном на девятилинейном пучке. Время занятия – 3 мин.

Решение.

$\mu = 20$ ч$^{-1}$, $A1 = \Lambda_1/\mu = 5$ Эрл, $A2 = \Lambda_2/\mu = 10$ Эрл,

$\Lambda_1 = 100$ ч$^{-1}$, $\lambda_1 = 100\ E_6(5) \approx 19$ ч$^{-1}$, $k_1 \approx 4$,

$\Lambda_2 = 200$ ч$^{-1}$, $\lambda_2 = 200\ E_9(10) = 55$ ч$^{-1}$, $k_2 \approx 3$,

$\Lambda_s = 174{,}2$ ч$^{-1}$, $\lambda_s = 74$ ч$^{-1}$, $k = 1{,}35$.

Суммарный поток – поток Эрланга 1,35-порядка с функцией распределения

$$F(t) = 1 - \frac{I(1{,}35; 174{,}2t)}{I(1{,}35; 0)} \exp(-174{,}2t).$$

ЗАДАЧА 14. Определить потери по времени, вызовам и нагрузке для $N = 5$, $V = 5$. Удельная нагрузка $a = 0{,}8$ Эрл.

Решение.

В нашем случае $N = V$, поэтому используем модель Энгсета.

$$A = \frac{a}{1-a} = \frac{0{,}8}{1-0{,}8} = 4 \text{ Эрл.}$$

$$P_t = \frac{C_5^5 4^5}{\sum_{i=0}^{5} C_5^i 4^i} = 0{,}327; \quad P_c = \frac{C_4^5 4^5}{\sum_{i=0}^{5} C_4^i 4^i} = 0, \quad P_y = \left(1 - \frac{5}{5}\right) P_t = 0.$$

ЗАДАЧА 15. Оператор использует учрежденческую АТС на $N = 50$ номеров через пучок $V = 30$ линий. Удельная нагрузка 0,5 Эрл. Вычислить потери по времени и вызовам.

Решение.

Пуассоновская нагрузка второго рода $A = a/(1 - a) = 1$ Эрл.

Потери по времени $P_c = E_{N,V}(A) = 0{,}044$.

Потери по вызовам $P_t = E_{N-1,V}(A) = 0{,}035$.

ЗАДАЧА 16. Оператор использует учрежденческую АТС на $N = 50$ номеров через пучок $V = 30$ линий. Измерения потерь по вызовам оказались равными 0,026. Оператор хочет снизить потери до значения не более 0,001. Сколько нужно добавить линий?

Решение.

Из уравнения $P_c = E_{N-1,V}(A)$ при известных значениях $P_c = 0{,}026$; $N = 50$, $V = 30$ находим пуассоновскую нагрузку второго рода $A = 0{,}95$ Эрл.

При известных значениях $P_c = 0{,}001$, $N = 50$, $A = 0{,}95$ Эрл находим, что уравнение $P_c = 0{,}001 \geq E_{N-1,V}(A)$ удовлетворяется при емкости пучка $V = 35$. Следовательно, нужно добавить пять линий.

5. СИСТЕМА С ОЖИДАНИЕМ БЕЗ ПОТЕРЬ (Blocked Calls Queued, BCQ)

5.1. Система M/M/V. Стационарные вероятности

В системе с ожиданием без потерь при недостатке обслуживающих приборов блокированные вызовы (пакеты) становятся в ***бесконечную*** очередь и обслуживаются в порядке поступления. Во многих практических случаях число мест ожидания довольно велико, поэтому результаты расчета такой системы вполне пригодны для инженерных оценок.

Диаграмма переходов Марковского процесса системы обслуживания M/M/V с бесконечным числом мест ожидания изображена на рис. 20.

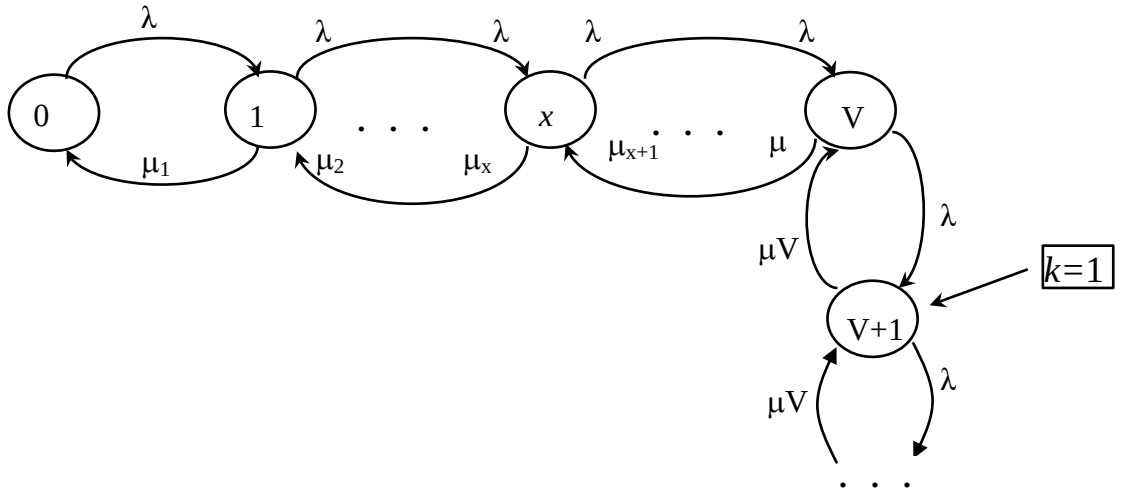


Рис. 20. Диаграмма переходов системы M/M/V

Модель справедлива при следующих предположениях:

– вызовы, поступающие на вход системы, образуют пуассоновский поток постоянной интенсивности с параметром λ;

– длительность занятия подчиняется экспоненциальному распределению с параметром μ (в системе коммутации пакетов $\mu = VEL/LEN$, где VEL – скорость передачи интерфейса, LEN – средняя длина пакета);

– любой из V выходов пучка доступен, когда он свободен, для любого поступающего вызова;

– вызов, не принятый к обслуживанию в момент занятости V линий пучка, поступает в бесконечную очередь и обслуживается в порядке поступления (дисциплина FIFO);

– исходной для расчета является поступающая нагрузка;

– система находится в стационарном режиме.

Используя уравнения стационарного процесса рождения и гибели (3.1), **при** $\boldsymbol{A} < \boldsymbol{V}$ имеем

$$[x]=\begin{cases}\dfrac{(\lambda/\mu)^x}{x!}[0], & 0\le x\le V,\\[2ex] \dfrac{(\lambda/\mu)^V}{V!}\left(\dfrac{\lambda}{V\mu}\right)^{x-V}[0], & x\ge V,\end{cases}$$

или, принимая $A = \lambda/\mu$,

$$[x]=\begin{cases}\dfrac{A^x}{x!}[0], & 0\le x\le V,\\[2ex] \dfrac{A^V}{V!}\left(\dfrac{A}{V}\right)^{x-V}[0], & x\ge V,\end{cases}$$

где [0] определяется из нормирующего условия

$$[0]^{-1}=\sum_{x=0}^{V}\frac{A^x}{x!}+\frac{A^V}{V!}\sum_{x=V+1}^{\infty}\left(\frac{A}{V}\right)^{x-V}\Bigg|_{A<V}=\sum_{x=0}^{V}\frac{A^x}{x!}+\frac{A^V}{V!}\frac{A}{V-A},$$

а вероятность состояния *{V}* равна

$$[V]=\frac{A^V}{V!}[0]=\frac{(V-A)E_V(A)}{V-A+A\cdot E_V(A)}.$$

Средняя длина очереди равна $Qw=\sum\limits_{k=V+1}^{\infty}(k-V)\cdot[k]=\dfrac{A\cdot V}{(V-A)^2}\cdot[V]$,

среднее число вызовов в обслуживании $Qs=A$,

среднее число вызовов в системе $Qe=Qw+Qs=\dfrac{A\cdot V}{(V-A)^2}\cdot[V]+A$.

Потери по времени равны вероятности того, что вызов не будет обслужен немедленно и встанет в очередь:

$$P_t=P(>0)=\sum_{x=V}^{\infty}[x]=\frac{A^V}{V!}[0]\sum_{x=V}^{\infty}\left(\frac{A}{V}\right)^{x-V}=$$
$$=\frac{V}{V-A}[V]=\frac{V\cdot E_V(A)}{V-A+A\cdot E_V(A)}<1. \qquad (5.1)$$

(5.1) – вторая формула Эрланга.

Потери по вызовам равны $P_c=[V+K]_{|K\to\infty}=0$.

За среднее время пребывания вызова в системе все ранее поступившие вызовы будут обслужены, т. е. покинут ее. Поэтому среднее число вызовов,

находящихся в системе Qe, равно произведению параметра обслуженного потока λ на среднее время пребывания вызова в системе te:

$$Qe = \lambda(tw + ts) = \lambda \cdot te. \qquad (5.2)$$

(5.2) – формула Литтла.

Используя формулу Литтла, находим:

– среднее время ожидания начала обслуживания последнего вызова в очереди

$$tw = \frac{Qw}{\lambda} = \frac{1}{\mu}\frac{V}{(V-A)^2}[V] = \frac{1}{\mu}\frac{V}{V-A}\frac{E_V(A)}{V-A+A\cdot E_V(A)},$$

– среднее время ожидания конца обслуживания (среднее время пребывания вызова в системе)

$$te = \frac{Qe}{\lambda} = \frac{Qw+Qs}{\lambda} = \frac{1}{\mu}\frac{V}{(V-A)^2}[V] + \frac{1}{\mu} = \frac{1}{\mu}\left\{\frac{V}{V-A}\frac{E_V(A)}{V-A+A\cdot E_V(A)} + 1\right\}.$$

При $A = V$ все вероятности состояний стремятся к нулю.

Потери по времени $P_t = \dfrac{V\cdot E_V(A)}{V-A+A\cdot E_V(A)_{|A=V}} = 1.$

Потери по вызовам $P_c = [V+K]_{|K\to\infty} = 0$. Стремление к нулю всех состояний означает неопределенность места ожидания, при этом время пребывания пакета в маршрутизаторе может оказаться недопустимо большим. За время tep в буфере маршрутизатора накопится $Qw = \lambda \cdot tep - A$ пакетов. Для ограничения времени пребывания пакетов значением $tep > 1/\mu$ следует ограничить объем буфера очереди до величины $R\max = Qw \cdot LENp = (\lambda \cdot tep - A) \cdot LENp$ байт, где $LENp$ – длина пакета без полей канального и физического уровней.

При $A > V$ стационарного процесса обслуживания вызовов нет: длина очереди будет монотонно стремиться к бесконечности, каждую единицу времени занятия очередь будет увеличиваться на $(A - V)$ вызовов, $P_t = P_c = 1$.

5.2. Система M/M/V. Функция распределения

Обозначим через $P(>t)$ вероятность ожидания начала обслуживания больше, чем t, а через $P_k(>t)$ – условную вероятность ожидания начала обслуживания при поступлении вызова в момент, когда в системе находятся k вызовов, из них V вызовов обслуживаются, $(k - V)$ вызовов ожидают обслуживания, а поступивший вызов встанет на $(k - V + 1)$-е место ожидания. По формуле полной вероятности

$$P(>t)=\sum_{k=V}^{\infty}[k]P_k(>t).$$

Вероятность $P_k(>t)$ есть вероятность того, что за время t произойдет не более 1, 2, ...(k – V) освобождений и вычисляется с использованием распределения Пуассона

$$P_k(>t)=\sum_{i=0}^{k-V}\frac{(V\mu t)^i}{i!}e^{-V\mu t}.$$

Поэтому

$$P(>t)=\sum_{k=V}^{\infty}[k]\sum_{i=0}^{k-V}\frac{(V\mu t)^i}{i!}e^{-V\mu t}=$$

$$=\frac{A^V}{V!}[0]e^{-V\mu t}\left(\sum_{k=V}^{\infty}(A/V)^{k-V}\sum_{i=0}^{k-V}\frac{(V\mu t)^i}{i!}\right)_{|V>A}=$$

$$=\frac{A^V}{V!}\frac{1}{1-A/V}[0]e^{-(V\mu-\lambda)t}=P(>0)e^{-(V\mu-\lambda)t}.$$

Функция распределения времени (ФРВ) ожидания начала обслуживания $F_w(t)$ за время t

$$F_w(t)=1-P(>t)=1-P(>0)e^{-(V\mu-\lambda)t}, \qquad (5.3)$$

а ФРВ самого обслуживания – $F_s(t)=1-e^{-\mu t}$, $\lambda>0$.

ФРВ конца обслуживания вызова находим, используя преобразование Лапласа:

$$f_w(z)=z\int_0^{\infty}F_w(t)e^{-zt}dt=\frac{V\mu-\lambda+z-zP(>0)}{V\mu-\lambda+z}, \qquad f_s(z)=z\int_0^{\infty}F_s(t)dt=\frac{\mu}{\mu+z},$$

$$F_e(z)=\frac{1}{z}\cdot\frac{V\mu-\lambda+z-zP(>0)}{V\mu-\lambda+z}\cdot\frac{\mu}{\mu+z},$$

$$F_e(t)=\left\{\begin{array}{ll}1+\dfrac{P(>0)}{V-A-1}[e^{-\mu(V-A)t}-e^{-\mu t}]-e^{-\mu t}, & 0<A\neq V-1,\\ 1-\mu tP(>0)e^{-\mu t}-e^{-\mu t}, & 0<A=V-1,\end{array}\right\} \qquad (5.4)$$

где $P(>0)=\dfrac{V\cdot E_V(A)}{V-A+A\cdot E_V(A)}$, $\qquad A=\dfrac{\lambda}{\mu}<V$.

5.3. Система M/M/1

Принимая в предыдущих формулах значение $V = 1$, $0 < A = a < 1$, получим стационарные вероятности состояния $\{x\}$

$$[0]^{-1} = \frac{1}{1-a}, \qquad [x] = a^x(1-a), \qquad [1] = a(1-a),$$

Вероятность ожидания начала обслуживания (потери по времени)

$$P_t = P(>0) = \frac{V}{V-A}[V]_{|V=1} = a,$$

– среднее число вызовов в очереди (место ожидания последнего вызова)

$$Qw = \sum_{k=V+1}^{\infty}(k-V)\cdot[k] = \frac{A\cdot V}{(V-A)^2}\cdot[V]_{|V=1} = \frac{a^2}{1-a};$$

– среднее число вызовов в обслуживании $Qs = a$;

– среднее число вызовов в системе

$$Qe = Qw + Qs = \frac{A\cdot V}{(V-A)^2}\cdot[V] + A_{|V=1} = \frac{a^2}{1-a} + a = \frac{a}{1-a};$$

– среднее время ожидания начала обслуживания

$$tw = \frac{Qw}{\lambda} = \frac{1}{\mu}\frac{V}{(V-A)^2}[V]_{|V=1} = \frac{1}{\mu}\frac{a}{1-a};$$

– среднее время пребывания вызова в системе

$$te = tw + ts = \frac{1}{\mu}\frac{a}{1-a} + \frac{1}{\mu} = \frac{1}{\mu}\frac{1}{1-a}.$$

ФРВ ожидания начала обслуживания $F_w(t) = 1 - a\cdot e^{-(1-a)\mu t}$; ФРВ конца обслуживания

$$F_e(t) = 1 - e^{-(1-a)\mu t}, \quad a > 0. \tag{5.5}$$

5.4. Система M/D/V

На полнодоступный пучок из V линий поступает простейший поток вызовов с параметром λ. Длительность занятия (обслуживания) постоянна и равна μ. Если в момент поступления вызова все линии заняты, то вызов ставится в ожидание и обслуживается в порядке поступления. Это имеет место в сетях АТМ, где используются пакеты постоянной длины.

Для определения вероятности того, что время ожидания начала обслуживания будет больше, чем t, воспользуемся тем, что за время $t_0 = 1/\mu$ система обслуживает V вызовов; следовательно, вызов будет находиться на ожидании более t только в том случае, если в системе на ожидании находится не менее $V\mu t$ вызов, т. е.

$$P(>t) = \sum_{k=V\mu t}^{\infty} [k] = \sum_{k=V\mu t}^{\infty} \frac{A^V}{V!}\left(\frac{A}{V}\right)^k [0] = \frac{E_V(A)}{1-\frac{A}{V}\left(1-E_V(A)\right)}\left(\frac{A}{V}\right)^{V\mu t} = P(>0)\left(\frac{A}{V}\right)^{V\mu t},$$

откуда ФРВ ожидания начала обслуживания (вероятность окончания времени ожидания начала обслуживания за время t)

$$F_w(t) = 1 - P(>t) = 1 - P(>0)\left(\frac{A}{V}\right)^{V\mu t}, \qquad 0 < A < V,$$

а ФРВ самого обслуживания

$$F_s(t) = \begin{cases} 0, & t < 1/\mu, \\ 1, & t \geq 1/\mu. \end{cases}.$$

ФРВ конца обслуживания вызова находим, используя преобразование Лапласа:

$$f_w(z) = z\int_0^{\infty} F_w(t)e^{-zt} = \frac{z - V\mu\ln(A/V) - zP(>0)}{z - V\mu\ln(A/V)},$$

$$f_s(z) = z\int_0^{\infty} F_s(t)e^{-zt} = z\int_{1/\mu}^{\infty} e^{-zt}dt = e^{-z/\mu}.$$

$$F_e(z) = \frac{1}{z}\cdot\frac{z - V\mu\ln(A/V) - zP(>0)}{z - V\mu\ln(A/V)}\cdot e^{-z/\mu}.$$

$$F_e(t) = 1 - P(>0)e^{-V\ln(A/V)}e^{V\mu t\ln(A/V)} = 1 - P(>0)\left(\frac{A}{V}\right)^{-V}\left(\frac{A}{V}\right)^{V\mu t} =$$

$$= 1 - P(>0)\left(\frac{A}{V}\right)^{V(\mu t - 1)},$$

где $P(>0) = \dfrac{V\cdot E_V(A)}{V - A\cdot\left(1-E_V(A)\right)}, \qquad A = \dfrac{\lambda}{\mu}, \quad t \geq 1/\mu.$

При $A > V$ стационарного процесса обслуживания вызовов нет: длина очереди будет монотонно стремиться к бесконечности.

5.5. Система M/D/1

Для однолинейной системы M/D/1 ($V = 1$, $0 < A/V = a < 1$, $P(>0) = a$)

$$F_w(t) = 1 - a \cdot a^{\mu t}, \qquad 0 < a < 1,$$

$$F_s(t) = \begin{cases} 0, & t < 1/\mu, \\ 1, & t \geq 1/\mu. \end{cases}$$

ФРВ окончания обслуживания вызова в системе M/D/1

$$F_e(t) = 1 - a \cdot a^{\mu t - 1} = 1 - a^{\mu t}, \qquad t \geq 1/\mu. \tag{5.6}$$

Заметим, что (5.6) имеет изображение

$$f_e(z) = z\int_0^\infty (1 - a^{\mu t}) \cdot e^{-zt} = z\int_0^\infty (1 - e^{\mu t \ln a}) \cdot e^{-zt} dt = 1 - \frac{z}{z - \mu \ln a} = \frac{-\mu \ln a}{z - \mu \ln a}.$$

Среднее время ожидания начала обслуживания

$$tw = -\frac{df_w(z)}{dz}_{|z=0} = -\frac{(1-a)(z - \mu \ln a) - (z - \mu \ln a - az)}{(z - \mu \ln a)^2}_{|z=0} =$$

$$= \frac{(1-a)\mu \ln a - \mu \ln a}{(\mu \ln a)^2} = -\frac{a}{\mu \ln a}.$$

Суммируя вышеизложенное для системы M/D/1 при $0 < a < 1$, получаем: вероятность постановки в очередь (потери по времени)

$$P(>0) = A = a = \lambda / \mu,$$

– среднее число вызовов в очереди (место ожидания последнего вызова)

$$Qw = \lambda \cdot tw = -\frac{a^2}{\ln a},$$

– среднее число вызовов в обслуживании

$$Qs = a,$$

– среднее число вызовов в системе

$$Qe = Qw + Qs = \frac{-a^2}{\ln a} + a = \frac{a \cdot \ln a - a^2}{\ln a},$$

– среднее время ожидания начала обслуживания

$$tw = \frac{Qw}{\lambda} = -\frac{a}{\mu}\frac{1}{\ln a},$$

– среднее время пребывания вызова в системе

$$te = tw + \frac{1}{\mu} = \frac{1}{\mu}\frac{\ln a - a}{\ln a}.$$

При $a > 1$ стационарного процесса обслуживания вызовов нет: длина очереди будет монотонно стремиться к бесконечности.

На рис. 21 приведены графики ФРВ конца обслуживания для экспоненциального (Fe exp) и постоянного (Fe const) времени обслуживания. Вычисления проводились по формулам (5.5) и (5.6) при $\lambda = 10$ с$^{-1}$ = 0,01 мс; ts = 7 мс; $V = 1$; $\mu = 1/ts$ = 0,1428 мс$^{-1}$; $a = \lambda/\mu$ = 0,07 Эрл.

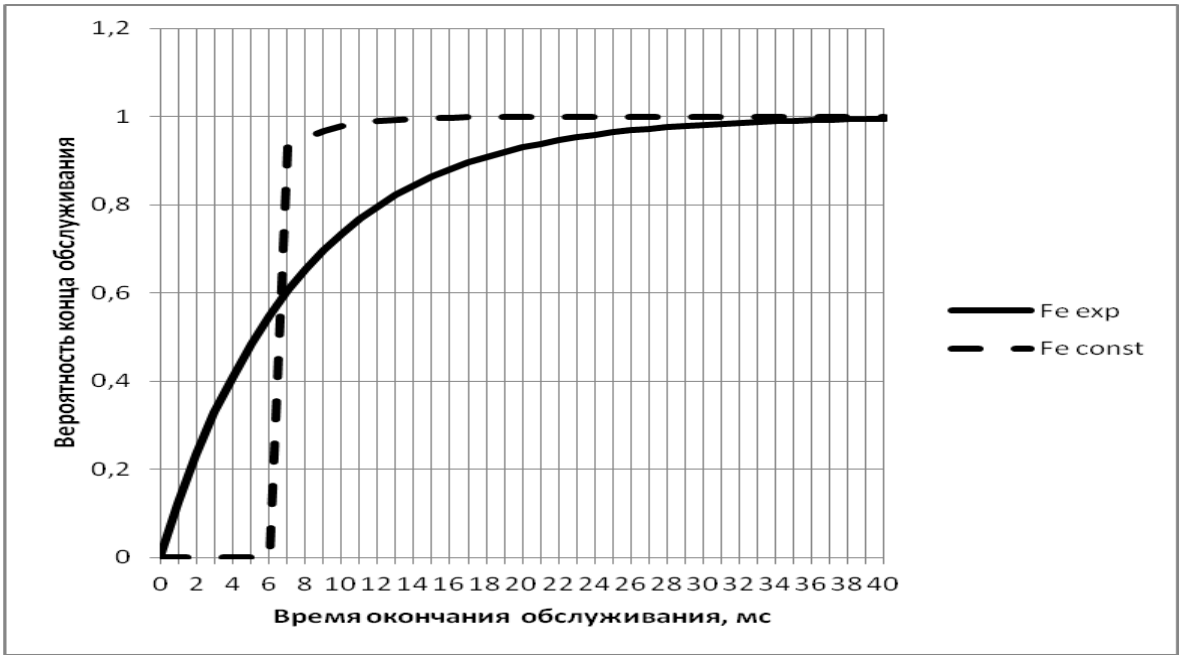


Рис. 21. ФРВ конца обслуживания для экспоненциального и постоянного времени обслуживания при $\lambda = 10$ с$^{-1}$ = 0,01 мс$^{-1}$; ts = 7 мс; $V = 1$

5.6. Однолинейная многофазная система обслуживания

Предположим, что вызов последовательно обслуживается цепью из S элементов вида M/M/1, каждый из которых часть поступающих вызовов обслуживает полностью, а другую часть вызовов обслуживает частично и передает их на дообслуживание в последующие элементы. Полностью обслуженные вызовы покидают систему. В последнем элементе все вызовы покидают систему. Такую структуру может иметь IP-сеть, в которой пакеты следуют через проходные локальные сети (Local Area Network,

LAN). Математически такая модель в теории телетрафика описывается однолинейной (V = 1) многофазной (решение задачи в несколько этапов – фаз) системой массового обслуживания.

Очевидно, наибольшие задержки будут иметь место для той части вызовов, которая проходит через все элементы системы. Для этой части потока ФРВ конца обслуживания вызова в системе имеет вид:

$$F_e(t) = L^{-1}[F_e(z)] = F_{1e}(t) * F_{2e}(t) * ... F_{Se}(t), \tag{5.7}$$

вычисление которой производится поэтапно

$$\begin{gathered} f_{ke}(z) = z\ L\left[F_{ke}(t)\right], \quad k = 1, 2, ...S, \\ f_e(z) = \prod_{k=1}^{S} f_{ke}(z), \quad F_e(z) = \frac{1}{z} f_e(z), \\ F_e(t) = L^{-1}\left[F_e(z)\right], \quad T_e = -\frac{df_e(z)}{dz}|_{z=0}, \end{gathered} \tag{5.8}$$

где $F_e(z)$ – изображение суммарной ФРВ конца обслуживания, L^{-1} – обратное преобразование Лапласа, $F_{ke}(t)$ – ФРВ конца обслуживания вызова k-м (k = 1,…S) элементом, * – символ свертки, $f_{ke}(t)$ и $f_{ke}(z)$ – плотности ФРВ конца обслуживания вызова k-м элементом и ее изображение, L – преобразование Лапласа, $F_e(t)$ и $F_e(z)$ – ФРВ конца обслуживания вызова k-м элементом и ее изображение, T_e – среднее время ожидания конца обслуживания вызова.

Рассмотрим однолинейную двухфазную систему на рис. 22. Первый элемент имеет производительность β_1 и обслуживает поток с параметром λ_1. Часть потока λ_{S1} он обслуживает самостоятельно, а оставшуюся часть ($\lambda_1 - \lambda_{S1}$) передает на дообслуживание во второй элемент.

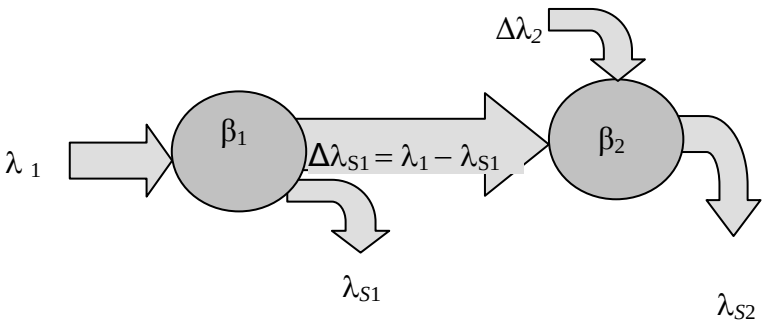


Рис. 22. Однолинейная двухфазная система

Второй элемент имеет производительность β_2 и обслуживает поток с параметром $\lambda_2 = \lambda_1 - \lambda_{S1} + \Delta\lambda_2$, образованный остаточным потоком

с первого элемента и потоком от собственных пользователей. Покажем, что при фиксированной сумме ($\beta_1 + \beta_2$) наивысшая пропускная способность общего потока достигается, если выполняется равенство $D_1 = (\beta_1 - \lambda_1) = D_2 = (\beta_2 - \lambda_2) = D$. Для доказательства ***Положения*** достаточно показать, что $F_{D_1=D_2}(t) - F_{D_1\neq D_2}(t) > 0$.

Используя (5.7) – (5.8), получим:

$$f_{1e}(z) = L\left[\left(\beta_1-\lambda_1\right)\exp\left\{-\left(\beta_1-\lambda_1\right)t\right\}\right] = \int_0^\infty \left(\beta_1-\lambda_1\right)e^{-\left(\beta_1-\lambda_1+z\right)t}dt = \frac{\beta_1-\lambda_1}{\beta_1-\lambda_1+z},$$

$$f_{2e}(z) = \frac{\beta_2-\lambda_2}{\beta_2-\lambda_2+z},$$

$$F_e(z) = \frac{1}{z}\frac{\left(\beta_1-\lambda_1\right)}{\left(\beta_1-\lambda_1+z\right)}\frac{\left(\beta_2-\lambda_2\right)}{\left(\beta_2-\lambda_2+z\right)} = \frac{D_1D_2}{z\left(D_1+z\right)\left(D_2+z\right)}.$$

$$F_{D_1=D_2}(t) = 1 - e^{-Dt} - Dte^{-Dt}, \qquad \text{при } D_1 = D_2,$$

$$F_{D_1\neq D_2}(t) = 1 - \frac{D_2}{D_2-D_1}e^{-D_1t} - \frac{D_1}{D_1-D_2}e^{-D_2t}, \quad \text{при } D_1 \neq D_2.$$

Положим $D_1 = D - \delta$, а $D_2 = D + \delta$. Тогда

$$F_{D_1=D_2}(t) - F_{D_1\neq D_2}(t) = \left[\frac{D}{2\delta}\left(e^{\delta t} - e^{-\delta t}\right) + \frac{1}{2}\left(e^{\delta t} + e^{-\delta t}\right) - (1+Dt)\right]e^{-Dt} =$$

$$= \left\{\frac{D}{\delta}\left[\frac{(\delta t)^1}{1!} + \frac{(\delta t)^3}{3!} + \ldots\right] + \left[\frac{(\delta t)^0}{0!} + \frac{(\delta t)^2}{2!} + \ldots\right] - (1+Dt)\right\}e^{-Dt} > \qquad (5.9)$$

$$> \left\{\frac{D}{\delta}\left[\frac{(\delta t)^1}{1!} + \frac{(\delta t)^3}{3!} + \ldots\right] - Dt\right\}e^{-Dt} = \left(\delta^2\frac{t^3}{3!} + \delta^4\frac{t^5}{5!} + \ldots\right)D\,e^{-Dt} > 0.$$

Положение доказано. Из него следует, что минимизация суммарной производительности однолинейной многофазной системы состоит в таком распределении мощностей (ресурсов) системы или поступающих потоков вызовов, которое приводит к выравниванию среднего времени ожидания конца обслуживания вызова между элементами системы. Это объясняется тем, что при неравенстве среднего времени ожидания элементы с меньшей производительностью перегружены, а с большей – простаивают в ожидании поступления следующего вызова. При равенстве времени ожидания моменты окончания обслуживания предыдущего вызова и поступления нового вызова совпадают, поэтому суммарное время обслуживания вызова минимально, а пропускная способность максимальна.

Проведем анализ суммарной производительности однолинейной двухфазной системы с однофазной системой по среднему времени ожидания конца обслуживания вызова, проходящего через обе фазы обслуживания.

Составим табл. 5.1 соответствия вероятностей временам окончания обслуживания первого потока согласно рис. 22.

Таблица 5.1

Зависимость среднего времени обслуживания распределения нагрузки

Вероятность окончания времени обслуживания вызова первого потока двухфазной системой	Значение среднего времени окончания обслуживания вызова первого потока двухфазной системой
λ_{S1}/λ_1	$(\beta_1-\lambda_1)^{-1}$
$(\lambda_1-\lambda_{S1})/\lambda_1$	$(\beta_1-\lambda_1)^{-1}+(\beta_2-\lambda_2)^{-1}$

Определим среднее время окончания обслуживания вызова как математическое ожидание:

$$T_e=\frac{\lambda_{S1}}{\lambda_1}\frac{1}{\beta_1-\lambda_1}+\frac{\lambda_1-\lambda_{S1}}{\lambda_1}\left[\frac{1}{\beta_1-\lambda_1}+\frac{1}{\beta_2-\lambda_1+\lambda_{S1}-\Delta\lambda_2}\right]\Bigg|_{\lambda_{S1}=\lambda_1}=\frac{1}{\beta_1-\lambda_1}.$$

Из анализа зависимости T_e от λ_{S1} видно, что во всем диапазоне λ_{S1} = 0, …λ_1 она представляет собой монотонно убывающую функцию, достигающую своего минимума, равного $(\beta_1-\lambda_1)^{-1}$ при $\lambda_{S1}=\lambda_1$.

Из вышеизложенного следует, что для достижения одной и той же пропускной способности многофазной однолинейной системе требуется более высокая суммарная производительность, чем в случае ее реализации в виде однофазной. Тем не менее, многофазные однолинейные системы широко применяются в сетях связи при поэтапном обслуживании вызовов через географически удаленные элементы сети (маршрутизаторы).

Из (5.9) следует, что ФРВ конца обслуживания вызова S-фазной однолинейной оптимальной системы, имеющей минимальную суммарную производительность, вычисляется по формуле

$$F_e(t)=1-\sum_{k=0}^{S-1}\frac{(D\cdot t)^k}{k!}e^{-D\,t}, \qquad (5.10)$$

где $D=(\mu-\lambda)$ – интенсивность прохождения одной фазы с учетом прохождения очередей, D^{-1} – среднее время прохождения вызова через одну фазу обслуживания (среднее время ожидания конца обслуживания вызова в одном сетевом элементе).

ЗАДАЧА 1. Сеть оператора связана с сетью Интернет интерфейсом со скоростью *VEL* = 8 Мбит/с = 1 Мбайт/с. Средняя длина пакета *LEN* = 500 байт. Оператор желает обеспечить среднюю задержку прохождения пакетов по интерфейсу не более *te* = 5 мс. Какой параметр потока должен быть оговорен в SLA, чтобы провайдер услуг мог обеспечить требования оператора?

Решение.

Интенсивность обслуживания пакетов μ = *VEL/LEN* = 2000 $с^{-1}$. Из уравнения $te = \frac{1}{\mu - \lambda}$ при известных значениях *te* = 0,005 с; μ = 2000 $с^{-1}$; находим λ = 1800 $с^{-1}$. Загрузка интерфейса составит *a* = λ/μ = 0,9 Эрл или 90 %.

ЗАДАЧА 2. Сеть оператора связана с сетью Интернет интерфейсом со скоростью *VEL* = 8 Мбит/с = 1 Мбайт/с. Средняя длина пакета *LEN* = 500 байт. Оператор желает обеспечить задержку по интерфейсу не более *te* = 5 мс для 95 % пакетов. Какой параметр потока должен быть оговорен в SLA, чтобы провайдер услуг мог обеспечить требования оператора?

Решение.

Интенсивность обслуживания пакетов μ = *VEL/LEN* = 2000 $с^{-1}$. Из уравнения $F_e(t) = 1 - e^{-(\mu-\lambda)\ t} = 0{,}95$ при известных значениях *te* = 0,005 с; μ = 2000 $с^{-1}$; находим λ = 1400 $с^{-1}$. Загрузка интерфейса составит *a* = λ/μ = = 0,7 Эрл или 70 %.

Ужесточение требований к качеству обслуживания по сравнению с задачей 1, естественно приводит к необходимости снижения параметра потока (или к расширению полосы передачи).

ЗАДАЧА 3. Сеть оператора состоит из двух локальных сетей N1 и N2 и имеет структуру и распределение потоков, изображенную на рис. 23.

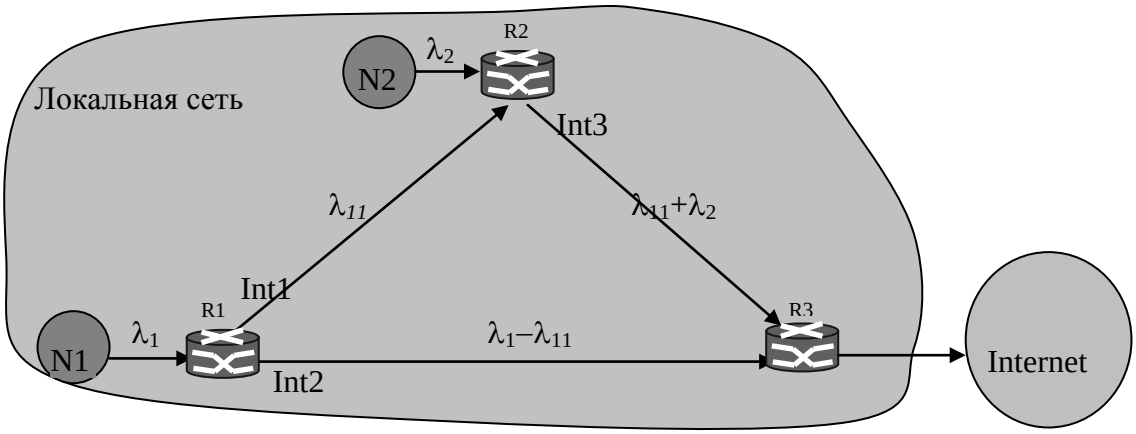


Рис. 23. К задаче 3

Имеются следующие значения параметров потоков в ЧНН: $\lambda_1 = 10$ с$^{-1}$, $\lambda_2 = 10$ с$^{-1}$. Скорости интерфейсов Int1, Int2, Int3 равны 2 Мбит/с, 1 Мбит/с, 2 Мбит/с, соответственно. Средняя длина пакета *LEN* = 1000 байт. Определить оптимальное распределение потока λ_1 в маршрутизаторе R1 между маршрутизаторами R2 и R3, чтобы обеспечивалось одинаковое качество обслуживания для всех пользователей.

Решение.

Согласно *Положению* должно выполняться следующие условия: $T1 = (\mu_1 - \lambda_{11})^{-1} + (\mu_3 - \lambda_{11} - \lambda_2)^{-1} = (\mu_2 - \lambda_1 + \lambda_{11})^{-1} = T2$, где $\mu_1 = VEL1/LEN$=250 с$^{-1}$, $\mu_2 = VEL2/LEN$ = 125 с$^{-1}$, $\mu_3 = VEL3/LEN$ = 250 c^{-1}. Результаты расчета приведены в табл. 5.2.

Таблица 5.2

Оптимальное распределение потоков

λ_{11}	0	1	2	3	4	5	6	7	8	9	10
T_1	8,16	8,2	8,23	8,26	8,30	8,33	8,37	8,4	8,44	8,47	8,51
T_2	8,69	8,62	8,54	8,47	8,40	8,33	8,26	8,19	8,13	8,06	8,0

Из данных таблицы становится очевидным, что результат достигается при $\lambda_1 = 5$ с$^{-1}$, т. е. поток от локальной сети N1 должен распределяться между интерфейсами Int1 и Int2 поровну.

6. СИСТЕМА С ОЖИДАНИЕМ И ПОТЕРЯМИ

6.1. Система M/M/V/K. Стационарные вероятности

В системе с ожиданием и потерями при недостатке обслуживающих приборов блокированные вызовы становятся в ***конечную*** очередь и обслуживаются в порядке поступления. Такой принцип обслуживания применяется в системе с коммутацией пакетов, например, в маршрутизаторах с ограниченной очередью при использовании алгоритма «с наибольшими усилиями» (Best Effort).

Диаграмма переходов Марковского процесса системы обслуживания M/M/V/K с конечным числом мест ожидания *K* изображена на рис. 24.

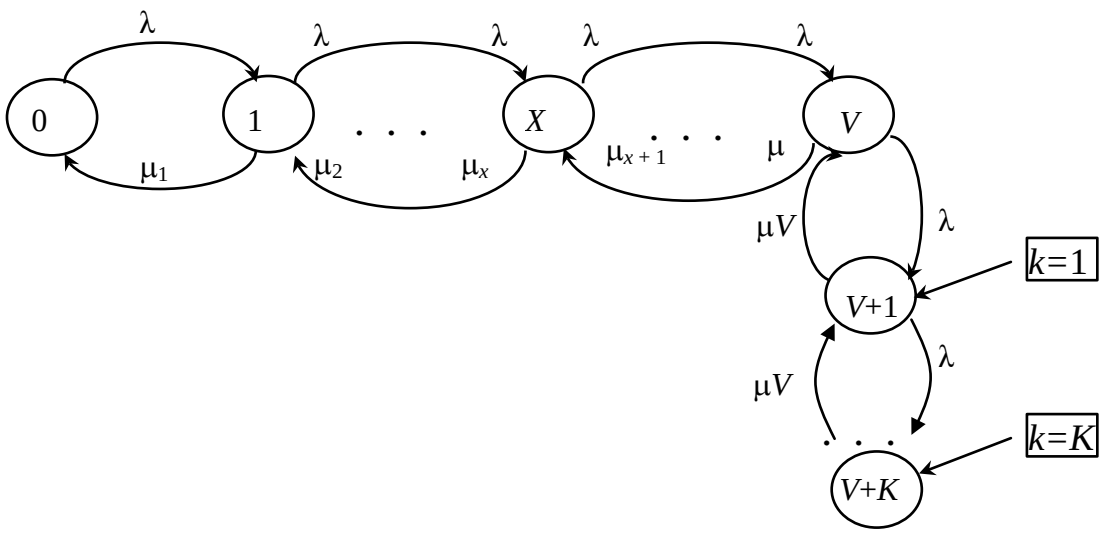


Рис. 24. Диаграмма переходов системы M/MV/K

Используя уравнения стационарного процесса рождения и гибели (3.1) **при** $\boldsymbol{A \neq V}$, имеем:

$$[x] = \begin{cases} \dfrac{A^x}{x!}[0], & 0 \le x \le V, \\ \dfrac{A^V}{V!}\left(\dfrac{A}{V}\right)^{x-V}[0], & V < x \le K, \end{cases}$$

где [0] определяется из нормирующего условия:

$$[0]^{-1} = \sum_{x=0}^{V} \frac{A^x}{x!} + \frac{A^V}{V!}\sum_{x=V+1}^{V+K}\left(\frac{A}{V}\right)^{x-V} = \sum_{x=0}^{V} \frac{A^x}{x!} + \frac{A^V}{V!}\frac{a(1-a^K)}{1-a}, \quad a = A/V.$$

Потери по времени равны вероятности того, что вызов не будет обслужен немедленно и встанет в очередь:

$$P_t = P(>0) = \sum_{x=V}^{V+K}[x] = \frac{A^V}{V!}\sum_{x=V}^{V+K}\left(\frac{A}{V}\right)^{x-V}[0] = \frac{A^V}{V!}\frac{1-a^{K+1}}{1-a}[0] =$$

$$= \frac{\left(1-a^{K+1}\right)E_V(A)}{1-a+a\left(1-a^K\right)E_V(A)} > 0.$$

Потери по вызовам наступают только в том случае, если система окажется в состоянии *{V + K}*

$$P_c = [V+K] = \frac{a^K(1-a)E_V(A)}{1-a+a\left(1-a^K\right)E_V(A)} > 0\,.$$

Обслуженная нагрузка

$$As = A\cdot\left(1-P_c\right) = A\frac{1-a+a\left(1-a^{K-1}\right)E_V(A)}{1-a+a\left(1-a^K\right)E_V(A)}\,.$$

Если число обслуживающих приборов равно нагрузке, т. е. ***A* = *V***, то

$$[0]^{-1} = \sum_{x=0}^{V}\frac{A^x}{x!} + \frac{A^V}{V!}\sum_{x=V+1}^{V+K}\left(\frac{A}{V}\right)^{x-V} = \sum_{x=0}^{V}\frac{A^x}{x!} + \frac{A^V}{V!}K.$$

Потери по времени равны вероятности того, что вызов не будет обслужен немедленно и встанет в очередь

$$P_t = P(>0) = \sum_{x=V}^{V+K}[x] = \frac{A^V}{V!}\sum_{x=V}^{V+K}\left(\frac{A}{V}\right)^{x-V}[0] = \frac{A^V}{V!}(K+1)[0] = \frac{(K+1)E_V(A)}{1+K\cdot E_V(A)} > 0\,.$$

Потери по вызовам наступают только в том случае, если система окажется в состоянии *{V + K}*:

$$P_c = [V+K] = \frac{E_V(A)}{1+K\cdot E_V(A)} > 0\,.$$

Обслуженная нагрузка

$$As = A\cdot(1-P_c) = A\frac{1+(K-1)E_V(A)}{1+K\cdot E_V(A)}\,.$$

На рис. 25 показано распределение вероятностей состояний для системы M/M/3/20 при *A* = 2/3/4 Эрл.

