ББК 32.811.2
Б53
УДК 621.391.28: 518.5

Рецензенты: д-р техн. наук профессор В.Н.Красюк,
кафедра радиотехнических комплексов ВКА им. А. Ф. Можайского

Бестугин и др.

Б53 Контроль и диагностирование телекоммуникационных сетей / А. Р. Бестугин, А. Ф. Богданова, Г. В. Стогов. – СПб: Политехника, 2003. 174 с.: ил.

ISBN 5–7325–0xxx–x

Книга посвящена проблемам разработки сетевых систем связи с учетом реальных условий функционирования. В ней рассматриваются вопросы моделирования, контроля и диагностирования при организации управления сетями связи, построенными на основе как низкоскоростных, так и высокоскоростных линий связи.

Книга адресована инженерам-разработчикам сетевых систем связи, научным работникам соответствующего направления, а также менеджерам проектов в области электронного инжиниринга.

ISBN 5–7325–0727–2

Содержание

ВВЕДЕНИЕ

Анализ особенностей отдельных сетей передачи информации, изучение требований пользователей, исследование тенденций развития новых информационных служб привели к выводу, что средства связи как составная часть инфраструктуры общества могут стать наиболее эффективными только при условии объединения (интеграции) всех сетей связи в единую систему.

В настоящее время широко развиваются спутниковые системы связи, которые в сочетании с наземными системами связи позволяют организовать информационные сети с большими техническими возможностями. Это и спутниковые системы связи и телевидения, и сети сбора экологической информации и предупреждения чрезвычайных ситуаций, и сети с использованием космических аппаратов (КА), ретрансляторов и т.д. В связи с этим особую актуальность приобретает проведение исследований в области создания сетей интегрального обслуживания, ориентированных на передачу цифровой информации, т.е. цифровых сетей интегрального обслуживания (ЦСИО).

С точки зрения расположения пользователей и размещения информационно-вычислительных ресурсов, а также принципов обмена информации в интегральной сети выделяют две подсети: терминальную (абонентскую) и базовую (магистральную). В связи с различным целевым назначением и существенными различиями в требованиях к этим сетям они могут рассматриваться отдельно. Базовая сеть включает в себя устройства коммутации и каналы связи, а также системы управления базовой сетью. Управление процессом обмена информации в ЦСИО тесно связано с использованием в сети методов (режимов) комбинации. К настоящему времени получили распространение два метода комбинации: комбинация каналов (КК) и комбинация сообщений (тактов) (КС, КТ). В свою очередь метод КТ применяется в двух разновидностях: комбинация датаграмм (ДГ) и комбинация виртуальных каналов (ВК).

Одним из наиболее общих требований, предъявляемых к сети, является обеспечение эффективного использования технических средств связи: каналов связи, центров коммутации сообщений и другого оборудования. Решение поставленной задачи возможно при обеспечении высокой надежности функционирования сети. Необходимым условием обеспечения надежности является введение в систему управления сетью подсистемы контроля и диагностирования. В большинстве работ, посвященных этой проблеме, вопросы проектирования подсистем контроля и диагностирования рассматриваются отдельно и без учета тех задач, которые стоят перед сетями. Вместе с тем разработка подсистем контроля и диагностирования сети является частью общей проблемы управления сетью. Причем значительная часть информации о состоянии сети используется как системой оперативного управления сетью, так и системой технического обслуживания. Это позволяет создать единую базу данных о состоянии сети.

Деятельность по созданию подсистемы контроля и диагностирования тесно связана с задачами проектирования самих сетей, поскольку синтез устройств контроля основывается на результатах исследования моделей сети. Однако традиционный подход к проектированию, базирующийся на схеме «задача−модель−алгоритм», в данном случае представляется малопривлекательным ввиду необозримого количества решаемых задач и их модификаций. Между тем разукрупнение исходной задачи и выделение составных задач и обеспечение возможности их решения в различных комбинациях, декомпозиция используемой сетевой модели, выделение ее предельных и вырожденных аналогов (например, модели при идеальной надежности, фиксированных маршрутах и т.п.) позволяют получать результаты, которые могут быть положены в основу при проектировании в качестве априорных данных о состоянии сети, а исследование моделей сети позволит получить прогнозные данные, используемые в дальнейшем в системах контроля.

В подготовке материала книги приняли участие канд. техн. наук А. Р. Бестугин, научный сотрудник В. Ф. Богданова, д-р техн. наук Г. В. Стогов.