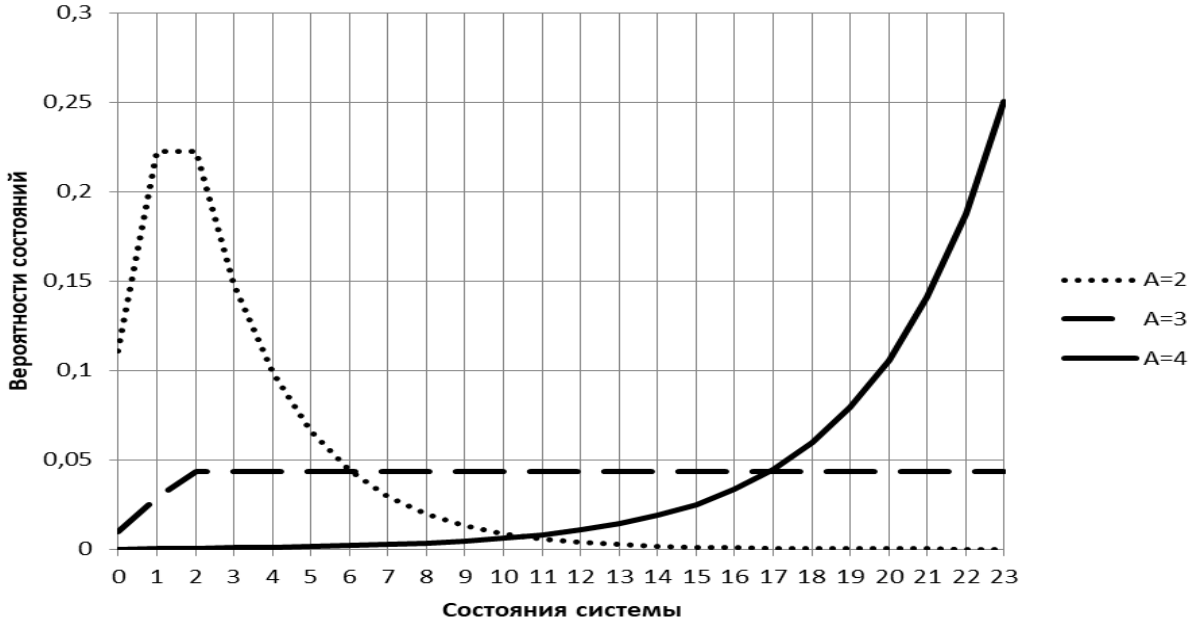


Рис. 25. Распределение вероятностей состояний
системы M/M/3/20

Для ограничения времени пребывания пакета в маршрутизаторе значением $tep > 1/\mu$ следует ограничить объем буфера очереди до величины $R\max = (\lambda \cdot tep - A) \cdot LENp$ байт.

6.2. Система M/M/1/K

Диаграмма переходов Марковского процесса системы обслуживания M/M/1/K изображена на рис. 26.

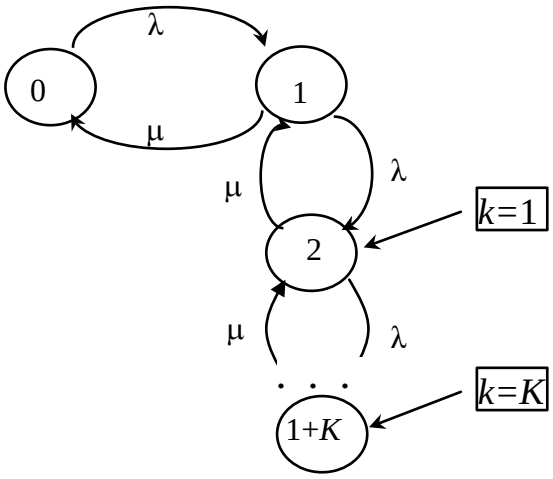


Рис. 26. Диаграмма переходов системы M/M/1/K

Стационарные вероятности состояния *{x}* при **0 < *A* = *a* ≠ 1**

$$[0]^{-1}=\sum_{x=0}^{V}\frac{A^x}{x!}+\frac{A^V}{V!}\sum_{x=V+1}^{V+K}\left(\frac{A}{V}\right)^{x-V}=\sum_{x=0}^{V}\frac{A^x}{x!}+\frac{A^V}{V!}\frac{\left(a-a^{K+1}\right)}{1-a}_{|V=1}=\frac{1-a^{K+2}}{1-a},$$

$$[x]=\frac{a^x(1-a)}{1-a^{K+2}},$$

вероятность постановки в очередь (потери по времени)

$$P_t=P(>0)=\sum_{x=V}^{V+K}[x]=\frac{A^V}{V!}\sum_{x=V}^{V+K}\left(\frac{A}{V}\right)^{x-V}[0]=\frac{A^V}{V!}\frac{1-a^{K+1}}{1-a}[0]_{|V=1}=\frac{a^{K+2}-a}{a^{K+2}-1}>0,$$

потери по вызовам

$$P_c=[1+K]=\frac{a^{K+1}(1-a)}{1-a^{K+2}}>0.$$

Обслуженная нагрузка

$$As=A\cdot(1-P_c)=A\frac{1-a^{K+1}}{1-a^{K+2}}.$$

При ***A* = *a* = 1**

$$\lim_{a\to1}[x]=\lim_{a\to1}a^x\frac{1-a}{1-a^{K+2}}=\lim_{a\to1}\frac{\left(a^x-a^{x+1}\right)}{\left(1-a^{K+2}\right)}=$$

$$=\frac{xa^{x-1}-(x+1)a^x}{-(K+2)a^{K+1}}=\frac{x-(x+1)}{-(K+2)}=\frac{1}{K+2},$$

поэтому потери по времени

$$P_t=P(>0)=\sum_{x=V}^{1+K}[x]=\frac{K+1}{K+2}>0,$$

потери по вызовам

$$P_c=[1+K]=\frac{1}{K+2}>0.$$

Обслуженная нагрузка

$$As=A\cdot\left(1-P_c\right)=A\frac{K+1}{K+2}.$$

На рис. 27 приведено распределение вероятностей состояний системы M/M/1/20 при удельных нагрузках *a* = 0,8; 1; 1,25.

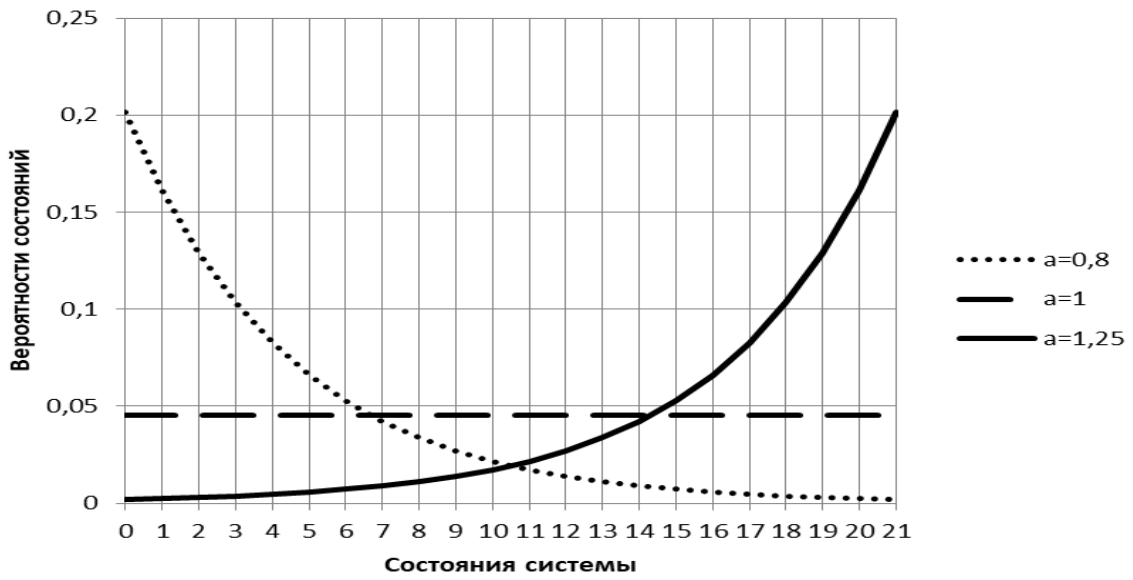


Рис. 27. Распределение вероятностей состояний системы M/M/1/20

Для ограничения времени пребывания пакета в маршрутизаторе значением *tep* следует ограничить объем буфера очереди до величины *R*max = (λ · *tep* – 1) · *LENp* байт, где *LENp* – длина пакета без полей канального и физического уровней.

ЗАДАЧА 1. Определить потери по вызовам для системы M/M/1/1. Удельная нагрузка *a* = 0,5 Эрл.

Решение.

$$P_c = [1+K] = \frac{a^{K+1}(1-a)}{1-a^{K+2}} = \frac{a^2(1-a)}{1-a^3} = 0{,}14.$$

ЗАДАЧА 2. Определить потери по вызовам для системы M/M/1/1. Удельная нагрузка *a* = 1 Эрл.

Решение.

$$P_c = \frac{1}{K+2} = 0{,}33.$$

7. СИСТЕМА С ОЖИДАНИЕМ И ПРИОРИТЕТАМИ

7.1. Схема прохождения вызова

Как уже указывалось, механизм обслуживания очередей маршрутизатора состоит из двух устройств – классификатора пакетов и планировщика очередей. Прохождение пакета на передачу изображено на рис. 28.

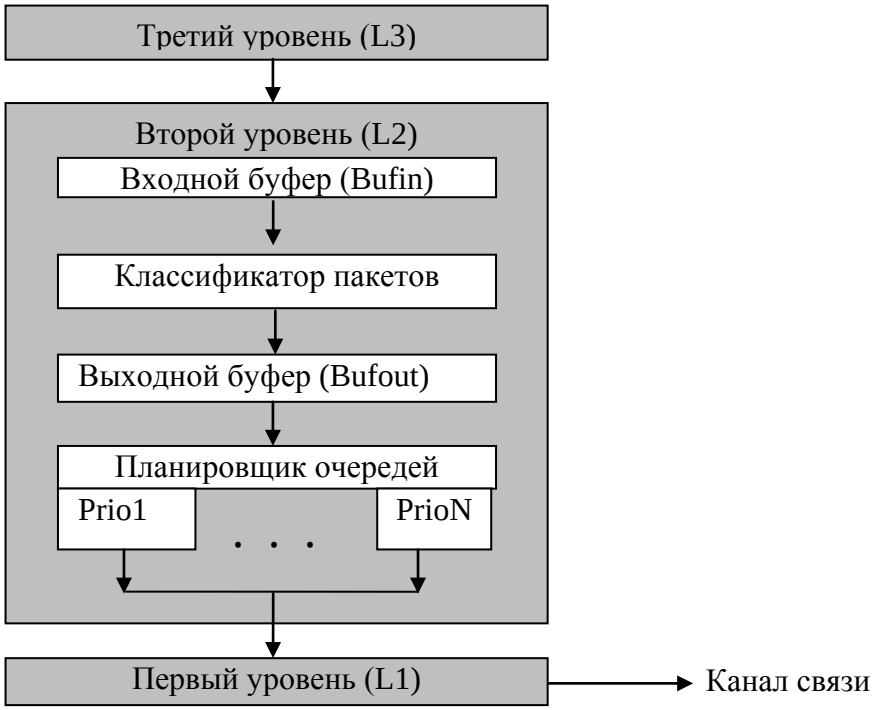


Рис. 28. Прохождение пакета на передачу

Классификатор пакетов (Packet Classifier) выполняет две функции: распознавание пакета и его маркировку (присвоение пакету определенного приоритета или класса). Он устанавливается в каждом пограничном маршрутизаторе. Функция распознавания зависит от различных параметров: IP-адреса, порта источника, поля протокола IP-пакета (ICMP, TCP, UDP, …).

Маркирование пакетов зависит от принятого механизма обеспечения качества обслуживания.

Маркированные пакеты помещаются в буфер Bufout для дальнейшей обработки планировщиком очередей.

Алгоритм функционирования классификатора пакетов изображен на рис. 29.

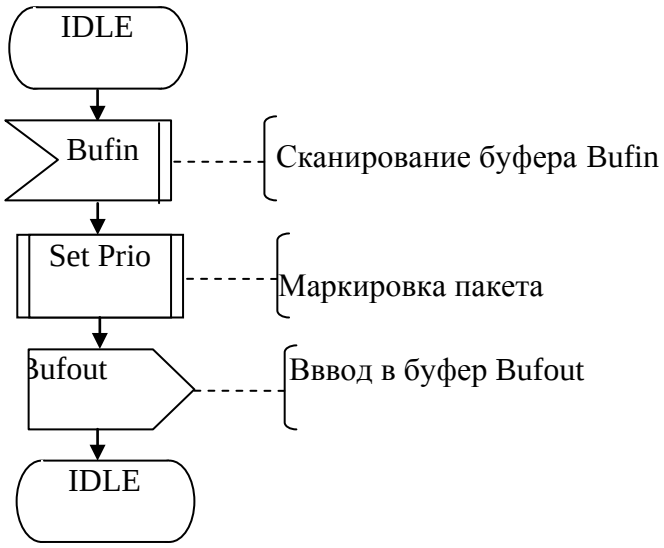


Рис. 29. Алгоритм функционирования классификатора пакетов

Планировщик очередей (Packet Scheduler) организует очереди и использует один из нижеперечисленных механизмов обслуживания очереди: абсолютного приоритета (Absolute Priority Queueing, APQ), взвешенного справедливого обслуживания классов (Class-Based Weighted Fair Queuing, CBWFQ), взвешенного справедливого обслуживания потока (Weighted Fair Queuing, WFQ).

7.2. Система с абсолютными приоритетами M/M/1/ / /NPQ

При использовании маршрутизатором механизма с абсолютными приоритетами (APQ) без прерывания обслуживания уже находящихся в обслуживании вызовов (Non-Preemptive Queueing, NPQ) создаются четыре класса (очереди): high, medium, normal, low.

Планировщик очередей всегда начинает обслуживание очереди с класса high без прерывания обслуживания уже находящихся в обслуживании вызовов. Только после того, как все вызовы (пакеты) этого класса будут обслужены, начнется обслуживание класса medium. Соответственно, передача класса low окажется возможной только после того, как будут обслужены все более приоритетные классы. Вполне возможно, что в моменты перегрузки сети вся скорость передачи интерфейса будет занята высокоприоритетными пакетами, а передача пакетов класса low будет происходить с большими задержками.

Алгоритм функционирования планировщика очередей, основанный на абсолютных приоритетах без прерывания вызовов, уже находящихся в обслуживании (Non-Preemptive Queueing) для двух потоков изображен на рис. 30.

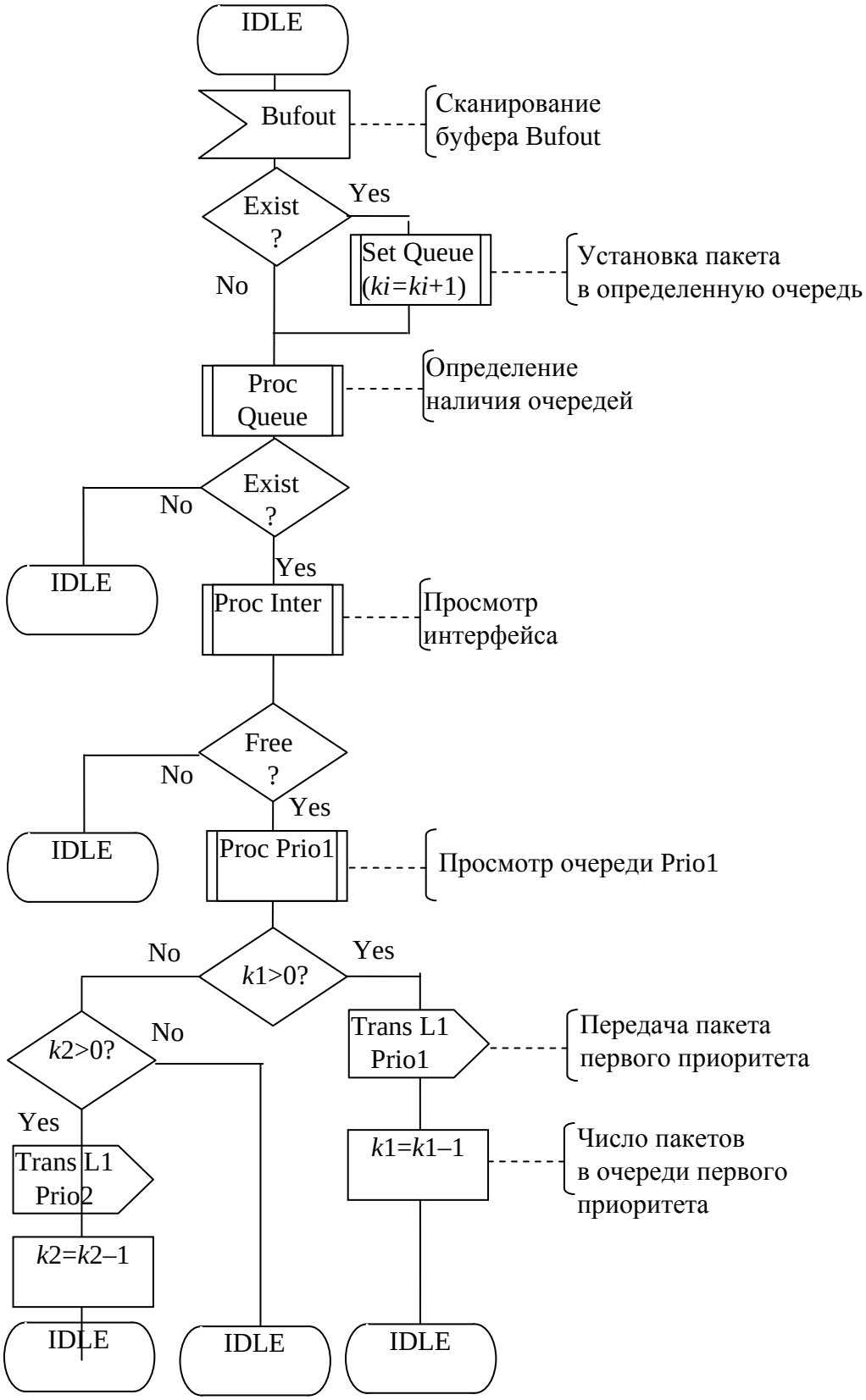


Рис. 30. Алгоритм функционирования планировщика очередей

Для системы M/M/1/ / /NPQ справедливы следующие предположения:

– на вход системы поступают четыре пуассоновских потока с интенсивностями $\lambda 1$, $\lambda 2$, $\lambda 3$, $\lambda 4$, соответственно;

– длительность занятия подчиняется экспоненциальному распределению с параметром $\mu = VEL/LEN$, где VEL – скорость интерфейса, LEN – средняя длина пакета;

– вызов, не принятый к обслуживанию в момент занятости линии пучка, поступает в свою бесконечную очередь. При освобождении линии поочередно обслуживаются все вызовы из первой очереди, затем – все вызовы из второй, третьей и четвертой очередей. Следовательно, вызовы четвертого потока обслуживаются только при отсутствии в очереди вызовов первого, второго и третьего потоков;

– пучок доступен, когда он свободен, для любого вызова;

– исходной для расчета является поступающая нагрузка;

– система находится в стационарном режиме.

Интенсивности поступающих нагрузок равны $a1 = \lambda 1/\mu$, $a2 = \lambda 2/\mu$, $a3 = \lambda 3/\mu$, $a4 = \lambda 4/\mu$. Если общая нагрузка $a = a1 + a2 + a3 + a4 < 1$, то все потоки будут обслужены.

Среднее время пребывания вызова в системе без приоритетов

$$te = \frac{1}{\mu(1-a)}, \qquad a < 1.$$

При $a > 1$ будут обслужены только первые j потоков, для которых $\sum_{i=1}^{j} ai < 1$.

Рассмотрим вначале систему с абсолютными приоритетами и прерыванием менее приоритетных вызовов M/M/1/ / /PPQ**.** В такой системе поступивший вызов более высокого приоритета прерывает обслуживание менее приоритетного вызова, который будет обслужен после более приоритетных вызовов.

Поскольку потоки низших приоритетов не влияют на потоки высших приоритетов, то для первого потока в системе M/M/1/ / /PPQ имеем:

– среднее число вызовов только первого потока в системе

$$Q1(PPQ) = Qe1(PPQ) = \frac{a1}{1-a1};$$

– среднее число вызовов первого потока в обслуживании

$$Qs1(PPQ) = a1;$$

– среднее число вызовов первого потока на ожидании

$$Qw1(PPQ) = Qe1(PPQ) - Qs1(PPQ) = a1\frac{1-(1-a1)}{1-a1} = \frac{a1^2}{1-a1};$$

– среднее время пребывания вызова первого потока в системе

$$te1(PPQ) = \frac{Qe1(PPQ)}{\lambda 1} = \frac{1}{\mu}\frac{1}{1-a1}, \qquad a1 < 1.$$

Для второго потока в системе M/M/1/ / /PPQ:

– среднее число вызовов только первого и второго потоков в системе

$$Q2(PPQ) = \frac{a1+a2}{1-a1-a2},$$

– среднее число вызовов второго потока в системе

$$Qe2(PPQ) = Q2(PPQ) - Q1(PPQ) = \frac{a2}{(1-a1-a2)(1-a1)};$$

– среднее число вызовов второго потока в обслуживании

$$Qs2(PPQ) = a2;$$

– среднее число вызовов второго потока на ожидании

$$Qw2(PPQ) = Qe2(PPQ) - Qs2(PPQ) = a2\frac{1-(1-a1-a2)(1-a1)}{(1-a1-a2)(1-a1)};$$

– среднее время пребывания вызова второго потока в системе

$$te2(PPQ) = \frac{Qe2(PPQ)}{\lambda 2} = \frac{1}{\mu(1-a1-a2)(1-a1)}, \qquad a1 + a2 < 1.$$

Для третьего потока в системе M/M/1/ / /PPQ:

– среднее число вызовов только первого, второго и третьего потоков в системе

$$Q3(PPQ) = \frac{a1+a2+\mathrm{a}3}{1-a1-a2-a3},$$

– среднее число вызовов третьего потока в системе

$$Qe3(PPQ) = Q3(PPQ) - Q2(PPQ) = \frac{a3}{(1-a1-a2-a3)(1-a1-a2)};$$

– среднее число вызовов третьего потока в обслуживании

$$Qs3(PPQ) = a3;$$

– среднее число вызовов третьего потока на ожидании

$$Qw3(PPQ) = Qe3(PPQ) - Qs3(PPQ) =$$
$$= a3\frac{1-(1-a1-a2-a3)(1-a1-a2)}{(1-a1-a2-a3)(1-a1-a2)};$$

– среднее время пребывания вызова третьего потока в системе

$$te3(PPQ) = \frac{Qe3(PPQ)}{\lambda 3} = \frac{1}{\mu(1-a1-a2-a3)(1-a1-a2)},$$
$$a1 + a2 + a3 < 1.$$

Для четвертого потока в системе M/M/1/ / /PPQ:

– среднее число вызовов первого, второго, третьего и четвертого потоков в системе

$$Q4(PPQ) = \frac{a1+a2+\mathrm{a}3+a4}{1-a1-a2-a3-a4},$$

– среднее число вызовов четвертого потока в системе

$$Qe4(PPQ) = Q4(PPQ) - Q3(PPQ) =$$
$$= \frac{a4}{(1-a1-a2-a3-a4)(1-a1-a2-a3)};$$

– среднее число вызовов четвертого потока в обслуживании

$$Qs4(PPQ) = a4;$$

– среднее число вызовов четвертого потока на ожидании

$$Qw4(PPQ) = Qe4(PPQ) - Qs4(PPQ) =$$
$$= a4\frac{1-(1-a1-a2-a3-a4)(1-a1-a2-a3)}{(1-a1-a2-a3-a4)(1-a1-a2-a3)};$$

– среднее время пребывания вызова четвертого потока в системе

$$te4(PPQ) = \frac{Qe4(PPQ)}{\lambda 4} = \frac{1}{\mu(1-a1-a2-a3-a4)(1-a1-a2-a3)},$$

$$a1 + a2 + a3 + a4 < 1.$$

Вернемся к системе M/M/1/ / /NPQ**.** Для этого воспользуемся следующими рассуждениями.

Среднее время пребывания вызова первого потока $te1(NPQ)$ увеличится на время окончания обслуживания любого низкоприоритетного вызова, т. е. на $(a2 + a3 + a4)/\mu$.

Среднее время пребывания вызова второго потока $te2(NPQ)$ увеличится на время окончания обслуживания третьего и четвертого потоков и уменьшится на время прерывания обслуживания первым потоком, т. е. на $(a3 + a4 - a1)/\mu$.

Среднее время пребывания вызова тртьего потока $te3(NPQ)$ увеличится на время окончания обслуживания четвертого потока и уменьшится на время прерывания обслуживания первым и вторым потоками, т. е. на $(a4 - a1 - a2)/\mu$.

Среднее время пребывания в системе вызова четвертого потока $te4(NPQ)$ уменьшится на время прерывания обслуживания первого, второго и третьего потоков, т. е. на $(a1 + a2 + a3)/\mu$. В итоге

$te1(NPQ) = te1(PPQ) + (a2 + a3 + a4)/\mu$;

$te2(NPQ) = te2(PPQ) + (a3 + a4 - a1)/\mu$;

$te3(NPQ) = te3(PPQ) + (a4 - a1 - a2)/\mu$;

$te4(NPQ) = te4(PPQ) - (a1 + a2 + a3)/\mu$.

Из приведенных формул становится очевидным, что в системе M/M/1/ / /NPP время окончания обслуживания самых приоритетных вызовов увеличилось, а самых неприоритетных – уменьшилось.

7.3. Система с относительными приоритетами M/M/1/ / /CBWFQ

При использовании механизма CBWFQ все пакеты, имеющие общий признак (IP-приоритет, поле DSCP и т. д.), относятся к одному из i $(i = 1,\ldots n)$ классов. В отличие от системы с абсолютными приоритетами администратор сети каждому классу в явном виде выделяет определенную долю полосы *bwi%* от скорости передачи интерфейса *VEL*.

Пакеты одного класса одинаково маркируются *классификатором пакетов (Packet Classifier)*, а *планировщик очередей (Packet Scheduler)* помещает их в соответствующую очередь. В дальнейшем планировщик циклически обходит все очереди и считывает из каждой очереди число пакетов, в соответствии с выделенной полосой.

Для системы M/M/1/ / /CBWFQ справедливы следующие предположения:

– на вход системы поступают n пуассоновских потоков различных классов с известными интенсивностями λi, $i = 1,...n$;

– длительность занятия для всех потоков подчиняется экспоненциальному распределению с параметром $\mu = VEL/LEN$, где VEL – скорость интерфейса, LEN – средняя длина пакета;

– вызов, не принятый к обслуживанию в момент занятости линии пучка, поступает в бесконечную очередь своего класса. При освобождении линии из i-ой очереди обслуживается $qi \le bwi\%/min(bwi\%)$ вызовов в соответствии с выделенной полосой;

– пучок доступен, когда он свободен, для любого вызова;

– исходной для расчета является поступающая нагрузка;

– система находится в стационарном режиме.

Этот механизм с относительными приоритетами используется IP-сетях с дифференцированным обслуживанием.

Алгоритм функционирования планировщика очередей, основанный на классах без прерывания вызовов уже находящихся в обслуживании вызовов (CBWFQ) для двух приоритетов, изображен на рис. 31.

Интенсивности поступающих нагрузок: $ai = \lambda i/\mu$, $a = \Sigma ai$. При $a < 1$ все потоки будут обслужены.

При $a > 1$ будут обслужены только первые j потоков, для которых $\sum_{i=1}^{j} ai < 1$.

Число вызовов i-го потока в системе с приоритетами прямо пропорционально поступающей нагрузке i-го потока ai и обратно пропорционально выделенной для этого потока полосе bwi. Поэтому чем меньше отношение $prio(i) = ai/bwi\%$, тем выше приоритет обслуживания i-го потока. Таким образом, номер 1 будет присвоен потоку с наивысшим значением $prio(i)$.

При расчете системы следуем следующим правилам.

1. Перенумеруем потоки в соответствии с приоритетом обслуживания $prio(i) = ai/bwi\%$, где $ai = \lambda i/\mu$ – поступающая нагрузка, $bwi\%$ – доля выделенной полосы для i-го приоритетного потока.

2. При равенстве $prio(i)$ потоки объединяются в один поток с суммированием нагрузок и полос. Каждый из потоков использует выделенную ему полосу и неиспользованную полосу более приоритетных потоков. Так первый поток с наивысшим приоритетом использует долю полосы $bw1\%$, первые два потока – ($bw1\% + bw2\%$) и т. д.

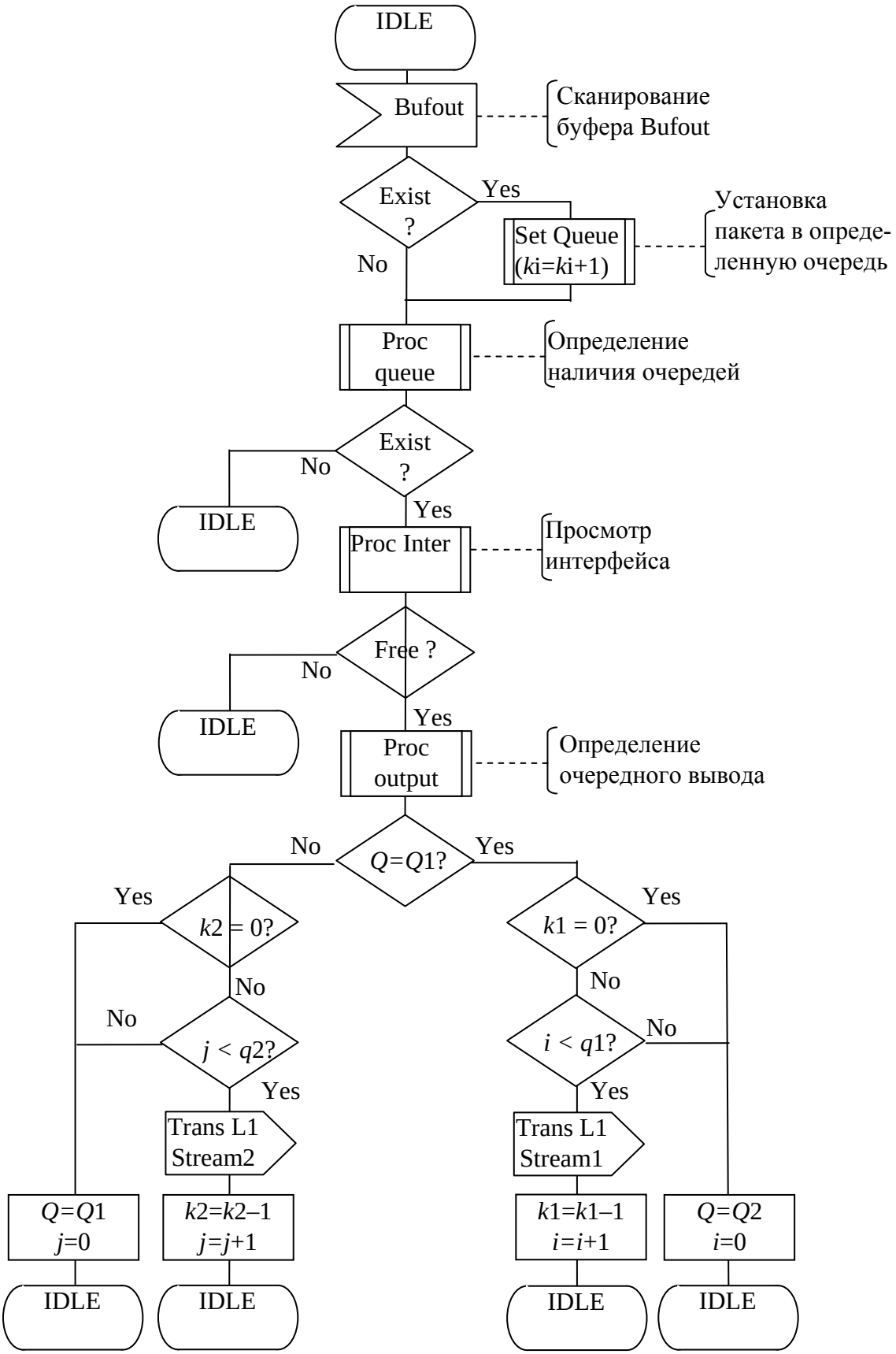


Рис. 31. Алгоритм функционирования планировщика очередей CBWFQ

Для первого потока в системе M/M/1/ / /CBWFQ имеем следующие значения:

– среднее число вызовов только первого потока в очереди

$$QW1(CBQ) = Qw1(CBQ) = \frac{a1^2}{(1-a1)\cdot bw1\%};$$

– среднее число вызовов первого потока в обслуживании

$$Qs1(CBQ) = a1;$$

– среднее число вызовов первого потока в системе

$$Qe1(CBQ) = Qw1(CBQ) + Qs1(CBQ);$$

– среднее время пребывания вызова первого потока в системе

$$te1(CB) = \frac{Qe1(CBQ)}{\lambda 1}, \qquad a1 < 1,\ bw1\% > 0.$$

Для второго потока в системе M/M/1/ / /CBWFQ:
среднее число вызовов только первого и второго потоков в очереди

$$QW2(CBQ) = \frac{(a1+a2)^2}{(1-a1-a2)\cdot(bw1\%+bw2\%)};$$

– среднее число вызовов только второго потока в очереди

$$Qw2(CBQ) = QW2(CBQ) - QW1(CBQ);$$

– среднее число вызовов второго потока в обслуживании

$$Qs2(CBQ) = a2;$$

– среднее число вызовов второго потока в системе

$$Qe2(CBQ) = Qw2(CBQ) + Qs2(CBQ);$$

– среднее время пребывания вызова второго потока в системе

$$te2(CBQ) = \frac{Qe2(CBQ)}{\lambda 2}, \qquad a1 + a2 < 1,\ bw2\% > 0.$$

Для третьего потока в системе M/M/1/ / /CBWFQ:

– среднее число вызовов только первого, второго и третьего потоков в очереди

$$QW3(CBQ) = \frac{(a1+a2+a3)^2}{(1-a1-a2-a3)\cdot(bw1\%+bw2\%+bw3\%)};$$

– среднее число вызовов только третьего потока в очереди

$$Qw3(CBQ) = QW3(CBQ) - QW2(CBQ);$$

– среднее число вызовов третьего потока в обслуживании

$$Qs3(CBQ) = a3;$$

– среднее число вызовов третьего потока в системе

$$Qe3(CBQ) = Qw3(CBQ) + Qs3(CBQ);$$

– среднее время пребывания вызова третьего потока в системе

$$te3(CBQ) = \frac{Qe3(CBQ)}{\lambda 3}, \qquad a1 + a2 + a3 < 1,\ bw3\% > 0.$$

. . .

7.4. Система с относительными приоритетами M/M/1/ / /WFQ

При использовании механизма WFQ все пакеты, имеющие общий признак (например, IP-адрес назначения), относятся к одному потоку.

Классификатор пакетов (Packet Classifier) каждому потоку присваивает вес в соответствии с соотношением $Wi = 4096/(IPi + 1)$, где IPi – значение поля IP-приоритета (3 бита) в поле TOS (8 бит) заголовка пакета i-го потока. *Планировщик очередей (Packet Scheduler)* пакеты одного потока помещает в индивидуальную очередь в порядке поступления со смещением на $Wi \cdot LEN$ байт. В результате i-му потоку достается часть общей полосы $bwi\% = (TOSi + 1)/\Sigma(TOSi + 1)$.

В дальнейшем за один период прохода всех очередей планировщик очередей *(Packet Scheduler)* считывает пакет с наименьшим местом ожидания. Упрощенно механизм WFQ можно представить как одну очередь, в которой место ожидания тем меньше, чем меньше вес и длина пакета.

Для системы M/M/1/ / /WFQ справедливы предположения:

– на вход системы поступают n пуассоновских потоков с известными интенсивностями λi, $i = 1,\ldots n$ и полями $TOSi$;

– длительность занятия для всех потоков подчиняется экспоненциальному распределению с параметром $\mu = VEL/LEN$;

– вызов, не принятый к обслуживанию в момент занятости линии пучка, поступает в бесконечную очередь своего потока. При освобождении линии из i-ой очереди обслуживается $qi \le TOSi + 1$ вызовов;

– пучок доступен, когда он свободен, для любого вызова;

– исходной для расчета является поступающая нагрузка;

– система находится в стационарном режиме.

Интенсивности поступающих нагрузок равны $ai = \lambda i/\mu$, $a = \Sigma ai$. При $a < 1$ все потоки будут обслужены.

Среднее время пребывания вызова в системе без приоритетов

$$te = \frac{1}{\mu(1-a)}, \qquad a < 1.$$

При расчете системы следуем следующим правилам.

1. Перенумеруем потоки в соответствии с приоритетом обслуживания $prio(i) = ai/bwi\%$, где $a\ i = \lambda i/\mu$ – поступающая нагрузка, $bwi\% = (TOSi + 1)/\Sigma(TOSi + 1)$ – доля выделенной полосы для i-го потока.

2. При равенстве $prio(i)$ потоки объединяются в один поток с суммированием нагрузок и полос.

3. Следуем методике расчета аналогично системе M/M/1/ / /CBWFQ.

ЗАДАЧА 1. На интерфейс маршрутизатора со скоростью передачи VEL = 10 Мбит/с в режиме M/M/1/ / /PPQ поступает четыре потока: речевых сообщений (VoIP) с параметром потока λv = 700 с$^{-1}$, сигнальных сообщений (Signaling) с параметром потока λs = 100 с$^{-1}$, информационных сообщений (WEB) с параметром потока λw = 1100 с$^{-1}$, почтовых сообщений (Mail) с параметром потока λm = 500 с$^{-1}$. Средняя длина пакета LEN = 500 байт. Определить среднее время ожидания конца обслуживания для каждого потока.

Решение.

Интенсивность обслуживания пакетов (вызова) $\mu = VEL/LEN$ = 2500 с$^{-1}$.

Поступающие нагрузки:

av = 0,28 Эрл, as = 0,04 Эрл, aw = 0,44 Эрл, am = 0,2 Эрл, a = 0,96 Эрл.

Среднее время ожидания конца обслуживания первого потока

$$tev = \frac{1}{\mu}\frac{1}{1-av} = 0{,}00055 \text{ с};$$

– среднее время ожидания конца обслуживания второго потока

$$tes = \frac{1}{\mu(1-av-as)(1-av)} = 0{,}00081 \text{ с};$$

– среднее время ожидания конца обслуживания третьего потока

$$tew = \frac{1}{\mu(1-av-as-aw)(1-av-as)} = 0{,}00245 \text{ с};$$

– среднее время ожидания конца обслуживания четвертого потока

$$tem = \frac{1}{\mu(1 - av - as - aw - am)(1 - av - as - aw)} = 0{,}04166 \text{ с.}$$

В системе без приоритетов

$$tev = tes = tew = tem = \frac{1}{\mu}\frac{1}{1 - a} = 0{,}01 \text{ с.}$$

На рис. 32 приведены ФРВ конца обслуживания $F(t) = 1 - exp(t/tei)$ для приведенных потоков.

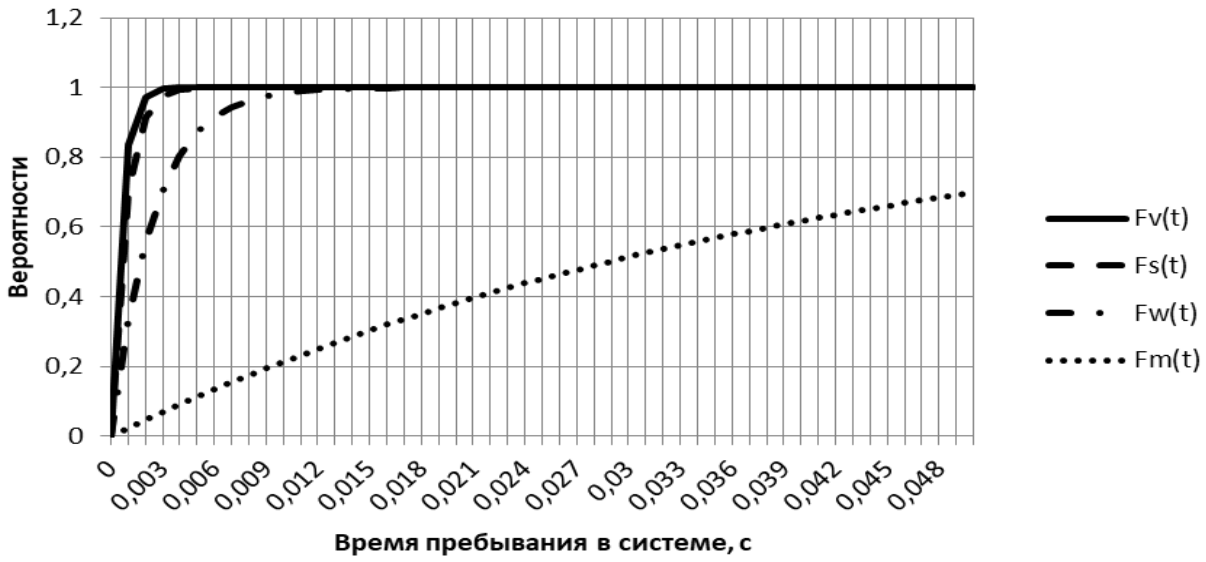


Рис. 32. К задаче 1

ЗАДАЧА 2. На интерфейс маршрутизатора со скоростью передачи *VEL* = 10 Мбит/с в режиме M/M/1/ / /CBWFQ поступает четыре потока: речевых сообщений (VoIP) с параметром потока λv = 700 с$^{-1}$ и выделенной полосой *bwv*% = 75 %, сигнальных сообщений (Signaling) с параметром потока λs = 100 с$^{-1}$ и выделенной полосой *bws%* = 1 %, информационных сообщений (WEB) с параметром потока λw = 1100 с$^{-1}$ и выделенной полосой *bww%* = 19 %, почтовых сообщений (Mail) с параметром потока λm = 500 с$^{-1}$ и выделенной полосой *bwm%* = 5 %. Средняя длина пакета *LEN* = 500 байт. Определить среднее время ожидания конца обслуживания для каждого потока.

Решение.

Интенсивность обслуживания пакетов (вызова) $\mu = VEL/LEN$ = 2500 с$^{-1}$.

Нумеруем потоки в соответствии с приоритетом обслуживания:

$$prio(1) = av/bwv\% = 0{,}37;\ prio(2) = aw/bww\% = 2{,}32;$$
$$prio(3) = as/bws\% = 4;\ prio(4) = am/bwm\% = 4.$$

Поступающие нагрузки с учетом объединения третьего и четвертого потоков:

$$av = 0{,}28 \text{ Эрл},\ aw = 0{,}44 \text{ Эрл},\ as + am = 0{,}24 \text{ Эрл},\ a = 0{,}96 \text{ Эрл}.$$

Для первого потока (VoIP)

$$QW1(CBQ) = Qw1(CBQ) = \frac{av^2}{(1-av)\cdot bwv\%} = 0{,}1451;$$
$$Qs1(CBQ) = av = 0{,}28;$$
$$Qe1(CBQ) = Qw1(CBQ) + Qs1(CBQ) = 0{,}4252;$$
$$te1(CBQ) = \frac{Qe1(CBQ)}{\lambda v} = 0{,}0006 \text{ с}.$$

Для второго потока (WEB)

$$QW2(CBQ) = \frac{(av+aw)^2}{(1-av-aw)\cdot(bwv\%+bww\%)} = 1{,}9696;$$
$$Qw2(CBQ) = QW2(CBQ) - QW1(CBQ) = 1{,}8244;$$
$$Qs2(CBQ) = aw = 0{,}44;$$
$$Qe2(CBQ) = Qw2(CBQ) + Qs2(CBQ) = 2{,}2644;$$
$$te2(CBQ) = \frac{Qe2(CBQ)}{\lambda w} = 0{,}002 \text{ с}.$$

Для объединенного третьего и четвертого потоков (Signaling + Mail)

$$QW3(CBQ) = \frac{(av+aw+as+am)^2}{(1-av-aw-as-am)\cdot(bwv\%+bww\%+bws\%+bwm\%)} = 23{,}04;$$
$$Qw3(CBQ) = QW3(CBQ) - QW2(CBQ) = 21{,}0704;$$
$$Qs3(CBQ) = as + am = 0{,}24;$$
$$Qe3(CBQ) = Qw3(CBQ) + Qs3(CBQ) = 21{,}3104;$$
$$te3(CBQ) = \frac{Qe3(CBQ)}{\lambda s + \lambda m} = 0{,}0355 \text{ с}.$$

На рис. 33 приведены ФРВ конца обслуживания $F(t) = 1 - exp(t/tei)$ для приведенных потоков.

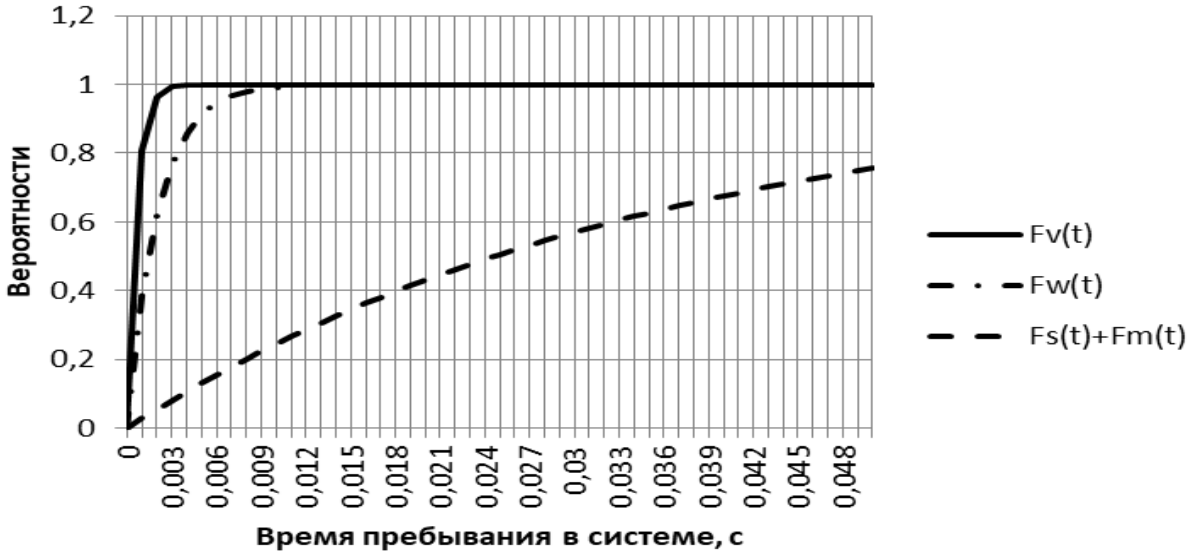


Рис. 33. К задаче 2

ЗАДАЧА 3. На интерфейс маршрутизатора со скоростью передачи *VEL* = 10 Мбит/с в режиме M/M/1/ / /WFQ поступает четыре потока: речевых сообщений (VoIP) с параметром потока λv = 700 $с^{-1}$ и полем *TOSv* = 5, сигнальные сообщения (Signaling) с параметром потока λs = 100 $с^{-1}$ и полем *TOSs* = 1, информационные сообщения (WEB) с параметром потока λw = 1100 $с^{-1}$ и полем *TOSw* = 2, почтовые сообщения (Mail) с параметром потока λm = 500 $с^{-1}$ и полем *TOSm* = 0. Средняя длина пакета *LEN* = 500 байт. Необходимо определить среднее время ожидания конца обслуживания для каждого потока.

Решение.

Поступающие нагрузки:

$$as = 0{,}04 \text{ Эрл}, \; av = 0{,}28 \text{ Эрл}, \; aw = 0{,}44 \text{ Эрл},$$
$$am = 0{,}2 \text{ Эрл}, \; a = 0{,}96 \text{ Эрл}.$$

Выделенные полосы: *bws%* = 2/12 = 0,166; *bwv%* = 6/12 = 0,5; *bww%* = 3/12 = 0,25; *bwm%* = 1/12 = 0,083;

Нумеруем потоки в соответствии с приоритетом обслуживания:

$$prio(1) = as/bws\% = 0{,}24; \; prio(2) = av/bwv\% = 0{,}56;$$
$$prio(3) = aw/bww\% = 1{,}76; \; prio(4) = am/bwm\% = 2{,}4.$$

Для первого потока (Signaling):

$$QW1(WFQ) = Qw1(WFQ) = \frac{as^2}{(1-as)\cdot bws\%} = 0{,}01;$$

$$Qs1(WFQ) = as = 0{,}04;$$

$$Qe1(WFQ) = Qw1(WFQ) + Qs1(WFQ) = 0{,}05;$$

$$te1(WFQ) = \frac{Qe1(WFQ)}{\lambda s} = 0{,}0005 \text{ с.}$$

Для второго потока (VoIP):

$$QW2(WFQ) = \frac{(as + av)^2}{(1 - as - av) \cdot (bws\% + bwv\%)} = 0{,}2259;$$

$$Qw2(WFQ) = QW2(WFQ) - QW1(WFQ) = 0{,}2159;$$

$$Qs2(WFQ) = av = 0{,}28;$$

$$Qe2(WFQ) = Qw2(WFQ) + Qs2(WFQ) = 0{,}4959;$$

$$te2(WFQ) = \frac{Qe2(WFQ)}{\lambda v} = 0{,}0007 \text{ с.}$$

Для третьего потока (WEB):

$$QW3(WFQ) =$$

$$\frac{\left(as + av + aw\right)^2}{\left(1 - as - av - aw\right) \cdot \left(bws\% + bwv\% + bww\%\right)} = 2{,}6255;$$

$$Qw3(WFQ) = QW3(WFQ) - QW2(WFQ) = 2{,}3996;$$

$$Qs3(WFQ) = aw = 0{,}44;$$

$$Qe3(WFQ) = Qw3(WFQ) + Qs3(WFQ) = 2{,}8396;$$

$$te3(WFQ) = \frac{Qe3(WFQ)}{\lambda w} = 0{,}0025 \text{ с.}$$

Для четвертого потока (Mail):

$$QW4(WFQ) =$$

$$\frac{\left(av + as + aw + am\right)^2}{\left(1 - av - as - aw - am\right) \cdot \left(bwv\% + bws\% + bww\% + bwm\%\right)} = 23{,}04;$$

$$Qw4(WFQ) = QW4(WFQ) - QW3(WFQ) = 20{,}4145;$$

$$Qs4(WFQ) = am = 0{,}2;$$

$$Qe4(WFQ) = Qw4(WFQ) + Qs4(WFQ) = 20{,}6145;$$

$$te4(WFQ) = \frac{Qe4(WFQ)}{\lambda m} = 0{,}0412 \text{ с.}$$

На рис. 34 приведены ФРВ конца обслуживания $F(t) = 1 - exp(t/tei)$ для приведенных потоков.

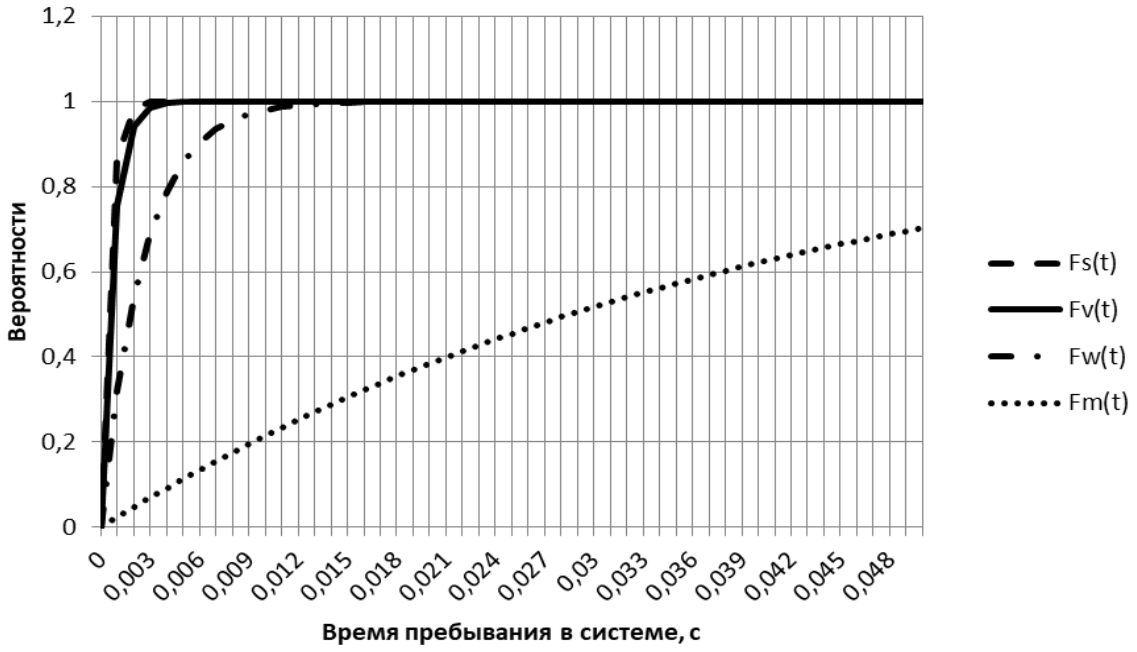


Рис. 34. К задаче 3

8. СИСТЕМА С ПОВТОРНЫМИ ВЫЗОВАМИ (Blocked Calls Repeated, BCR)

Причинами повторных вызовов в системе с коммутацией каналов являются блокировка вызова, ошибка в наборе номера, занятость абонента и безответность абонента. Повторные вызовы создают дополнительную неоплачиваемую нагрузку как на каналы связи, так и на управляющие устройства, поскольку начисление платы за разговор начинается с момента ответа вызываемого абонента. Особенно сильно повторные вызовы загружают управляющие устройства, так как они обрабатываются так же, как и обычные вызовы.

Возникновение повторных вызовов в значительной мере связано с малой информативностью абонентов о состоянии ресурсов в сети связи и отсутствием в системе дополнительных видов обслуживания. Например, если абонент отключен от сети за неуплату, то при использовании системой только акустических тональных сигналов все поступающие на него вызовы завершаются подачей акустического сигнала «Занято». Вызывающий абонент, предполагая, что вызываемый абонент действительно занят, будет безуспешно создавать повторные неоплачиваемые вызовы.

Изменить ситуацию помогает выдача автоинформатором соответствующего голосового уведомления. В рассматриваемом выше случае подача вызывающему абоненту голосового уведомления типа «Вызываемый абонент не обслуживается» освобождает его от последующих повторных вызовов на этот номер.

Еще одним действенным способом уменьшения числа повторных вызовов для абонентов цифровой сети интегрального обслуживания является применение дополнительной услуги «Завершение соединения к занятому абоненту» (Call Completion to Busy Subscriber – CCBS), когда вызываемая сторона информирует вызывающую сторону об освобождении абонента.

Число повторных вызовов абсолютно настойчивого абонента оценивается средним числом попыток M на один успешный вызов.

Для модели с повторными вызовами справедливы следующие предположения:

– вызовы, поступающие на вход системы, образуют пуассоновский поток постоянной интенсивности с параметром λ;

– длительность обслуживания вызова управляющим устройством подчиняется экспоненциальному распределению с параметром μ;

– вызов, не принятый к обслуживанию в момент занятости V линий пучка, удаляется из системы и практически сразу вновь поступает на вход системы;

– любой из V выходов пучка доступен, когда он свободен, для любого поступающего вызова;

– исходной для расчета является обслуженная нагрузка $A_S < V$;

– система находится в стационарном режиме.

Для пояснения функционирования системы с коммутацией каналов с повторными вызовами рассмотрим простую схему рис. 35, а, которая имеет два источника нагрузки и один обслуживающий прибор (канал связи).

На рис. 35, б изображен возможный сценарий поступления вызовов от двух источников.

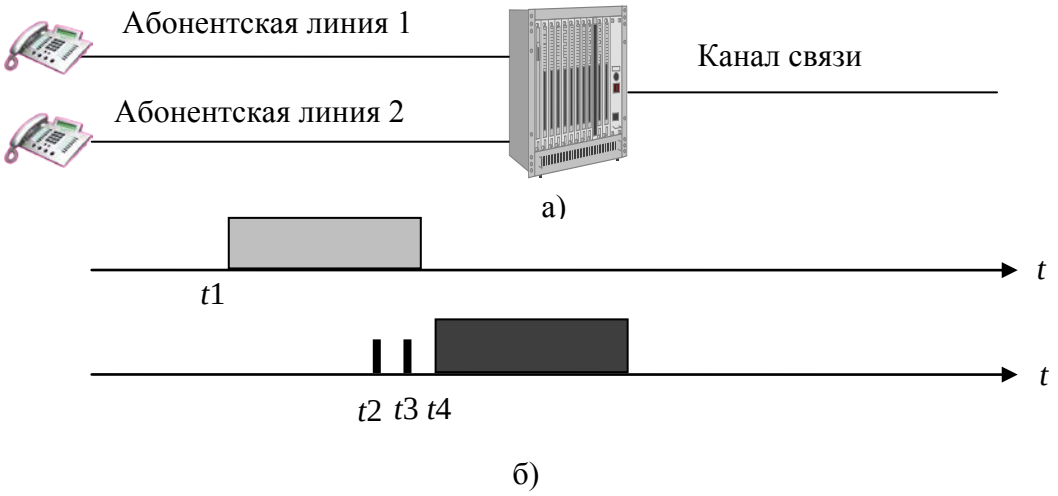


Рис. 35. Система с повторными вызовами:
а) функциональная схема; б) диаграмма поступления вызовов

В момент времени $t1$ от первого источника поступил вызов. Во время занятия линии первым вызовом второй источник дважды безуспешно пытался установить соединение в моменты времени $t2$ и $t3$. Только третья попытка от второго источника в момент времени $t4$ оказалась успешной.

По факту поступило четыре вызова, два из них обслужено, два потеряно. Казалось бы, для устранения потерь двух вызовов нужно добавить еще две линии и в сумме следует иметь три линии. Но на самом деле два источника никогда не могут занять более двух линий. Моменты времени $t2$, $t3$ и $t4$ относятся к одному и тому же единственному вызову, а причиной ошибочного умозаключения стали повторные вызовы.

Как видим, процесс обслуживания в системе с повторными вызовами ничем не отличается от процесса обслуживания в системе с явными потерями. В обоих случаях в момент занятости всех линий пучка поступивший вызов теряется. Однако системы существенно отличаются уровнем потерь: если в системе с явными потерями норма потерь не превышает 0,005 [8], то в системе с повторными вызовами потери могут быть на порядок выше. Причина состоит в том, что один и тот же вызов в системе с явными поте-

рями обслуживается один раз, а в системе с повторными вызовами – многократно, что создает иллюзию высокой поступающей нагрузки.

На рис. 36 приведены схемы распределения нагрузки в системе с потерями и с повторными вызовами. Приняты следующие обозначения: КП – коммутационное поле, A_0 – первичная поступающая нагрузка, A_S – обслуженная нагрузка, A_L – потерянная нагрузка, A_R – повторная нагрузка.

Из физического смысла следует, что в стационарном режиме в системе с повторными вызовами обслуженная и первичная поступающая нагрузки совпадают, а консолидированная нагрузка $B = A_0 + A_R$, образованная первичной поступающей нагрузкой и повторной нагрузкой, вычисляется из трансцендентного уравнения $A_S = B\left(1 - E_V(B)\right)$ при известной обслуженной нагрузке A_S.

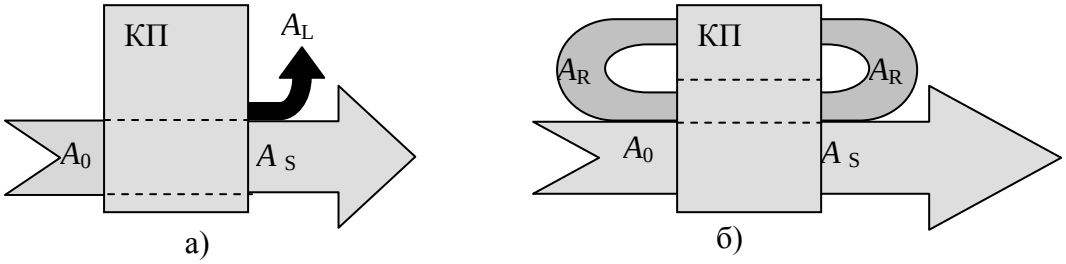


Рис. 36. Схема распределения нагрузки:
а) в системе с потерями; б) с повторными вызовами

Для расчета потерь в системе с повторными вызовами можно использовать формулу Эрланга, но в качестве исходных данных следует использовать обслуженную, а не поступающую (консолидированную) нагрузку.

Например, пусть пучок $V = 30$ линий обслуживает нагрузку $A_S = 26{,}02$ Эрл. Из уравнения $26{,}2 = B\left(1 - E_{30}(B)\right)$ находим консолидированную нагрузку $B = 30$ Эрл, а по уровню потерь $E_{30}(30) = 0{,}13 >> 0{,}005$ классифицируем тип системы как систему с повторными вызовами. Если принять в качестве поступающей нагрузки консолидированную нагрузку, то при заданном уровне потерь 0,005 следует иметь $V = 44$ линии, так как $E_{44}(30) \le 0{,}005$. На самом же деле требуется иметь $V = 39$ линий, так как $E_{39}(26{,}02) \le 0{,}005$.

Вычислим изображение ФРВ конца обслуживания вызовов управляющим устройством, имеющее интенсивность обслуживания μ:

$$F_e(z) = \sum_{k=0}^{\infty} E_V{}^k(B) \cdot \left\{1 - E_V(B)\right\} \cdot \frac{\mu^{k+1}}{z(z+\mu)^{k+1}} =$$

$$= \frac{1 - E_V(B)}{z \cdot E_V(B)} \sum_{k=0}^{\infty} E_V{}^{k+1}(B) \frac{\mu^{k+1}}{(z+\mu)^{k+1}} = \frac{1}{z} \frac{\mu\left(1 - E_V(B)\right)}{z + \mu\left(1 - E_V(B)\right)},$$

оригиналом которой является

$$F_e(t) = 1 - \exp\left\{-\mu\left(1 - E_V(B)\right)t\right\}.$$

Среднее время обслуживания повторного вызова управляющим устройством $TR = \left(\mu \cdot \left(1 - E_V(B)\right)\right)^{-1}$, а одиночного вызова – $T1 = (\mu)^{-1}$, поэтому среднее число попыток на один успешный вызов

$$M = \frac{TR}{T1} = \frac{1}{1 - E_V(B)} = \frac{B}{A_S}.$$

Потери по времени означают, что вызов не будет обслужен немедленно, и абонент повторит попытку. При $As < V$ потери по времени $P_t = E_V(B)$, а потери по вызовам $P_c = 0$.

Для систем с явными потерями $E_V(B) \leq 0{,}005$, поэтому $M \leq 1{,}005$.

ЗАДАЧА 1. Покажите, что при низкой вероятности потерь система с повторными вызовами «превращается» в систему с явными потерями.

Решение.

В системе с повторными вызовами при низкой вероятности потерь повторная нагрузка стремится к нулю, консолидированная нагрузка стремится к обслуженной. Среднее число попыток $M = B/A_s = \left(A_s + A_R\right)/A_{s|A_R \to 0} \to 1$, что является признаком системы с явными потерями.

ЗАДАЧА 2. Четырехлинейная система обслуживает нагрузку 3 Эрл. Классифицировать тип системы коммутации и характеристики нагрузки.

Решение.

Из уравнения $A_S = B\left(1 - E_V(B)\right)$ при A_S = 3 Эрл; V = 4, находим B = 5 Эрл, $E_4(5)$ = 0,398.

Поскольку вероятность потерь 0,39 >> 0,005, классифицируем тип системы как систему с повторными вызовами.

Повторная нагрузка $A_R = B - A_S$ = 2 Эрл. Среднее число попыток на один успешный вызов $M = \frac{B}{A_S} = \frac{5}{3} = 1{,}66.$

Потери по времени $P_t = E_4(5)$ = 0,398. Потери по вызовам P_c = 0.

ЗАДАЧА 3. Однолинейная система обслуживает нагрузку 0,839 Эрл. Классифицировать тип системы коммутации и характеристики нагрузки.

Решение.

Из уравнения $A_S = B\left(1 - E_V(B)\right)$ при A_S = 0,839 Эрл; V = 1, находим B = 5,247 Эрл, E_1(5,245) = 0,839.

Поскольку вероятность потерь 0,839 >> 0,005, классифицируем тип системы как систему с повторными вызовами.

Повторная нагрузка $A_R = B - A_S$ = 4,408 Эрл.

Среднее число попыток на один успешный вызов $M = \dfrac{B}{A_S} = \dfrac{5,247}{0,839} = 6,25.$

Потери по времени $P_t = E_1$(5,247). Потери по вызовам P_c = 0.

9. СИСТЕМА С КОЛЛИЗИЕЙ (Collisions Calls Repeated, CCR)

Коллизии имеют место в локальных сетях с коммутацией пакетов Ethernet. При возникновении коллизии повторная попытка соединения осуществляется через случайный момент времени, вычисляемый по формуле $trep = t_{512} \cdot k$, где t_{512} – длительность промежутка в 512 тактовых интервалов. Значение k равновероятно принимает целочисленные значения в диапазоне ($0–2^n$), где n равно номеру повторной попытки, если номер попытки не больше 10. С десятой попытки значение n не увеличивается и остается равным 10. Например, для Ethernet со скоростью 10 Мбит/с длительность промежутка в 512 тактовых интервалов t_{512} = 51,2 мкс, максимальная задержка первой повторной попытки составит $trep1$ = = 51,2 мкс · 2 = 0,1 мс, а максимальная задержка шестнадцатой попытки составит $trepmax$ = 51,2 мкс · 2^{10} = 52,4288 мс.

Для модели с коллизией примем следующие предположения:

– вызовы, поступающие на вход системы, образуют пуассоновский поток постоянной интенсивности с параметром λ;

– длительность занятия подчиняется экспоненциальному распределению с параметром μ;

– любой из V выходов пучка доступен, когда он свободен, для любого поступающего вызова;

– вызов, не принятый к обслуживанию в момент занятости V линий пучка, становится на ожидание в виртуальной входной очереди. Если вызов в очереди единственный, то при освобождении любой линии он обслуживается, в противном случае все вызовы из очереди удаляются и через случайный интервал времени вновь поступают на вход системы;

– исходной для расчета является консолидированная нагрузка B;

– система находится в стационарном режиме.

Рассмотрим систему без потерь – бесконечный пучок линий. Определим вероятность поступления на него за время занятия более V вызовов (формула Пуассона), когда в системе наступает коллизия

$$P = \sum_{k=V+1}^{\infty} \frac{B^k}{k!} e^{-B} = 1 - \sum_{k=0}^{V} \frac{B^k}{k!} e^{-B}.$$

Определим долю нагрузки, обслуженную V линиями пучка, когда нагрузка обслуживается без коллизии:

$$A_S = B\,(1 - P) = B \sum_{k=0}^{V} \frac{B^k}{k!} e^{-B},$$

которая достигает своего максимального значения при выполнении равенства $\frac{dA_S}{dB} = 0$.

Находим производную

$$\frac{dA_S}{dB} = \sum_{k=0}^{V} \frac{B^k}{k!} e^{-B} - B\frac{B^V}{V!} e^{-B} = \left(\sum_{k=0}^{V} \frac{B^k}{k!} - (V+1)\frac{B^{V+1}}{(V+1)!} \right) e^{-B},$$

из которой находим условие обеспечения максимальной пропускной способности V линий пучка, когда они обслуживают максимальную долю поступающей нагрузки:

$$\sum_{k=0}^{V} \frac{B_{\max}{}^k}{k!} = (V+1)\frac{B_{\max}{}^{V+1}}{(V+1)!}$$

или, иначе:

$$B_{\max} \cdot E_V(B_{\max}) = 1. \qquad (9.1)$$

Условие (9.1) позволяет рассчитать максимальную консолидированную нагрузку $B_{\max}$, при которой система с коллизией обслуживает вызовы без потерь.

Принимая в (9.1) V = 1 (сеть Ethernet), получаем: $B_{\max}^2 - B_{\max} - 1 = 0$, откуда максимальное значение консолидированной нагрузки $B_{\max} = \frac{1+\sqrt{5}}{2} = 1{,}618$ Эрл, при этом находим максимальную обслуженную нагрузку $A_{S\max} = A_{0\max} = B_{\max}\left(e^{-B_{\max}} + B_{\max}e^{-B_{\max}}\right) = 0{,}839$ Эрл.

В однолинейной системе с коллизией без потерь при первичной поступающей нагрузке $A_0 < 0{,}839$ Эрл существует стационарный процесс обслуживания вызовов. В такой системе справедливы равенства: обслуженная (она же первичная) нагрузка $A_S = A_0 = B\left(e^{-B} + Be^{-B}\right)$, нагрузка коллизии $A_C = B - A_S$, среднее число попыток на один успешный вызов $M = B/A_S$. При $A_0 \geq 0{,}839$ Эрл стационарного процесса обслуживания вызовов нет: нагрузка коллизии A_C начинает монотонно увеличиваться, обслуженная нагрузка A_S – стремится к нулю, а число попыток на один успешный вызов M стремится к бесконечности. Ниже в табл. 9.1 приведены некоторые значения B, A_S, A_C, M для однолинейной системы с коллизией.

При максимальной обслуженной нагрузке $A_{S\max} = 0{,}839$ Эрл число попыток на один успешный вызов $M = 1{,}926 \approx 2$.

Для предотвращения самопроизвольного переходного процесса, приводящего к полному прекращению обслуживания, в сети Ethernet каждый

узел ограничивает число повторных вызовов до 16, при этом консолидированная нагрузка достигает значения 4,48 Эрл.

Таблица 9.1

Параметры однолинейной системы с коллизией

B	0,10	0,30	0,50	1,00	1,30	1,50	**1,618**	2.00	4,48	5,00	10,00
A_0	0,10	0,29	0,45	0,74	0,82	0,83	**0,839**	*A*0 > 0,839			
A_S	0,10	0,29	0,45	0,74	0,82	0,83	**0,839**	0,81	0,28	0,20	0,005
A_C	0,00	0,01	0,05	0,26	0,48	0,67	**0,779**	1,19	4,20	4,80	9,995
M	1,00	1,03	1,09	1,36	1,59	1,79	**1,926**	2,46	16,0	24,7	2002

Среднее время разговора *ts* = 3 мин = 180 с. При использовании кодека G.711 без подавления пауз (*tcod* = 30 мс, *tdec* = 0 мс, *bw* = 84,8 кбит/с = = 10,6 Кбайт/с, *LEN* = 318 байт) средний размер «разговорного» файла составляет $Q \le bw{\cdot}ts$ = 1,9 Мбайт. Таким образом, размер «разговорного» файла такой же, как в приведенной ниже задаче 1, но он передается в течение 180 с отдельными фрагментами длиной 318 байт через фиксированные равные промежутки времени 30 мс. Поэтому вычисление задержек и максимального количества пользователей вычисляется иначе.

Допустим пользователь совершает три вызова в час ($\alpha = 3$ ч$^{-1}$) со средней продолжительностью разговора *ts* 3 мин (*ts* = 3 мин). Пользователи подсоединены к концентратору нагрузки сети Ethernet 10 bt (скорость передачи *VEL* = 10 Мбит/с = 1,25 Мбайт/с) и используют кодек G.711 (*tcod* = 30 мс, *tdec* = 0 мс, *bw* = 10,6 Кбайт/с, *LEN* = 318 байт). Тогда параметр потока вызовов от одного пользователя $\lambda = \alpha \cdot ts/tcod$ = $= 3 \cdot 180 \cdot 1000/3600 \cdot 30$ с$^{-1}$=5 с$^{-1}$, время передачи одного фрагмента разговора (пакета) *tpac* = *LEN/VEL* = 318/1,25 · 1000000 с = 0,254 мс, удельная нагрузка одного активного пользователя *apac* = $\lambda \cdot$ *tpac* = 0,0013 Эрл, задержка повторной попытки $trep1 = t_{512} \cdot 2^{M-1}$=51,2 мкс · 2 = 0,1 мс. Максимальное число пользователей *N* = *As/apac* = 0,839/0,0013 = 645 абонентов. Максимальное время передачи фрагмента с учетом повторных попыток составит *tfrag* = *trep*1 + *tpac* = 0,1 мс + 0,254 мс = 0,354 мс.

На самом деле максимальное число пользователей (абонентов) *N* следует уменьшить, по крайней мере, в два раза из-за наличия в сети Ethernet трудно поддающихся учету служебных сообщений протоколов DHCP, ARP и др.

ЗАДАЧА 1. Определить максимальное число компьютеров *n* в концентраторе нагрузки сети Ethernet 10 bt (скорость передачи *VEL* = 10 Мбит/с = = 1,25 Мбайт/с, максимальная длина фрагмента *MTU* = 1500 байт), при ко-

тором обеспечивается максимум обслуженной нагрузки. Среднее число запросов с одного компьютера составляет α = 5 вызовов в час, средняя длина передаваемого файла *LEN* = 1,9 Мбайт.

Решение.

Параметр потока заявок λ = α · *LEN/MTU*, время передачи одного фрагмента *tpac* = *MTU/VEL*, нагрузка от одного компьютера $a = \lambda \cdot tpac = \frac{\alpha \cdot LEN}{VEL} = \frac{5 \cdot 1{,}9}{1{,}25 \cdot 3600} = 0{,}0021$ Эрл. Максимальное число компьютеров $N = \frac{A_s}{a} = \frac{0{,}839}{0{,}0021} \approx 400.$

Время передачи фрагмента без повторных попыток *tpac* = *MTU/VEL* = = 1,2 мс. Среднее время передачи фрагмента с учетом повторных попыток составит *tfrag* = *trep* + *tpac* = $t_{512} \cdot 2^{M-1}$ + *tmtu* ≈ 51,2 мкс · 2 + 1,2 мс ≈ 0,1 мс + + 1,2 мс = 1,3 мс. Общее количество фрагментов *n* = *LEN/MTU* = 1266. Среднее время передачи всего файла *te* = *n* · *tfrag* = 1,64 с.

10. ОБЕСПЕЧЕНИЕ КАЧЕСТВА ОБСЛУЖИВАНИЯ В ИНТЕРНЕТЕ

Сеть с коммутацией пакетов Интернета предназначена для обслуживания различных видов связи (речь, данные, видео). Обслуживание вызова сетью состоит в доставке пакетов от отправителя до получателя. В процессе прохождения пакета по сети связи он по различным причинам может быть утерян, искажен или задержан на случайное время. ***Под качеством обслуживания в сети с коммутацией пакетов понимается вероятность доставки пакета в заданный пункт (или пункты) назначения за нормируемое время.*** Обычно аппаратной задержкой пренебрегают из-за ее малого значения по сравнению с задержкой при прохождении очереди, которая зависит от скорости передачи интерфейса.

Нормы качества обслуживания в IP-сети прописаны в Рекомендациях Y.1540 и Y.1541 [16, 17].

Рассмотрим типовую схему обслуживания вызовов маршрутизатором (рис. 37).

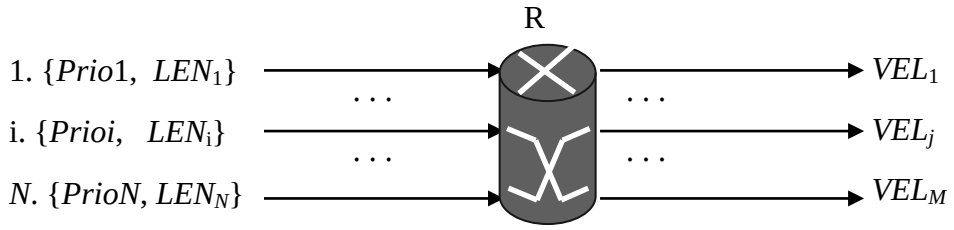


Рис. 37. Схема обслуживания вызовов маршрутизатором

Каждая заказанная услуга порождает поток пакетов (вызовов). Пакет характеризуется своим приоритетом *Prio* (Priority) и длиной *LEN*. На каждый из *N* входов (интерфейсов) маршрутизатора поступает множество пакетов, принадлежащих разным услугам.

Каждый пакет может потребовать соединения с одним или с несколькими из *M* выходов (интерфейсов) маршрутизатора, причем *j*-й выход ($j = 1, \ldots M$) имеет скорость передачи *VELj*. Если в момент поступления пакета имеется требуемый свободный выход, то пакет обслуживается немедленно, в противном случае он обслуживается с учетом своего приоритета, принятым механизмом обслуживания очереди.

Наиболее широко применяются следующие механизмы обслуживания очередей: взвешенного справедливого обслуживания потока (Weighted Fair Queuing, WFQ), взвешенного справедливого обслуживания классов (Class-Based Weighted Fair Queuing, CBWFQ), абсолютного приоритетного обслуживания (Absolute Priority Queueing, APQ).

10.1. Общие принципы формирования пакетов

Мультимедийные приложения сети следующего поколения предполагают использование одной из триады услуг (Triple Service) или их различных комбинаций:

- передача данных (текста, факсимиле, неподвижных изображений);
- передача видео (подвижных изображений);
- передача аудио (речи, музыки).

Изначально сеть Интернет проектировалась только для передачи данных и не была рассчитана на передачу приложений в реальном масштабе времени, каковыми являются приложения видео и аудио. Поэтому возникла необходимость дополнить базовые протоколы Интернета дополнительными специализированными протоколами.

Во многих случаях исходная информация является аналоговой. Преобразование аналогового сигнала в цифровой сигнал может осуществляться различными способами. Преобразование аналогового речевого сигнала обычно включает в себя четыре этапа.

1. *Ограничение полосы (Band Limitation). Выбирается эффективный спектр речевого сигнала, находящийся в диапазоне от 300 до 3400 Гц.*

2. *Выборка (Sampling).* Для правильного воспроизведения аналогового сигнала достаточно передавать его отдельные значения, между которыми исходный аналоговый сигнал изменяется монотонно. Поскольку наибольшие изменения создает самая высокая частота исходного аналогового сигнала, то частота квантования (выборки) выбирается равной удвоенному значению наивысшей частоты исходного аналогового сигнала, т. е. не менее 6800 Гц, в практических реализациях она принята равной 8000 раз в секунду.

3. *Кодирование (Coding).* Каждое значение выборки измеряется и кодируется в двоичном коде длиной 8 бит (1 байт). Первый бит определяет знак выборки (0 – отрицательное значение, 1 – положительное). Следующие три бита определяют сегмент, т. е. часть масштабной области, в которой находится измеренная выборка (000 – младший сегмент, 111 – старший). Последние четыре бита определяют один из 16-ти шагов внутри сегмента (0000 – первый шаг, 1111 – последний). Погрешность измерения порождает шумы квантования. Таким образом, для речевого сигнала полоса передачи равна $8000\ \text{с}^{-1} \cdot 8$ бит = 64 000 бит/с, или 64 кбит/с.

4. *Сжатие цифрового сигнала.* Эта операция производится в отношении группы оцифрованных выборок и не является обязательной, если в сети используются высокоскоростные каналы связи. Однако некоторые части сети могут иметь гораздо более низкую скорость передачи. В этих случаях сжатие цифрового сигнала становится необходимым. Следует отметить,

что сжатие сигнала на передаче всегда связано с понижением качества его воспроизведения на приеме. Для оценки передачи качества голосового сигнала ITU-T (Рекомендация P.800) использует шкалу средней экспертной оценки (Mean Opinion Scores, MOS), приведенную в табл. 10.1.

Таблица 10.1

Шкала MOS

MOS-оценка	Описание качества
5	Междугородное качество, соответствующее качеству передачи по медным физическим проводам
4	Качество, близкое к междугородному, с почти незаметными искажениями
3	Ощутимые искажения, приводящие к повторению части сказанного
2	Искажения весьма раздражают
1	Невозможно вести разговор

Группа оцифрованных выборок (со сжатием или без) объединяются в поле данных, обрамляются заголовками стека протоколов и в виде пакета передаются в канал связи. На приеме восстановление аналогового сигнала производится в обратном порядке.

10.2. Формирование речевых пакетов. Рекомендации G.711, G.729, G.723, G.726

В соответствии с Рекомендацией G.711 время между соседними выборками равно 125 мкс, поэтому в поле данных пакета (Payload) за 10, 20 или 30 мс помещается соответственно 80, 160 или 240 байт. При прохождении по стеку протоколов добавляются заголовки RTP (12 байт), UDP (8 байт), IP (20 байт) и байты канального уровня (8 байт для PPP или 14 байт для сети Ethernet). При использовании Ethernet с учетом преамбулы и межпакетного интервала в сумме добавляется 78 байт (рис. 38).

Preamble (8 байт)	Ethernet (14 байт)	IP, UDP, RTP (40 байт)	Payload (*PLD*)	CRC (4 байта)	Pause (12 байт)

Рис. 38. Поля речевого пакета (Ethernet)

Общая характеристика пакета для приведенных случаев дана в табл. 10.2.

Таблица 10.2

Параметры пакетов с кодеком G.711

Тип кодека	Алгоритм кодирования	Полоса передачи речи, кбит/с	Время формирования пакета *tcod*, мс	Общая длина пакета, байт (*LEN*)	Суммарная полоса передачи, кбит/с (*bw*)	MOS
G.711	PCM	64	10 (80 байт) 20 (160 байт) 30 (240 байт)	80+78=158 160+78=238 240+78=318	126,4 95,2 84,8	4,3 4,3 4,3

Рекомендация G.729A использует алгебраический линейный предиктивный кодер связанной структуры (Conjugate-Structure Algebraic-Code-Excited Linear-Prediction, CS-ACELP). Для сжатия речевых пакетов применяются специализированные цифровые процессоры (Digital Signal Processor, DSP) высокой производительности, использующие различные алгоритмы сжатия. В качестве исходных данных используются блоки 10 миллисекундных выборок размером 80 байт, а на выходе процессор формирует пакет длиной 10 байт. Время обработки блока процессором (look-ahead) – 5 мс, общее время формирования пакета 15 мс. Общая характеристика пакета дана в табл. 10.3.

Таблица 10.3

Параметры пакетов с кодеком G.729A

Тип кодека	Алгоритм кодирования	Полоса передачи речи, кбит/с	Время формирования пакета *tcod*, мс	Общая длина пакета, байт (*LEN*)	Суммарная полоса передачи, кбит/с (*bw*)	MOS
G.729A	CS-ACELP	8	15 (10 байт)	10+78=88	70,4	3,9

Рекомендация G.723.1A использует два алгоритма: многоимпульсное квантование с максимальным правдоподобием (Multipulse Maximum Likelihood Quantization, MP-MLQ) или алгебраическое линейное предиктивное кодирование (Algebraic-Code-Excited Linear-Prediction, CS-ACELP). В качестве исходных данных используются блоки 30 миллисекундных выборок размером 240 байт. В первом случае на выходе процессор формирует пакет длиной 192 бит (24 байта), во втором – 160 бит (20 байт). Время обработки блока процессором составляет 7,5 мс, общее время формирования пакета – 37,5 мс. Общая характеристика пакета для приведенных случаев дана в табл. 10.4.

Таблица 10.4

Параметры пакетов с кодеком G.723.1A

Тип кодека	Алгоритм кодирования	Полоса передачи речи, кбит/с	Время формирования пакета *tcod*, мс	Общая длина пакета, байт (*LEN*)	Суммарная полоса передачи, кбит/с (*bw*)	MOS
G.723.1A	MP-MLQ	6,3	37,5 (24 байта)	24+78=102	27,2	3,9
G.723.1A	ACELP	5,3	37,5 (20 байт)	20+78=98	26,1	3,6

Для еще большей экономии полосы передачи кодеки G.711, G.729, G.723 используют процедуру подавления пауз в разговоре. На передаче детектор активности (Voice Activity Detection, VAD) обнаруживает паузы в речи путем сравнения энергии сигнала, поступившего от источника, с энергией сигнала, поступившего с генератора шума. При обнаружении длительной паузы (длительностью примерно 10 мс) передатчик начинает посылать пакеты (Silence Insertion Descriptor, SID), имеющие гораздо меньшую длину (4 байта), чем обычные речевые пакеты. Приемник, получив такой пакет, включает местный генератор комфортного шума (Comfort Noise Generator, CNG), чтобы снять у абонента ощущение обрыва канала связи. Процедура подавления пауз в разговоре позволяет уменьшить нагрузку на 35–45 %.