Авторы признательны д-ру техн. наук В. Н. Красюку, взявшему на себя труд рецензирования рукописи. Его рекомендации учтены при доработке содержания книги.

Книга предназначена для специалистов, связанных с проектированием и эксплуатацией телекоммуникационных сетей. Она будет полезна студентам и аспирантам, специализирующимся в области средств и систем передачи данных.

ГЛАВА 1
КОНТРОЛЬ И ДИАГНОСТИРОВАНИЕ СЕТИ СВЯЗИ

1.1. Задачи систем контроля и управления

Широкое развитие систем спутниковой связи определяет одновременную работу большого числа земных станций и обслуживающих областей, широкие возможности маневрирования в сетях связи в ситуациях, когда возникают перегрузки или аварии на отдельных участках сети. Кроме того, спутниковые сети каналов обладают низкой стоимостью, упрощением проблем маршрутизации и изменения структуры сети.

Таким образом, если в недавнем прошлом спутниковые системы связи обеспечивали передачу 10–15% общего объема информации, то сейчас они будут передавать до 70–80% [1].

Для обеспечения надежной работы спутниковых сетей связи организуются *центры управления сетью* (ЦУС) и *центры технической эксплуатации* (ЦТЭ).

Функционирование этих центров невозможно без процессов измерения, сбора и обработки контрольной информации.

ЦУС обеспечивают оперативное управление средствами и потоками сообщений в условиях изменяющейся ситуации с целью удовлетворения требований по качеству обслуживания потоков информации и достижения максимальной пропускной способности сети. ЦТЭ увеличивают бесперебойное функционирование сети, осуществляют технический контроль и диагностирование отказов элементов сети. ЦУС работают в тесной взаимосвязи с ЦТЭ, используя единую систему контроля элементов сети и сбора служебной информации.

В процессе функционирования элементы сети контролируются и управляются системой оперативного управления сетью, одним из составных элементов которой является ЦУС. Во время же профилактических работ или при отказах они поступают в ведение системы технического обслуживания, в состав которой входит ЦТЭ.

В существующих сетях связи можно выделить следующие методы технического обслуживания: профилактический, восстановительный и статистический [4]. Каждый из них имеет определенные преимущества и недостатки перед другими, поэтому используются различные сочетания методов. Однако в связи с повышением надежности все большее предпочтение в современных сетях связи получает *статистический* метод обслуживания, суть которого состоит в том, что ремонтно-восстановительные работы начинаются после того, как качество функционирования достигло критического значения. Элементы сети подвергаются техническому диагностированию для получения информации о состоянии элементов сети. При появлении признаков снижения качества функционирования они, как правило, выводятся из рабочей конфигурации для восстановления работоспособности. Данная методика позволяет

исключить многие виды дефектов, которые обычно возникают при профилактическом обслуживании в связи с демонтажем и другими работами, а в сети и ее элементах допустимо некоторое число неисправностей, не приводящих к прекращению правильного функционирования благодаря наличию видов избыточности.

Целесообразность применения статистического метода технического обслуживания в сетях определяется в основном двумя факторами: развитой системой контроля и диагностирования и использованием в элементах сети высоконадежной элементной базы.

Функционирование сети связи происходит в условиях постоянного воздействия различного рода возмущений, что приводит к выходу из строя УК и каналов связи, возникновению ошибок в передаваемых сообщениях, к случайному характеру циркулирующих потоков информации. В этих условиях задача контроля и управления сетью заключается в обеспечении передачи максимального количества информации с требуемым качеством. Качество связи практически полностью определяют три важнейшие свойства систем связи – точность, надежность и верность доставки информации [5]. Несмотря на то, что они имеют различную физическую природу, эти свойства могут быть объединены в рамках одной модели благодаря тому, что при каких бы условиях не проводилась доставка информации и какие бы требования к ней не предъявлялись, каждая реализация процесса доставки информации может быть описана практически исчерпывающим образом продолжительностью данного процесса и его структурой (соотношением между временем собственно передачи информации и временем непроизводительных затрат). Так под качеством передаваемых сообщений за определенный интервал времени с требуемым качеством понимается *производительность* сети, а под максимально возможной производительностью – *пропускная способность* сети. Она зависит как от структуры сети, интенсивности потоков сообщений, требований к качеству их обслуживания, так и в значительной степени от эффективности контроля и управления сетью [1]. Поскольку пропускная способность сети зависит от контроля и управления сетью, то следует различать потенциальную и реализованную пропускную способности.