Рекомендация G.726 предлагает использовать следующие кодеки (табл. 10.5).

Таблица 10.5

Параметры пакетов с кодеком G.726

Тип кодека	Алгоритм кодирования	Полоса передачи речи, кбит/с	Время формирования пакета *tcod*, мс	Общая длина пакета, байт (*LEN*)	Суммарная полоса передачи, кбит/с (*bw*)	MOS
G.726	ADPCM	40	20 (100 байт)	100+78=178	71,2	4,1
G.726	ADPCM	32	20 (80 байт)	80+78=158	63,2	4,1
G.726	ADPCM	24	20 (60 байт)	60+78=138	55,2	4,0
G.726	ADPCM	16	20 (40 байт)	40+78=118	47,2	4,0

Последние два случая (с полосой 24 и 16 кбит/с) использовать не рекомендуется.

Формирование пакетов видео приведено в Рекомендациях H.261, H.262 и H.263. Рекомендации предназначены для передачи подвижных

изображений телевизионного стандарта с разверткой 525 или 625 строк со скоростью развертки от 10 до 75 кадров в секунду. Длина кадра различна, полоса передачи зависит от размера изображения и от скорости развертки и расположена от 0,064 до 100 Мбит/с. Рекомендации описывают несколько опций сжатия, позволяющие понизить суммарную полосу передачи.

Современные форматы (H.264, MPEG-4) используют полосу 0,128 – 40 Мбит/с.

На рис. 39 показаны места возникновения задержек при прохождении пакета в IP-сети.

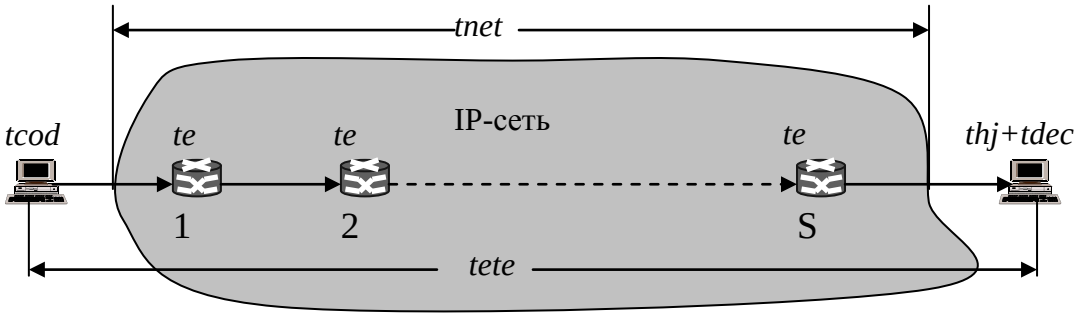


Рис. 39. Возникновение задержек в IP-сети

Приняты следующие обозначения:

- *tcod* – задержка кодера при формировании пакета;
- *te* – средняя задержка в одном маршрутизаторе,
- *thj* – задержка хоста для компенсации сетевого джиттера;
- *tdec* – задержка декодирования пакета;
- *S* – число проходных маршрутизаторов в сети между пользователями;
- *tnet* – задержка прохождения пакета в сети;
- $tete = tcod + S \cdot te + thj + tdec$ – задержка пакета «из конца в конец» при проходе S однотипных маршрутизаторов.

Качество обслуживания в Интернете оценивается заданной вероятностью доставки пакета «из конца в конец» за определенное время. Величина задержки нормируется Рекомендациями Y.1540 [16] и Y.1541 [17].

ЗАДАЧА 1. Назовите обязательные и опциональные операции при формировании аудио- и видео пакетов.

Ответ. Обязательные операции – ограничение полосы, выборка, кодирование, опциональная – сжатие.

ЗАДАЧА 2. Какой транспортный протокол применяется для переноса информации пользователя в реальном масштабе времени?

Ответ. UDP.

ЗАДАЧА 3. Поясните значение полос передачи второй строки табл. 10.2 и первой строки табл. 10.4.

Ответ. В первом случае каждые 20 мс передается пакет длиной 238 байт. Следовательно, за 1 с передается 238 · 1000/20 байт = 1190 байт = = 95200 бит. Во втором случае за 30 мс передается пакет длиной 102 байта. Следовательно, за 1 с передается 102 · 1000/30 байт = 3400 байт = 27200 бит.

ЗАДАЧА 4. Возможен ли обмен информацией (например, разговор) между пользователями, использующими различные типы кодеков?

Ответ. Передающая и приемная части должны использовать один и тот же тип кодека.

ЗАДАЧА 5. Чему равно время декодирования пакета при использовании кодека G.711?

Ответ. Нулю.

11. СРАВНЕНИЕ СЕТЕЙ КОММУТАЦИИ КАНАЛОВ И ПАКЕТОВ

Сети коммутации каналов и пакетов имеют различные критерии оценки качества обслуживания.

Эффективность функционирования сети с коммутацией каналов оценивается ее пропускной способностью, под которой понимается нагрузка, обслуженная с заданным качеством (потерями). В общем случае соединение проходит через несколько сетевых элементов (узлов), и суммарные сетевые потери («из конца в конец») зависят от числа узлов, участвующих в соединении. Сетевые потери нормируются для самого «длинного» соединения. В «длинном» соединении участвует максимальное число проходных узлов между абонентами: оконечная станция (ОС) – узловая станция (УС) – междугородная станция (АМТС) – узел автоматической коммутации (УАК) – международный центр коммутации (МЦК) – МЦК – УАК – АМТС – УС – ОС. Итого – 10 сетевых элементов. Исходя из нормы сетевых потерь, на каждом сетевом элементе потери нормируются таким образом, чтобы их сумма не превосходила сетевые потери. Для «длинного» соединения потери будут максимальными, при меньшем числе проходных станций потери снимаются. Для каждого сетевого элемента при нормированной вероятности потерь $E_V(A)$ и известной поступающей нагрузке A по формуле Эрланга находят необходимое число каналов

$$[V] = E_V(A) = \frac{A^V}{V!}\left[\sum_{i=0}^{V}\frac{A^i}{i!}\right]^{-1}. \qquad (11.1)$$

Например, при норме сетевых потерь $P = 0{,}05$ вероятность потерь на каждом из $S = 10$ проходных узлов не должна превышать $E_V(A) \approx P/S = 0{,}005$. Если на всех участках используется ИКМ-тракт ($V = 30$ каналов), то максимальная поступающая нагрузка в соответствии с (11.1) не должна превышать $A = 19$ Эрл. При типовой абонентской нагрузке $a0 = 0{,}1$ Эрл это соответствует $N = A/a0 = 190$ обслуживаемым абонентам «из конца в конец».

Эффективность функционирования сети коммутации пакетов оценивается заданной вероятностью доставки пакета «из конца в конец» за нормируемое время.

В [17] указано, что хорошее качество обслуживания речевых пакетов обеспечивается, если на коротких промежутках времени (меньше минуты) максимальная сетевая задержка «из конца в конец» не превосходит значения *tsum* = 150 мс с вероятностью не более 0,001. На длинных промежутках времени допускается вероятность потерь 0,05. Очевидно, в качестве нормируемого параметра для сети с коммутацией пакетов является вероятность доставки пакета «из конца в конец» с вероятностью не менее 99,9 %

за время не более *tsum* = 150 мс. В качестве проходных сетевых элементов в сети коммутации пакетов используются маршрутизаторы. При расчете качества обслуживания будем учитывать только скорость передачи интерфейса, пренебрегая аппаратной задержкой, которую вносит программное обеспечение маршрутизаторов при решении о перемещении пакета с одного интерфейса на другой.

Рассмотрим характеристики сети коммутации пакетов, имеющей те же скорости интерфейса и число проходных сетевых элементов, что и система коммутации каналов. Сравним их с характеристиками сети коммутации каналов, численные значения которой были приведены выше.

11.1. Сеть с интегрированным обслуживанием

Сеть коммутации пакетов с интегрированным обслуживанием (***IntServ***) в части функционирования сетевых элементов аналогична (системе M/M/1/0) – системе коммутации каналов с явными потерями. Операторы сетей в каждом маршрутизаторе резервируют суммарную полосу передачи *BW*, *которая ограничивает максимальное число одновременно установленных приоритетных соединений*. По запросу пользователя с помощью протокола RSVP каждый проходной маршрутизатор выделяет такому соединению часть *bw* от суммарной полосы *BW* и индивидуальную очередь. При занятом интерфейсе поступивший пакет ставится на ближайшее (первое) место ожидания в своем буфере очереди (механизм WFQ). За один период прохода всех очередей пакет обязательно будет передан, в буфере очереди будет находиться не более одного пакета. Максимальный объем буфера для одного соединения равен длине пакета *LENp* без полей канального и физического уровней. Максимальный джиттер образуется при считывании пакета из последней очереди. За счет такой организации обслуживания в системе с интегрированным обслуживанием речевым пакетам обеспечивается малое значение сетевой задержки, т. е. гарантируется высокое качество обслуживания.

Предположим, что речевым приложениям выделено k = 75 % скорости передачи интерфейса (*VEL* = 2048 кбит/с), т. е. $BW = 0{,}75 \cdot VEL =$ = 1536 кбит/с = 192 Кбайт/с. За счет приоритетного обслуживания речевым пакетам практически достается вся полоса *VEL* = 2048 кбит/с.

При использовании кодека G.711 (*tcod* = 10 мс, *tdec* = 0 мс, *bw* = = 126,4 кбит/с, *LEN* = 158 байт, *LENp* = 120 байт, *PLD* = 80 байт) максимальное число одновременных соединений составит $V = INT(BW/bw) = 12$. Подставляя в (11.1) значения $E_V(A) = P/S = 0{,}005$; $V = 12$ находим нагрузку приоритетных соединений $A = 5{,}3$ Эрл. Используя исходные данные для приоритетных соединений последовательно находим:

$V = INT(BW/bw) = 12$ – максимальное число приоритетных соединений в маршрутизаторе;

$\lambda = V \cdot bw/LEN = 1200$ с$^{-1}$ – максимальная интенсивность поступления пакетов от всех приоритетных абонентов в ЧНН;

$\lambda_0 = 1/tcod = 100$ с$^{-1}$ – параметр потока от активного абонента;

$\mu = VEL/LEN = 1620$ с$^{-1}$ – интенсивность обслуживания пакетов интерфейсом маршрутизатора;

$a = \lambda/\mu = 0{,}74$ Эрл – нагрузку на интерфейсе маршрутизатора;

$te = (\mu - \lambda)^{-1} = 2{,}38$ мс – задержку в одном маршрутизаторе с учетом прохождения очередей;

$tej = te = 2{,}38$ мс – джиттер маршрутизатора;

$thj = \sqrt{\frac{S}{S-1}}tej = 2{,}5$ мс – джиттер хоста;

$tnet = tsum - tcod - tdec - thj = 137{,}5$ мс – допустимое время прохождения пакета по сети;

$tete = tcod + tdec + S \cdot te + thj = 36{,}3$ мс – время прохождения пакета «из конца в конец»;

$R\max = LENp = 120$ байт – емкость буфера очереди маршрутизатора для одного приоритетного соединения;

$hbuf = 2 \cdot thj \cdot \lambda_0 \cdot PLD + PLD = 121$ байт – объем приемного буфера хоста;

$N = A/a0 = 53$ – число обслуживаемых абонентов с гарантированным качеством обслуживания.

При использовании кодека G.711 ($tcod = 20$ мс, $tdec = 0$ мс, $bw = 95{,}2$ кбит/с, $LEN = 238$ байт, $LENp = 200$ байт, $PLD = 160$ байт) максимальное число одновременных соединений составит $V = INT(BW/bw) = 16$. Подставляя в (11.1) значения $E_V(A) = P/S = 0{,}005$; $V = 16$ находим нагрузку приоритетных соединений $A = 8{,}1$ Эрл. Используя исходные данные для приоритетных соединений последовательно находим

$V = INT(BW/bw) = 16$ – максимальное число приоритетных соединений в маршрутизаторе;

$\lambda = V \cdot bw/LEN = 801$ с$^{-1}$ – максимальная интенсивность поступления пакетов от всех приоритетных абонентов в ЧНН;

$\lambda_0 = 1/tcod = 50$ с$^{-1}$ – параметр потока от активного абонента;

$\mu = VEL/LEN = 1075$ с$^{-1}$ – интенсивность обслуживания пакетов интерфейсом маршрутизатора;

$a = \lambda/\mu = 0{,}74$ Эрл – нагрузку на интерфейсе маршрутизатора;

$te = (\mu - \lambda)^{-1} = 3{,}65$ мс – задержку в одном маршрутизаторе с учетом прохождения очередей;

$tej = te = 3{,}63$ мс – джиттер маршрутизатора;

$thj = \sqrt{\frac{S}{S-1}}tej$ = 3,84 мс – джиттер хоста;

$tnet = tsum - tcod - tdec - thj$ = 126,2 мс – допустимое время прохождения пакета по сети;

$tete = tcod + tdec + S \cdot te + thj$ = 60,1 мс – время прохождения пакета «из конца в конец»;

$R\max = LENp$ = 200 байт – емкость буфера очереди маршрутизатора для одного приоритетного соединения;

$hbuf = 2 \cdot thj \cdot \lambda_0 \cdot PLD + PLD$ = 222 байт – объем приемного буфера хоста;

$N = A/a0$ = 81 – число обслуживаемых абонентов с гарантированным качеством обслуживания.

При использовании кодека G.711 ($tcod$ = 30 мс, $tdec$ = 0 мс, bw = 84,8 кбит/с, LEN = 318 байт, $LENp$ = 280 байт, PLD = 240 байт) максимальное число одновременных соединений составит $V = INT(BW/bw)$ = 18. Подставляя в (11.1) значения $E_V(A) = P/S$ = 0,005; V = 18, находим нагрузку приоритетных соединений A = 9,6 Эрл. Используя исходные данные для приоритетных соединений последовательно находим:

$V = INT(BW/bw)$ = 18 – максимальное число приоритетных соединений в маршрутизаторе;

$\lambda = V \cdot bw/LEN$ = 600 $с^{-1}$ – максимальная интенсивность поступления пакетов от всех приоритетных абонентов в ЧНН;

$\lambda_0 = 1/tcod$ = 33,3 $с^{-1}$ – параметр потока от активного абонента;

$\mu = VEL/LEN$ = 805 $с^{-1}$ – интенсивность обслуживания пакетов интерфейсом маршрутизатора;

$a = \lambda/\mu$ = 0,74 Эрл – нагрузку на интерфейсе маршрутизатора;

$te = (\mu - \lambda)^{-1}$ = 4,87 мс – задержку в одном маршрутизаторе с учетом прохождения очередей;

$tej = te$ = 4,87 мс – джиттер маршрутизатора;

$thj = \sqrt{\frac{S}{S-1}}tej$ = 5,14 мс – джиттер хоста;

$tnet = tsum - tcod - tdec - thj$ =114,8 мс – допустимое время прохождения пакета по сети;

$tete = tcod + tdec + S \cdot te + thj$ = 83,9 мс – время прохождения пакета «из конца в конец»;

$R\max = LENp$ = 280 байт – емкость буфера очереди маршрутизатора для одного приоритетного соединения;

$hbuf = 2 \cdot thj \cdot \lambda_0 \cdot PLD + PLD$ = 322 байт – объем приемного буфера хоста;

$N = A/a0 = 96$ – число обслуживаемых абонентов с гарантированным качеством обслуживания.

При использовании кодека G.729 ($tcod$ = 15 мс, $tdec$ = 5 мс, bw = 70,4 кбит/с, LEN = 88 байт, $LENp$ = 50 байт, PLD = 10 байт) максимальное число одновременных соединений составит $V = INT(BW/bw) = 21$. Подставляя в (11.1) значения $E_V(A) = P/S = 0{,}005$; $V = 21$, находим нагрузку приоритетных соединений $A = 12$ Эрл. Используя исходные данные для приоритетных соединений, последовательно находим:

$V = INT(BW/bw) = 21$ – максимальное число приоритетных соединений в маршрутизаторе;

$\lambda = V \cdot bw/LEN = 2100$ с$^{-1}$ – максимальная интенсивность поступления пакетов от всех приоритетных абонентов в ЧНН;

$\lambda_0 = 1/tcod = 66{,}6$ с$^{-1}$ – параметр потока от активного абонента;

$\mu = VEL/LEN = 2909$ с$^{-1}$ – интенсивность обслуживания пакетов интерфейсом маршрутизатора;

$a = \lambda/\mu = 0{,}72$ Эрл – нагрузку на интерфейсе маршрутизатора;

$te = (\mu - \lambda)^{-1} = 1{,}23$ мс – задержку в одном маршрутизаторе с учетом прохождения очередей;

$tej = te = 1{,}23$ мс – джиттер маршрутизатора;

$thj = \sqrt{\dfrac{S}{S-1}}tej = 1{,}3$ мс – джиттер хоста;

$tnet = tsum - tcod - tdec - thj = 128{,}7$ мс – допустимое время прохождения пакета по сети;

$tete = tcod + tdec + S \cdot te + thj = 33{,}6$ мс – время прохождения пакета «из конца в конец»;

$R\max = LENp = 50$ байт – емкость буфера очереди маршрутизатора для одного приоритетного соединения;

$hbuf = 2 \cdot thj \cdot \lambda_0 \cdot PLD + PLD = 12$ байт – объем приемного буфера хоста;

$N = A/a0 = 119$ – число обслуживаемых абонентов с гарантированным качеством обслуживания.

При использовании кодека G.723 ($tcod$ = 37,5 мс, $tdec$ = 7,5 мс, bw = 26,1 кбит/с, LEN = 98 байт, $LENp$ = 60 байт, PLD = 20 байт) максимальное число одновременных соединений составит $V = INT(BW/bw) = 58$. Подставляя в (11.1) значения $E_V(A) = P/S = 0{,}005$; $V = 58$, находим нагрузку приоритетных соединений: $A = 43{,}0$ Эрл. Используя исходные данные для приоритетных соединений, последовательно находим

$V = INT(BW/bw) = 58$ – максимальное число приоритетных соединений в маршрутизаторе;

$\lambda = V \cdot bw/LEN = 1930$ с$^{-1}$ – максимальная интенсивность поступления пакетов от всех приоритетных абонентов в ЧНН;

$\lambda_0 = 1/tcod = 26{,}6$ с$^{-1}$ – параметр потока от активного абонента;

$\mu = VEL/LEN = 2612$ с$^{-1}$ – интенсивность обслуживания пакетов интерфейсом маршрутизатора;

$a = \lambda/\mu = 0{,}74$ Эрл – нагрузку на интерфейсе маршрутизатора;

$te = (\mu - \lambda)^{-1} = 1{,}46$ мс – задержку в одном маршрутизаторе с учетом прохождения очередей;

$tej = te = 1{,}46$ мс – джиттер маршрутизатора;

$thj = \sqrt{\frac{S}{S-1}}tej = 1{,}55$ мс – джиттер хоста;

$tnet = tsum - tcod - tdec - thj = 103{,}4$ мс – допустимое время прохождения пакета по сети;

$tete = tcod + tdec + S \cdot te + thj = 61{,}2$ мс – время прохождения пакета «из конца в конец»;

$R\max = LENp = 60$ байт – емкость буфера очереди маршрутизатора для одного приоритетного соединения;

$hbuf = 2 \cdot thj \cdot \lambda_0 \cdot PLD + PLD = 22$ байт – объем приемного буфера хоста;

$N = A/a0 = 430$ – число обслуживаемых абонентов с гарантированным качеством обслуживания.

Таким образом, сети с коммутацией пакетов с интегрированным обслуживанием обеспечивают качество обслуживания, сравнимое с сетью коммутации каналов. Однако по абонентской емкости они имеют преимущество перед сетями с коммутацией каналов только при использовании кодека G.723 со сжатием информации.

11.2. Сеть с дифференцированным обслуживанием

Сеть коммутации пакетов с дифференцированным обслуживанием (***DiffServ***) в части функционирования сетевых элементов аналогична системе коммутации с ожиданием (системе M/M/1). Каждый проходной маршрутизатор резервирует суммарную полосу передачи *BW*, *которая ограничивает максимальную интенсивность нагрузки приоритетных потоков*. Равно приоритетные потоки имеют общую полосу *BW* и общую для них очередь (механизм CBWFQ).

При занятом интерфейсе поступивший пакет ставится в конец очереди своего класса. За один период прохода всех очередей из каждой очереди считываются пакеты, стоящие первыми в очереди. Объем считанных пакетов из каждой очереди прямо пропорционален ширине выделенной для

этого приоритета полосы. Не обслуженные пакеты будут считаны при следующем проходе очередей. При этом условии в системе с дифференциальным обслуживанием достигается приемлемое значение сетевой задержки и потери пакетов, т. е. приемлемое качество обслуживания.

Расчет параметров сети сводится к подбору такого максимального значения интенсивности потока λ, чтобы при заданных значениях скорости передачи интерфейса *VEL*, общей выделенной полосы *BW*, доли полосы приложения *bw%*, параметрах кодека *tcod/tdec/LEN*, числе проходных маршрутизаторов *S* время доставки пакета «из конца в конец» за время *tsum* = 150 мс было не менее 0,999, т. е. удовлетворяло неравенству (5.10)

$$F_e(tnet) = 1 - \sum_{k=0}^{S-1} \frac{\left((\mu - \lambda) * tnet\right)^k}{k!} e^{-(\mu-\lambda)\cdot tnet} \geq 0{,}999,$$

где $\mu = VEL/LEN$, $tnet = tsum - tcod - tdec - thj$, $thj = \sqrt{\frac{S}{S-1}}tej$.

Предположим, что речевым приложениям выделено k = 75 % от скорости передачи интерфейса (*VEL* = 2048 кбит/с), т. е. $BW = 0{,}75 \cdot VEL$ = = 1536 кбит/с = 192 Кбайт/с. Оставшимся мультимедийным приложениям гарантированно достается полоса *(VEL – BW)* = 512 кбит/с.

При использовании кодека G.711 (*tcod* = 10 мс, *tdec* = 0 мс, *LEN* = 158 байт, *LENp* = 120 байт, *PLD* = 80 байт) для приоритетных соединений последовательно находим следующие значения:

$\lambda = 1448$ с$^{-1}$ – максимальную интенсивность поступления пакетов;

$\lambda_0 = 1/tcod = 100$ с$^{-1}$ – параметр потока от активного абонента;

$\mu = VEL/LEN = 1620$ с$^{-1}$ – интенсивность обслуживания пакетов интерфейсом маршрутизатора;

$a = \lambda/\mu = 0{,}89$ Эрл – нагрузку на интерфейсе маршрутизатора;

$te = \frac{1}{\mu}\left(\frac{a}{(1-a)\cdot k} + 1\right)$ = 7,56 мс – среднее время прохождения пакета через маршрутизатор с вероятностью ≥ 0,999;

$tej = te$ = 7,56 мс – джиттер маршрутизатора;

$thj = \sqrt{\frac{S}{S-1}}tej$ = 7,97 мс – джиттер хоста;

$tnet = tsum - tcod - tdec - thj$ = 132 мс – допустимое время прохождения пакета по сети;

$tep = tnet/S$ = 13,2 мс – предельное время прохождения пакета через маршрутизатор;

$tete = tcod + tdec + S \cdot te + thj = 93{,}6$ мс – время прохождения пакета «из конца в конец» с вероятностью ≥0,999;

$R\max = BW \cdot tep \cdot LENp/LEN = 1925$ байт – емкость буфера очереди маршрутизатора с учетом джиттера;

$hbuf = 2 \cdot thj \cdot \lambda_0 \cdot PLD + PLD = 208$ байт – емкость приемного буфера хоста;

$N = \lambda/(a0 \cdot \lambda_0) = 145$ – число обслуживаемых абонентов.

При использовании кодека G.711 ($tcod = 20$ мс, $tdec = 0$ мс, $LEN = 238$ байт, $LENp = 200$ байт, $PLD = 160$ байт) для приоритетных соединений последовательно находим следующие данные:

$\lambda = 891$ с$^{-1}$ – максимальную интенсивность поступления пакетов;

$\lambda_0 = 1/tcod = 50$ с$^{-1}$ – параметр потока от активного абонента;

$\mu = VEL/LEN = 1075$ с$^{-1}$ – интенсивность обслуживания пакетов интерфейсом маршрутизатора;

$a = \lambda/\mu = 0{,}82$ Эрл – нагрузку на интерфейсе маршрутизатора;

$te = \frac{1}{\mu}\left(\frac{a}{(1-a)\cdot k}+1\right) = 6{,}9$ мс – среднее время прохождения пакета через маршрутизатор с вероятностью ≥ 0,999;

$tej = te = 6{,}9$ мс – джиттер маршрутизатора;

$thj = \sqrt{\frac{S}{S-1}}tej = 7{,}28$ мс – джиттер хоста;

$tnet = tsum - tcod - tdec - thj = 122{,}7$ мс – допустимое время прохождения пакета по сети;

$tep = tnet/S = 12{,}3$ мс – предельное время прохождения пакета через маршрутизатор;

$tete = tcod + tdec + S \cdot te + thj = 96{,}4$ мс – время прохождения пакета «из конца в конец» с вероятностью ≥0,999;

$R\max = BW \cdot tep \cdot LENp/LEN = 1980$ байт – емкость буфера очереди маршрутизатора с учетом джиттера;

$hbuf = 2 \cdot thj \cdot \lambda_0 \cdot PLD + PLD = 276$ байт – емкость приемного буфера хоста;

$N = \lambda/(a0 \cdot \lambda_0) = 178$ – число обслуживаемых абонентов.

При использовании кодека G.711 ($tcod = 30$ мс, $tdec = 0$ мс, $LEN = 318$ байт, $LENp = 280$ байт, $PLD = 240$ байт) для приоритетных соединений последовательно находим:

$\lambda = 605$ с$^{-1}$ – максимальную интенсивность поступления пакетов;

$\lambda_0 = 1/tcod = 33{,}3$ с$^{-1}$ – параметр потока от активного абонента;

$\mu = VEL/LEN = 805$ с$^{-1}$ – интенсивность обслуживания пакетов интерфейсом маршрутизатора;

$a = \lambda/\mu = 0{,}75$ Эрл – нагрузку на интерфейсе маршрутизатора;

$te = \frac{1}{\mu}\left(\frac{a}{(1-a)\cdot k}+1\right) = 6{,}26$ мс – среднее время прохождения пакета через маршрутизатор с вероятностью ≥ 0,999;

$tej = te = 6{,}26$ мс – джиттер маршрутизатора;

$thj = \sqrt{\frac{S}{S-1}}tej = 6{,}6$ мс – джиттер хоста;

$tnet = tsum - tcod - tdec - thj = 113{,}4$ мс – допустимое время прохождения пакета по сети;

$tep = tnet/S = 11{,}3$ мс – предельное время прохождения пакета через маршрутизатор;

$tete = tcod + tdec + S \cdot te + thj = 99{,}2$ мс – время прохождения пакета «из конца в конец» с вероятностью ≥ 0,999;

$R\max = BW \cdot tep \cdot LENp/LEN = 1917$ байт – емкость буфера очереди маршрутизатора с учетом джиттера;

$hbuf = 2 \cdot thj \cdot \lambda_0 \cdot PLD + PLD = 346$ байт – емкость приемного буфера хоста;

$N = \lambda/(a0 \cdot \lambda_0) = 181$ – число обслуживаемых абонентов.

При использовании кодека G.729 ($tcod = 15$ мс, $tdec = 5$ мс, $LEN = 88$ байт, $LENp = 50$ байт, $PLD = 10$ байт) для приоритетных соединений последовательно находим:

$\lambda = 2724$ с$^{-1}$ – максимальную интенсивность поступления пакетов;

$\lambda_0 = 1/tcod = 66{,}6$ с$^{-1}$ – параметр потока от активного абонента;

$\mu = VEL/LEN = 2909$ с$^{-1}$ – интенсивность обслуживания пакетов интерфейсом маршрутизатора;

$a = \lambda/\mu = 0{,}93$ Эрл – нагрузку на интерфейсе маршрутизатора;

$te = \frac{1}{\mu}\left(\frac{a}{(1-a)\cdot k}+1\right) = 7{,}1$ мс – среднее время прохождения пакета через маршрутизатор с вероятностью ≥ 0,999;

$tej = te = 7{,}1$ мс – джиттер маршрутизатора;

$thj = \sqrt{\frac{S}{S-1}}tej = 7{,}48$ мс – джиттер хоста;

$tnet = tsum - tcod - tdec - thj = 122{,}5$ мс – допустимое время прохождения пакета по сети;

$tep = tnet/S = 12{,}2$ мс – предельное время прохождения пакета через маршрутизатор;

$tete = tcod + tdec + S \cdot te + thj = 98{,}4$ мс – время прохождения пакета «из конца в конец» с вероятностью ≥ 0,999;

$R\max = BW \cdot tep \cdot LENp/LEN = 1336$ байт – емкость буфера очереди маршрутизатора с учетом джиттера;

$hbuf = 2 \cdot thj \cdot \lambda_0 \cdot PLD + PLD = 20$ байт – емкость приемного буфера хоста;

$N = \lambda/(a0 \cdot \lambda_0) = 409$ – число обслуживаемых абонентов.

При использовании кодека G.723 ($tcod = 37{,}5$ мс, $tdec = 7{,}5$ мс, $LEN = 98$ байт, $LENp = 60$ байт, $PLD = 20$ байт) для приоритетных соединений последовательно находим следующие значения:

$\lambda = 2383$ с$^{-1}$ – максимальную интенсивность поступления пакетов;

$\lambda_0 = 1/tcod = 26{,}6$ с$^{-1}$ – параметр потока от активного абонента;

$\mu = VEL/LEN = 2612$ с$^{-1}$ – интенсивность обслуживания пакетов интерфейсом маршрутизатора;

$a = \lambda/\mu = 0{,}91$ Эрл – нагрузку на интерфейсе маршрутизатора;

$te = \frac{1}{\mu}\left(\frac{a}{(1-a)\cdot k}+1\right) = 5{,}7$ мс – среднее время прохождения пакета через маршрутизатор с вероятностью ≥ 0,999;

$tej = te = 5{,}7$ мс – джиттер маршрутизатора;

$thj = \sqrt{\frac{S}{S-1}}tej = 6{,}0$ мс – джиттер хоста;

$tnet = tsum - tcod - tdec - thj = 99$ мс – допустимое время прохождения пакета по сети;

$tep = tnet/S = 9{,}9$ мс – предельное время прохождения пакета через маршрутизатор;

$tete = tcod + tdec + S \cdot te + thj = 108$ мс – время прохождения пакета «из конца в конец» с вероятностью ≥0,999;

$R\max = BW \cdot tep \cdot LENp/LEN = 1164$ байт – емкость буфера очереди маршрутизатора с учетом джиттера;

$hbuf = 2 \cdot thj \cdot \lambda_0 \cdot PLD + PLD = 27$ байт – емкость приемного буфера хоста;

$N = \lambda/(a0 \cdot \lambda_0) = 894$ – число обслуживаемых абонентов.

Таким образом, сети с коммутацией пакетов с дифференцированным обслуживанием обеспечивают приемлемое качество обслуживания при большей абонентской емкости по сравнению с системой интегрированного обслуживания.

ЗАДАЧА 1. Что ограничивают в системе с интегрированным обслуживанием и в системе с дифференцированным обслуживанием для достижения требуемого качества обслуживания?

Ответ. В системе с интегрированным обслуживанием ограничивают максимальное число соединений, а в системе с дифференцированным обслуживанием – максимальную нагрузку.

ЗАДАЧА 2. Какая система проще в реализации – с интегрированным или дифференцированным обслуживанием?

Ответ. С дифференцированным обслуживанием.

ЗАДАЧА 3. В системе с интегрированным обслуживанием речевым приложениям на маршрутизаторе выделено 80 % скорости передачи интерфейса (VEL = 10 Мбит/с). Используется кодек G.711 ($tcod$ = 10 мс). Определить максимальное число соединений в маршрутизаторе с гарантированным обслуживанием.

Решение.

Общая резервируемая полоса передачи равна $BW = 0{,}8 \cdot VEL = 8000$ кбит/с, резервируемая полоса одним пользователем равна $bw = 126{,}4$ кбит/с. Максимальное число соединений в маршрутизаторе с гарантированным обслуживанием $V = INT(BW/bw) = 63$.

ЗАДАЧА 4.

На какую минимальную интенсивность поступающих речевых пакетов рассчитывается маршрутизатор?

Решение.

От одного активного («говорящего») пользователя пакет поступает каждые $tcod$ с. Поэтому интенсивность поступающих речевых пакетов от одного активного пользователя равна $\lambda_0 = 1/tcod$. Например, при использовании кодека G.711 ($tcod$ = 10 мс) $\lambda_0 = 100$ с$^{-1}$. Однако, если пользователь осуществляет 1 вызов в час с длительностью разговора ts = 3 мин, то его параметр сетевого потока в ЧНН всего лишь $\lambda = \lambda_0 \cdot ts/tcod = 3 \cdot 60 \cdot 100/3600$ с$^{-1}$ = 5 с$^{-1}$. Поэтому расчетный суммарный поток вызовов, поступающий на маршрутизатор, рассчитывается на пиковую нагрузку и не может быть менее λ_0 вызовов.

ЗАДАЧА 5. Как вычисляется среднее время прохождения пакетов приоритетного потока через маршрутизатор при использовании механизма CBWFQ, если потоку выделено bw% от скорости передачи интерфейса?

Решение.

В соответствии с разделом 7.3 для приоритетного потока имеем

$$Qw = \frac{a^2}{(1-a)\cdot bw\%},$$

$$Qs = a,$$

$$Qe = \frac{a^2}{(1-a)\cdot bw\%} + a,$$

$$te = \frac{Qe}{\lambda} = \frac{1}{\mu}\left(\frac{a}{(1-a)\cdot bw\%} + 1\right).$$

ЗАДАЧА 6. Как вычисляется джиттер маршрутизатора?

Решение.

Среднее время прохождения пакетов приоритетного потока через маршрутизатор составляет *te*. Минимальное время прохождения пакетов приоритетного потока через маршрутизатор равно нулю. Следовательно, джиттер маршрутизатора – $tej = te - 0 = te$.

ЗАДАЧА 7. Как вычисляется джиттер хоста?

Решение.

Средняя задержка в обслуживании пакета маршрутизатором при наличии очереди равна *te*. Джиттер маршрутизатора $tej = te$.

Джиттер хоста (сетевой джиттер) определяется как стандартное отклонение от среднего S ($S > 1$) измерений:

$$thj = \sqrt{\frac{S}{S-1}}tej.$$

ЗАДАЧА 8. Как вычисляется объем приемного буфера хоста?

Решение.

Во время разговора полезная нагрузка *PLD* поступает со скоростью λ_0 с$^{-1}$. За время сетевого джиттера *thj* поступает $\lambda_0 \cdot thj$ дополнительных пакетов. С учетом упреждения/отставания прихода пакетов и пакета в обслуживании объем буфера приемного хоста $hbuf = 2 \cdot \lambda_0 \cdot thj \cdot PLD + PLD$ байт.

ЗАДАЧА 9. В каком случае IP-сеть имеет преимущество перед сетью с коммутацией каналов по числу обслуживаемых пользователей?

Ответ. IP-сеть имеет преимущество перед сетью с коммутацией каналов по числу обслуживаемых пользователей только при использовании кодеков со сжатием. При этом качество воспроизведения речи за счет задержки и потери части пакетов ухудшается.

ЗАДАЧА 10. Может ли пользователь в системе с интегрированным обслуживанием получить отказ в резервировании ресурсов?

Ответ. Обычно число пользователей или ресурсы сетевого элемента рассчитываются таким образом, чтобы на 99,5 % обеспечивалось резервирование ресурсов. В сети с коммутацией каналов с вероятностью 0,005 пользователь получает отказ в обслуживании, а в сети с коммутацией пакетов с той же вероятностью – отказ в резервировании ресурсов.

ЗАДАЧА 11. Может ли задержка пакета «из конца в конец» в системе с интегрированным обслуживанием превысить нормированное значение при зарезервированных ресурсах?

Ответ. Нет.

ЗАДАЧА 12. Может ли задержка пакета «из конца в конец» в системе с дифференцированным обслуживанием превысить нормированное значение?

Ответ. Обычно число пользователей или ресурсы сети рассчитывается таким образом, чтобы на 99,9 % обеспечивалось нормированное время доставки пакета «из конца в конец» за 150 мс. В сети с коммутацией каналов при установленном соединении пользователь всегда получает гарантированное качество, а в сети с коммутацией пакетов – с вероятностью 0,001 задержка может превысить нормированное значение, и качество обслуживания не обеспечивается.

12. РАСЧЕТ ОТДЕЛЬНЫХ ХАРАКТЕРИСТИК МАРШРУТИЗАТОРА

Нормальная работа маршрутизатора обеспечивается двумя главными механизмами: механизмом (алгоритмом) выбора наилучшего маршрута и механизмом предотвращения перегрузки сети. При вычислении наилучшего маршрута в IP-сети используется суммарная метрика, представляющая собой сумму отдельных метрик последовательного пути «из конца в конец». Отдельная метрика отображает ценность отрезка пути между сетевыми элементами (маршрутизаторами); более предпочтительный путь имеет меньшую метрику. В качестве метрики удобно использовать время доставки. Среднее время окончания конца обслуживания пакета однолинейным пучком (одним интерфейсом) вычисляется по формуле

$$te = \frac{1}{\mu - \lambda} = \frac{1}{\mu} \cdot \frac{1}{1-a} = \frac{LEN}{VEL} \cdot \frac{1}{1-a_{|a=0}} = \frac{1}{\mu} = \frac{LEN}{VEL},$$

где $\mu = VEL/LEN$ – интенсивность обслуживания; VEL – скорость передачи интерфейса; LEN – средняя длина пакета; λ – интенсивность поступления пакетов; $a = \lambda/\mu$ – удельная загрузка интерфейса. В [15] предлагается вычислять метрику как число $M = INT(100000000/VEL)$, равное задержке (в секундах) обслуживания пакета длиной 100000000 бит при отсутствии очереди ($a = 0$). Поскольку пакеты Ethernet не превышают 1500 байт, то такая метрика определяет **не абсолютную**, а **условную статическую** задержку обслуживания пакета интерфейсом. Так, для интерфейса со скоростью VEL = 100 Мбит/с и выше $M = 1$, для интерфейса с VEL = 2048 кбит/с $M = 48$, для интерфейса с VEL = 64 кбит/с $M = 1562$ и т. д.

Если смежные маршрутизаторы сети связаны между собой одним интерфейсом, то для последовательного пути следования пакета средние значения задержек суммируются, соответственно суммарная метрика вычисляется как сумма метрик последовательного пути. Так, например, если пункт назначения проходит через $S = 2$ последовательных маршрутизатора со скоростями интерфейсов 32 и 8 Мбит, соответственно, то суммарная метрика вычисляется как

$$M = \sum_{j=1}^{S} M_j = M_1 + M_2 = INT\left(\frac{100000000}{32000000}\right) + INT\left(\frac{100000000}{8000000}\right) = 3 + 12 = 15.$$

Большинство существующих маршрутизаторов при вычислении метрики либо совсем не учитывают нагрузку на интерфейсе (OSPF), либо используют ее ***косвенно*** (EIGRP), что приводит к большой погрешности вычисления метрики в моменты пика нагрузки, когда значение метрики может измениться в несколько раз. Это существенный недостаток такого

подхода: метрика ***обязательно*** должна динамично изменяться в соответствии с поступающей нагрузкой и иметь реальный физический смысл (параметрическое измерение).

Механизм предотвращения перегрузки сети разработан для TCP-соединений. Он не чувствителен для UDP-соединений, которые используются для доставки пакетов в реальном времени (речь и видео). Поскольку доля таких приложений и, соответственно, соединений непрерывно возрастает, необходимость разработки механизма предотвращения перегрузки сети для UDP-соединений очевидна.

12.1. Расчет и использование метрик с учетом нагрузки на интерфейсе

Учет метрики особенно критичен для приложений реального времени (речь и видео), которые используют один из классов сетевого качества обслуживания: интегрированное обслуживание (IntServ) [12] или дифференцированное обслуживание (DiffServ) [14].

В обоих случаях из общей скорости передачи интерфейса VEL для приложений реального времени выделяется часть полосы BW ≤ VEL.

Предлагается следующий метод расчета абсолютной динамической метрики, учитывающей задержку обслуживания пакета интерфейсом с учетом нагрузки в системе дифференцированного обслуживания.

1. В начале интервала измерения T сбрасываются (устанавливаются в ноль) все счетчики.

2. При поступлении каждого пакета в течение измеряемого интервала времени T инкриминируются два счетчика:

– счетчик числа поступивших пакетов (значение L);

– счетчик суммарного числа обслуженных байтов (значение $LENS$).

3. В конце периода измерений подсчитывается:

– нагрузка на интерфейсе маршрутизатора $a = LENS/(VEL \cdot T)$;

– средняя длина пакета $LEN = LENS/L$;

– метрика интерфейса (при $BW \approx VEL$)

$$M = INT\left(\frac{1000000}{\mu - \lambda}\right) = \frac{1000000}{\mu} \cdot \frac{1}{1-a} = \frac{1000000 \cdot LEN}{VEL \cdot (1-a)} = \frac{1000000}{\mu s},$$

численно равная задержке (в микросекундах) обслуживания пакета длиной LEN бит на интерфейсе со скоростью передачи VEL бит/с при удельной загрузке интерфейса a Эрл. Здесь $\mu = VEL/LEN$ – интенсивность обслуживания пакетов интерфейсом при отсутствии нагрузки ($a = 0$), μs – реальная интенсивность обслуживания.

Например, при малой нагрузке ($a \approx 0$) и длине пакета LEN = 576 байт для интерфейса со скоростью передачи VEL = 100 Мбит/с M = 46 [мкс],

для интерфейса с VEL = 2048 кбит/с M = 2250 [мкс], для интерфейса с VEL = 64 кбит/с M = 72000 [мкс] и т. д.

Если пункт назначения проходит через S = 2 последовательных маршрутизатора со скоростями передачи интерфейсов $VEL1$ = 32 Мбит/с и $VEL2$ = 8 Мбит/с, соответственно, то при длине пакета LEN = 576 байт = = 4608 бит и нагрузке на каждом интерфейсе a = 0,5 Эрл

– метрика интерфейса первого маршрутизатора

$$M1 = (1\,000\,000 \cdot LEN) / VEL1 \cdot (1 - a) = 576 \cdot 8 / 16 = 288 \text{ [мкс]};$$

– метрика интерфейса второго маршрутизатора

$$M1 = (1\,000\,000 \cdot LEN) / VEL2 \cdot (1 - a) = 576 \cdot 8 / 4 = 1152 \text{ [мкс]};$$

– суммарная метрика с учетом нагрузки

$$MS = \sum_{j=1}^{S} Mj = M1 + M2 = 288 + 1152 = 1440 \text{ [мкс]};$$

– интенсивность доставки пакета через два маршрутизатора с учетом нагрузки $\mu s = \dfrac{1\,000\,000}{MS} = 694 \text{ с}^{-1}$.

Последнее означает, что два последовательных интерфейса со скоростями 32 и 8 Мбит/с при загруженности на 50 % эквивалентны одному незагруженному виртуальному интерфейсу со скоростью передачи $VELS = \mu s \cdot LEN = 694 \text{ с}^{-1} \cdot 4608$ бит ≈3,2 Мбит/с.

Интервал измерения нагрузки T критичен только для приложений реального времени, поэтому можно воспользоваться подходом измерения речевой нагрузки, используемом в цифровых сетях интегрального обслуживания (ISDN). По определению значение нагрузки численно равно числу вызовов, поступивших за время занятия (за среднее время разговора). В эталонной модели ISDN [8] удельная нагрузка от абонента принята равной 0,1 Эрл при параметре потока вызовов 4 ч^{-1}. Отсюда следует, что среднее время занятия равно 90 с и время измерения нагрузки для каждого интерфейса (измеряемый интервал времени T) следует выбрать равным 90 с.

Рассылка значений метрики другим маршрутизаторам осуществляется протоколами маршрутизации вида OSPF [13] только при значительном изменении метрики (не менее, чем в два раза).

Рассмотрим влияние метрики на алгоритм выбора наилучшего пути. Для этого обратимся к фрагменту сети (рис. 40) и определим стратегию выбора пути от маршрутизатора R1 к маршрутизатору R3 при различных значениях нагрузки.

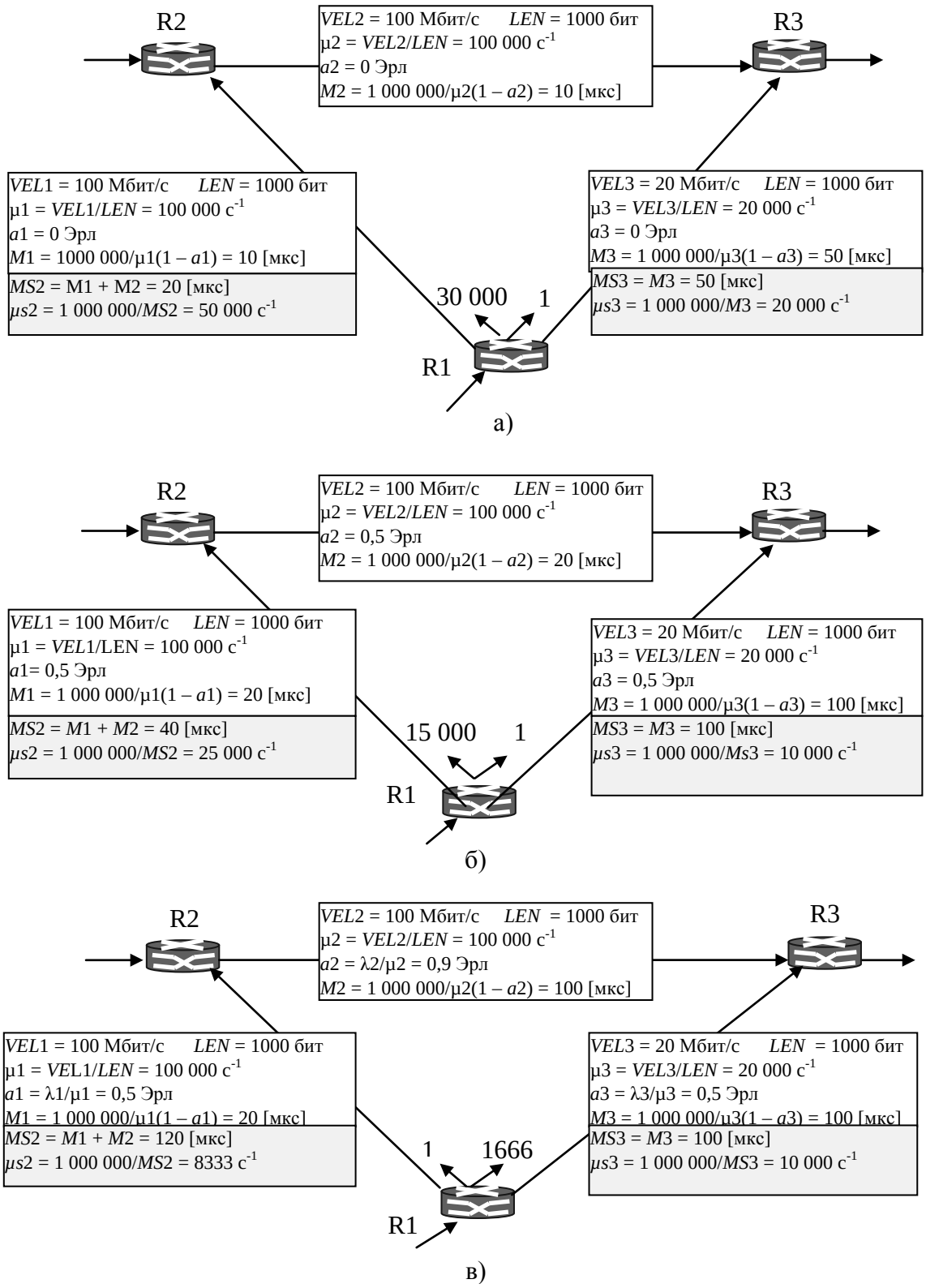


Рис. 40. Распределение пакетов маршрутизатором R1
между альтернативными путями:
а) при малой нагрузке; б) средней нагрузке; в) высокой нагрузке

Пусть от маршрутизатора R1 к маршрутизатору R2 имеется интерфейс со скоростью $VEL1$ = 100 Мбит/с, от маршрутизатора R2 к маршрутизатору R3 – интерфейс со скоростью $VEL2$ = 100 Мбит/с, от маршрутизатора R1 к R3 – интерфейс со скоростью $VEL3$ = 20 Мбит/с. Длину пакета для простоты примем равной LEN = 125 байт = 1000 бит.

При малой нагрузке $a1 = a2 = a3 \approx 0$ Эрл (рис. 40, а) практически все пакеты от маршрутизатора R1 к маршрутизатору R3 будут следовать по пути R1–R2–R3, поскольку метрика этого пути

$$MS2 = M1 + M2 =$$

$$= \frac{1\,000\,000 * LEN}{VEL1} \cdot \frac{1}{1-a1} + \frac{1\,000\,000 * LEN}{VEL2} \cdot \frac{1}{1-a2} = 20 \text{ [мкс]}$$

меньше метрики прямого маршрута R1–R3 ($MS3 = M3$ = 50 [мкс]). Интенсивность доставки пакетов по пути R1–R2–R3 составит $\mu s2$ = 1 000 000/MS2 = 50 000 $с^{-1}$, а интенсивность доставки пакетов по пути R1–R3 равна $\mu s3$ = 1 000 000/MS3 = 20 000 $с^{-1}$. В соответствии с *Положением* разд. 5.6 ($\mu s2 - \mu s3 = \lambda 2 - \lambda 3$) каждую секунду на пакет, отправленный по пути R1–R3, маршрутизатор R1 по пути R1–R2–R3 должен отправлять ($\mu s2 - \mu s3$) = 50 000 – 20 000 = 30 000 пакетов.

При средней нагрузке $a1 = a2 = a3$ = 0,5 Эрл (рис. 40, б) опять практически все пакеты от маршрутизатора R1 к маршрутизатору R3 будут следовать по пути R1–R2–R3, поскольку метрика этого пути $MS2 = M1 + M2$ = 40 [мкс] меньше метрики прямого маршрута R1–R3 ($MS3 = M3$ = 100 [мкс]). Интенсивность доставки пакетов по пути R1–R2–R3 равна $\mu s2$ = 1 000 000/MS2 = 25 000 $с^{-1}$, а интенсивность доставки пакетов по пути R1–R3 равна $\mu s3$ = 1 000 000/MS3 = 10 000 $с^{-1}$. В соответствии с *Положением* разд. 5.6 каждую секунду пакет, отправленный по пути R1–R3, маршрутизатор R1 по пути R1–R2–R3 должен отправлять ($\mu s2 - \mu s3$) = 25 000 – 10 000 = 15 000 пакетов.

При перегрузке интерфейса на участке R2–R3 $a1 = a3$ = 0,5 Эрл, $a2$ = 0,9 Эрл (рис. 40, в) метрика пути R1–R2–R3 станет равной $MS2 = M1 + M2$ = 120 [мкс], а метрика пути R1–R3 останется равной $MS3$ = 100 [мкс]. Интенсивность доставки пакетов по пути R1–R2–R3 равна $\mu s2$ = 1 000 000/MS2 = 8 333 $с^{-1}$, а интенсивность доставки пакетов по пути R1–R3 равна $\mu s3$ = 1 000 000/MS3 = 10 000 $с^{-1}$. Поэтому маршрутизатор R1 должен произвести динамическое перераспределение нагрузки, а именно: в соответствии с *Положением* разд. 5.6 каждую секунду на пакет, отправленный по пути R1–R2–R3, маршрутизатор R1 по пути R1–R3 должен отправлять ($\mu s3 - \mu s2$) = 10 000 – 8 333 = 1667 пакетов.

Как видно из данного примера смена маршрута произошла при изменении метрики перегруженной части сети в *MS*2(a2 = 0,9 Эрл)/*MS*2(a2 = = 0 Эрл) = 120/20 = 6 раз.

12.2. Расчет метрики параллельных интерфейсов

Иногда между смежными маршрутизаторами используется несколько параллельных интерфейсов (составной интерфейс) в режиме разделения нагрузки. Пусть путь следования пакета до следующего маршрутизатора образован V параллельными интерфейсами (V-линейным пучком), и каждый из этих интерфейсов имеет интенсивность обслуживания $\mu = VEL/LEN$, где VEL – скорость передачи интерфейса, LEN – средняя длина пакета. Воспользовавшись (5.3), определим изображение плотности ФРВ ожидания начала обслуживания $f_w(z)$

$$F_w(t) = 1 - P(>0)e^{-(V\mu-\lambda)t}, \quad f_w(z) = L\left[F_w(t)\right] = \frac{V\mu - \lambda + z - zP(>0)}{V\mu - \lambda + z},$$

и плотность ФРВ обслуживания $f_s(z)$

$$F_s(t) = 1 - e^{-\mu t}, \qquad f_s(z) = \frac{\mu}{\mu + z}.$$

Находим изображение плотности ФРВ конца обслуживания

$$f_e(z) = f_w(z) * f_s(z) = \frac{V\mu^2 - \lambda\mu + z\mu - z\mu P(>0)}{V\mu^2 - \lambda\mu + z\mu + V\mu z - \lambda z + z^2}$$

и среднее время окончания конца обслуживания пакета V-линейным пучком

$$TEV = -\frac{df_e(z)}{dz}\Big|_{z=0} = \frac{\mu P(>0) + V\mu - \lambda}{\mu(V\mu - \lambda)} = \frac{1}{V\mu - \lambda}\left\{P(>0) + V - A\right\}.$$

Заменим V-линейный пучок однолинейным, в котором скорость интерфейса равна суммарной скорости всех скоростей передачи параллельных интерфейсов. Используя результаты разд. 5.2, находим среднее время окончания конца обслуживания пакета однолинейным пучком $TE1 = \dfrac{1}{V\mu - \lambda}$.

Анализируя разность

$$\frac{1}{V\mu - \lambda}\left\{P(>0) + V - A - 1\right\} = \frac{1}{V\mu - \lambda}\left\{(V - A) - \left(1 - P(>0)\right)\right\} > 0,$$

становится очевидно, что *V*-линейный пучок всегда имеет меньшую пропускную способность, чем однолинейный пучок, имеющий скорость интерфейса, равную суммарной скорости всех интерфейсов *V*-линейного пучка.

При малой и средней нагрузках очереди на интерфейсе имеют небольшую длину. На рис. 41, а показано обслуживание девяти пакетов одиночным интерфейсом со скоростью *VEL* и нагрузкой 0,5 Эрл, а на рис. 41, б – обслуживание такого же числа пакетов трехлинейным пучком со скоростью интерфейса *VEL*/3 каждый.

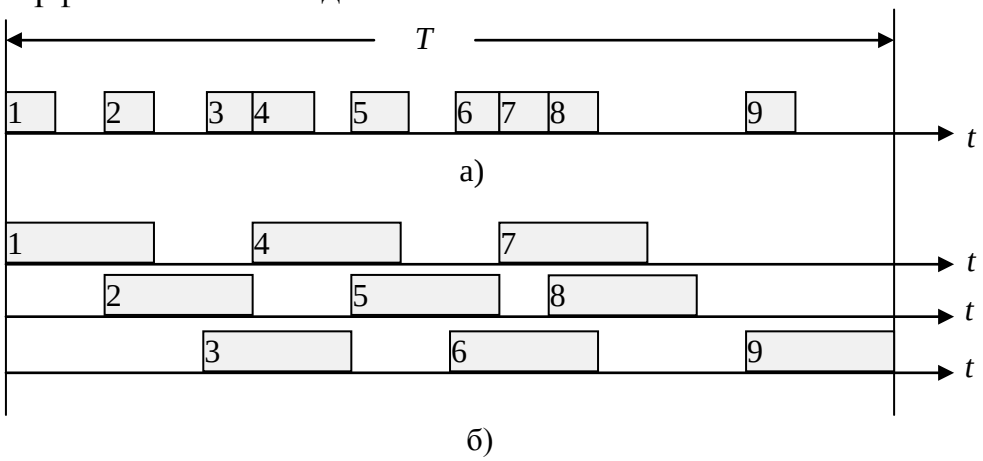


Рис. 41. Соотношение метрик 1:3 при небольших очередях:
а) для одиночного интерфейса; б) трех параллельных интерфейсов

Из рисунка видно, что время обслуживания пакета одиночным интерфейсом в три раза выше, т. е. соотношение метрик составляет 1 : 3.

Однако в моменты всплеска нагрузки, когда кратковременно образуются большие очереди, соотношение метрик меняется. Сказанное иллюстрирует рис. 42, где среднее время обслуживания девяти пакетов в однолинейном пучке со скорость интерфейса *VEL* составляет 5 единиц времени (рис. 42, а), а в трехлинейном пучке со скоростью интерфейса *VEL*/3 каждый – 6 единиц (рис. 42, б), что означает соотношение метрик 5 : 6, хотя время окончания обслуживания всех девяти пакетов в обоих случаях одинаково.

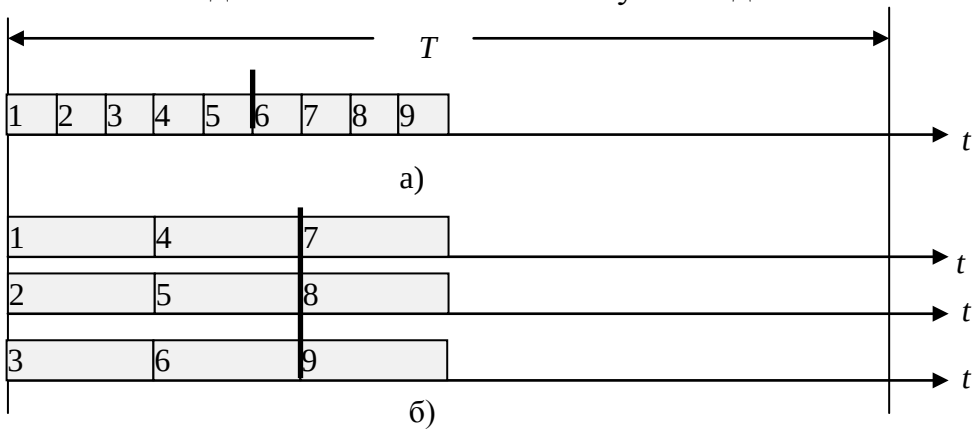


Рис. 42. Соотношение метрик 5 : 6 при больших очередях:
а) для одиночного интерфейса; б) трех параллельных интерфейсов

12.3. Расчет буфера памяти маршрутизатора для UDP-соединений

Каждое новое поколение технологии коммутации имеет тенденцию к расширению услуг, предоставляемым абонентам, с одновременным понижением качества обслуживания в части воспроизведения исходной информации реального времени. Для оценки качества передачи голосового сигнала Рекомендация ITU-T P.800 [10] использует шкалу средней экспертной оценки (Mean Opinion Scores, MOS), приведенную в табл. 12.1.

Таблица 12.1

Шкала MOS

MOS-оценка	Описание качества
5	Междугородное качество, соответствующее качеству передачи по медным физическим проводам
4	Качество, близкое к междугородному, с почти незаметными искажениями
3	Ощутимые искажения, приводящие к повторению части сказанного
2	Искажения весьма раздражают
1	Невозможно вести разговор

Эта таблица характеризует сходство исходной речевой информации с информацией на приеме, полученной после ее прохождения по сети связи. Оценка MOS для сетей с различной технологией коммутации представлена в табл. 12.2.

Таблица 12.2

MOS для различных сетей связи

MOS-оценка	Тип сети связи
5	ТФОП – Телефонная сеть связи общего пользования (PSTN – Public Switching Telephone Network)
4,3	ЦСИО – цифровая сеть интегрального обслуживания (ISDN – Integrated Services Digital Network)
3,6–3,9	Internet – Сеть пакетной коммутации

Вновь разрабатываемые сети связи не должны существенно ухудшать характеристики, применяемые в существующих сетях. IP-сеть является преемником цифровой сети интегрального обслуживания (ЦСИО), для которой нормы качества обслуживания при установлении соединений прописаны в Рекомендации Q.543 [8].

Сетевое качество обслуживания нормируется для самого «длинного» соединения. Для эталонной модели ISDN в «длинном» соединении участвует максимальное число проходных узлов между абонентами: оконечная станция (ОС) – узловая станция (УС) – междугородная станция (АМТС) – узел автоматической коммутации (УАК) – международный центр коммутации (МЦК) – МЦК – УАК – АМТС – УС – ОС. Итого – 10 сетевых элементов.

Задержка сигнального сообщения «из конца в конец» при установлении внутреннего или входящего соединения к абоненту – момент времени между окончанием приема последней цифры и моментом подключения зуммерного сигнала «Контроль посылки вызова» характеризуется табл. 12.3:

Таблица 12.3

Нормы задержки при установлении соединения

Нормы задержки	Нагрузка типа А	Нагрузка типа В
Среднее значение (не более), мс	650	1000
Достижение ФРВ окончания обслуживания на уровне 0,95 (не более), мс	900	1600

По аналогии при установлении TCP-соединения в IP-сети через S = 10 проходных маршрутизаторов и нагрузке типа В среднее значение задержки сигнального сообщения на одном маршрутизаторе не должно превышать 100 мс, а критическое значение задержки равно 160 мс. Это достаточно большие временные значения, поэтому проблем в части перегрузки интерфейса они не создают и обеспечивают качество не хуже, чем в ISDN.

Для доставки информации пользователя в реальном времени используются системы интегрированного или дифференцированного обслуживания, а нормы качества обслуживания приведены в Рекомендациях Y.1540 и Y.1541 [16, 17].

В системе интегрированного обслуживания при занятом интерфейсе маршрутизатора поступивший пакет ставится на ближайшее (первое) место ожидания в своем индивидуальном буфере очереди. За один период прохода всех очередей пакет обязательно будет передан, в буфере очереди всегда будет находиться не более одного пакета. Поэтому в системе интегрированного обслуживания проблем с перегрузкой интерфейса для приоритетных соединений не существует.

В системе дифференцированного обслуживания при занятом интерфейсе маршрутизатора поступивший пакет ставится в конец приоритетной очереди. За один период прохода всех приоритетных и не приоритетных очередей из каждой очереди считываются пакеты, стоящие первыми. Объем считанных пакетов из каждой очереди прямо пропорционален ширине вы-

деленной для этого приоритета полосы. При высокой интенсивности поступления приоритетных потоков в буфере очереди могут оставаться необслуженные пакеты, которые будут обслужены при следующем проходе очередей. Максимальный объем буфера очереди в маршрутизаторе для приоритетных пакетов выбирается исходя из допустимой задержки в маршрутизаторе, так как слишком большая задержка приведет к задержке воспроизведения всех последующих пакетов в приемном устройстве.

Рассмотрим процесс воспроизведения UDP-пакетов на приеме (рис. 43), где светло-серым цветом помечены пакеты, пришедшие с отставанием от среднего времени доставки сетью *tnet*, а темно-серым – с упреждением.

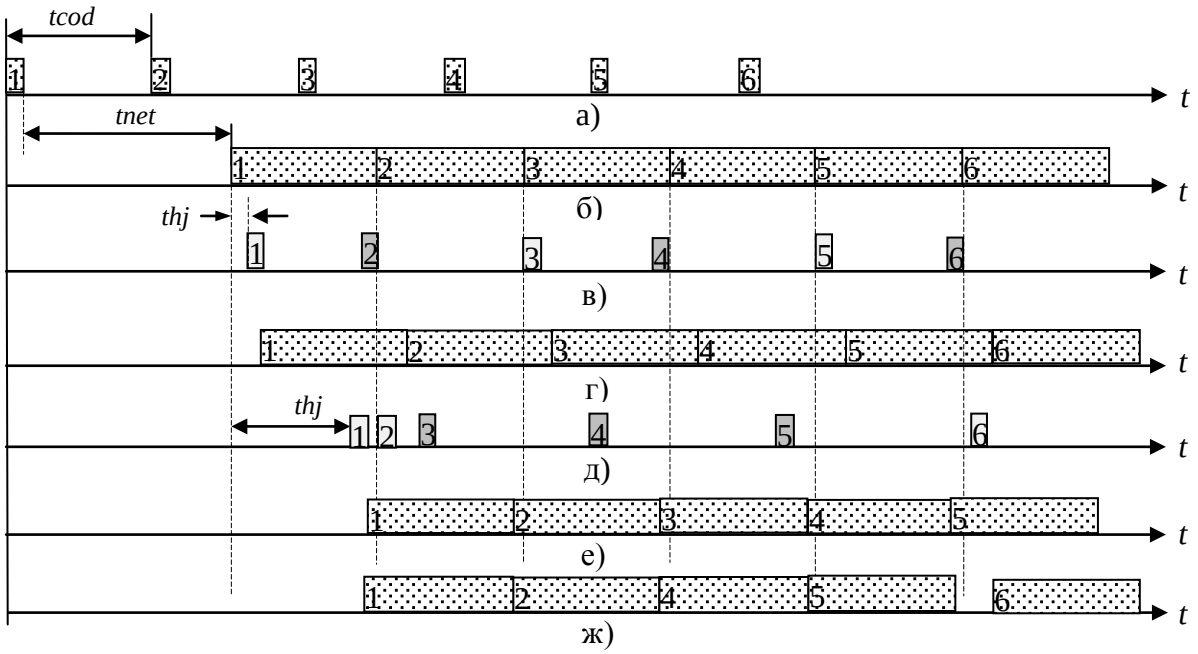


Рис. 43. Процесс воспроизведения информации на приеме
а) порядок следования пакетов на передаче;
б) порядок воспроизведения пакетов (*PLD*) на приеме при отсутствии джиттера;
в) порядок поступления пакетов на приеме при небольшом джиттере;
г) порядок воспроизведения пакетов на приеме при небольшом джиттере;
д) порядок поступления пакетов на приеме при большом джиттере;
е) порядок воспроизведения при большом большом буфере;
ж) порядок воспроизведения при ограниченном буфере

При отсутствии джиттера сети (рис. 43, б) объем приемного буфера хоста равен полезной нагрузке ($hbuf = PLD$), все пакеты воспроизводятся полностью с задержкой *tnet*.

При небольшом джиттере сети *thj* (рис. 43, в) для полного воспроизведения всех речевых пакетов необходимо иметь небольшой приемный буфер ($hbuf \geq PLD$) для компенсации задержки *thj*, все пакеты воспроизводятся полностью с задержкой $tnet + thj$ (рис. 43, г).

При большом джиттере сети *thj* (рис. 43, д) для полного воспроизведения всех речевых пакетов необходимо иметь большой приемный буфер (*hbuf* > *PLD*) для компенсации задержки *thj*, все пакеты воспроизводятся полностью с задержкой *tnet* + *thj* (рис. 43, е).

Для воспроизведения речи в реальном масштабе времени объем приемного буфера ограничивают, при этом из-за запоздавшего пакета (пакет 1 на рис. 43, ж) некоторые последующие пакеты (пакет 3) в приемный буфер (на два пакета) не помещаются и не воспроизводятся.

Поэтому одной из задач сети является ограничение объема буфера маршрутизатора для уничтожения невоспроизводимых пакетов (пакет 3 на рис. 43, ж).

Для удаления невоспроизводимых пакетов из IP-сети можно воспользоваться алгоритмом произвольного раннего обнаружения перегрузки (Random Early Detection, RED) для TCP-соединений, когда маршрутизатор ограничивает длину очереди. Вероятность отбрасывания пакета из очереди вычисляется по формуле [1]:

$$prob = \frac{Q - R\min}{R\max - R\min},$$

где Q – текущий размер очереди, Rmin – значение длины очереди буфера, с которого начинает действовать алгоритм RED, Rmax – максимальное значение длины очереди в байтах.

При малом значении очереди (Q<Rmin) отбрасывания пакетов не происходит, при превышении средней длины очереди допустимого значения (Rmin < Q < Rmax) вероятность отбрасывания пакетов линейно возрастает, а при достижении очереди своего максимального значения (Q = Rmax) отбрасываются все вновь поступившие пакеты.

Алгоритм RED может быть применим и для приоритетной очереди с UDP-соединениями. Для TCP-соединений алгоритм RED обеспечивает защиту от перегрузки для всей сети (от источника до получателя пакетов). Для UDP-соединений алгоритм RED обеспечивает защиту от перегрузки только части сети (от того маршрутизатора, где произошло отбрасывание пакета до получателя пакетов).

В табл. 12.4 приведены сетевые характеристики эталонной модели IP-сети разд. 11.2 при использовании различных типов кодеков при $Fe(tnet) \geq 0{,}999$ для системы с дифференцированным обслуживанием.

Как видно из таблицы при использовании различных кодеков значения $R\min = BW \cdot te \cdot LENp/LEN$ и $R\max = BW \cdot tep \cdot LENp/LEN$ значительно разнятся.

Таблица 12.4

Сетевые характеристики сети для различных кодеков

Характеристики	G.711(10)	G.711(20)	G.711(30)	G.729	G.723
tcod + *tdec*, мс	10	20	30	20	45
LEN, байт	158	238	318	88	98
λ, $с^{-1}$	1448	891	605	2724	2383
λ_0, $с^{-1}$	100	50	33,3	66,6	26,6
$\mu = BW/LEN$, $с^{-1}$	1620	1075	805	2909	2612
$a = \lambda/\mu$, Эрл	0,89	0,83	0,75	0,93	0,91
te, мс	7,56	6,91	6,26	7,09	**5,7**
$thj = te \cdot \sqrt{S}/(S - 1)$, мс	7,97	7,28	6,6	7,4	**6**
$tnet = tsum - tcod - tdec - thj$, мс	132	122,7	113	122,5	**99,9**
$tep = tnet/S$, мс	13,2	12,2	11,3	12,2	**9,9**
PLD, байт	80	160	240	10	20
$Rmin = BW \cdot te \cdot LENp/LEN$, байт	1104	1115	1056	774	668
$Rmax = BW \cdot tep \cdot LENp/LEN$, байт	1925	1979	1917	1336	1163
$hbuf = 2 \cdot thj \cdot \lambda_0 \cdot PLD + PLD$, байт	208	276	346	20	26

В IP-сети могут использоваться различные типы кодеков. Поскольку маршрутизатор никогда не анализирует поле данных типа используемого кодека, то при расчете параметров маршрутизатора из табл. 12.4 следует ориентироваться на самое малое время прохождения через маршрутизатор *te* = 5,7 мс, *tep* = 9,9 мс и средние значения длин пакетов *LENm* и *LENpm*. Приложению хоста параметры кодека известны, поэтому текущий объем буфера *hbuf* приложения хоста индивидуален и вычисляется на основании временных штампов (RTP Timestamp) протокола RTP.

В табл. 12.5 приведены пересчитанные *предельные* общие сетевые характеристики сети для различных кодеков. Курсивом выделены изменяемые параметры в зависимости от выделенной полосы *BW*, а жирным шрифтом – неизменные параметры сети для всех кодеков. Значения *hbuf* округлены до наибольшего целого, кратного байту.

Таблица 12.5

Пересчитанные характеристики сети для различных кодеков

Характеристики	G.711(10)	G.711(20)	G.711(30)	G.729	G.723
tcod + *tdec*, мс	**10**	**20**	**30**	**20**	**45**
BW, байт/с	*192 000*				
LENm, байт	**180**				
LENpm, байт	**142**				
λ_0, $с^{-1}$	**100**	**50**	**33,3**	**66,6**	**26,6**
te, мс	**5,7**				
thj, мс	**6**				
tnet, мс	**99,9**				
tep, мс	**9,9**				
PLD, байт	**80**	**160**	**240**	**10**	**20**
Rmin = 0,00449 · *BW*, байт	*862*				
Rmax = 0,00781 · *BW*, байт	*1500*				
hbuf = 2 · *thj* · λ_0 · *PLD* + *PLD*, байт	**176**	**256**	**336**	**18→24**	**26→32**

ЗАДАЧА 1. Какая Рекомендация нормирует время доставки голосового сигнала «из конца в конец» в IP–сети?

Ответ. Рекомендация ITU–T G.114.

ЗАДАЧА 2. Какая Рекомендация нормирует качество обслуживания сетевого элемента в системе коммутации каналов?

Ответ. Рекомендация ITU–T Q.543.

ЗАДАЧА 3. Какие Рекомендации нормируют качество обслуживания сети в системе коммутации пакетов?

Ответ. Рекомендации Y.1540 и Y.1541.

ЗАДАЧА 4. Почему в сети PSTN качество воспроизведения голосового сигнала выше, чем в IP-сети?

Ответ. В сети PSTN происходит минимальное преобразование исходной аналоговой информации, только ограничение спектра.

В IP-сети происходит существенное преобразование исходной информации: ограничение спектра, выборка, кодирование, сжатие, частичная потеря информации. В результате на приеме теряется сходство с исходным аналоговым сигналом.

ЗАДАЧА 5. Почему в сети PSTN не реализованы услуги, существующие в сети пакетной коммутации?

Ответ. При существующем этапе развитии техники реализация некоторых услуг (например, электронная почта) в сети PSTN обходится намного дороже, чем в сети пакетной коммутации.

ЗАДАЧА 6. Какой транспортный протокол используется в IP-сети для установления соединений?

Ответ. TCP.

ЗАДАЧА 7. Какой транспортный протокол используется в IP-сети для доставки пакетов пользователя «из конца в конец»?

Ответ. UDP.

ЗАДАЧА 8. Дайте определение джиттера сети.

Ответ. Джиттер – это стандартное отклонение задержки от среднего значения при доставке пакета «из конца в конец».

ЗАДАЧА 9. Почему хост имеет ограниченный объем буфера?

Ответ. Чем больше объем буфера хоста, тем больше задержка в воспроизведении содержимого пакетов. Для воспроизведения в реальном масштабе времени приходится ограничивать его объем. При этом запоздавшие пакеты приводят к неполному воспроизведению последующих, и качество воспроизведения речи ухудшается.

ЗАДАЧА 10. Как вычисляется объем буфера маршрутизатора $R\min$ для приоритетных соединений, с которого начинает действовать алгоритм RED?

Решение.

Число байтов, принятых маршрутизатором за время te, равно $BW \cdot te$, включая байты канального и физического уровней. В буфере маршрутизатора будет храниться только $R\min = BW \cdot te \cdot LENpm/LENm$ байт = $= 0{,}00449 \cdot BW$ байт.

Соответственно, $R\max = BW \cdot tep \cdot LENpm/LENm$ байт $= 0{,}00781 \cdot BW$ байт.

ЗАДАЧА 11. Какие особенности следует учитывать на хосте при приеме потокового видео?

Ответ. При приеме потокового видео критичным является джиттер сети, а среднее время доставки не столь критично. Чтобы компенсировать джиттер сети, буфер хоста должен накопить достаточное количество пакетов. При этом возникает постоянная задержка, и начало воспроизведения практически может достигать несколько секунд (рис. 43, е).

13. РАСПРЕДЕЛЕНИЕ КАНАЛОВ СВЯЗИ НА «ПОСЛЕДНЕЙ МИЛЕ СЕРВЕРА»

Условно трафик в IP-сетях можно разделить на симметричный и ассиметричный. К первым относится голосовые соединения (VoIP), ко вторым – все остальные.

Асимметричность обмена данными происходит по очень простой причине. Компьютер подключается к сети не только для обмена данными с другим таким же компьютером, но и для того, чтобы использовать информационные ресурсы сети, превращаясь в потребителя информации. Пользователь компьютера посылает простые короткие запросы, а сервер в ответ возвращает ему большой объем данных. Обмен информацией становится сильно асимметричным.

Многие статистические данные показывают, что:

1) UDP-трафик на порядок меньше TCP–трафика.

2) UDP-трафик в целом симметричен.

3) TCP-трафик является сильно асимметричным, причем объем входящих данных от сети в 2.8 раза превосходит исходящий от абонента.

Большой объем трафика от сервера к получателю может привести к ее длительной задержке. Между тем величина задержки нормируется Рекомендациями Y.1540 [16] и Y.1541 [17].

В соответствии с Рекомендацией Y.1540, если от источника пакет отправлен в момент времени $t1$, а получателем принят в момент $t2$, то передаваемый по IP-сети пакет может быть либо успешно доставлен, либо поврежден, доставлен не по адресу, утерян.

Успешно доставленным пакетом считается неповрежденный пакет, доставленный по адресу за нормированное время $T\max \geq (t2 - t1)$.

Поврежденным пакетом считается пакет, доставленный по адресу за нормированное время $T\max$, содержательная часть которого искажена.

Пакетом, доставленным не по адресу, считается пакет, принятый от постороннего источника.

К утерянным пакетам причисляют пакеты, не дошедшие до получателя или дошедшие с опозданием $(t2 - t1) \geq T\max$.

Рекомендация Y.1540 определяет следующие основные параметры доставки пакетов «из конца в конец»:

– задержка доставки пакетов (IP packet transfer delay, IPTD). Она определяется как разность $(t2 - t1) \leq T\max$, где $t1$ – время отправки от источника, $t2$ – время прибытия пакета в пункт назначения, $T\max$ – допустимое время доставки, по истечение которого пакет считается потерянным. Среднее время доставки пакетов определяется как среднее арифметическое совокупности пакетов за время наблюдения не менее 1 мин;

– вариация задержки пакетов (IP packet delay variation, IPDV), иначе – джиттер. Она определяется как разность между реальной величиной задержки x_k и эталонной (по существу средней) величиной задержки доставки пакета $d_{1,2}$: $v_k = x_k - d_{1,2}$. Положительные значения джиттера образуются в случае, если время доставки пакета больше эталонной задержки, отрицательные – если меньше. Компенсация джиттера осуществляется буфером приемного хоста;

– коэффициент ошибочных пакетов (IP packet error ratio, IPER). Определяется как отношение числа поврежденных пакетов к суммарному числу поврежденных пакетов и успешно доставленных пакетов;

– коэффициент потери пакетов IP (IP packet loss ratio, IPLR). Определяется как отношение числа утерянных пакетов к суммарному числу переданных пакетов.

Рекомендация Y.1541 определяет численные значения параметров для различных услуг, существующих в настоящий момент. Численные значения рассмотренных параметров приведены в табл.13.1.

Таблица 13.1

Численные значения параметров доставки

Параметр	Описание	Классы качества обслуживания					
		Класс 0	Класс 1	Класс 2	Класс 3	Класс 4	Класс 5
IPTD	Верхняя граница среднего значения (Прим.1)	100 мс	400 мс	100 мс	400 мс	1 с	U
IPDV	Разность (IPTDmax – IPTDmin) с квантилем 0,999 (Прим. 2)	50 мс	50 мс	U	U	U	U
IPLR	Верхняя граница	0,001	0,001	0,001	0,001	0,001	U
IPER	Верхняя граница	0,0001	0,0001	0,0001	0,0001	0,0001	U

Примечания:
U означает «не специфицировано».
Прим. 1. Определяется как среднее арифметическое совокупности пакетов. В некоторых случаях это значение для классов 0 и 2 не всегда достижимо. С другой стороны, не исключается возможность меньшего значения.
Прим. 2. Джиттер IPTD (отклонение от среднего значения 100 мс) в 99,9 % не должен превышать 50 мс в течение коротких интервалов времени (не более 1 мин для класса 0). Для длинных интервалов времени в этом же случае используется квантиль 95 % (Appendix II).

Рекомендация Y.1541 устанавливает соответствие между классом качества обслуживания и наиболее подходящими ему приложениями:

- класс 0 – приложения реального времени, чувствительные к джиттеру, характеризуемые высоким уровнем интерактивности (VoIP, видеоконференции через наземные каналы связи);
- класс 1 – приложения реального времени, чувствительные к джиттеру, характеризуемые невысоким уровнем интерактивности (потоковое Video, видеоконференции через спутниковые каналы связи);
- класс 2 – транзакции данных, характеризуемые высоким уровнем интерактивности (например, сигнализация);
- класс 3 – обмен данными с низким уровнем интерактивности (Web-сайты, электронная почта);
- класс 4 – приложения, допускающие низкий уровень потерь (короткие транзакции, большие массивы данных);
- класс 5 – традиционные применения сетей IP.

По мере усовершенствования сетей будут вводиться и новые классы 6 и 7 (табл. 13.2), которые в настоящее время являются временными и требуют дальнейшего изучения.

Таблица 13.2

Временные численные значения параметров доставки

Параметр	Описание	Классы качества обслуживания	
		6	7
IPTD	Верхняя граница среднего значения	100 мс	400 мс
IPDV	Верхняя граница разности (IPTD – mean IPTD) с квантилем 0,999	50 мс	
IPLR	Верхняя граница	$1x10^{-5}$	
IPER	Верхняя граница	$1x10^{-6}$	

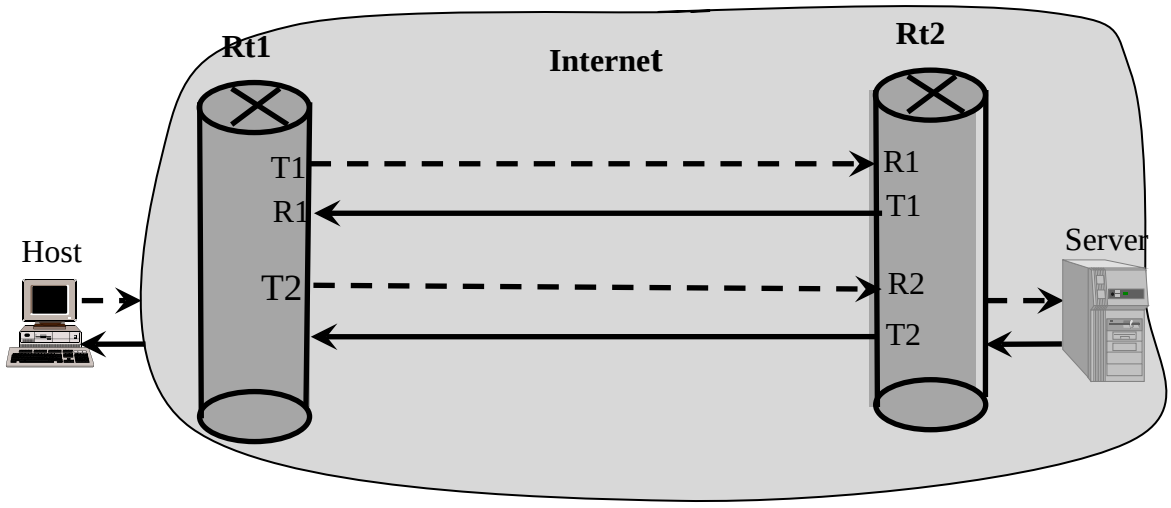
Перекос трафика на абонентском участке «последней миле абонента» компенсируется использованием технологии ADSL (Asymmetric Digital Subscriber Line — асимметричная цифровая абонентская линия), в которой общая доступная полоса пропускания канала распределена между исходящим и входящим трафиками асимметрично. Полоса пропускания от сети к пользователю значительно превосходит полосу от пользователя к сети связи.

В ADSL общая полоса пропускания делится на две части: частоты от 26 кГц до 138 кГц отведены исходящему потоку данных (от пользователя к сети), а частоты от 138 кГц до 1,1 МГц — входящему (от сети к пользователю).

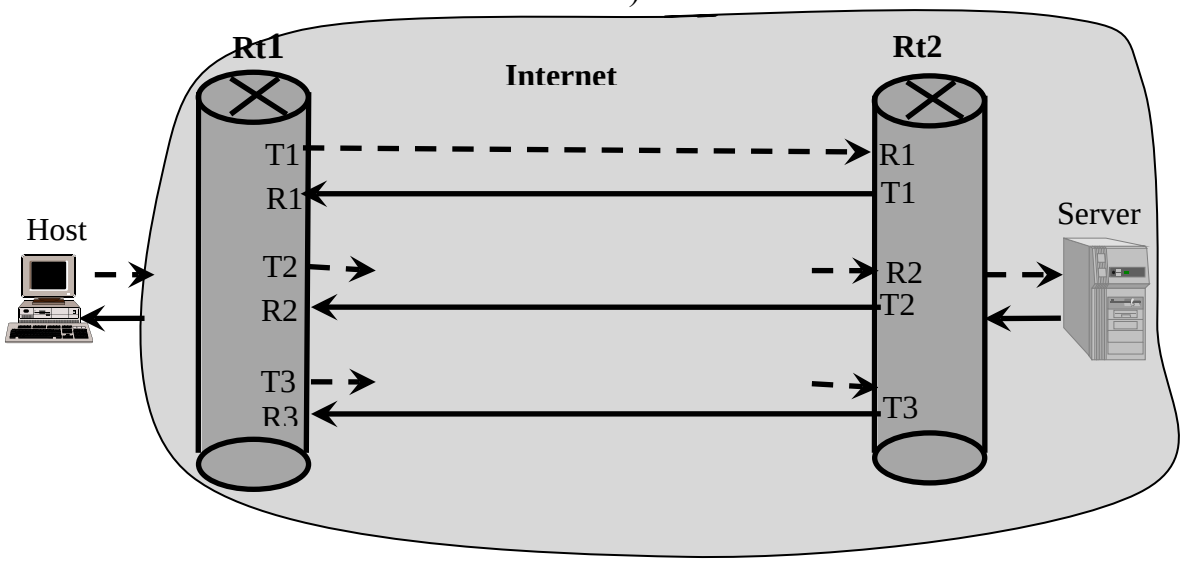
Перекос трафика на участке «последняя миля сервера» (предпоследний – последний маршрутизатор) можно компенсировать перераспределением каналов связи, как показано на рис. 44.