Потенциальная пропускная способность определяется в предположении идеальной системы контроля и управления, а *реализованная* – для реальной системы, требующей накладных расходов. Последняя не позволяет учесть все многообразие факторов воздействия на сеть.

При отсутствии контроля и с увеличением нагрузки пропускная способность резко уменьшается, особенно в условиях нестационарного характера нагрузки на сеть. Чем более совершенна система контроля и управления сетью, тем ближе реализованная пропускная способность к потенциальной.

В процессе функционирования сети связи необходимо обеспечивать требуемое *качество соединений. Контроль соответствия количественных*

параметров заключается как в непосредственной оценке критерия правильного функционирования, так и по результатам функционального диагностирования нижеописанных уровней доставки [6]. Критерием правильного функционирования для любого режима доставки является время безошибочной доставки сообщения.

Контроль времени доставки сообщения производится от момента времени, когда первый знак вводится впервые в сеть отправителя, до момента выдачи получателю последнего знака корректного сообщения.

Контроль безошибочности включает проверку корректности формата, проверку отсутствия в принимаемых данных искажений, вставок, выпадений знаков или группы знаков, проверку отсутствия потерь, размножений и засылок не по адресу.

Несоответствие заранее заданным количественным параметрам соединения (наличие нарушения) вызывает переход либо к процедурам управления структурой (*восстановление правильного функционирования*), либо к завершению данного функционирования (*частичный отказ*) с последующими переходами к процедурам технического обслуживания (*восстановление работоспособности*).

Управление структурой заключается в переходе от пораженной и/или поврежденной структуры, не обеспечивающей текущих требований к качеству соединения, к структуре, отвечающей этим требованиям, в предположении того, что такой переход, по сути, является либо заменой отказавшего оборудования на работоспособное, либо перераспределением имеющихся ресурсов: уменьшением состава выполняемых функций, реконфигурацией первичной и/или вторичной сети.

Реконфигурация первичной сети сводится к удалению или добавлению новых линий связи (метод замены линий). Реконфигурация вторичной сети сводится либо к предоставлению дополнительных ресурсов (при установлении факта такого старения сообщения, когда изменение приоритета в обслуживании не обеспечивает своевременную доставку), либо к выбору нового маршрута передачи, т.е. наилучшего в данный момент пути $П_{ij}$ из совокупности путей W_{ij} в соответствии со следующими принципами:

1) выбирать наиболее короткий маршрут по числу приемов и физической длине

$$l_{\mu_{ij}} = \min_{П_{ij} \in W_{ij}} \left(l_{П_{ij}} \right),$$

но не более $l_{доп}$;

2) выбирать маршрут с максимально свободными ресурсами

$$C_{\mu_{ij}} = \max_{П_{ij} \in W_{ij}} \left(C_{П_{ij}} \right) = \max \left[\min_{\vartheta \in П_{ij}} \left(C_{\vartheta} \right) \geq C_{S_{ij}}^{P} \right].$$

При выборе маршрута предлагается руководствоваться следующими ограничениями:

1) не использовать обходные маршруты в цепях, непосредственно соединенных с узлом адресатом, при работоспособной линии связи между ними;

2) не использовать маршрут, возвращающий сообщение в узел, из которого оно поступило (явление “кинг-конг”).

При возможности выбора маршрута, удовлетворяющего необходимым требованиям, производится переход к процедурам завершения с последующим переходом к процедурам технического обслуживания.

1.2. Контроль и диагностирование. Общетехнические понятия

Контроль технического состояния, в соответствии с определением, рекомендованным государственным стандартом, – это определение вида технического состояния (ТС) объекта.

Под *техническим состоянием* понимается совокупность подверженных изменению в процессе производства или эксплуатации свойств объекта, характеризуемых в определенный момент времени признаками, установленными технической документацией на этот объект [7].

Видом ТС является такая категория ТС, которая характеризуется соответствием или несоответствием качества объекта определенными технической документацией на этот объект.

Зависимости между входными, выходными и внутренними переменными объекта, записываемыми функционально (операторами и т.п.), можно поставить в соответствие фазовое пространство технических состояний, присущих данному объекту. Причем каждому попарно-различному сочетанию значений указанных переменных соответствует определенная точка этого пространства. Все множество точек фазового пространства установленной функцией качества можно разбить на два или более подмножеств.