При симметричном распределении каналов число каналов связи в восходящем (от пользователей к серверу) и нисходящем (от сервера к пользователям) направлениях одинаково. На рис. 44, а показано распределение четырех каналов связи в соотношении 2↑/2↓.

Асимметричное распределение каналов (рис. 44, б) достигается перенаправлением каналов связи и введением дополнительных портов маршрутизаторов R3 и T3. На рис. 44, б показано распределение четырех каналов связи в соотношении 1↑/3↓.



а)



б)

Рис. 44. Использование четырех каналов связи на «последней миле сервера»
а) при симметричном распределении каналов;
б) при асимметричном распределении каналов

Маршрутизатор Rt1 в направлении пользователи – сервер использует только приемные порты R2 и R3, передающие порты T2 и T3 блокируются.

Маршрутизатор Rt2 в направлении сервер – пользователи использует только передающие порты T2 и T3, приемные порты R2 и R3 блокируются.

Эффект повышения качества обслуживания достигается за счет такого перераспределения каналов связи, чтобы восходящему трафику (к серверу) выделялось меньшее число каналов связи, чем нисходящему трафику (от сервера).

Экономический эффект достигается тем, что стоимость дополнительных портов маршрутизаторов значительно ниже стоимости канала связи.

Для оценки предложенного решения без учета надежности рассмотрим качество обслуживания на «последней миле сервера» для обоих вариантов при одних и тех же значениях трафика. Оценку будем производить по функции распределения времени ожидания конца обслуживания, используя формулу (5.4):

$$F_e(t) = \begin{cases} 1 + \dfrac{P(>0)}{V - A - 1}[e^{-\mu(V-A)t} - e^{-\mu t}] - e^{-\mu t}, & 0 < A \neq V - 1 \\ 1 - \mu t P(>0) e^{-\mu t} - e^{-\mu t}, & A = V - 1 \end{cases}, \qquad (13.1)$$

где

$$P(>0) = \frac{V \cdot E_V(A)}{V - A \cdot \left(1 - E_V(A)\right)}, \qquad A = \frac{\lambda}{\mu}.$$

Пусть скорость каждого интерфейса в обеих направлениях *VEL* = = 10 Мбит/с. Каждый из *N* = 1 000 000 пользователей осуществляет сеанс связи один раз в 10 мин, создавая общий поток к серверу λ = 1666,6 пакетов/с. Длина пакета запроса к серверу *LEN_U* = 200 байт, на который сервер отвечает пакетом длиной *LEN_S* = 1400 байт.

Симметричное распределение каналов 2↑/2↓.

- В направлении пользователи – сервер:
  - число каналов связи $V = 2$;
  - интенсивность обслуживания пакетов $\mu = VEL/LEN_U = 6250$ с$^{-1}$;
  - нагрузка от пользователей к серверу $A = \lambda/\mu = 0{,}2666$ Эрл;
  - $P(>0) = 0{,}0313$.
- В направлении сервер – пользователи:
  - число каналов связи $V = 2$;
  - интенсивность обслуживания пакетов $\mu = VEL/LEN_S = 892{,}85$ с$^{-1}$;
  - нагрузка от сервера к пользователям $A = \lambda/\mu = 1{,}866$ Эрл;
  - $P(>0) = 0{,}9011$.

Подставляя эти значения в (13.1), построим графики как на рис. 45 ФРВ конца обслуживания в направлении пользователи – сервер *Fus(t)* и ФРВ конца обслуживания в направлении сервер – пользователи *Fsu(t).*

Рис. 45. Функции распределения времени конца обслуживания пакета на «последней миле сервера» при симметричном распределении каналов

Ассимметричное распределение каналов 1↑/3↓.

- В направлении пользователи – сервер:
- число каналов связи $V = 1$;
- интенсивность обслуживания пакетов $\mu = VEL/LEN_U = 6250$ с$^{-1}$;
- нагрузка от пользователей к серверу $A = \lambda/\mu = 0{,}2666$ Эрл;
- $P(>0) = 0{,}2666$.
- В направлении сервер – пользователи:
- число каналов связи $V = 3$;
- интенсивность обслуживания пакетов $\mu = VEL/LEN_S = 892{,}85$ с$^{-1}$;
- нагрузка от сервера к пользователям $A = \lambda/\mu = 1{,}866$ Эрл;
- $P(>0) = 0{,}3837$.

Подставляя эти значения в (13.1), построим графики как на рис. 46 ФРВ конца обслуживания в направлении пользователи – сервер *Fus(t)* и ФРВ конца обслуживания в направлении сервер – пользователи *Fsu(t).*

Рис. 46. Функции распределения времени конца обслуживания пакета на «последней миле сервера» при асимметричном распределении каналов

Сравнивая рис. 45 и 46, можно сделать следующие выводы.

1. В направлении пользователи – сервер качество обслуживания практически не изменилось.

2. В направлении сервер – пользователи ФРВ конца обслуживания на уровне 0,95 снизилось с 26 мс до 6 мс. Таким образом, качество обслуживания значительно улучшилось.

Для обеспечения надежности следует использовать следующие схемы асимметричного распределения каналов: 2↑/4↓, 2↑/6↓, 3↑/5↓, 2↑/8↓, 3↑7↓, 4↑/6↓, …

14. ОЦЕНКА НАДЕЖНОСТИ СЕТЕВЫХ ЭЛЕМЕНТОВ

Под системой технического обслуживания понимается комплекс мер со стороны оператора для поддержания сети связи в работоспособном состоянии. Известны две ***системы технического обслуживания (*СТО*)***: децентрализованная и централизованная.

При ***децентрализованной системе технического обслуживания*** на каждом узле коммутации имеется свой обслуживающий персонал, который занимается обнаружением и устранением неисправностей. Накопления неисправностей не происходит, т. е. при возникновении неисправности обслуживающий персонал сразу же приступает к ее устранению. Такая система широко применялась в электромеханических узлах коммутации, в которых неисправности возникали достаточно часто, что не позволяло оставлять оборудование без присмотра на какой-либо длительный срок.

При ***централизованной системе технического обслуживания*** только на одном (главном) узле коммутации организуется центр технического обслуживания (ЦТО), где размещаются обслуживающий персонал и несколько ремонтно-восстановительных бригад. При этом подопечные соседние узлы коммутации не имеют постоянного обслуживающего персонала и функционируют в необслуживаемом режиме, без постоянного присутствия обслуживающего персонала. Накопление неисправностей на подопечных необслуживаемых узлах контролируется дистанционно из ЦТО; при достижении критического состояния наблюдаемого необслуживаемого узла на него из ЦТО выезжает одна из ремонтно-восстановительных бригад.

Переход к централизованной системе технического обслуживания стал возможным благодаря двум факторам: резко возросшей надежности сетевых элементов и применению хорошо отработанных протоколов дистанционного контроля (например, SNMP) [11].

Надежность системы оценивается ***коэффициентом готовности,*** под которым понимается доля времени, в течение которого вызовы обслуживаются с заданным качеством. На коэффициент готовности оборудования решающее влияние оказывают два параметра: среднее время наработки оборудования на отказ $T\xi$ и время его восстановления Tr.

Среднее время наработки на отказ $T\xi$ (MTBF – Mean Time Between Failures) равно промежутку времени от начала функционирования в исправном состоянии до момента отказа. Величина ξ, обратная времени наработки на отказ, называется **интенсивностью отказов**.

Время восстановления Tr включает в себя три составляющие: время обнаружения неисправности, время прибытия ремонтно-восстановительной

бригады и время замены неисправного оборудования. Время обнаружения неисправности зависит от глубины программно-аппаратной проработки СТО в части локализации неисправности; время прибытия ремонтно-восстановительной бригады – от децентрализованной или централизованной СТО; время замены имеет практически постоянное значение. Величина r, обратная времени восстановления, называется **интенсивностью восстановления**.

Нормируемые Рекомендацией ITU–T E.550 [6] параметры системы коммутации – Tr = 0,4 ч простоя в год ($T\xi$ = 8760 ч), что соответствует коэффициенту готовности $K = T\xi/(T\xi + Tr) \approx 0{,}99995$. Коэффициент неготовности $\underline{K} = (1 - K)$ определяет потери по времени из-за неисправности оборудования.

В обеих системах технического обслуживания для повышения надежности может использоваться один из двух способов резервирования.

Функционально–задействованное резервирование (режим разделения нагрузки, load-sharing redundancy) характеризуется тем, что резервное оборудование участвует в обслуживании вызовов вместе с основным оборудованием и обслуживает половину поступающей нагрузки. Основное и резервное оборудование имеют ***независимые базы данных*** установленных соединений. Выход из строя основного *или* резервного оборудования приводит к потере половины установленных соединений. Оставшееся исправное оборудование продолжает обслуживание всей поступающей нагрузки, но с пониженным качеством. Выход из строя основного *и* резервного оборудования приводит к отказу системы. Этот способ резервирования наиболее характерен для систем с коммутацией пакетов. В сети коммутации пакетов применяют многократное резервирование, когда исходящая нагрузка распределяется между несколькими интерфейсами маршрутизатора (между несколькими последующими маршрутизаторами).

Функционально-незадействованное резервирование (режим замещения, hot-stand-by redundancy) характеризуется тем, что резервное оборудование постоянно отслеживает и копирует состояние основного оборудования, ***имея одну и ту же базу данных*** установленных соединений. Выход из строя основного или резервного оборудования не приводит к снижению качества обслуживания. Выход из строя основного и резервного оборудования приводит к отказу всей системы. Этот способ резервирования применяется тогда, когда потеря ранее установленных соединений приводит к весьма негативным последствиям и применяется, например, для резервирования коммутационных полей в системах с коммутацией каналов. В случае сложности или дороговизны основного оборудования многократное резервирование, как правило, не применяется.

При воздействии на систему нескольких типов неисправностей с интенсивностями отказов/восстановлений $\{\xi_1, r_1\}$, $\{\xi_2, r_2\}$, … $\{\xi_m, r_m\}$ коэффициент готовности вычисляется по формуле $K = r/(\xi + r)$, где $\xi = 1/T\xi$ – суммарная интенсивность отказов, $r = 1/Tr$ – суммарная интенсивность восстановления. Последние определяются из равенств:

$$\xi = \xi_1 + \xi_2 + ... + \xi_m = 1/T\xi_1 + 1/T\xi_2 + ... + 1/T\xi_m,$$

$$r = p_1 r_1 + ... + p_m r_m = \frac{\xi_1}{\xi_1 + \xi_2 + ... + \xi_m} r_1 + ... + \frac{\xi_m}{\xi_1 + \xi_2 + ... + \xi_m} r_m.$$

ЗАДАЧА 1. Статистически замерено, что устройство выходит из строя по причине перенапряжения в среднем через 400 ч, а по причине износа элементов – через 500 ч. Определить суммарное время наработки на отказ.

Решение.

Интенсивность отказов вследствие перенапряжения $\xi 1 = (400\ ч)^{-1} = 2{,}5 \cdot 10^{-3}\ ч^{-1}$.

Интенсивность отказов по причине старения элементов $\xi 2 = (500\ ч)^{-1} = 2 \cdot 10^{-3}\ ч^{-1}$.

Суммарная интенсивность отказов $\xi = \xi 1 + \xi 2 = 4{,}5 \cdot 10^{-3}\ ч^{-1}$.

Суммарное время наработки на отказ $T\xi = 1/\xi = 222$ ч.

ЗАДАЧА 2. Статистически замерено, что устройство выходит из строя по причине перенапряжения в среднем через 400 ч, а из–за старения элементов – через 500 ч. Неисправность первого типа устраняется через 5 ч, неисправность второго типа устраняется через 10 ч. Определить коэффициент готовности системы.

Решение.

Интенсивность отказов по причине перенапряжения $\xi 1 = (400\ ч)^{-1} = 2{,}5 \cdot 10^{-3}\ ч^{-1}$.

Интенсивность отказов по причине старения элементов $\xi 2 = (500\ ч)^{-1} = 2 \cdot 10^{-3}\ ч^{-1}$.

Суммарная интенсивность отказов $\xi = \xi 1 + \xi 2 = 4{,}5 \cdot 10^{-3}\ ч^{-1}$.

Суммарное время наработки на отказ $T\xi = 1/\xi = 222$ ч.

Интенсивность восстановления отказов по причине перенапряжения $r1 = (5\ ч)^{-1} = 0{,}2\ ч^{-1}$.

Интенсивность восстановления отказов по причине старения элементов $r2 = (10\ ч)^{-1} = 0{,}1\ ч^{-1}$.

Суммарная интенсивность восстановления неисправностей

$$r = p_1 r_1 + p_2 r_2 = \frac{\xi_1}{\xi_1 + \xi_2} r_1 + \frac{\xi_2}{\xi_1 + \xi_2} r_2 = \frac{2,5}{4,5} \cdot 0,2 + \frac{2}{4,5} \cdot 0,1 = 0,15 = 150 \cdot 10^{-3} \text{ ч}^{-1}.$$

Тогда коэффициент готовности

$$K = r / (\xi + r) = \frac{150}{4,5 + 150} = 0,97 .$$

ЗАДАЧА 3. Программный коммутатор (Soft Switch) без дублирования имеет время наработки на отказ три месяца (0,25 года). Применяется система централизованного обслуживания, время устранения неисправности – 1,5 ч. Определить коэффициент готовности.

Решение.

Интенсивность отказов $\xi = (0,25 \text{ года})^{-1} = 4 \text{ год}^{-1} = 4,56 \cdot 10^{-4} \text{ ч}^{-1}$.

Интенсивность восстановления $r = (1,5 \text{ ч})^{-1} = 0,66 \text{ ч}^{-1}$.

Коэффициент готовности $K = r / (\xi + r) = 0,9993$.

ЗАДАЧА 4. Определить коэффициент готовности предыдущего примера при функционально-незадействованном дублировании программного коммутатора.

Решение.

Интенсивность отказов $\xi = (0,25 \text{ года})^{-1} = 4,56 \cdot 10^{-4} \text{ ч}^{-1}$.

Интенсивность восстановления $r = (1,5 \text{ ч})^{-1} = 0,66 \text{ ч}^{-1}$.

Интенсивность нагрузки неисправностей $B = \xi / r = 6,9 \cdot 10^{-4}$ Эрл.

Коэффициент неготовности дублированной схемы численно равен вероятности отказа двух элементов (4.2):

$$\underline{K} = E_2(B) = \frac{B^2 / 2}{1 + B + B^2/2} = 4,55 \cdot 10^{-7}.$$

Коэффициент готовности $K = 1 - \underline{K} = 0,9999997$.

Среднее время выхода из строя в обслуживаемом режиме

$$T_2 = \frac{1}{\xi \; E_1(B)} = \frac{1 + B}{\xi \; B} = 3\,180\;428 \text{ ч} = 362 \text{ года} .$$

15. ПЕРЕХОДНЫЕ ВЕРОЯТНОСТИ СОСТОЯНИЙ В ПОЛНОДОСТУПНОМ ПУЧКЕ

Стационарный процесс характеризуется неизменностью значений вероятности на фиксированном промежутке времени, независимо от расположения этого промежутка времени на оси времени, поэтому вероятности таких событий не зависят от времени. Подавляющую долю времени сетевые элементы функционируют в стационарном режиме, что характерно для ранее рассмотренных моделей.

Вместе с тем имеются процессы, вероятность реализации которых зависит от времени. Они получили название **переходных процессов**.

Переходные процессы возникают при резком изменении какого-либо параметра сетевого элемента: интенсивности потока поступающих вызовов, числа обслуживающих приборов, полосы пропускания либо при смене основного маршрута следования соединения и т. д.

При неизменности параметров системы с течением времени переходный процесс заканчивается стационарным процессом. Таким образом, стационарному процессу предшествует переходный процесс.

Ниже рассматриваются переходные процессы, которые имеют место в централизованных (необслуживаемых) системах технического обслуживания.

Переходный процесс рождения хорошо описывает накопление неисправностей на необслуживаемых узлах коммутации, а процесс гибели – их устранение во время посещения необслуживаемого узла ремонтно–восстановительной бригадой.

Переходный процесс рождения и гибели возникает при первоначальном запуске системы, при изменении входящего потока вызовов, вызванного перераспределением нагрузки на сети, при перегрузках сети и т. д. Для нахождения переходных вероятностей в полнодоступном пучке используются соответствующие дифференциальные уравнения процесса рождения и гибели при условии неравенства нулю производных тех же вероятностей.

15.1. Процесс рождения

Диаграмма переходов Марковского процесса рождения, описывающая процесс накопления неисправности на необслуживаемом узле, изображена на рис. 47.

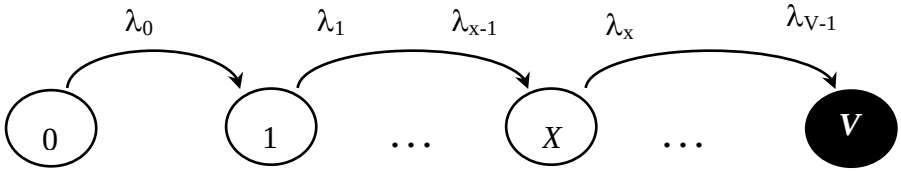


Рис. 47. Диаграмма переходов процесса рождения

Для нее справедлива следующая система дифференциальных уравнений

$$\begin{aligned} &\frac{dp_0(t)}{dt} = -\lambda_0 p_0(t), && x = 0, \\ &\frac{dp_x(t)}{dt} = \lambda_{x-1} p_{x-1}(t) - \lambda_x(t), && x = 1,\dots V-1, \\ &\frac{dp_V(t)}{dt} = \lambda_{V-1} p_{V-1}(t), && x = V. \end{aligned} \qquad (15.1)$$

где $p_x(t)$ – вероятность того, что в момент времени t занято x линий: $x = 0,\dots V$;

$\frac{dp_x(t)}{dt}$ – производные от вероятности $p_x(t)$;

λ_x – параметр потока отказов в состоянии $\{x\}$ с x занятыми линиями: $x = 0,\dots V$;

$\lambda_V = 0$;

V – число линий (обслуживающих приборов) в системе.

Система (15.1) тривиально получается из системы дифференциальных уравнений рождения и гибели с учетом того, что $\mu_x = 0$, $x = 0,\dots V$.

Система (15.1) имеет единственное решение при заданном исходном состоянии пучка. Пусть исходное состояние пучка – {0}, тогда функция распределения $F_x(t)$ находится как вероятность того, что за время t будет занято не менее x линий пучка

$$F_x(t) = \begin{cases} (1-e^{-\lambda_0 t}) * \dots * (1-e^{-\lambda_{x-1} t}), & \lambda_0 \neq \dots \neq \lambda_{x-1}, \quad x = 0,1,2,\dots V, \\ (1-e^{-\lambda t}) * \dots * (1-e^{-\lambda t}), & \lambda_0 = \dots = \lambda_{x-1} = \lambda, \quad x = 0,1,2,\dots V, \end{cases} \qquad (15.2)$$

где * – символ свертки.

Находим изображение (15.2):

$$F_x(t) = \begin{cases} \dfrac{1}{z} \displaystyle\prod_{i=0}^{x-1} \frac{\lambda_i}{z+\lambda_i} & \lambda_0 \neq \dots \neq \lambda_{x-1}, \quad x = 0,1,2,\dots V, \\ \dfrac{1}{z}\left(\dfrac{\lambda}{z+\lambda}\right)^V, & \lambda_0 = \dots = \lambda_{x-1} = \lambda, \quad x = 0,1,2,\dots V. \end{cases}$$

А затем – его оригинал:

$$F_x(t)=\begin{cases}1-\sum\limits_{i=0}^{x-1}e^{-\lambda_i t}\prod\limits_{\substack{i=0\\ j\neq i}}^{x-1}\dfrac{\lambda_j}{\lambda_j-\lambda_i} & \lambda_0\neq\ldots\neq\lambda_{x-1},\quad x=0,1,2,\ldots V,\\ 1-\sum\limits_{i=0}^{x-1}\dfrac{(\lambda t)^i}{i!}e^{-\lambda t}, & \lambda_0=\ldots=\lambda_{x-1}=\lambda,\quad x=0,1,2,\ldots V.\end{cases}$$

Вероятности достижения пучка состояния $\{x\}$ за время t равны $p_x(t) = F_x(t) - F_{x+1}(t)$, откуда для примитивного потока вызовов ($\lambda_x = (N - x)\alpha$, α – параметр потока неисправностей, воздействующий на один элемент):

$$p_x(t)=\left|\begin{array}{ll}C_V^x(e^{\alpha t}-1)^x e^{-V\alpha t}, & x=0,\ldots V-1,\\ F_V(t)=(1-e^{-\alpha t})^V, & x=V,\end{array}\right. \qquad (15.3)$$

со средним временем достижения состояния $\{V\}$

$$T_V=\int_0^\infty t dF_V(t)=V\sum_{k=0}^{V-1}(-1)^k C_{V-1}^k\frac{1}{(k+1)^2\alpha},$$

а для простейшего потока вызовов (λ – суммарный поток неисправностей)

$$p_x(t)=\left|\begin{array}{ll}\dfrac{(\lambda t)^x}{x!}e^{-\lambda t}, & x=0,\ldots V-1,\\ F_V(t)=1-\sum\limits_{i=0}^{V-1}\dfrac{(\lambda t)^i}{i!}e^{-\lambda t}, & x=V,\end{array}\right. \qquad (15.4)$$

со средним временем достижения состояния $\{V\}$

$$T_V=\int_0^\infty t dF_V(t)=\int_0^\infty\frac{(\lambda t)^V}{(V-1)!}e^{-\lambda t}dt=\frac{V}{\lambda}.$$

Зависимости переходных вероятностей $p_x(t)$ процесса рождения за наблюдаемое время t приведены в виде графиков на рис. 48.

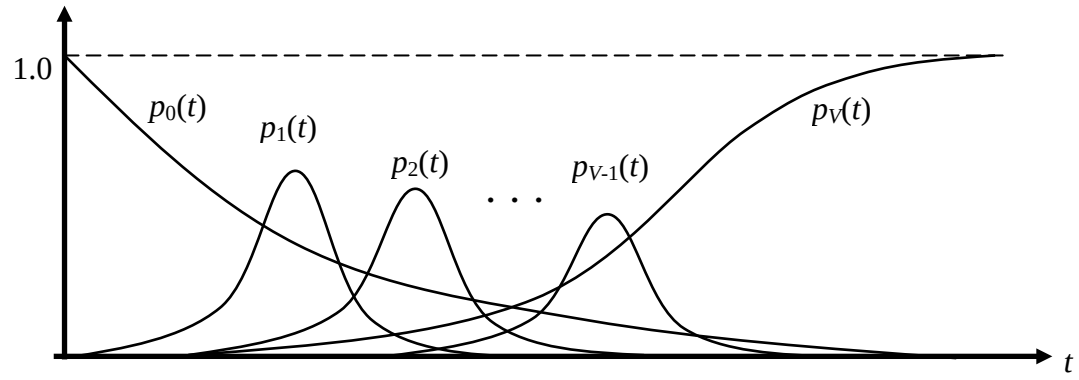


Рис. 48. Вероятности процесса рождения

Выражения (15.3) и (15.4) могут быть использованы и для случая, когда процесс рождения начинается с произвольного состояния k, $k = 0,...V$. В этом случае следует заменить $p_k(t)$ на $p_0(t)$, $p_{k+1}(t)$ на $p_1(t)$ и т. д.

Стационарное состояние пучка процесса рождения – $\{V\}$.

15.2. Процесс рождения и гибели

Диаграмма переходов Марковского процесса рождения и гибели изображена на рис. 17 (стр. 38) и описывается следующей системой дифференциальных уравнений

$$\begin{aligned} \frac{dp_0(t)}{dt} &= -\lambda_0 p_0(t) + \mu_1 p_1(t), && x = 0, \\ \frac{dp_x(t)}{dt} &= \lambda_{x-1} p_{x-1}(t) - \left(\lambda_x + \mu_x\right) p_x(t) + \mu_{x+1} p_{x+1}, && x = 1,...V-1, \\ \frac{dp_V(t)}{dt} &= \lambda_{V-1} p_{V-1}(t) - \mu_V p_V(t), && x = V, \end{aligned} \qquad (15.5)$$

где μ_x – параметр потока освобождения в состоянии $\{x\}$ с x занятыми линиями: $x = 0,...V$, $\mu_0 = 0$.

Система (15.5) имеет единственное решение при заданной матрице исходного состояния пучка $\mathbf{I_v} = |i_x|$, $x = 0,...V$.

Общее решение (15.5) представляется в виде суммы частных решений:

$$p_x(t) = C(0,x)e^{-\gamma_0 t} + C(1,x)e^{-\gamma_1 t} + ... + C(V,x)e^{-\gamma_V t} = \sum_{k=0}^{V} C(k,x)e^{\gamma_k t}, \quad x = 0,...V,$$

где γ_k – k-й корень характеристического уравнения, $C(k, x)$ – постоянный коэффициент, соответствующий k и x.

Воспользуемся тем, что в (15.5) $C(k, x)$ не зависят от t и найдем значения $p_x(t)$ и $dp_x(t)/dt$ в момент времени $t = 0$:

$$p_x(t) = \sum_{k=0}^{V} C(k,x), \qquad \frac{dp_x(t)}{dt}_{|t=0} = \sum_{k=0}^{V} \gamma_k C(k,x), \qquad x = 0,...V\ .$$

Любое частное решение должно удовлетворять (15.5). Подставим в (15.5) одно из них. Получим следующую систему линейных уравнений:

$$\begin{gathered} -\left(\lambda_0 + \gamma_k\right) C(k,0) + \mu_1 C(k,1) = 0, \qquad k = 0,1,...V, \quad x = 0, \\ \lambda_{x-1} C(k,x-1) - \left(\lambda_x + \mu_x + \gamma_k\right) C(k,x) + \mu_{x+1} C(k,x+1) = 0, \\ k = 0, 1, ...V, \quad x = 1, ...V-1, \end{gathered} \qquad (15.6)$$

$$\lambda_{V-1}C(k,V-1)+(\mu_V+\gamma_k)C(k,V)=0, \qquad k=0,1,...V, \quad x=0.$$

Суммируя в (15.6) порознь правые и левые части системы, приходим к основополагающему равенству для нахождения корней:

$$\sum_{x=0}^{V}\gamma_k C(k,x)=0, \qquad k=0,...V. \tag{15.7}$$

Первый корень $\gamma_0 = 0$ определяет решение (15.5) в режиме стационарного равновесия: $C(0,0) = [0]$, $C(0,x) = [x]$. Последующие корни при $\gamma_k \neq 0$ находятся из равенства

$$\sum_{x=0}^{V}C(k,x)=0, \qquad k=1,...V, \tag{15.8}$$

причем вследствие (15.6) $C(k,x)$ образуют рекуррентную последовательность:

$$C(k,x)=\frac{(\lambda_{x-1}+\mu_{x-1}+\gamma_k)C(k,x-1)-\lambda_{x-2}C(k,x-2)}{\mu_x}, \tag{15.9}$$
$$k=1,...V, \quad x=1,...V,$$

где $\mu_0 = \lambda_{-1} = C(k, -1) = 0$. Подбором γ_k из (15.9) при начальном приближении $C^*(k, 0) = 1$ итерационно вычисляются $C^*(k, x)$, $k = 1,...V$, $x = 0,...V$, отличающиеся от $C(k, x)$ постоянным множителем A_k, добиваясь выполнения равенства (15.8).

Истинные значения $C(k, x) = A_k \cdot C^*(k, x)$, $k = 1,...V$ получаем из решения системы линейных уравнений вида $\mathbf{I}_V = A_k\, C^*(k, x)$ относительно A_k, $k = 0,\ldots V$, где $\mathbf{I}_V = |i_x|$ – матрица исходного состояния пучка,

$$C^*(k,x)=\begin{vmatrix} C^*(0,0) & C^*(1,0) & ...C^*(V,0) \\ C^*(0,1) & C^*(1,1) & ...C^*(V,1) \\ & . \quad . \quad . & \\ C^*(0,V) & C^*(1,V) & ...C^*(V,V) \end{vmatrix}$$ – матрица коэффициентов,

$\boldsymbol{A}_k = |A_k|$ – искомая матрица постоянных коэффициентов.

Например, если переходный процесс начинается с состояния {0}, то в матрице $\mathbf{I}_V$ принимаем $i_0 = 1$, $i_1 = i_2 = \ldots = i_V = 0$ и, пользуясь формулой Крамера, находим:

$$A_k = \frac{\begin{vmatrix} C^*(0,0) & C^*(1,0) & \dots & 1 & \dots & C^*(V,0) \\ C^*(0,1) & C^*(1,1) & \dots & 0 & \dots & C^*(V,1) \\ & & & \dots & & \\ C^*(0,V) & C^*(1,V) & \dots & 0 & \dots & C^*(V,V) \end{vmatrix}}{\begin{vmatrix} C^*(0,0) & C^*(1,0) & \dots & C^*(k,0) & \dots & C^*(V,0) \\ C^*(0,1) & C^*(1,1) & \dots & C^*(k,1) & \dots & C^*(V,1) \\ & & & \dots & & \\ C^*(0,V) & C^*(1,V) & \dots & C^*(k,V) & \dots & C^*(V,V) \end{vmatrix}}, k = 0,...V. \quad (15.10)$$

Стационарная вероятность состояния *{x}*, $x = 0,\dots V$ пучка в режиме рождения и гибели – $C(0, x) = [x]$. Графики зависимости $p_x(t)$ приведены на рис. 49.

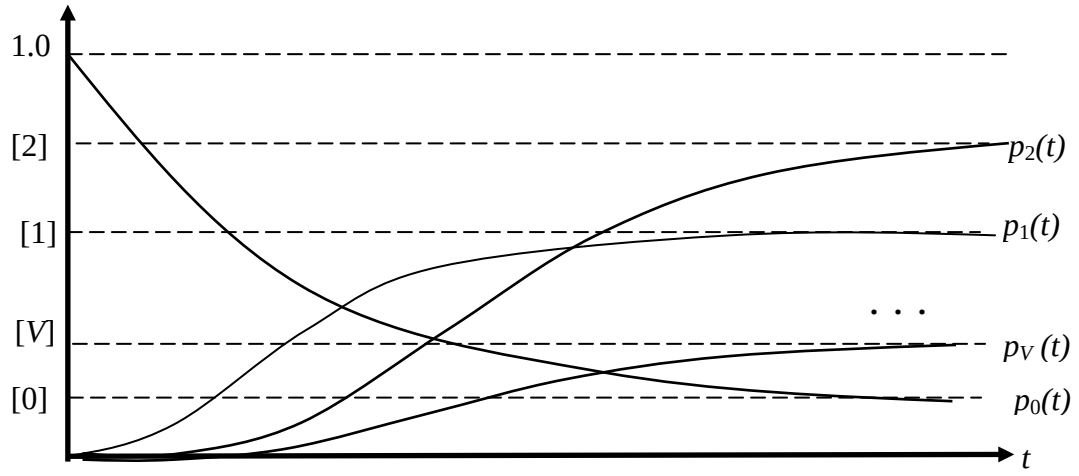


Рис. 49. Переходные вероятности процесса рождения и гибели

15.3. Процесс гибели

С момента прибытия ремонтно-восстановительной бригады на необслуживаемый узел (например, маршрутизатор) начинается устранение неисправностей с одновременным появлением новых неисправностей – процесс рождения и гибели. Этот процесс, как показано выше, описывается достаточно сложными дифференциальными уравнениями. Однако если за короткое время восстановления пренебречь достаточно малой вероятностью возникновения новых неисправностей, ситуация сводится к значительно более простому процессу – процессу гибели.

Диаграмма переходов Марковского процесса гибели, описывающая процесс устранения неисправностей ремонтно-восстановительной бригадой на необслуживаемом узле, изображена на рис. 50.

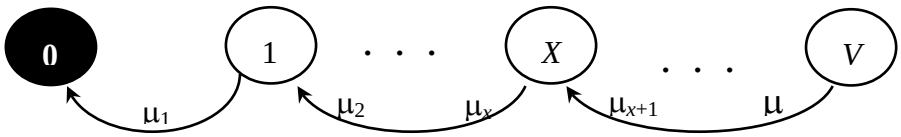


Рис. 50. Диаграмма переходов процесса гибели

Для нее справедлива следующая система дифференциальных уравнений:

$$\begin{aligned} &\frac{dp_0(t)}{dt} = \mu_1 p_1(t), && x = 0, \\ &\frac{dp_x(t)}{dt} = -\mu_x p_x(t) + \mu_{x+1} p_{x+1}, && x = 1,...V-1, \\ &\frac{dp_V(t)}{dt} = -\mu_V p_V(t), && x = V. \end{aligned} \qquad (15.11)$$

которая тривиально получается из системы дифференциальных уравнений рождения и гибели с учетом того, что $\lambda_x = 0$, $x = 0,...V$ (V – число неисправных элементов).

Система (15.11) имеет единственное решение при заданном исходном состоянии пучка и решается аналогично системе процесса рождения. Пусть исходное состояние пучка – $\{V\}$, тогда вероятность $p_{V-x}(t)$ достижения пучка состояния $\{V - x\}$ за время t для примитивного потока (в случае одновременного устранения неисправностей V ремонтно-восстановительными бригадами, μ – интенсивность восстановления неисправности одной ремонтно-восстановительной бригадой)

$$p_{V-x}(t) = \left| \begin{aligned} &C_V^x \left(e^{\mu t} - 1\right)^x e^{-V\mu t}, && x = 0,...V-1, \\ &\left(1 - e^{-\mu t}\right)^V, && x = V, \end{aligned} \right. \qquad (15.12)$$

а для простейшего потока вызовов (в случае устранения неисправностей одной ремонтно-восстановительной бригадой, μ – интенсивность восстановления неисправности одной ремонтно-восстановительной бригадой)

$$p_{V-x}(t) = \left| \begin{aligned} &\frac{(\mu t)^x}{x!} e^{-\mu t}, && x = 0,...V-1, \\ &1 - \sum_{i=0}^{V-1} \frac{(\mu t)^i}{i!} e^{-\mu\, t}, && x = V. \end{aligned} \right. \qquad (15.13)$$

Выражения (15.12) и (15.13) могут быть использованы и для случая, когда процесс рождения начинается с произвольного состояния k, $k = 0,...V$. В этом случае следует заменить $p_k(t)$ на $p_V(t)$, $p_{k-1}(t)$ на $p_{V-1}(t)$ и т. д.

Стационарное состояние пучка процесса гибели – {0}. Графики зависимости $p_{V-x}(t)$ приведены на рис. 51.

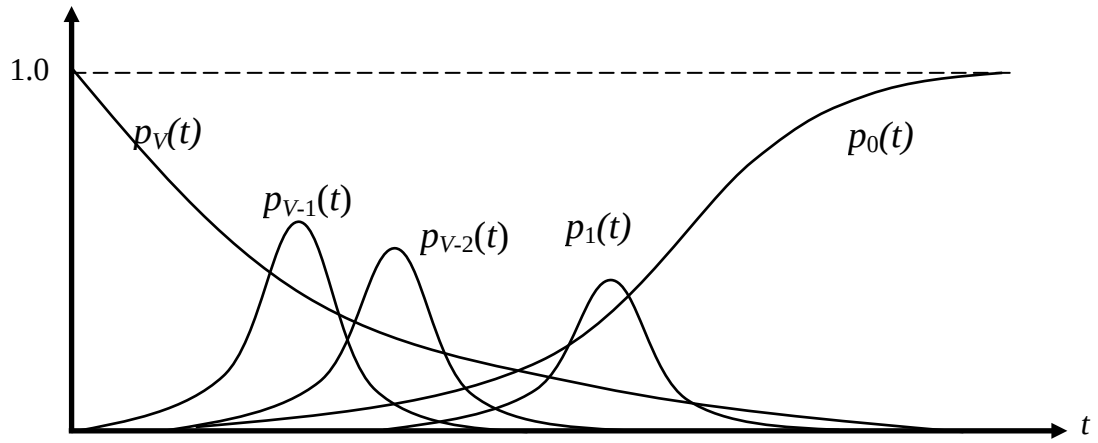


Рис. 51. Вероятности процесса гибели

ЗАДАЧА 1. В сети имеется дублированный необслуживаемый программный коммутатор в режиме с функционально-незадействованным резервированием. Среднее время безотказной работы одного коммутатора – 3 месяца (0,25 года). Считая поток неисправности примитивным, определить вероятности а) безотказной работы; б) перехода на резерв; в) выхода из строя в необслуживаемом режиме в течение одного месяца, трех месяцев; 0,5 года; одного года.

Решение.

Параметр потока неисправностей (отказов) $\xi = (0{,}25 \text{ года})^{-1} = 4 \text{ год}^{-1}$.

Согласно (15.3) вероятность безотказной работы $p_0(t) = e^{-2\xi t}$, вероятность перехода на резерв $p_1(t) = 2\left(e^{-\xi t} - e^{-2\xi t}\right)$, вероятность выхода из строя $p_2(t) = \left(1 - e^{-\xi t}\right)^2$, среднее время выхода из строя $T_2 = 2\sum_{k=0}^{1}(-1)^k C_{V-1}^k \frac{1}{(k+1)^2 \xi} = 1{,}5\xi^{-1}$. Подставляя значения t = 0,08 года; 0,25 года; 0,5 года; 1 год, получим вероятности (табл.15.1).

Таблица 15.1

Вероятности событий дублированной системы

Вероятности:	Время работы			
	0,08 года	0,25 года	0,5 года	1 год
Безотказной работы	0,5273	0,1353	0,0183	0,0003
Перехода на резерв	0,3977	0,4650	0,2340	0,0359
Выхода из строя	0,0749	0,3996	0,7476	0,9637
Среднее время выхода из строя в необслуживаемом режиме – T_2 = 3/8 года = 4,5 месяца. В обслуживаемом режиме – 362 года (см. задачу 4 разд. 14).				

ЗАДАЧА 2. На полнодоступный трехлинейный пучок, находящийся в момент времени $t = 0$ в состоянии {0}, начинает поступать примитивный поток вызовов от $N = V$ источников нагрузки с параметром свободного источника $\alpha = 2$ ч$^{-1}$. Время обслуживания вызова распределено по экспоненциальному закону с параметром $\mu = 20$ ч$^{-1}$. Определить время T, при котором с вероятностью не менее 0,99 процесс обслуживания вызовов можно считать стационарным.

Решение.

Подставляя значения $\alpha = 2$ ч$^{-1}$, $\mu = 20$ ч$^{-1}$, $V = 3$ в (15.9) и пользуясь (15.8), находим корни $\gamma_1 = -22{,}0$ ч$^{-1}$, $\gamma_2 = -44{,}1$ ч$^{-1}$, $\gamma_3 = -66{,}1$ ч$^{-1}$ и коэффициенты:

$$p_0(t) = 1{,}00 + 1{,}00e^{-22.0t} + 1{,}00e^{-44.1t} + 1{,}00e^{-66.1t};$$
$$p_1(t) = 0{,}300 - 0{,}800e^{-22.0t} - 1{,}904e^{-44.1t} - 3{,}005e^{-66.1t};$$
$$p_2(t) = 0{,}030 - 0{,}189e^{-22.0t} + 0{,}807e^{-44.1t} + 3{,}013e^{-66.1t};$$
$$p_3(t) = 0{,}001 - 0{,}011e^{-22.0t} + 0{,}097e^{-44.1t} - 1{,}008e^{-66.1t}.$$

Для нахождения постоянных коэффициентов привлекаем (15.10):

$$p_0(t) = 0{,}7514 + 0{,}2252e^{-22.0t} + 0{,}0229e^{-44.1t} + 0{,}0005e^{-66.1t};$$
$$p_1(t) = 0{,}2254 - 0{,}1802e^{-22.0t} - 0{,}0437e^{-44.1t} - 0{,}0015e^{-66.1t};$$
$$p_2(t) = 0{,}0225 - 0{,}0425e^{-22.0t} + 0{,}0185e^{-44.1t} + 0{,}0015e^{-66.1t};$$
$$p_3(t) = 0{,}0007 - 0{,}0025e^{-22.0t} + 0{,}0023e^{-44.1t} - 0{,}0005e^{-66.1t}.$$

По условию задачи процесс можно считать стационарным, если $p_0(t) - [0] < 0{,}01$, что выполняется при $t > 0{,}15$ ч = 9,0 мин.

ЗАДАЧА 3. В сети неисправны два маршрутизатора. Время устранения неисправности в одном маршрутизаторе с учетом прибытия ремонтно-восстановительной бригады $T = 1{,}5$ ч. Определить время восстановления двух маршрутизаторов одной бригадой с вероятностью не менее 0,99.

Решение.

Интенсивность восстановления $r = T^{-1}$ 0,66 ч$^{-1}$.

Используя (15.13), убеждаемся, что $p_0(t) = 1 - e^{-rt} - r \cdot t \cdot e^{-rt} > 0{,}99$ при $t > 10{,}2$ ч.

16. КЛАСС ОБСЛУЖИВАНИЯ СЕТЕВЫХ ЭЛЕМЕНТОВ

16.1. Понятие о классе обслуживания

Рассмотренные выше параметры качества обслуживания (Quality of Service, QoS) используются ***для расчета*** качества обслуживания в сети и предназначены в основном для разработчиков и проектировщиков сетей. С другой стороны, обслуживающему персоналу необходим набор параметров ***для контроля*** текущего качества обслуживания отдельного сетевого элемента в ЧНН на основании измерений параметров нагрузки. Эти параметры относятся к классу обслуживания (Grade of Service, GoS) сетевого элемента и используются обслуживающим персоналом для расчета его ключевых показателей эффективности (Key Performance Indicator, KPI). Модель для расчета данных KPI в узле коммутации приведена на рис. 52.

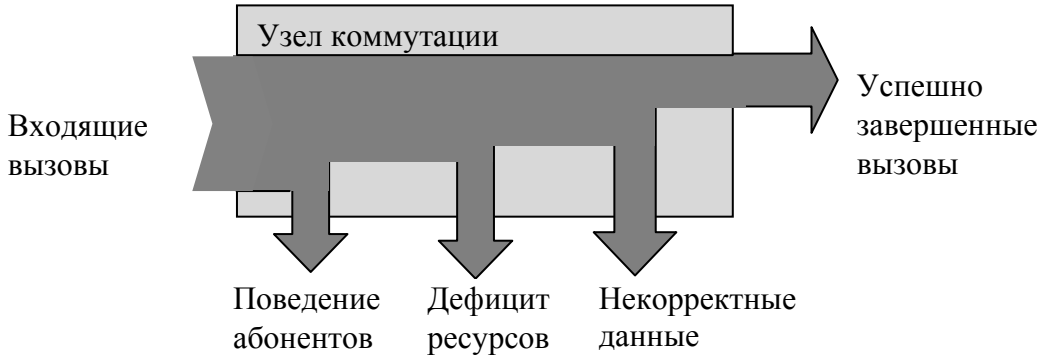


Рис. 52. Модель для расчета KPI

Вызовы, поступающие на обслуживание в узел коммутации (входящие вызовы), делятся на успешно завершенные и не принятые к обслуживанию. Причины появления необслуженных вызовов:

– поведение абонентов (отказа от вызова при неполном наборе номера, неправильно набранный номер, . . .);

– дефицит ресурсов узла коммутации (например, нехватке исходящих каналов связи);

– некорректные данные (ошибки в функционировании системы).

Статистические данные (показания счетчиков) для расчета KPI снимаются на фоне обслуживания реального трафика (вызовов) по директивам оператора сети. Запись на диск статистических данных производится каждые 15 мин (900 с). Обработка записанных на диск статистических данных производится отдельным оборудованием – сетевым анализатором.

Набор параметров и формулы для расчета KPI зависят от типа сетевого элемента, вида связи, типа подсистемы и специфичны для оператора сети.

Слежение за статистическими данными нагрузки рекомендуется производить еженедельно. Дело в том, что на протяжении жизненного цикла узла коммутации параметры нагрузки значительно меняются. Причинами их изменения являются отказ части абонентов от услуг связи в связи со сменой оператора сети; подключение новых абонентов; введение в сети новой услуги; изменение тарифов; изменение маршрутов соединения и т. д.

16.2. Виды связи

Рекомендация Q.544 [9] определяет следующие виды трафика в узле коммутации каналов (рис. 53).

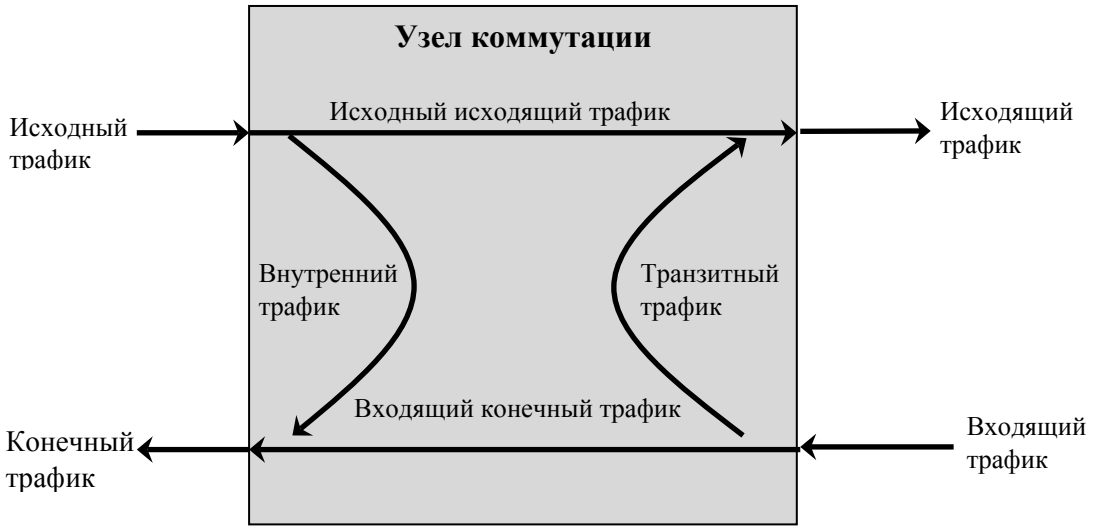


Рис. 53. Виды связи в узле коммутации

Исходный трафик (Originating traffic) инициируется оконечными устройствами (ОУ) и состоит из исходного исходящего и внутреннего трафика.

Исходный исходящий трафик (Originating outgoing traffic) инициируется оконечными устройствами (ОУ) и требует соединения с абонентом другого узла.

Исходящий трафик (Outgoing traffic) образуется собственными вызовами, уходящими в другие узлы, и транзитным трафиком. Транзитный трафик часто представляет собой нагрузку от сетей связи других операторов, не имеющих разветвленной транспортной сети и вынужденных арендовать транспортные каналы других операторов.

Входящий трафик (Terminating traffic) инициируется всеми вызовами, поступающими от других узлов, и состоит из входящего конечного и транзитного трафика.

Входящий конечный трафик (Incoming terminating traffic) инициируется всеми вызовами, поступающими от других узлов, и требует соединения с оконечными устройствами узла.

Конечный трафик (Terminating traffic) состоит из входящего конечного и внутреннего трафика.

Классификация трафика имеет значение по двум причинам.

Во-первых, она необходима для расчета требуемого оборудования, так как каждый вид трафика использует определенный тип оборудования.

Во-вторых, каждый вид трафика (сетевого соединения) имеет свою стоимость, что важно учитывать при финансовых расчетах между различными операторами сети.

16.3. Параметры и формулы для комплексного контроля KPI

Набор параметров и формулы для расчета KPI специфичны для оператора сети. Временная диаграмма установления исходного речевого соединения по общему каналу сигнализации (CCS) в центре коммутации мобильных абонентов (MSC) изображена на рис. 54.

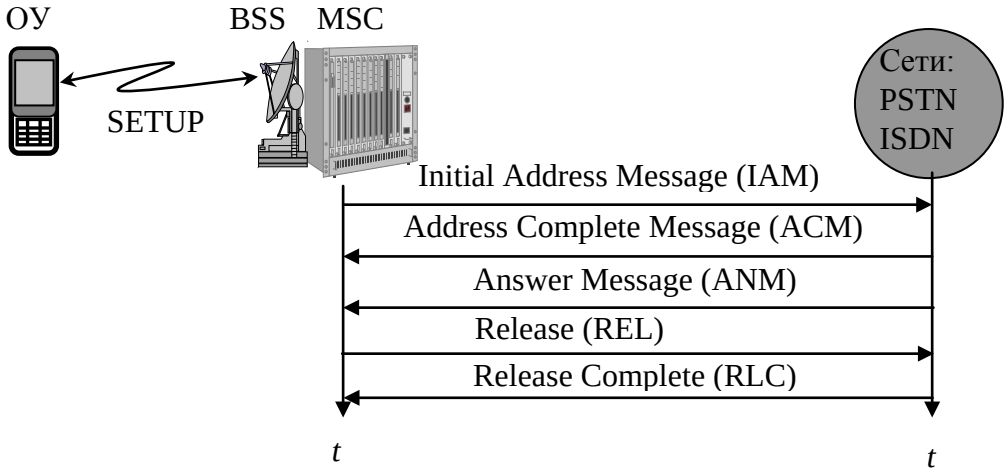


Рис. 54. Временная диаграмма прохождения исходящего вызова

Для расчета KPI исходного трафика (Mobile Originating Call, MOC) речевых соединений в центре коммутации мобильных абонентов (MSC) GSM D900 компания SIEMENS использует следующие статистические данные счетчиков.

- *MOC_SCR* – доля успешных исходных соединений в центре коммутации мобильных абонентов GSM D900.
- *SP_CALLS_W_ANS_MOC* – число исходных вызовов, на которые получен ответ. Срабатывание счетчика при получении сигнального сообщения ANM.

- *BS_REL_W_ANS_MOC* – число исходных отвеченных вызовов, разрушенных по причине неисправности радиоканала. Срабатывание счетчика при получении сигнала неисправности от радиоподсистемы.
- *UNS_MS_BUSY_MOC* – число разрушенных исходных вызовов из-за занятости вызываемого абонента. Срабатывание счетчика при приеме сигнального сообщения REL после приема ACM.
- *CALL_REL_bf_ALERT_MOC* – число разрушенных исходных вызовов до полного набора цифр вызываемого абонента. Срабатывание счетчика при передаче сигнального сообщения REL до приема ACM.
- *CALL_REL_ALERT_MOC* – число разрушенных исходных вызовов после полного набора номера до ответа вызываемого абонента. Срабатывание счетчика при передаче сигнального сообщения REL после приема ACM.
- *CALL_TIOUT_ALERT_MOC* – число разрушенных исходных вызовов из-за истечения таймера ожидания ответа вызываемого абонента. Срабатывание счетчика при приеме сигнального сообщения RLC после приема ACM и срабатывания таймера.
- *CALL_ATTS_MOC* – число поступивших исходных вызовов. Срабатывание счетчика при приеме сигнального сообщения SET UP.

Эти статистические данные счетчиков используются для расчета доли успешно обслуженных исходных вызовов системой (*MOC_SCR*):

$$MOC_SCR = \frac{\begin{array}{c} SP_CALLS_W_ANS_MOC + \\ + BS_REL_W_ANS_MOC + \\ + UNS_MS_BUSY_MOC + \\ + CALL_REL_bf_ALERT_MOC + \\ + CALL_REL_ALERT_MOC + \\ + CALL_TIOUT_ALERT_MOC \end{array}}{CALL_ATTS_MOC} 100\ \%$$

В числитель приведенной формулы помещаются все случаи успешного завершения соединения с точки зрения системы коммутации. Некоторые вызовы, завершившиеся неуспешно с точки зрения абонента, с точки зрения системы коммутации считаются успешными. Например, все вызовы на занятого абонента (*UNS_MS_BUSY_MOC*) с точки зрения системы коммутации считаются успешными, так как система отработала корректно и претензий к ней быть не должно. То же самое относится к счетчикам *CALL_REL_bf_ALERT_MOC* (неполный набор номера), *CALL_REL_ALERT_MOC* (преждевременный отказ от соединения), *CALL_TIOUT_ALERT_MOC* (отсутствие абонента).

Нормированное значение *MOC_SCR* составляет не менее 95 %, в случае его снижения принимаются меры для его повышения.

Ниже в табл. 16.1 приведена реальная статистика для расчета KPI исходного трафика (статистика за 900 с).

Таблица 16.1

Ключевые показатели эффективности MOC

SP_CALLS_W_ANS_MOC	9138
BS_REL_W_ANS_MOC	124
UNS_MS_BUSY_MOC	890
CALL_REL_bf_ALERT_MOC	649
CALL_REL_ALERT_MOC	4281
CALL_TIOUT_ALERT_MOC	133
CALL_ATTEMPTS_MOC	15626
MOC_SCR	97,4 %

Проведем анализ полученной статистики.

Доля успешно обслуженных исходных вызовов составляет 97,4 %, что выше нормированного значения (не менее 95 %). Доля исходных вызовов, завершившихся разговором (*SP_CALLS_W_ANS_MOC*), составляет 58,5 %.

Временная диаграмма установления конечного речевого соединения по общему каналу сигнализации (CCS) в центре коммутации мобильных абонентов (MSC) изображена на рис. 55.

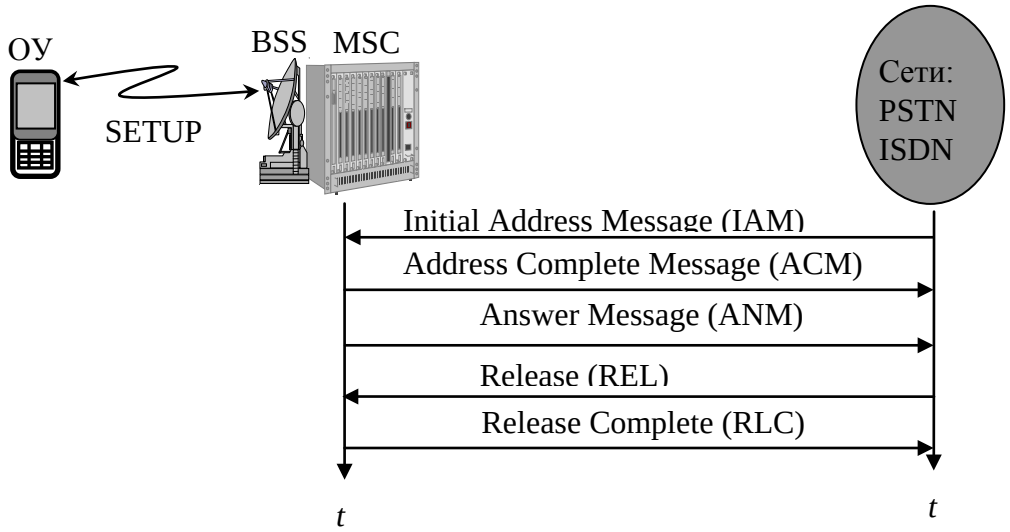


Рис. 55. Временная диаграмма прохождения входящего вызова

Для расчета KPI конечного трафика (Mobile Terminating Call, MTC) речевых соединений в центре коммутации мобильных абонентов (MSC) GSM D900 используются следующие статистические данные.

- *MTC_SCR* – доля успешных конечных соединений в центре коммутации мобильных абонентов GSM D900.
- *SP_CALLS_W_ANS_MTC* – число конечных вызовов, на которые получен ответ. Срабатывание счетчика при передаче сигнального сообщения ANM.
- *BS_REL_W_ANS_MTC* – число конечных отвеченных вызовов, разрушенных по причине неисправности радиоканала. Срабатывание счетчика: при получении сигнала неисправности от радио подсистемы.
- *UNS_MS_BUSY_MTC* – число разрушенных конечных вызовов из-за занятости вызываемого абонента. Срабатывание счетчика: при передаче сигнального сообщения REL после приема IAM.
- *CALL_REL_bf_ALERT_MTC* – число разрушенных вызовов А-абонентом до полного набора цифр вызываемого абонента. Срабатывание счетчика: при приеме сигнального сообщения REL до передачи ACM.
- *CALL_REL_ALERT_MTC* – число разрушенных вызовов А-абонентом до ответа вызываемого абонента. Срабатывание счетчика при приеме сигнального сообщения REL после передачи ACM.
- *UNS_MS_BLOCKED_MTC* – число разрушенных вызовов В-абонентом из-за его блокировки администратором сети или в результате заявленной услуги «Черный список». Срабатывание счетчика: при передаче сигнального сообщения REL после приема IAM в результате анализа базы данных абонента.
- *CALL_ATTS_MTC* – число поступивших конечных вызовов. Срабатывание счетчика: при приеме сигнального сообщения IAM.

Эти статистические данные используются для расчета доли успешно обслуженных конечных вызовов системой (*MTC_SCR*):

$$MTC_SCR = \frac{\begin{gathered} SP_CALLS_W_ANS_MTC + \\ + BS_REL_W_ANS_MTC + \\ + UNS_MS_BUSY_MTC + \\ + CALL_REL_bf_ALERT_MTC + \\ + CALL_REL_ALERT_MTC + \\ + UNS_MS_BLOCKED_MTC \end{gathered}}{CALL_ATTS_MTC} 100\ \%$$

Ниже в табл. 16.2 приведена реальная статистика для расчета KPI конечного трафика (за 900 с).

Таблица 16.2

Ключевые показатели эффективности МТС

SP_CALLS_W_ANS_MTC	23594
BS_REL_W_ANS_MTC	313
UNS_MS_BUSY_MTC	3057
CALL_REL_bf_ALERT_MTC	895
CALL_REL_ALERT_MTC	7247
UNS_MS_BLOCKED_MTC	0
CALL_ATTEMPTS_MTC	35532
MOC_SCR	98,8 %

Проведем анализ полученной статистики.

Доля успешно обслуженных вызовов составляет 98,8 %, что выше нормированного значения (не менее 95 %). Доля конечных вызовов, завершившихся разговором (*SP_CALLS_W_ANS_MTC*), составляет 66,3 %.

Нормированное значение *MOC_SCR* и *MTC_SCR* – не менее 95 %. В случае их низких значений производится поиск причины путем анализа KPI других подсистем: пучков каналов, центрального процессора, подсистемы сигнализации, радио подсистемы и т. д.

16.4. Параметры и формулы для контроля KPI пучка каналов

Для расчета KPI пучка двухсторонних каналов речевых соединений в центре коммутации мобильных абонентов (MSC) GSM D900 используются следующие данные статистики за 900 с.

- *TRUNKS_NUMBER* – число каналов в пучке.
- *TRUNKS_BLOCKED* – число блокированных (неисправных) каналов.
- *TRUNKS_AVAILABLE* – число доступных (исправных) каналов, которые были задействованы при снятии статистических данных.
- *CALL_ATTEMPTS_INCOMING* – параметр потока входящих вызовов (число поступивших вызовов от сети за время наблюдения). Срабатывание счетчика: при приеме каждого сигнального сообщения IAM.
- *CALL_ATTEMPTS_OUTGOING* – параметр потока исходящих вызовов (число поступивших вызовов от собственных абонентов за время наблюдения). Срабатывание счетчика: при приеме каждого сигнального сообщения SETUP.
- *CALL_ATTEMPTS_TOTAL* = (*CALL_ATTEMPTS_INCOMING* + + *CALL_ATTEMPTS_OUTGOING*) – параметр общего потока вызовов (общее число вызовов обслуженных системой за время наблюдения).

- *TRAFFIC_VOLUME_INCOMING* – суммарное время занятия входящими вызовами (в секундах). Срабатывание счетчика: при приеме каждого сигнального сообщения IAM. Остановка счетчика при приеме сигнального сообщения RLC, соответствующего сообщению IAM.
- *TRAFFIC_VOLUME_OUTGOING* – суммарное время занятия исходящими вызовами (секунд). Срабатывание счетчика: при приеме каждого сигнального сообщения SETUP. Остановка счетчика: при приеме сигнального сообщения RLC, соответствующего сообщению SETUP.
- *TRAFIC_VOLUME_TOTAL=(TRAFFIC_VOLUME_INCOMING + + TRAFFIC_VOLUME_OUTGOING)* – суммарное время занятия всеми вызовами (в секундах).
- *CALL_CARRIED_INCOMING* – число входящих вызовов, принятых к обслуживанию. Срабатывание счетчика: при передаче каждого сигнального сообщения SETUP.
- *CALL_CARRIED_OUTGOING* – число исходящих вызовов, принятых к обслуживанию. Срабатывание счетчика: при передаче каждого сигнального сообщения IAM.
- *CALL_CARRIED_TOTAL = CALL_CARRIED_INCOMING + + CALL_CARRIED_OUTGOING* – общее число вызовов, принятых к обслуживанию.

Эти статистические данные используются для расчета следующих KPI.

Входящая обслуженная нагрузка (в Эрлангах) вычисляется как отношение суммарного времени занятия входящими вызовами (в секундах) к времени наблюдения (15 мин = 900 с):

$$TRAFFIC_CARRIED_INCOMING = \frac{TRAFFIC_VOLUME_INCOMING}{900}.$$

Исходящая обслуженная нагрузка (Эрл) вычисляется как отношение суммарного времени занятия исходящими вызовами (в секундах) к времени наблюдения (15 мин = 900 с):

$$TRAFFIC_CARRIED_OUTGOING = \frac{TRAFFIC_VOLUME_OUTGOING}{900}.$$

Общая обслуженная нагрузка (Эрл) вычисляется как отношение суммарного времени занятия всеми вызовами (в секундах) к времени наблюдения (15 мин = 900 с):

$$TRAFFIC_CARRIED_TOTAL = \frac{TRAFFIC_VOLUME_TOTAL}{900}.$$

Удельная обслуженная нагрузка на один канал (Эрл) вычисляется как отношение общей нагрузки (Эрл) к числу доступных (исправных) каналов, которые были задействованы при снятии статистики:

$$TRAFFIC_PER_TRUNK = \frac{TRAFFIC_CARRIED_TOTAL}{TRUNKS_AVAILABLE}.$$

Среднее время занятия вычисляется как отношение суммарного времени занятия всеми вызовами (с) к общему числу вызовов, принятых к обслуживанию:

$$HOLDING_TIME = \frac{TRAFFIC_VOLUME_TOTAL}{CALL_CARRIED_TOTAL}.$$

Вероятность потерь вычисляется как дополнение до 1 отношения числа исходящих вызовов, принятых к обслуживанию, к числу поступивших вызовов от собственных абонентов:

$$LOSS = \left(1 - \frac{CALL_CARRIED_OUTGOING}{CALL_ATTEMPT_OUTGOING}\right)100\,\%.$$

Нормируемая вероятность потерь (*LOSS*) для исходящего соединения – не более 0,01 % = 0,001.

В том случае, если вероятность потерь превысит нормируемое значение, следует увеличить емкость пучка (*TRUNKS_NUMBER*) так, чтобы выполнялось данное неравенство:

$$E_{TRUNKS_NUMBER}\left(TRAFFIC_CARRIED_TOTAL\right) < 0,001,$$

где $E_{TRUNKS\ NUMBER}$ (*TRAFFIC_CARRIED_TOTAL*) – формула Эрланга.

Ниже в табл. 16.3 приведена реальная статистика (за 900 с) для расчета KPI двухстороннего пучка каналов.

Таблица 16.3

Ключевые показатели эффективности каналов пользователей

TRUNKS_NUMBER	1040
TRUNKS_BLOCKED	0
TRUNKS_AVAILABLE	**1040**
CALL_ATTEMPTS_INCOMING	10970
CALL_ATTEMPTS_OUTGOING	4347
CALL_ATTEMPTS_TOTAL	15317
TRAFFIC_VOLUME_INCOMING (*с*)	554388
TRAFFIC_VOLUME_OUTGOING (*с*)	368196
TRAFIC_VOLUME_TOTAL (*с*)	922584

Окончание табл. 16.3

CALL_CARRIED_INCOMING	10970
CALL_CARRIED_OUTGOING	4344
CALL_CARRIED_TOTAL	15314
TRAFFIC_CARRIED_INCOMING (*Эрл*)	615,99
TRAFFIC_CARRIED_OUTGOING (*Эрл*)	409,1
TRAFFIC_CARRIED_TOTAL (*Эрл*)	**1025,09**
TRAFFIC_PER_TRUNK (*Эрл*)	0,98
HOLDING_TIME (*с*)	60,23
LOSS	**0,0007**

Проведем анализ полученной статистики.

Общие потери вызовов составляют $E_{1040}(1025{,}09) = 0{,}016$. Потери же исходящих вызовов *LOSS* = 0,0007. Последнее означает, что подавляющая часть вызовов на двухстороннем пучке теряется на входящей стороне и требуется расширение пучка до 1125 каналов.

16.5. Параметры и формулы для контроля KPI управляющего устройства

Расчет KPI управляющего устройства базируется на Рекомендации ITU-T Q.543, которая устанавливает среднюю нагрузку на процессор не более 0,7 Эрл = 700 мЭрл.

Допускается (но не рекомендуется) увеличение средней нагрузки до 800 мЭрл.

При использовании многопроцессорных систем средняя нагрузка увеличивается в соответствующее количество раз. Например, для четырех процессорного (четырех ядерного) управляющего устройства нагрузка на него может быть увеличена до 2,8 Эрл.

16.6. Параметры и формулы для контроля KPI сигнальных каналов

Сигнальные каналы в современных системах коммутации строятся на технологии коммутации пакетов. Для уменьшения времени прохождения пакетов нагрузка на один сигнальный канал не должна превышать 400 мЭрл. С целью повышения надежности функционирования пучок сигнальных каналов в общем случае содержит несколько сигнальных каналов. При выходе из строя одного из сигнальных каналов общая нагрузка равномерно распределяется между остальными исправными каналами пучка так,

Таблица 16.4

Нагрузка на один сигнальный канал

Число каналов в пучке	Максимальная нагрузка на каждый канал, мЭрл
1	400
2	200
3	266
4	300
5	320
6	333
7	342
8	350

чтобы индивидуальная нагрузка не превышала бы 400 мЭрл. Поэтому максимальная нагрузка на один сигнальный канал зависит от емкости пучка следующим образом (табл. 16.4).

Для расчета KPI пучка сигнальных каналов (SL) со скоростью передачи 64 кбит/с = 64 000 бит/с каждый используются следующие статистические данные счетчиков (статистика за 900 с):

R_MSU – число принятых сигнальных сообщений;

R_Bytes – общее число принятых байтов;

T_MSU – число переданных сигнальных сообщений;

T_Bytes – общее число переданных байтов.

Эти статистические данные используются для расчета следующих KPI.

Интенсивность поступления сигнальных сообщений вычисляется как отношение числа принятых сообщений времени наблюдения (15 мин = 900 с):

$$R_MSU\ /\ T = \frac{R_MSU}{900}.$$

Средняя длина принятых сигнальных сообщений вычисляется как отношение общего числа принятых байтов к числу принятых сигнальных сообщений:

$$R_Bytes\ /\ MSU = \frac{R_Bytes}{R_MSU}.$$

Нагрузка на принимаемый сигнальный канал вычисляется как отношение числа принятых битов к общему числу битов, принятых за время наблюдения (15 мин = 900 с):

$$R_Load = \frac{R_Bytes * 8}{64000 * 900}.$$

Интенсивность передачи сигнальных сообщений вычисляется как отношение числа переданных сообщений к времени наблюдения (15 мин = 900 с):

$$T_MSU\ /\ T = \frac{T_MSU}{900}.$$

Средняя длина переданных сигнальных сообщений вычисляется как отношение общего числа переданных байтов к числу принятых сигнальных сообщений:

$$T_Bytes / MSU = \frac{T_Bytes}{T_MSU}.$$

Нагрузка на передающий сигнальный канал вычисляется как отношение числа переданных битов к общему числу битов, переданных за время наблюдения (15 мин = 900 с):

$$T_Load = \frac{T_Bytes * 8}{64\,000 * 900}.$$

Ниже в табл. 16.5 приведена реальная статистика для расчета KPI пучка из четырех сигнальных каналов (за 900 с).

Таблица 16.5

Ключевые показатели эффективности сигнальных каналов

N SL	*R_MSU* шт	*R_Bytes* байт	*R_MSU* $с^{-1}$	*R_Bytes/ MSU*	*R_Load* мЭрл	*T_MSU* шт	*T_Bytes* байт	*T_MSU* $с^{-1}$	*T_Bytes/ MSU*	*T_Load* мЭрл
0	5929	934062	6,58	157,5	129	5753	582324	6,39	101	80
1	5839	930424	6,48	159,3	129	5733	578877	6,37	101	80
2	5870	926360	6,52	157,8	128	5736	575107	6,37	100	79
3	5594	875472	6,21	156,5	121	5855	588273	6,50	100	81
Med	5808	916579	6,45	157,7	126,75	5769	581145	6,4	100,5	80

Проведем анализ полученной статистики.

Средняя нагрузка на один приемный сигнальный канал составляет *Med R_Load* = 126,75 мЭрл, а на передающий – *Med T_Load* = 80 мЭрл. Оба параметра не превышают нагрузку 300 мЭрл, допустимую в четырехлинейном пучке (см. табл. 16.4). Подсистема сигнализации функционирует нормально.

ЗАДАЧА 1. Какая Рекомендация описывает распределение трафика внутри сетевого элемента?

Ответ. Рекомендация ITU-T Q.544.

ЗАДАЧА 2. Какой принцип расчета ключевых показателей эффективности (Key Performance Indicator, KPI) использует метод GoS?

Ответ. Ключевой показатель эффективности для каждой подсистемы рассчитывается как отношение числа успешно обслуженных вызовов подсистемой к общему числу поступивших вызовов.

ЗАДАЧА 3. Поясните отличие расчета параметров качества обслуживания методом QoS (Quality of Service) от GoS (Grade of Service).

Ответ. В формулах расчета качества обслуживания методом QoS используются произвольные исходные данные, а в формулах расчета методом GoS – исходные данные статистических измерений.

ЗАДАЧА 4. Поясните взаимосвязь методов QoS и GoS.

Ответ. Формулы расчета по GoS используются для оценки существующего качества обслуживания на основе статистических данных. Если измеренное качество обслуживания не соответствует нормам, то с использованием формул моделей QoS рассчитываются требуемые параметры (число каналов, скорость интерфейса, производительность управляющих устройств и т. д.).

ЗАДАЧА 5. Нормированная нагрузка на один сигнальный канал не должна превышать 400 мЭрл. Какова максимальная нагрузка на один сигнальный канал в пучке из четырех сигнальных каналов?

Решение.

При выходе из строя одного сигнального канала общая нагрузка на пучок сигнальных каналов не должна превышать $3 \cdot 400$ мЭрл = 1200 мЭрл. В нормальном состоянии при исправности всех четырех каналов нагрузка на один сигнальный канал не должна превышать 1200 мЭрл/4 = 300 мЭрл.

ЗАДАЧА 6. Как статистические данные используются при расчете исходящей нагрузки на интерфейсе маршрутизатора?

Решение.

Пусть скорость интерфейса равна VEL, время наблюдения – T.

Измеряется общий объем переданных байтов $LENS$ за время наблюдения.

Тогда нагрузка на исходящем интерфейсе $a = LENS/(VEL \cdot T)$.

Например, пусть по интерфейсу со скоростью VEL = 10 Мбит/с за время наблюдения T = 10 с переданы пакеты общим объемом 10^7 байт.

Тогда нагрузка на исходящем интерфейсе $a = LENS/(VEL \cdot T)$ = 0,8 Эрл.

17. ОБНАРУЖЕНИЕ КОМПРОМЕТАЦИИ ПЕРСОНАЛЬНЫХ ДАННЫХ

Обеспечение безопасности функционирования сети предусматривает как защиту персональных данных пользователя, так и обнаружение их компрометации, т. е. их раскрытие и использование злоумышленником.

Защита персональных данных пользователя осуществляется путем аутентификации и шифрования перед отправкой по сети связи. Здесь же рассматривается обнаружение компрометации персональных данных на примере сети связи мобильных абонентов стандарта GSM. Эта задача решается путем анализа определенных сетевых событий, которые рассмотрены ниже.

До установления любого соединения мобильная станция должна пройти процедуру регистрации (аутентификации).

В сети мобильных абонентов GSM различают несколько видов регистрации.

Начальная регистрация (IMSI Attach) используется мобильной станцией при первоначальном включении или прибытии из другой сети. В любом случае SIM-карта содержит код прежнего местоположения, который сравнивается с кодом текущего местоположения, принимаемого из радиоканала. На рис. 56 приведена временная диаграмма регистрации мобильного абонента в сети. Приняты следующие обозначения.

MS – мобильная станция (**M**obile **S**tation). Компонентами мобильной станции (**M**obile **S**tation, **MS**) являются мобильное оборудование (**M**obile **E**quipment, **ME**) и модуль идентификации абонента (**S**ubscriber **I**dentification **M**odule, **SIM**). Посредством сменяемого мобильного оборудования пользователь осуществляет сеансы связи, а идентификация пользователя осуществляется посредством данных, хранящихся в модуле идентификации абонента (на SIM-карте). Эти данные записываются на SIM-карту оператором сети при заключении контракта на обслуживание абонента и содержат в себе сетевой номер (MS ISDN), международный идентификатор мобильного абонента (International Mobile Subscriber Identity – IMSI), временный идентификатор мобильного абонента (Temporary Mobile Subscriber Identity, TMSI), постоянные ключи, алгоритмы идентификации, шифрования информации и др.

IMSI – международный идентификатор мобильного абонента (**I**nternational **M**obile **S**ubscriber **I**dentity). Содержит сетевой адрес **HLR**.

TIMSI – временный идентификатор мобильного абонента (Temporary Mobile Subscriber Identity). Содержит сетевой адрес **VLR**.

BSS – подсистема базовых станций (**B**ase **S**tation **S**ystem).

MSC – центр коммутации мобильных абонентов (**M**obile services **S**witching **C**entre), обеспечивающий физические соединения между мобильными станциями в пределах радиопокрытия своей зоны.

VLR – гостевой регистр (**V**isitor **L**ocation **R**egister), однозначно ассоциирован с одним MSC и хранит копии базы данных мобильных абонентов, которые временно пребывают в зоне обслуживания MSC. Практически часто реализуются в виде единого устройства с MSC и обозначаются как MSC/VLR.

HLR – домашний регистр (**H**ome **L**ocation **R**egister), является постоянным хранилищем оригиналов базы данных мобильных абонентов своего оператора и высылает ее по запросу VLR. Заносятся оператором HLR при заключении договора об обслуживании.

AuC – центр аутентификации (**A**uthentication **C**entre – **AuC**) однозначно ассоциирован с HLR и содержит данные о секретных ключах и алгоритмах шифрования. При каждом сеансе связи мобильный абонент должен пройти процедуру идентификации, Параметры идентификации (случайное число и ответ на него) генерируются AC и высылаются в VLR по его запросу.

Проще говоря, происходят следующие события.

1. Мобильная станция посылкой сообщения Location Update Request (LOC_UPD_REQ) инициирует в MSC/VLR запрос на регистрацию. В качестве собственного идентификатора MS использует IMSI.

2. VLR по адресу IMSI адресуется в HLR (OP = SAI) и запрашивает у него идентификационные параметры (триплеты). В качестве своего идентификатора VLR использует сетевой номер (VLR number). HLR запрашивает триплеты у AuC, получает несколько триплетов (обычно 5) и отсылает их VLR. Триплет содержит случайное число и ответ на него.

3. VLR выбирает один из триплетов и инициирует процедуру аутентификации (посылает MS случайное число и сравнивает его с истинным ответом).

4. При успешной аутентификации MS и VLR, используя постоянный ключ и случайное число из триплета, генерируют и согласовывают временный ключ шифрования на данный сеанс связи.

5. При успешном исходе аутентификации VLR сообщением Operation Code=updateLocation (OP=UPL) запрашивает у HLR базу данных (профайл) абонента. В ответ HLR высылает запрошенные данные (OP=insertSubscriberData) и запоминает номер VLR. С этого момента HLR известно местоположение собственного мобильного абонента.

6. VLR запоминает абонентские данные, присваивает TMSI (псевдоним абонента) взамен IMSI, отсылает TMSI мобильной станции.

7. VLR завершает регистрацию (Location Update Accept – LOC_UPD_ACC). Мобильная станция записывает TMSI и код нового местоположения в SIM-карту.

8. MSC/VLR высылает команду BSS о завершении регистрации.

9. BSS высылает команду о MS о завершении сеанса связи и получает подтверждение.

10. BSS рапортует MSC/VLR об освобождении радиоканала.

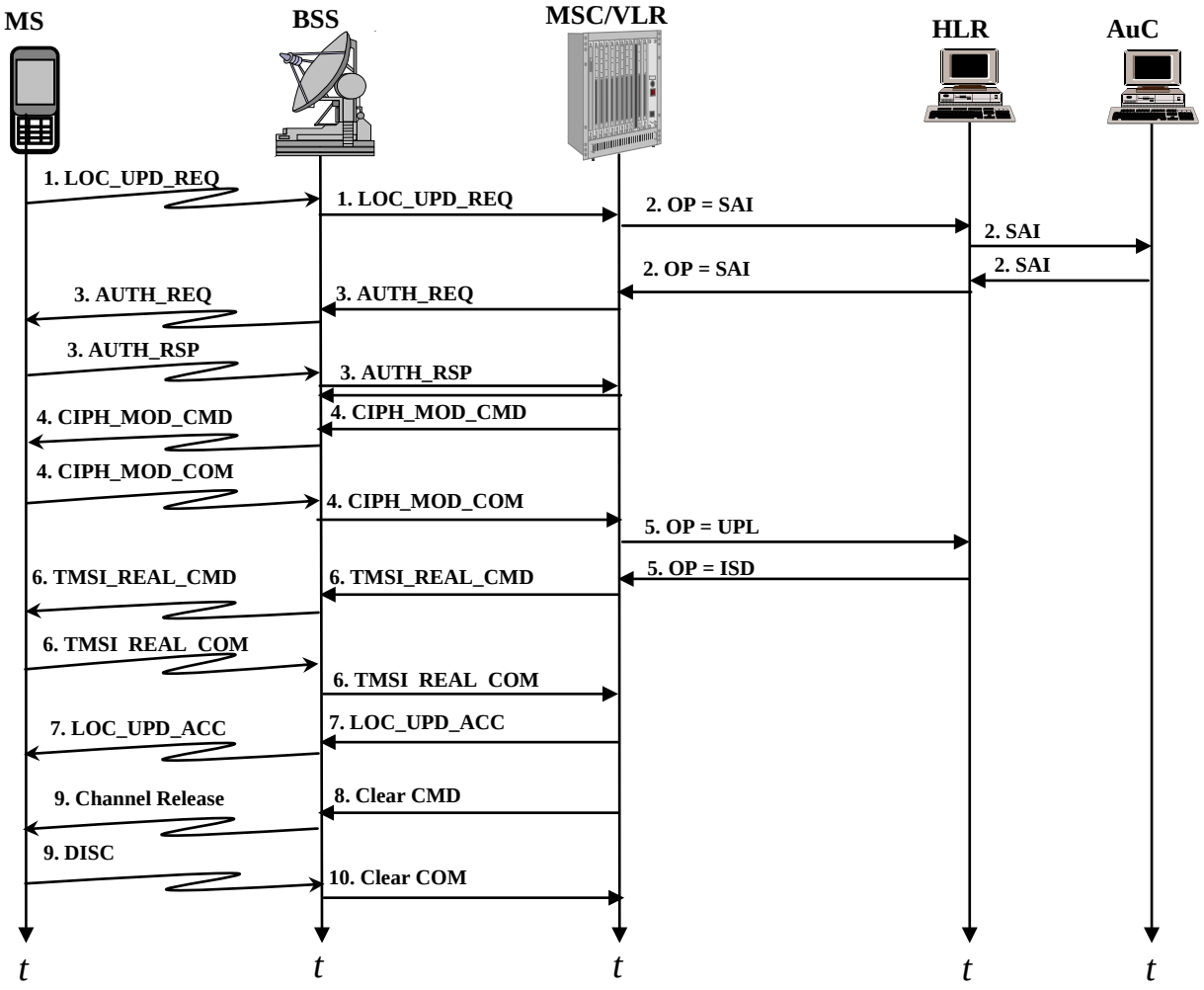


Рис. 56. Начальная регистрация (IMSI Attach)

В дальнейшем мобильная станция производит *периодическую регистрацию* (период устанавливается оператором в диапазоне 6–480 мин), подтверждая свое присутствие в сети. В этом случае VLR, используя TMSI и новый триплет, производит те же операции регистрации, кроме пункта 5, т. е. не обращается к HLR. При каждом сеансе связи VLR использует новый триплет, при их истощении VLR запрашивает у HLR новую партию.

Предотвращение угроз основывается на анализе данных, которые поступают со счетчиков сетевых устройств. Счетчик представляет собой программный элемент сетевого устройства, который срабатывает при действиях, запрограммированных в его коде. Можно выделить следующие виды счетчиков: счетчики, регистрирующие данные MSC, VLR, HLR, AuC и многие другие. Для реализации алгоритмов потребуется ограниченный набор счетчиков, основные из которых перечислены ниже.

- IDENT_REQ_CORR_FAILURE представляет собой счетчик VLR, который регистрирует количество несовпадений TMSI, полученных от абонента, с TMSI, назначенным VLR.
- INTW_BSS_IMSI_UNKNOWN_IN_HLR является счетчиком VLR, описывающим количество несовпадений IMSI, полученных от абонента с IMSI, записанным в HLR.
- RECVD_UPD_LOC_MSG_VLR_HPLMN является счетчиком HLR и показывает, сколько раз, где и в какое время абонент регистрировался в пределах домашней сети.
- RECVD_UPD_LOC_MSG_VLR_VPLMN является счетчиком HLR и показывает, сколько раз, где и в какое время абонент регистрировался в пределах гостевой сети.

Анализируя и обрабатывая данные, которые поступают с этих счетчиков, можно предотвращать угрозы, связанные с неправомерным использованием личных идентификационных данных абонентов.

17.1. Использование чужих персональных данных

В случае утечки или похищения персональных данных происходит следующая последовательность событий.

1. Легитимный абонент успешно проходит процесс регистрации, используя свои идентификационные данные IMSI. Пройдя процесс аутентификации, он получает из пула новый TMSI_a, который будет храниться как в мобильной станции, так и в VLR оператора сети.

2. Злоумышленник, используя клонированную SIM‑ карту с теми же идентификационными данными, тоже проходит процесс регистрации, используя IMSI. Его мобильная станция посылает свои идентификационные данные, сеть инициализирует процедуру начальной регистрации. В этот момент срабатывает счетчик IDENT_REQ_CORR_FAILURE, так как возникла ошибка аутентификации, связанная с тем, что абонент не использовал ранее выданный TMSI_a. Однако сеть регистрирует абонента и выдает злоумышленнику TMSI_b, тогда как старый TMSI_a возвращается в пул.

3. Когда наступит время легитимному абоненту проходить процедуру перерегистрации, TMSI_a не совпадет с TMSI_b, хранящимся в сети опера-

тора. Сеть оператора опять инициализирует процедуру регистрации по IMSI и выдаст абоненту TMSI_c. В этот момент срабатывает счетчик IDENT_REQ_CORR_FAILURE, так как повторно возникла ошибка аутентификации, связанная с тем, что TMSI_a абонента не совпадает с TMSI_b, хранящимся в сети оператора.

4. Когда наступит время злоумышленнику проходить процедуру перерегистрации, его TMSI_b не совпадет с TMSI_c, хранящимся в сети оператора. Произойдет новое срабатывание счетчика IDENT_REQ_CORR_FAILURE.

Таким образом, значение счетчика IDENT_REQ_CORR_FAILURE будет монотонно увеличиваться, что позволит обнаружить и принять необходимые меры.

Единичное срабатывание счетчика может быть результатом ошибки системы, однако превышение определенного предела дает уверенность в том, что данный случай является не ошибкой системы, а инцидентом компрометации персональных данных (SIM-карты).

Дальнейшие действия будут направлены на то, чтобы уведомить клиента о том, что данные его карты скомпрометированы, заблокировать SIM-карту и сформировать отчет об инциденте, которым займется служба безопасности оператора сети.

17.2. Неизвестные идентификационные данные

Сценарии этих событий описывают угрозы, связанные с попыткой незаконного проникновения в сеть и использования ее ресурсов. Это происходит, например, при использовании злоумышленником похищенного устройства или в случае изготовления нелегитимных SIM-карт. Для фиксирования инцидента используются данные счетчика INTW_BSS_IMSI_UNKNOWN_IN_HLR. Сценарий срабатывания этого счетчика следующий.

1. Нелегальная мобильная станция инициирует в MSC/VLR запрос на регистрацию. В качестве собственного идентификатора MS использует IMSI.

2. VLR по адресу IMSI адресуется в HLR и запрашивает у него идентификационные параметры (триплеты). HLR не находит в своей базе данных параметров, соответствующих IMSI, и высылает отказ в регистрации.

3. Происходит срабатывание счетчика VLR INTW_BSS_IMSI_UNKNOWN_IN_HLR.

4. VLR высылает абоненту отказ в регистрации.

Срабатывание счетчика может происходить по ряду причин, в том числе:

– повреждение персональных данных (SIM–карты);

– использование мобильного устройства, уже заявленного как похищенное, и удаленного из HLR;

– попытка проникновения в сеть нелегитимного абонента.

В первом случая легитимный абонент не сможет пройти процедуру аутентификации и лишается пользования услугами связи. В таком случае абоненту следует обратиться к администрации сети и устранить проблему.

Последний случай требует более тщательного рассмотрения, не исключено, что злоумышленники производят нерабочие SIM-карты и распространяют их среди обычных пользователей. Также возможен сбой в аппаратуре, храненящей эту информацию.

17.3. Одновременная регистрация в разных сетях

Если в течение короткого времени происходит регистрация абонента из различных удаленных друг от друга сетей, то по показаниям счетчиков HLR RECVD_UPD_LOC_MSG_VLR_HPLMN и RECVD_UPD_LOC_MSG_VLR_VPLMN можно определить не только компрометацию персональных данных, но и местоположение злоумышленника.