В зависимости от применяемых критериев качества могут быть рассмотрены следующие подмножества точек пространства:

- подмножество точек, составляющее состояние исправности ($\Omega_и$);
- работоспособности ($\Omega_р$);
- функционирования ($\Omega_ф$);
- неисправности ($\Omega_{ни}$);
- неработоспособности ($\Omega_{нр}$);
- нефункционирования ($\Omega_{нф}$).

В соответствии с принятым способом разбиения точек фазового пространства на подмножества можно определить следующие, различающие эти подмножества, процессы:

- контроль ТС (различение $\Omega_и$, $\Omega_р$, $\Omega_ф$, $\Omega_{ни}$, $\Omega_{нр}$, $\Omega_{нф}$);

– контроль неисправности (различение $\Omega_{и}$, $\Omega_{ни}$);
– контроль работоспособности (различение $\Omega_{р}$, $\Omega_{нр}$);
– контроль функционирования (различение $\Omega_{ф}$, $\Omega_{нф}$).

Техническое диагностирование – это процесс определения объекта с определенной точностью [2]. Техническое диагностирование может быть:

– законченным самостоятельным процессом при исследовании объекта с неустановленными заранее значениями показателей его исправности, работоспособности или правильного функционирования, а также при поиске дефектов;

– частным процессом при контроле ТС или при прогнозировании ТС объекта.

Конечным этапом диагностирования является получение технического диагноза. Поскольку для контроля исправности, работоспособности или правильного функционирования объекта необходимо знание его фактического ТС. Контроль ТС всегда предполагает техническое диагностирование.

Ввиду того что определение процессов контроля и диагностирования базируется на понятии технического состояния, рассмотрим понятие ТС применительно к сетям связи.

1.3. Техническое состояние сети связи

Для определения видов технического состояния сети связи исходя из общетехнических понятий технических состояний объекта вводятся вторичные понятия [3].

Для доставки информации в задаваемых пользователями режимах необходимо иметь определенный ресурс сети. *Ресурсы сети* – это совокупность физических и логических средств, необходимых для выполнения функций.

В целях сравнения однотипных ресурсов вводится понятие “единичный ресурс”. *Единичный ресурс* – объем ресурсов, определенный минимальным объемом выполняемой функции для данной системы (например, единичным объемом памяти, единичной пропускной способностью).

Единичное соединение $S_{ij}(i, j = 1, 2, \ldots, N, i \neq j$, где N – число пользователей) – последовательная совокупность единичных ресурсов сети, способная реализовать все функции процесса доставки данных от i-го к j-му пользователю с заданными показателями назначения для p-го регламентированного режима, $p = \overline{1, P}$, где P – количество регламентных режимов. Соединение характеризуется следующими параметрами:

а) длиной

$$l\left(S_{ij}^{P}\right) = \sum_{q-1}^{k} \left(\vartheta_{q}\right) \in S_{ij}^{P} ,$$

где $l(\vartheta)$ – длина линии связи между переприемными узлами; k – ранг соединения (число линий связи, входящих в данное соединение; тогда единичное соединение – это соединение ранга 1);

б) временем существования

$$t\left(S_{ij}^{P}\right)=t_{\text{уст}}\left(S_{ij}^{P}\right)+t_{\text{сохр}}\left(S_{ij}^{P}\right)+t_{\text{зав}}\left(S_{ij}^{P}\right);$$

в) пропускной способностью $\mu\left(S_{ij}^{P}\right)$;

г) приоритетностью обслуживания;

д) возможностью вещания (многоадресностью);

е) дискретностью (прерывистостью) ввода;

ж) надежностью – возможностью реализации и возможностью восстановления соединения.

Вероятность реализации соединения есть вероятность того, что ресурсы, выделенные для этого соединения в заданном интервале времени, способны безошибочно доставлять данные в p-м режиме с пропускной способностью $\mu\left(S_{ij}^{P}\right)$, т.е.

$$Q\left(S_{ij}^{P}\right)=Q_{\mu}^{P}Q_{B}^{P},$$

где Q_{μ}^{P} – вероятность сохранения значений пропускной способности соединения за время $t\left(S_{ij}^{P}\right)$; Q_{B}^{P} – вероятность безошибочной доставки в p-м режиме.