Хотя пользователь использует одни и те же постоянные ключи аутентификации и шифрования, трудность их вскрытия состоит в том, что при каждом сеансе связи мобильный пользователь получает новый псевдоним и использует новый временный ключ шифрования. Временный ключ шифрования создается путем перемешивания битов постоянного ключа и сменяемого случайного числа. Поэтому злоумышленнику необходимо отслеживать перемещение пользователя (его псевдоним), что крайне затруднительно из–за смены местоположения мобильного абонента.

Для стационарных абонентов для повышения криптостойкости желательно попеременно использовать несколько SIM-карт с различными постоянными ключами аутентификации и шифрования.

ЗАКЛЮЧЕНИЕ

В учебном пособии приведены основные методы расчета систем распределения информации с различной технологией коммутации.

Понятие о качестве обслуживания в сетях коммутации каналов и пакетов (QoS) дано в **разделе 1**.

В **разделе 2** приведены параметры потоков вызовов и времени обслуживания, имеющие случайное или детерминированное распределение.

Понятие о нагрузке и описание различных технологий коммутации дано в **разделе 3**. Здесь же имеется вывод вероятностей стационарного процесса рождения и гибели, а также классификация различных систем коммутации.

Раздел 4 содержит методы расчета качества обслуживания в системах с коммутацией каналов, все еще в достаточном количестве присутствующие у операторов сетей связи. Наряду с классическими формулами Эрланга и Энгсета приведены их интегральные представления, использующие нецелочисленные значения исходных данных.

Расчету качества обслуживания в системах пакетной коммутации с неограниченным числом мест ожидания посвящен **раздел 5**. Исследуются их вероятностно-временные характеристики со случайным и постоянным временем обслуживания. Результаты применимы для расчета пропускной способности маршрутизаторов в IP-сетях.

В **разделе 6** приводится расчет качества обслуживания в системах пакетной коммутации с ограниченным числом мест ожидания. Результаты применимы для расчета пропускной способности контроллеров в IP-сетях.

Приложения реального времени, такие, как VoIP, имеют приоритет обслуживания по сравнению с другими приложениями. Системы с различными приоритетами (Absolute Priority Queueing, Class-Based Weighted Fair Queuing, Weighted Fair Queuing) исследуются в **разделе 7**.

Описание систем с повторными вызовами и коллизией приведено в **разделах 8** и **9**.

Общие принципы формирования речевых пакетов и процессы при их доставке по IP-сети описаны в **разделе 10**.

С недостатками и преимуществами системы коммутации пакетов по сравнению с системой коммутации каналов в части приложений реального времени можно ознакомиться в **разделе 11**.

Раздел 12 содержит расчет отдельных характеристик маршрутизатора. Дано понятие о метрике и ее влиянии на выбор маршрута доставки пакетов по IP-сети. Дан расчет емкости буфера маршрутизатора для ограничения нагрузки сети пакетами с недопустимой задержкой.

Использование асимметрии трафика «на последней миле сервера» для оптимального распределения каналов связи исследуется в **разделе 13**. Приведено сравнение вероятностно–временных характеристик при симметричном и асимметричном распределении каналов.

Описание надежностных характеристик, систем технического обслуживания, способов повышения надежности систем коммутации приведены в **разделе 14**.

Раздел 15 описывает переходные процессы в системах коммутации, предшествующие стационарному состоянию. Показано применение расчетов для оценки вероятностно–временных характеристик необслуживаемых систем коммутации.

Методы контроля качества обслуживания персоналом приведены в **разделе 16**. Здесь даны формулы комплексной оценки класса обслуживания (GoS) для расчета качества обслуживания (QoS) на основе статистических данных.

В **разделе 17** рассматривается обнаружение компрометации персональных данных на примере сети связи мобильных абонентов стандарта GSM.

Приложение содержит ряд программ на языке C++ для вычисления приведенных формул.

Материалы учебного пособия использовались автором для чтения лекций студентам старших курсов по дисциплинам «Теория массового обслуживания», «Теория телетрафика» и «Обеспечение качества обслуживания в инфокоммуникационных сетях». Автор надеется, что пособие окажется полезным для научных работников, аспирантов, специалистов и обслуживающего персонала телекоммуникационных систем.

Электронную версию учебного пособия можно найти в библиотеке СПбГУТ на сайте www.sut.ru.

Автор выражает благодарность рецензентам за ценные советы и замечания, способствовавшие устранению недостатков данной работы.

СПИСОК ЛИТЕРАТУРЫ

1. *Вегешна, Шринивас*. Качество обслуживания в IP–сетях : пер. с англ. / Шринивас Вегешна. – Москва : Издательский дом «Вильямс», 2003. – 368 с.

2. *Кожанов, Ю. Ф*. Расчет и проектирование электронных АТС. Справочник / Ю. Ф. Кожанов. – Москва : Радио и связь, 1991. – 144 с.

3. *Кожанов, Ю. Ф*. Протоколы и интерфейсы в цифровой сети с коммутацией каналов / Ю. Ф. Кожанов. – Санкт-Петербург : Искусство России, 2002. – 256 с.

4. *Кожанов, Ю. Ф*. Сети следующего поколения. Теория и практика : учебное пособие / Ю. Ф. Кожанов, М. О. Колбанев : Санкт-Петербург : ГУАП, 2010. – 270 с.

5. *Лившиц, Б. С.* Теория телефонных и телеграфных сообщений / Б. С. Лившиц, Я. В. Фидлин, А. Д. Харкевич. – Москва : Связь, 1971. – 304 с.

6. ITU-T Recommendation E.550. GRADE–OF–SERVICE AND NEW PERFORMANCE CRITERIA UNDER FAILURE CONDITIONS IN INTERNATIONAL TELEPHONE EXCHANGES.

7. ITU-T Recommendation G.114. International telephone connections and circuits – General Recommendations on the transmission quality for an entire international telephone connection.

8. ITU-T Recommendation Q.543. DIGITAL EXCHANGE PERFORMANCE DESIGN OBJECTIVES.

9. ITU-T Recommendation Q.544. DIGITAL EXCHANGE MESUREMENTS.

10. ITU-T Recommendation P.800. METHODS FOR SUBJECTIVE DETERMINATION OF TRANSMISSION QUALITY.

11. RFC 1155. Structure and identification of management information for TCP/IP-based internets. M.T. Rose, K. McCloghrie. May–01–1990.

12. RFC 1633. Integrated Services in the Internet Architecture: an Overview. R. Braden, D. Clark, S. Shenker. June 1994.

13. RFC 2328. OSPF Version 2. J. Moy. April 1998.

14. RFC 2474. Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers. K. Nichols, S. Blake, F. Baker, D. Black. December 1998.

15. RFC 4750. OSPF Version 2 Management Information Base. D. Joyal, P. Galecki, S. Giacalone. December 2006.

16. ITU-T Recommendation Y.1540. Internet protocol data communication service – IP packet transfer and availability performance parameters.

17. ITU-T Recommendation Y.1541. Network performance objectives for IP-based Services.

Приложение 1

Программа вычисления формулы Пуассона

Описание:

Программа вычисляет вероятность поступления K вызовов за время T при поступающем параметре потока вызовов L и наиболее вероятное число поступивших вызовов.

L и T – действительные положительные числа, K – положительное целое.

Формула Пуассона (раздел 2):

$$\pi_K(L;T) = \frac{(L \cdot T)^K}{K!} e^{-LT}.$$

Текст программы:

```
#include <iostream>
#include <windows.h>
#include <cmath>
using namespace std;
int main()
{
system("chcp 1251 >nul");
float  l, t, k, p;
cout <<"Введите значение параметра потока L= "; cin >>l;
cout <<"Введите значение времени наблюдения T= "; cin >>t;
cout <<"Введите значение числа вызовов K= "; cin >>k;
p=exp(–l*t);
for (int i=1; i<=k; i++)
  {
  p=l*t*p/i;
  }
cout <<"Вероятность поступления K вызовов = "<<p<<endl;
cout<<"Математическое ожидание числа вызовов="<<l*t<<endl;
system("pause");
return 0;
}
```

Приложение 2

Программа вычисления формулы Эрланга (система M/M/V/0)

Описание.

Программа вычисляет вероятность потерь и среднее число занятых приборов при заданных значениях поступающей нагрузки A и числе обслуживающих приборов V.

A и V – действительные положительные числа.

Интегральная формула Эрланга (разд. 4.2):

$$[V] = E_V(A) = \frac{A^V}{\int\limits_0^\infty (A+z)^V e^{-z} dz}.$$

Текст программы:

```
#include <iostream>
#include <windows.h>
#include <cmath>
using namespace std;
int main()
{
system("chcp 1251 >nul");
float  A, vmin, V, predel, z, dz, iva, y, p, e;
cout <<"Интенсивность нагрузки A= "; cin >>A;
cout <<"Число обслуживающих приборов V= "; cin >>V;
vmin=V–int(V);
if (vmin>0)
  {
  predel=100*(vmin+1); dz=0.0001; iva=0;
  for (z=0.0001; z<=predel; z=z+dz)
     {y=vmin*log(A+z)–z; p=dz*exp(y);
     iva=iva+p; e=pow(A, vmin)/iva;}
  }
else {e=1;}
for (float i=vmin+1; i<=V; i++) {e=(A*e)/(i+A*e);}
cout <<"Вероятность потерь= "<<e<<endl;
cout <<"Среднее число занятых приборов = "<<A*(1–e)<<endl;
system("pause");
return 0;
}
```

Приложение 3

Программа классификации суммарного потока вызовов

Описание.

Программа вычисляет коэффициент просеивания суммарного потока вызовов k, образованного двумя потоками. Первый поток имеет параметр $L1$ и коэффициент просеивания $k1$, второй – параметр $L2$ и коэффициент просеивания $k2$.

$L1$, $L2$, $k1$, $k2$ – действительные положительные числа.

Коэффициент просеивания суммарного потока вызовов (разд 4.4):

$$k = \frac{\Lambda_S}{\lambda_S} - 1,$$

где

$$\Lambda_s = \frac{\sum_{i=1}^{n} \lambda_i \Lambda_i}{\sum_{i=1}^{n} \lambda_i}, \qquad \lambda_S = \sum_{i=1}^{n} \lambda_i .$$

Текст программы:

```
#include <iostream>
#include <windows.h>
#include <cmath>
using namespace std;
int main()
{
system("chcp 1251 >nul");
float  l1, l2, k1, k2, L1, L2, Ls, k;
cout<<"Параметр первого просеянного потока L1= "; cin >>l1;
cout<<"Параметр второго просеянного потока L2= "; cin >>l2;
cout<<"Коэффициент просеивания первого потока k1= "; cin>>k1;
cout<<"Коэффициент просеивания второго потока k2= "; cin>>k2;
L1=(k1+1)*l1;
L2=(k2+1)*l2;
Ls=l1*L1/(l1+l2)+l2*L2/(l1+l2);
k=Ls/(l1+l2)–1;
cout<<"Параметр суммарного просеянного потока ls="<<(l1+l2)<<endl;
cout<<"Суммарный коэффициент просеивания K = "<<k<<endl;
system("pause");
return 0;
}
```

Приложение 4

Программа вычисления формулы Энгсета (система M/M/V/0/N)

Описание.

Программа вычисляет вероятность потерь по времени при заданных значениях удельной нагрузки a, обслуживающих приборов V и источников нагрузки N.

a, V, N – действительные положительные числа.

Интегральная формула Энгсета (разд. 4.6):

$$[V] = E_{N,V}(A) = \frac{A^V}{\int\limits_0^\infty \left(1 + A - e^{-\frac{z}{N-V}}\right)^V e^{-z} dz},$$

где $A = a/(1 - a)$.

Текст программы:

```
#include <iostream>
#include <windows.h>
#include <cmath>
using namespace std;
int main()
{
system("chcp 1251 >nul");
float  a, A, vmin, v, predel, z, dz, sum, y, p, e, n;
cout <<"Число источников нагрузки N = "; cin >>n;
cout <<"Число обслуживающих приборов V = "; cin >>v;
cout <<"Удельная нагрузка a = "; cin >>a;
A=a/(1–a); vmin=v–int(v);
if (vmin>0)
  {predel=100*(vmin+1); dz=0.0001; sum=0;
  for (z=0.0001; z<=predel; z=z+dz)
     {y=vmin*log(1+A–exp(–z/(n–vmin)))–z; p=dz*exp(y);
     sum=sum+p; e=pow(A, vmin)/sum;}
  }
else {e=1;}
for (float i=vmin+1; i<=v; i++)
{e=(A*(n–i+1)*e)/(i+A*(n–i+1)*e);}
cout <<"Вероятность потерь по времени = "<<e<<endl;
system("pause");
return 0;
}
```

Приложение 5

Программа вычисления функции распределения системы M/M/V

Описание.

Программа вычисляет вероятность окончания обслуживания вызова V-линейным пучком за время t при заданной интенсивности поступления вызовов λ, интенсивности обслуживания μ.

V – положительное целое, λ, μ, $t > 0$.

Формула M/M/V (разд. 5.2):

$$F_e(t) = \begin{cases} 1 + \dfrac{P(>0)}{V - A - 1}\left[e^{-\mu(V-A)t} - e^{-\mu t}\right] - e^{-\mu t}, & 0 < A \neq V - 1 < V, \\ 1 - \mu t\ P(>0)e^{-\mu t} - e^{-\mu t}, & A = V - 1, \end{cases}$$

$$P(>0) = \frac{V \cdot E_V(A)}{V - A \cdot \left(1 - E_V(A)\right)}, \qquad A = \frac{\lambda}{\mu}$$

Текст программы:

```
#include <iostream>
#include <windows.h>
#include <cmath>
using namespace std;
float p(float l, float m, float V)
    {
    float e=1, A=l/m, p0;
    for (float i=1; i<=V; i++) {e=(A*e)/(i+A*e);}
    p0=V*e/(V–A*(1–e));
    return (p0);
    }
int main()
{
system("chcp 1251 >nul");
float l, m, V, e, p0, p00, fe, t, A;
cout <<"Параметр потока l= "; cin >>l;
cout<<"Интенсивность обслуживания m= "; cin>>m;
cout <<"Число обслуживающих приборов V= "; cin >>V;
cout <<"Время окончания обслуживания t= "; cin >>t;
A=l/m; p0=p(l, m, V); p00=p0/(V–A–1);
if (A!=(V–1))   {fe=1+p00*(exp(–m*(V–A)*t)–exp(–m*t))–exp(–m*t);}
else          {fe=1–m*t*p0*exp(–m*t)–exp(–m*t);}
cout <<"Вероятность окончания обслуживания F(t)= "<<fe<<endl;
system("pause");
return 0;
}
```

Приложение 6

Программа вычисления функции распределения системы M/D/V

Описание.

Программа вычисляет вероятность окончания обслуживания вызова V-линейным пучком за время t при заданной интенсивности поступления вызовов λ, интенсивности обслуживания μ.

V – положительное целое, λ, μ, $t > 0$.

Формула M/D/V (разд. 5.4):

$$F_e(t) = 1 - P(>0)\left(\frac{A}{V}\right)^{V(\mu t - 1)}, \quad P(>0) = \frac{V \cdot E_V(A)}{V - A \cdot \left(1 - E_V(A)\right)}, \quad A = \frac{\lambda}{\mu}, \quad t \geq 1/\mu.$$

Текст программы:

```
#include <iostream>
#include <windows.h>
#include <cmath>
using namespace std;
float p(float l, float m, float V)
    {
     float e=1, A=l/m, p0;
     for (float i=1; i<=V; i++) {e=(A*e)/(i+A*e);}
     p0=V*e/(V–A*(1–e));
     return (p0);
     }
int main()
{
system("chcp 1251 >nul");
float l, m, V, p0, fe, t, A, a, k;
cout <<"Параметр потока l= "; cin >>l;
cout<<"Интенсивность обслуживания m= "; cin>>m;
cout <<"Число обслуживающих приборов V= "; cin >>V;
cout <<"Время окончания обслуживания t= "; cin >>t;
A=l/m; a=A/V; p0=p(l, m, V); k=V*(m*t–1);
if (t>1/m) {fe=1–p0*pow(a, k);}
else {fe=0;}
cout <<"Вероятность окончания обслуживания F(t)= "<<fe<<endl;
system("pause");
return 0;
}
```

Приложение 7

Программа расчета системы M/M/1/ / /PPQ

Описание.

Программа вычисляет среднее время пребывания вызова *i*-го потока в системе с абсолютными приоритетами и прерыванием менее приоритетных вызовов.

m, *Li* – действительные положительные числа.

Формулы из разд. 7.2.

Текст программы:

```
#include <iostream>
#include <windows.h>
#include <cmath>
using namespace std;
int main()
{
system("chcp 1251 >nul");
int n=4; float L[5]={0}, ai[10]={0}, a=0; int m;
cout <<"Интенсивность обслуживания (пакетов в сек) m="; cin >>m;
float Li; cout <<"Параметры потоков (пакетов в сек):"<<endl;
for (int i=1; i<=n; i++)  {cin>>Li; L[i]=Li;}
for (int i=1; i<=n; i++) {ai[i]=L[i]/m; a=a+ai[i];}  //нагрузки
float tei[5]={1E+10,1E+10,1E+10,1E+10};    //начальные значения tei
if (ai[1]>0 && ai[1]<1)   {tei[1]=1/(m*(1–ai[1]));}
if (ai[1]==0) {tei[1]=0;}
if (ai[1]+ai[2]>0 && ai[1]+ai[2]<1)   {tei[2]=1/(m*(1–ai[1]–ai[2])*(1–ai[1]));}
if (ai[2]==0) {tei[2]=0;}
if (ai[1]+ai[2]+ai[3]>0 && ai[1]+ai[2]+ai[3]<1)
  {tei[3]=1/(m*(1–ai[1]–ai[2]–ai[3])*(1–ai[1]–ai[2]));}
if (ai[3]==0) {tei[3]=0;}
if (ai[1]+ai[2]+ai[3]+ai[4]>0 && ai[1]+ai[2]+ai[3]+ai[4]<1)
  {tei[4]=1/(m*(1–ai[1]–ai[2]–ai[3]–ai[4])*(1–ai[1]–ai[2]–ai[3]));}
if (ai[4]==0) {tei[4]=0;}
for (int i=1; i<=n; i++)
   {
    cout<<"Время окончания обслуживания "<<i<<"–го потока = ";
    cout<<tei[i]<<endl;
   }
return 0;
}
```

Приложение 8

Программа расчета системы M/M/1/ / /CBWFQ

Описание.

Программа вычисляет среднее время пребывания вызова *i*-го потока в системе CBWFQ с относительными приоритетами с объединением потоков.

m, *Li*, *bwi*% – действительные положительные числа.

Формулы из разд. 7.3.

Текст программы:

```
#include <iostream>
#include <windows.h>
#include <cmath>
using namespace std;
int main()
{
system("chcp 1251 >nul");
int n; cout <<"Число потоков (n<=10) n="; cin>>n;
int num[11]={0};
for (int i=0; i<=n; i++) {num[i]=i;}
float m; cout <<"Интенсивность обслуживания m="; cin >>m;
float L[11]={0};
float Li; cout <<"Параметры потоков Li="<<endl;
for (int i=1; i<=n; i++)  {cin>>Li; L[i]=Li;}
float bwi, bw[10]={0}; cout <<"Доля выделенной полосы для i–го потока bwi%
(в  %)="<<endl;
for (int i=1; i<=n; i++)  {cin>>bwi; bw[i]=bwi/100;}
float a[11]={0};
for (int i=1; i<=n; i++)  {a[i]=L[i]/m;}    //расчет нагрузки
float pi[11]={0};
for (int i=1; i<=n; i++) {pi[i]=a[i]/bw[i];}  //расчет приоритетов
float temp;
for (int i=n–1; i>=0; i––)   //сортировка потоков по приоритетам
   {
   for (int j=i; j<n; j++)
      {
       if(pi[j]>pi[j+1])
         {
          temp=num[j]; num[j]=num[j+1]; num[j+1]=temp;
          temp=L[j]; L[j]=L[j+1]; L[j+1]=temp;
          temp=bw[j]; bw[j]=bw[j+1]; bw[j+1]=temp;
          temp=a[j]; a[j]=a[j+1]; a[j+1]=temp;
          temp=pi[j]; pi[j]=pi[j+1]; pi[j+1]=temp;
         }
      }
   }
```

```
float A[11]={0}, BW[10]={0}; //суммарная нагрузка и суммарная полоса
for (int i=1; i<=n; i++)  {A[i]=A[i–1]+a[i]; BW[i]=BW[i–1]+bw[i];}
float QW[11]={0};    //смммарная очередь
for (int i=1; i<=n; i++)  {QW[i]=(A[i]*A[i])/((1–A[i])*BW[i]);}
float te[11]={0};
for (int i=1; i<=n; i++)
   {
    te[i]=((QW[i]–QW[i–1])+a[i])/L[i];
    cout<<"Время окончания обслуживания "<<num[i]<<"–го потока = ";
    cout<<te[i]<<endl;
   }
system("pause");
return 0;
}
```

Приложение 9

Программа расчета системы M/M/1/ / /WFQ

Описание.

Программа вычисляет среднее время пребывания вызова *i*-го потока в системе WFQ с относительными приоритетами с объединением потоков.

m, *Li*, *TOSi* – действительные положительные числа.

Формулы из разд. 7.4.

Текст программы:

```
#include <iostream>
#include <windows.h>
#include <cmath>
using namespace std;
int main()
{
system("chcp 1251 >nul");
int n; cout <<"Число потоков (n<=10) n="; cin>>n;
int num[11]={0};
for (int i=0; i<=n; i++) {num[i]=i;}
float m; cout <<"Интенсивность обслуживания m="; cin >>m;
float L[11]={0};
float Li; cout <<"Параметры потоков Li="<<endl;
for (int i=1; i<=n; i++)  {cin>>Li; L[i]=Li;}
float tosi, TOS[10]={0}, S_TOS=0; cout <<"Значения поля TOSi="<<endl;
for (int i=1; i<=n; i++)  {cin>>tosi; TOS[i]=tosi+1; S_TOS=S_TOS+tosi+1;}
float bw[11]={0};
for (int i=1; i<=n; i++)  {bw[i]=TOS[i]/S_TOS;}  //расчет полосы
float a[11]={0};
for (int i=1; i<=n; i++)  {a[i]=L[i]/m;}     //расчет нагрузки
float pi[11]={0}; //расчет приоритетов
for (int i=1; i<=n; i++) {pi[i]=a[i]/bw[i];}
float temp;
for (int i=n−1; i>=0; i−−)  //сортировка
   {
   for (int j=i; j<n; j++)
      {
       if(pi[j]>pi[j+1])
         {
          temp=num[j]; num[j]=num[j+1]; num[j+1]=temp;
          temp=L[j]; L[j]=L[j+1]; L[j+1]=temp;
          temp=bw[j]; bw[j]=bw[j+1]; bw[j+1]=temp;
          temp=a[j]; a[j]=a[j+1]; a[j+1]=temp;
          temp=pi[j]; pi[j]=pi[j+1]; pi[j+1]=temp;
         }
      }
```

```
   }
float A[11]={0}, BW[11]={0}; //суммарная нагрузка и суммарная полоса
for (int i=1; i<=n; i++)  {A[i]=A[i–1]+a[i]; BW[i]=BW[i–1]+bw[i];}
float QW[11];  //сммарная очередь
for (int i=1; i<=n; i++)  {QW[i]=(A[i]*A[i])/((1–A[i])*BW[i]);}
float te[11]={0};
for (int i=1; i<=n; i++)
   {
    te[i]=((QW[i]–QW[i–1])+a[i])/L[i];
    cout<<"Время окончания обслуживания "<<num[i]<<"–го потока = ";
    cout<<te[i]<<endl;
   }
system("pause");
return 0;
}
```

Приложение 10

Программа расчета системы с повторными вызовами

Описание.

Программа вычисляет консолидированную нагрузку *B* и число попыток *M* = *B*/*As* на один успешный вызов при известной обслуженной нагрузке *As* и числе приборов *V*.

As, *V* – действительные положительные числа.

Формула (разд. 8)

$$A_S = B\left(1 - E_V(B)\right)$$

Текст программы:

```
#include <iostream>
#include <windows.h>
#include <cmath>
using namespace std;
int main()
{
system("chcp 1251 >nul");
float as, v, b=0, evb, a=0;
cout<<"Обслуженная нагрузка As = "; cin >>as;
cout<<"Число обслуживающих приборов V= "; cin >>v;
while (a<as)
   {
    evb=1;
    for (int i=1; i<=v; i++)
       {evb=(b*evb)/(i+b*evb);}
    a=b*(1–evb);
    b=b+0.01;
   }
cout<<"Консолидированная нагрузка B="<<b<<endl;
cout<<"Число попыток на один вызов M="<<b/as<<endl;
system("pause");
return 0;
}
```

Приложение 11

Программа расчета системы с интегрированным обслуживанием

Описание.

Программа вычисляет параметры системы с интегрированным обслуживанием в соответствии с разд. 11.1, где *tcod* – время кодирования пакета, *tdec* – время декодирования пакета, *VEL* – скорость передачи интерфейса, *k* – доля приоритетных соединений, *bw* – полоса передачи приоритетных приложений, *LEN* – средняя длина пакета, *PLD* – полезная нагрузка, *s* – число проходных маршрутизаторов.

Текст программы:

```
#include <iostream>
#include <windows.h>
#include <cmath>
double f(double pot, double v)
    {
    double A=0, da=0.1, e=0;
    while (e<pot)
        {
        e=1; A=A+da;
        for (int i=1; i<=v; i++) {e=(A*e)/(i+A*e);}
        }
    return (A);
    }
using namespace std;
int main()
{
system("chcp 1251 >nul");
double tcod, tdec, VEL, k, bw, LEN, s, PLD;
cout<<"Время кодирования пакета tcod (мс) = "; cin >>tcod;
cout<<"Время декодирования пакета tdec (мс)= "; cin >>tdec;
cout<<"Скорость передачи интерфейса VEL (кбит/с)= "; cin >>VEL;
cout<<"Доля полосы всех приоритетных соединений ( %)= "; cin >>k;
cout<<"Полоса приоритетного соединения bw (кбит/с)=  "; cin >>bw;
cout<<"Средняя длина пакета LEN (байт) = "; cin >>LEN;
cout<<"Полезная нагрузка PLD (байт)="; cin>> PLD;
cout<<"Число проходных маршрутизаторов S (S>1) = "; cin >>s;
cout<<endl<<"================================="<<endl;
tcod=tcod/1000; tdec=tdec/1000; VEL=1000*VEL; k=k/100;
double BW=VEL*k; bw=1000*bw; LEN=8*LEN, PLD=8*PLD;
double v=INT(BW/bw);
cout<<"Максимальное число приоритетных соединений V= "<<v<<endl;
double L=v*bw/LEN;
```

```
cout<<"Максимальная интенсивность поступления пакетов от всех абонентов в ЧНН (в сек) L= "<<L<<endl;
double L0=1/tcod;
cout<<"Интенсивность поступления пакетов от абонента (в сек) L0= "<<L0<<endl;
double m=VEL/LEN;
cout<<"Интенсивность обслуживания пакетов маршрутизатором (в сек) M= "<<m<<endl;
double a=L/m;
cout<<"Нагрузка на интерфейсе маршрутизатора (Эрланг) a= "<<a<<endl;
double te=1/(m–L);
cout<<"Задержка в маршрутизаторе (мс) te= "<<(1000*te)<<endl;
double tej=te;
cout<<"Джиттер маршрутизатора (мс) tej= "<<tej*1000<<endl;
double st=s/(s–1), thj=tej*pow(st,0.5);
cout<<"Джиттер хоста (мс) thj= "<<thj*1000<<endl;
double tnet=0.15–tcod–tdec–thj;
cout<<"Допустимое время прохождения пакета по сети (мс) tnet= "<<tnet*1000<<endl;
double tete=tcod+tdec+s*te+thj;
cout<<"Задержка их конца–в–конец (мс) tete= "<<tete*1000<<endl;
double Rmax=LEN–304;  //Минус физический и канальный уровни
cout<<"Емкость буфера очереди маршрутизатора для 1 соединения (байт) Rmax= "<<Rmax/8<<endl;
double hbuf=2*thj*L0*PLD+PLD;
cout<<"Емкость буфера очереди приложения (байт) hbuf= "<<hbuf/8<<endl;
double pot=0.05/s;
double A=f(pot, v);
cout<<"Число обслуживаемых абонентов = "<<A/0.1<<endl;
system("pause");
return 0;
}
```

Приложение 12

Программа расчета системы с дифференцированным обслуживанием

Описание.

Программа вычисляет параметры системы с дифференцированным обслуживанием в соответствии с разд. 11.2, где *tcod* – время кодирования пакета, *tdec* – время декодирования пакета, *VEL* – скорость передачи интерфейса, *k* – доля полосы для всех приоритетных соединений, *LEN* – средняя длина пакета, *PLD* – полезная нагрузка, *s* – число проходных маршрутизаторов.

Текст программы:

```
#include <iostream>
#include <windows.h>
#include <cmath>
using namespace std;
int main()
{
system("chcp 1251 >nul");
double tcod, tdec, VEL, k, LEN, s,PLD;
cout<<"Время кодирования пакета tcod (мс) = "; cin >>tcod;
cout<<"Время декодирования пакета tdec (мс)= "; cin >>tdec;
cout<<"Скорость передачи интерфейса VEL (кбит/с)= "; cin >>VEL;
cout<<"Доля полосы для всех приоритетных соединений k (в  %)= "; cin >>k;
cout<<"Средняя длина пакета LEN (байт) = "; cin >>LEN;
cout<<"Полезная нагрузка PLD (байт)="; cin>> PLD;
cout<<"Число проходных маршрутизаторов S = "; cin >>s;
cout<<endl<<"==============================="<<endl;
tcod=tcod/1000; tdec=tdec/1000; VEL=VEL*1000; k=k/100;
double BW=VEL*k; LEN=LEN*8, PLD=PLD*8;
double l=0, m, te, a, tej, thj, tnet, dt, st, fet=0;
while (fet<=0.001)
    {
     m=VEL/LEN;     //интенсивность обслужвания маршрутизатором
     a=l/m;
     te=(a/((1–a)*k)+1)/m; //задержка в маршрутизаторе с очередями
     tej=te;       //джиттер маршрутизатора
     st=pow(s/(s–1),0.5);
     thj=te*st;     //джиттер хоста
     tnet=0.15–tcod–tdec–thj; //среднее время прохождения по сети
     dt=(m–l)*tnet;
     double w=1, sum=1;
     for (int x=1; x<=s–1; x++) {w=w*dt/x; sum=sum+w;}
     fet=sum*exp(–dt);
     l=l+0.1;
```

```
    }
    cout<<"Максимальная интенсивность поступления пакетов (в сек) L= "<<l<<endl;
    double L0=1/tcod;
    cout<<"Интенсивность поступления пакетов от абонента (в сек) L0= "<<L0<<endl;
    cout<<"Интенсивность обслуживания интерфейсом (в сек) M= "<<m<<endl;
    cout<<"Нагрузка на интерфейсе маршрутизатора (Эрланг) a= "<<a<<endl;
    cout<<"Время прохождения маршрутизатора с вероятностью >0.999 (мс) te= "<<te*1000<<endl;
    cout<<"Джиттер маршрутизатора (мс) tej= "<<tej*1000<<endl;
    cout<<"Джиттер хоста (мс) thj= "<<thj*1000<<endl;
    cout<<"Допустимое время прохождения пакета по сети (мс) tnet= "<<tnet*1000<<endl;
    double tep=tnet/s;
    cout<<"Предельное время прохождения пакета через маршрутизатор (мс) tep= "<<tep*1000<<endl;
    double tete=tcod+tdec+s*te+thj;
    cout<<"Время прохождения пакета их конца–в–конец с вероятностью >0.999 (мс) tete= "<<tete*1000<<endl;
    double Rmax=BW*tep*(LEN–304)/LEN;
    cout<<"Емкость буфера очереди маршрутизатора с учетом джиттера (байт) Rmax= "<<Rmax/8<<endl;
    double hbuf=2*thj*L0*PLD+PLD;
    cout<<"Емкость приемного буфера хоста (байт) hbuf= "<<hbuf/8<<endl;
    double N=l/(0.1*L0);
    cout<<"Число обслуживаемых абонентов N= "<<N<<endl;
    cout<<endl<<endl;
    system("pause");
    return 0;
    }
```

Приложение 13

Программа расчета коэффициента готовности

Описание.

Программа вычисляет коэффициент готовности прибора при воздействии на него нескольких типов неисправностей.

Формула (разд. 14)

$$K = r/(\xi + r),$$

где

$$\xi = \xi_1 + \xi_2 + \ldots + \xi_m,\ r = \frac{\xi_1}{\xi_1 + \xi_2 + \ldots + \xi_m} r_1 + \ldots + \frac{\xi_m}{\xi_1 + \xi_2 + \ldots + \xi_m} r_m,$$

$\xi_i = 1/T\xi_i$, $r_i = 1/Tr_i$, $T\xi_i$ – среднее время появления i-й неисправности,

Tr_i – время восстановления (устранения) i-й неисправности.

$T\xi_i$, Tr_i – действительные положительные числа.

Текст программы:

```
#include <iostream>
#include <windows.h>
#include <cmath>
using namespace std;
int main()
{
system("chcp 1251 >nul");
int n; cout<<"Число неисправностей = "; cin >>n;
float sumte=0, te, masste[n], tr, masstr[n], r=0, k;
for (int i=1; i<=n; i++)
   {cout<<"Время появления "<<i<<"–ой неисправности=";
    cin>>te; te=1/te; masste[i]=te; sumte=sumte+masste[i];
    cout<<"Время устранения "<<i<<"–ой неисправности=";
    cin>>tr; tr=1/tr; masstr[i]=tr;};
for (int i=1; i<=n; i++)
    {r=r+masste[i]*masstr[i]/sumte;}
k=r/(sumte+r);
cout<<"Коэффициент готовности = "<<k<<endl;
system("pause");
return 0;
}
```

Приложение 14

Преобразование Лапласа – Стилтьеса

В тексте монографии многократно используется преобразование Лапласа. Ниже, практически без изменений, приводится описание преобразования Лапласа – Стилтьеса, взятое из монографии Н. А. Соколова «Задачи планирования сетей электросвязи».

Преобразование Лапласа устанавливает однозначную связь между функциями действительной – $F(t)$ и комплексной $\varphi(s)$ переменных. Эти функции называются оригиналом и изображением соответственно. Для оригиналов, существующих только в области неотрицательных значений аргумента (например, времени), целесообразно использовать одностороннее преобразование Лапласа:

$$\varphi(s) = \int_0^\infty F(t)e^{-st}dt\,. \qquad (\text{П}.1)$$

Функция $\varphi(s)$ комплексной переменной $s = \sigma + i\omega$ позволяет найти оригинал по формуле

$$F(t) = \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} \varphi(s)e^{st}ds\,. \qquad (\text{П}.2)$$

Последнее соотношение известно в математике как формула Римана – Меллина. Его также называют обратным преобразованием Лапласа.

В формулах (П.1) и (П.2) используются интегралы Римана. Во многих случаях приходится оперировать с дискретными функциями. Тогда предпочтительнее становится интеграл Стилтьеса, позволяющий унифицировать тип распределения случайной величины. Интеграл Стилтьеса определяется для двух функций: интегрируемой – $\varphi(s)$ и интегрирующей – $\theta(x)$:

$$\int_0^\infty \varphi(s)d\theta(x)\,.$$

Интеграл Римана представляет собой частный случай интеграла Стилтьеса, когда $\theta(x) = x$. Преобразование Лапласа – Стилтьеса для функции $F(t)$ определяется следующим образом

$$\psi(s) = \int_0^\infty dF(t)e^{-st}\,. \qquad (\text{П}.3)$$

Если изображения (П.1) и (П.3) существуют, то они связаны между собой очевидным соотношением:

$$\psi(s) = s\varphi(s).$$

Интерес к преобразованию Лапласа – Стилтьеса объясняется тем, что оно позволяет упростить исследование некоторых функций и вычисление параметров распределения. Следует выделить ряд свойств, которые характерны для преобразования Лапласа – Стилтьеса.

1. Преобразование Лапласа – Стилтьеса от производной функции $F(t)$ определяется по формуле

$$\frac{dF(t)}{dt} \to s\Psi(s) - s\Psi(0),$$

где → означает соответствие.

2. С помощью дифференцирования изображения можно получить r-й начальный момент распределения:

$$t^{(r)} = (-1)^r \frac{d\Psi(s)}{ds^r}_{|s=0}.$$

Это значительно упрощает вычисление математического ожидания (среднего значения) случайной величины:

$$t^{(1)} = M[t] = -\frac{d\Psi(s)}{ds}_{|s=0}.$$

3. Изображение функции распределения двух независимых случайных величин $\psi_1(s)$ и $\psi_2(s)$ определяется с помощью произведения их изображений, так называемая теорема о свертке:

$$\Psi(s) = \Psi_1(s)\Psi_2(s).$$

4. Сдвиг аргумента у оригинала на величину τ соответствует такому изменению изображения:

$$F(t-\tau) \to \Psi(s)e^{-s\tau}.$$

Из всех перечисленных свойств, присущих преобразованию Лапласа – Стилтьеса, следует выделить свертку их оригиналов, что позволяет отказаться от вычисления многократного интеграла. Поэтому вычисление свертки функций при помощи преобразований является одним из самых эффективных операций среди прочих возможностей.

Использование преобразования Лапласа – Стилтьеса в большинстве случаев можно представить в виде алгоритма, показанного на рис. П1. Этот алгоритм универсален для решения многих практических задач.

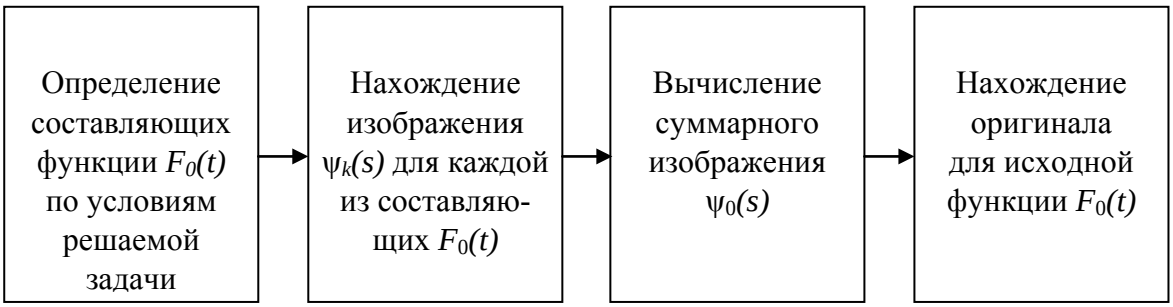


Рис. П1. Алгоритм использования преобразования Лапласа – Стилтьеса

Функция $F_0(t)$ представляет собой выражение, которое следует определить для искомого результата. Выбор этой функции можно рассматривать как постановку задачи. Далее находятся изображения каждой из составляющих этой функции $\psi_\kappa(s)$. Изображение может быть получено с помощью П.3 или с помощью таблиц преобразования Лапласа. Затем находится суммарное изображение $\psi_0(s)$ искомой функции $F_0(s)$. В завершение находится оригинал искомой функции $F_0(t)$ с использованием таблиц или иным способом.

Для вычисления оригинала $F_0(t)$ по известному изображению $\psi_0(s)$ часто применяют теорему разложения, предложенную Хевисайдом. Такой способ применим, если суммарное изображение $\psi_0(s)$ является рациональной дробью следующего вида:

$$\psi_0(s) = \frac{\varphi_1(s)}{\varphi_2(s)}. \qquad (П.4)$$

Числителем и знаменателем выражения (П.4) служат два полинома. Существенно то, что степень полинома числителя должна быть меньше степени полинома знаменателя. Оригинал искомой функции $F_0(t)$ находим следующим образом:

$$F_0(t) = \sum_{k=1}^{n} \frac{\varphi_1(s_k)(s - s_k)}{\varphi_2(s_k)} \, e^{s_k t}.$$

где n – число корней в полиноме $\varphi_2(s_k)$.

Такое преобразование справедливо в том случае, если полином $\varphi_2(s)$ не имеет кратных корней. В противном случае оригинал получить сложнее. В этом случае помогают таблицы нахождения оригиналов преобразования Лапласа.

Пусть имеется некоторая функция $F(t)$. Тогда математическое ожидание $M[t]$ и дисперсия $D[t]$ этой функции вычисляются таким образом:

$$M[t]=\int_0^\infty t dF(t)=\int_0^\infty t f(t)dt\,, \qquad (П.5)$$

$$D[t]=\int_0^\infty \left(t-M[t]\right)^2 dF(t)=\int_0^\infty \left(t-M[t]\right)^2 f(t)dt=$$

$$\int_0^\infty t^2 f(t)dt-2M[t]\int_0^\infty t f(t)dt+M^2[t]\int_0^\infty f(t)dt= \qquad (П.6)$$

$$=\int_0^\infty t^2 f(t)dt-2M^2[t]+M^2[t]=\int_0^\infty t^2 f(t)dt-M^2[t],$$

где $f(t)$ – плотность функции $F(t)$.

Вычислим преобразование Лапласа для функции $f(t)$:

$$f(s)=\int_0^\infty f(t)e^{-st}dt=\int_0^\infty f(t)\left[1-\frac{(st)^1}{1!}+\frac{(st)^2}{2!}-\frac{(st)^3}{3!}+\ldots\right]dt=$$

$$=\int_0^\infty f(t)dt-\frac{s^1}{1!}\int_0^\infty t f(t)dt+\frac{s^2}{2!}\int_0^\infty t^2 f(t)dt-\frac{s^3}{3!}\int_0^\infty t^3 f(t)dt+\ldots= \qquad (П.7)$$

$$=1-\frac{s^1}{1!}t^{(1)}+\frac{s^2}{2!}t^{(2)}-\frac{s^3}{3!}t^{(3)}+\ldots$$

где $t^{(i)}$ – i-й момент преобразования Лапласа.

Сравнивая (П.7) и (П.5) видим, что математическое ожидание функции $F(t)$ вычисляется как

$$M[t]=-t^{(1)}=-\frac{df(s)}{ds}_{|s=0}\,,$$

а дисперсия

$$D[t]=t^{(2)}-M^2[t]=\frac{d^2 f(s)}{ds^2}_{|s=0}-M^2[t]\,.$$

Кожанов Юрий Федорович

ТЕОРИЯ ТЕЛЕТРАФИКА

Учебное пособие

Редактор *С. В. Бардина*

Компьютерная верстка *Н. А. Ефремовой*

План издания 2020 г., п. 26

Подписано к печати 27.04.2020
Объем 12,75 печ. л. Тираж 26 экз. Заказ 1080

Редакционно-издательский отдел СПбГУТ
193232 СПб., пр. Большевиков, 22

Отпечатано в СПбГУТ