Безошибочность является сложным свойством и характеризуется возможностью появления:

- искажения символов;
- вставок и выпадений (потерь) символа или группы символов;
- засылок не по адресу символа, группы символов или сообщения.

Вероятность восстановления соединения есть вероятность того, что через время $t \le t_{B}^{p}$ пропускная способность соединения вновь достигает значения $\mu\left(S_{ij}^{P}\right)$ и/или исчезнут ошибки в доставляемых данных.

Пусть Π_{ij} (или совокупность соединений) – общая минимальная совокупность ресурсов сети, позволяющая организовать несколько соединений в любом необходимом сочетании их видов между i-м и j-м корреспондирующими пользователями, т.е.

$$\Pi_{ij}=\bigcup_{p=1}^{P}S_{ij}^{p}.$$

Заметим, что путь характеризуется теми же параметрами, что и соединение. В силу надежности ресурсов сети, а также в силу необходимости обеспечения заданной вероятности своевременной доставки i-м и j-м пользователям кроме основного предусматривают резервные пути, составляющие вместе совокупность путей. Совокупность путей W_{ij} –

совокупность всех существующих или возможных путей (с учетом перекроссировки) между *i*-м и *j*-м корреспондирующими пользователями

$$W_{ij} = \bigcup_{q=1}^{m} \Pi_{ij}^{q},$$

где m – число путей.

W_{ij} характеризуется числом путей и вероятностью существования хотя бы одного работоспособного пути между *i*-м и *j*-м пользователями.

Минимальное число независимых путей между двумя любыми пользователями называется *связностью сети*, т.е.

$$n = \min_{i,\, j \in l} W_{ij}.$$

Исправность совокупности W_{ij} характеризуется наличием всех работоспособных Π_{ij}. Отказ любого Π_{ij} приводит к переходу в состояние неисправности. *Работоспособность* W_{ij} характеризуется наличием хотя бы одного работоспособного Π_{ij}, при частичном отказе которого W_{ij} переходит в неработоспособное состояние, а при полном – в предельное состояние. *Работоспособность* Π_{ij} характеризуется наличием всех работоспособных соединений. Отказ любого соединения приводит к переходу Π_{ij} в состояние неработоспособности. Отказ последнего из существующих соединений пути (полный отказ пути) приводит к переходу Π_{ij} в предельное состояние.

Работоспособность S_{ij}^{p} характеризуется наличием всех единичных соединений, отказ любого из них приводит к переходу S_{ij}^{p} в состояние неработоспособности.

Исправность цифровой сети связи (*ЦСС*) – состояние сети, которое характеризуется наличием всех исправных совокупных путей.

Неисправность ЦСС – состояние сети, при котором отказано хотя бы одно соединение, а событие – повреждение сети.

Работоспособность ЦСС – состояние сети, которое характеризуется наличием между всеми парами корреспондирующих пользователей работоспособных совокупностей путей.

Неработоспособность ЦСС – состояние сети, при котором отказывает хотя бы одна из совокупностей путей, а событие – частичный отказ.

Состояние ЦСС, при котором отказывают все соединения, является *предельным*, а событие – *полный отказ* сети.

В отличие от рассмотренных технических состояний, которые характеризуют способность сети выполнять возложенные на нее задачи с заданным качеством и не зависят от наличия сообщений на входе сети, состояние правильного функционирования зависит от вида и количества режимов, установленных пользователями.

Правильное функционирование ЦСС – состояние сети, которое характеризуется наличием соединений между теми пользователями, которые корреспондируют в текущий момент времени.

Состояние сети, при котором между корреспондирующими в текущий момент времени пользователями отсутствует хотя бы одно работоспособное соединение, является состоянием *неправильного функционирования сети.*

Располагая строгим определением видов технических состояний сети и используя хорошо развитый математический аппарат описания состояния сети вероятностными графами, можно достаточно просто вычислить вероятность нахождения сети в одном из технических состояний. При проектировании сети связи разработчику необходимо учитывать состояние сети как с позиций обслуживающего персонала, так и с позиций пользователей.

Дадим определение видов технических состояний сети с позиций *i*-го пользователя, корреспондирующего с *j*-м пользователем ($i, j = 1, 2, \ldots, N, i \neq j$, где N – число пользователей).

Исправность сети – состояние сети, которое характеризуется наличием всех m возможных путей Π_{ij} между *i*-м и *j*-м пользователями, т.е. совокупностей путей

$$W_{ij} = \bigcup_{q=1}^{m} \Pi_{ij}^{q}.$$

Неисправность сети – состояние сети, при котором неработоспособно хотя бы одно соединение S_{ij}^{p} ($p = 1, 2, \ldots, P$, где P – число регламентированных режимов доставки сообщений).

Повреждение сети – состояние, приводящее к переходу из исправного состояния в неисправное.

Работоспособность сети – состояние, характеризующееся наличием хотя бы одного пути Π_{ij} между *i*-м и *j*-м пользователями.

Неработоспособность сети – состояние, характеризующееся отказом последнего Π_{ij}^{q} из всех q ($q = 1, 2, \ldots, m$) возможных путей между *i*-м и *j*-м пользователями.

Событие, приводящее к переходу из работоспособного состояния в неработоспособное, называется отказом сети.

Состояние правильного функционирования по установленному соединению – состояние сети, при котором безошибочно и своевременно обеспечивается доставка данных в определенном режиме p ($p = 1, 2, \ldots, P$) и в данный момент времени t.

Состояние неисправного функционирования – состояние сети, при котором не обеспечивается в данный момент времени безошибочная и/или своевременная доставка данных определенного режима p.

Событие, приводящее к переходу из состояния правильного функционирования в состояние неправильного функционирования, будем называть *нарушением сети*.

Поскольку для любого пользователя, как уже отмечалось, наибольший интерес представляет нахождение сети в работоспособном состоянии, что является необходимым и достаточным условием для обеспечения

правильного функционирования во всех режимах, то будем анализировать вид технического состояния сети с этих позиций, т.е. классифицировать состояние сети на два класса (подмножества) состояний: работоспособность или неработоспособность сети.

1.4. Диагностирование

С целью поиска места и причины отказа в сложных системах применяются методы технической диагностики, которые позволяют локализовать отказы с точностью до отдельного блока, а в некоторых случаях – до элемента платы.

К основным этапам процесса относится создание концептуальной и математической модели объекта, что позволяет в дальнейшем производить сравнение фактического состояния с предполагаемым в соответствии с выбранными решающими правилами; анализ развития контролируемых параметров и создание диагностической модели; измерение и оценка диагностических признаков, определение решающих правил; обучение распознающего устройства; экзамены распознающего устройства.

Пусть задано множество $C = \{c_i\}_{i=0}^{m}$ различных классов состояний системы и c_i (i ≥ 1, m) – классы неработоспособных состояний системы. При этом устройство распознавания производит распознавание отказа, т.е. одного из классов неработоспособных состояний системы, характеризующихся системой признаков. Распознавание состояний системы ввиду случайного отказа носит вероятностный характер. Техническое состояние рассматривается как совокупность свойств объекта, характеризующихся в определенный момент времени определенными признаками.

Каждое статистическое состояние (ТС) системы

$$Z = (Z_1, Z_2, \ldots, Z_n)^{\mathrm{T}}$$

является некоторой функцией работоспособности в пространстве параметров состояний (относящихся к классу состояний c_i):

$$Z \in c_i \leftrightarrow \bigcap_{j=1}^{i} \left(y_i \in \left| y_{ij\text{н}} y_{ij\text{в}} \right| \right).$$

Информацию о текущем состоянии получают путем измерения выходных сигналов y_i в выбранных контрольных точках.

Существуют различные способы отбора признаков диагностирования: по стоимости затрат на диагностирование, по объему информации о состоянии объекта, по величине корреляции признака и отказа.

В качестве одного из подходов к проблеме выбора состава контролируемых параметров может рассматриваться многомерная задача оптимизации совокупности контролируемых параметров системы. В качестве целевой функции при оптимизации может использоваться критерий средневзвешенной достоверности контроля, учитывающий показатели

весомости отказов отдельных блоков и узлов контролируемой системы, определяемые экспертным путем, надежностные характеристики, а также структура самой системы. Ограничениями являются затраты ресурсов, необходимых для осуществления контроля [13]. Такая оптимизационная задача может решаться комбинированным методом. Вначале нужно определить базовый набор параметров рекуррентным методом, а затем методом ветвей и границ – эффективное улучшение базового набора.

Следующим этапом является выбор наиболее приемлемого метода диагностирования, реализующего данный способ оценки состояния. Для этой цели могут быть использованы алгоритмы распознавания, которые позволяют по известным распределениям состояний и признаков состояний принимать решение о диагнозе.

Основная задача диагностики – определение класса состояний c_i в текущий момент времени – осуществляется по решающему правилу

$$z \in c_i \leftrightarrow \rho(y, y_i) = \max_{k=1,m} \rho(y, \tilde{y}_k),$$

где $\rho(y, \tilde{y}_k)$ – мера сходства сравниваемых секторов.

Достоверность процесса диагностирования отказов оценивается средней вероятностью правильного диагноза

$$P_{\text{пр.д.}} = 1 + \{a_{\text{д}} P_0(T) + \beta_{\text{д}} [1 - P_0(T)]\},$$

где $P_0(T)$ – вероятность отсутствия отказа за период диагностирования T; $a_{\text{д}}$ – вероятность ложной тревоги; $\beta_{\text{д}}$ – вероятность пропуска отказа.

Литература

1. **Аринов М. Н., Присяжнюк С. П., Шарифов Р. А.** Контроль и управление в сетях передачи данных с коммутацией пакетов. – Ташкент: Фан, 1988. – 160 с.
2. **ГОСТ 20911–75.** Техническая диагностика. Основные термины и определения.
3. **Захаренко Г. П.** Эксплуатация цифровых сетей связи: Учебное пособие / Министерство промышленности средств связи. Часть II.– М.: Ин-т повышения квалификации руководящих работников и специалистов, 1986. – 40 с.
4. **Агаян А. А., Захаренко Г. П.** Оптимизация структур цифровых сетей связи и технического обслуживания. — М.: Ин-т повышения квалификации руководящих работников и специалистов, 1987. – 39 с.
5. **Губин Н. М., Матлин Г. М.** Качество связи. Теория и практика. – М.: Радио и связь, 1986. – 272 с.

6. **Захаренко Г. П., Иванов В. К.** Эксплуатация цифровых сетей связи. Часть II. Основные задачи, понятия, определения. – М.: Ин-т повышения квалификации руководящих работников и специалистов, 1986. – 40 с.
7. **ГОСТ 19919–74.** Контроль автоматизированный технических состояний изделий авиационной техники. Основные понятия и определения.

ГЛАВА 2
ОБЩИЙ ПОДХОД К ПОСТРОЕНИЮ МОДЕЛИ СЕТИ СВЯЗИ

2.1. Концептуальная модель сети

Разработка комплексной модели сети с точки зрения системного подхода является очень сложной и актуальной проблемой. Комплексная модель сети необходима не только для определения оптимальной модели структуры сети связи, оптимального комплекса технических средств и алгоритмов функционирования, но должна также учитывать надежностные характеристики, управление сетью и систему технического обслуживания сети связи в целом, неотъемлемой частью которой является система контроля технического состояния сети связи.

Создание цифровой сети интегрального обслуживания (ЦСИО) требует немалых затрат, складывающихся из затрат на ее проектирование, реализацию и эксплуатацию. Можно сказать, что проблема стоимости представляет собой компромисс между стоимостью создания сети и стоимостью ее эксплуатации. Однако экономия средств на проектирование сети может привести к серьезным негативным последствиям во время ее работы. Так, при проектировании пакетной радиосети DARPA [28] дополнительные финансовые и временные затраты на обеспечение дистанционного диагностирования и технического обслуживания самоокупаются в несколько раз.

Процесс моделирования сети в зависимости от имеющейся априорной информации можно разделить на три этапа: концептуальное моделирование, логическая модель, математическая модель.

На этапе *концептуального моделирования* обычно производится описание модели. При этом учитывается уже имеющиеся модели сети, а также цели создания сети, требования к характеристикам, учет ограничений и выбор критериев оценки эффективности системы, позволяющей сравнивать различные модели сети.

Следующим этапом в рамках концептуальной модели является формализованное описание, позволяющее выделить структурное (морфологическое) описание сети и на его основе провести декомпозицию системы на ряд более простых функциональных модулей, процессов, блоков. Таким образом, в результате концептуального моделирования может быть получен ряд более простых моделей, по отношению к которым также может быть применена вышеизложенная методика или получено математическое описание модели.

Применение методологии общей теории систем позволяет рассматривать задачи синтеза и анализа открытых информационных сетей как части всего жизненного цикла, включающего этапы проектирования, внедрения и эксплуатации [25]. На рис. 2.1 представлена структурная схема жизненного цикла.

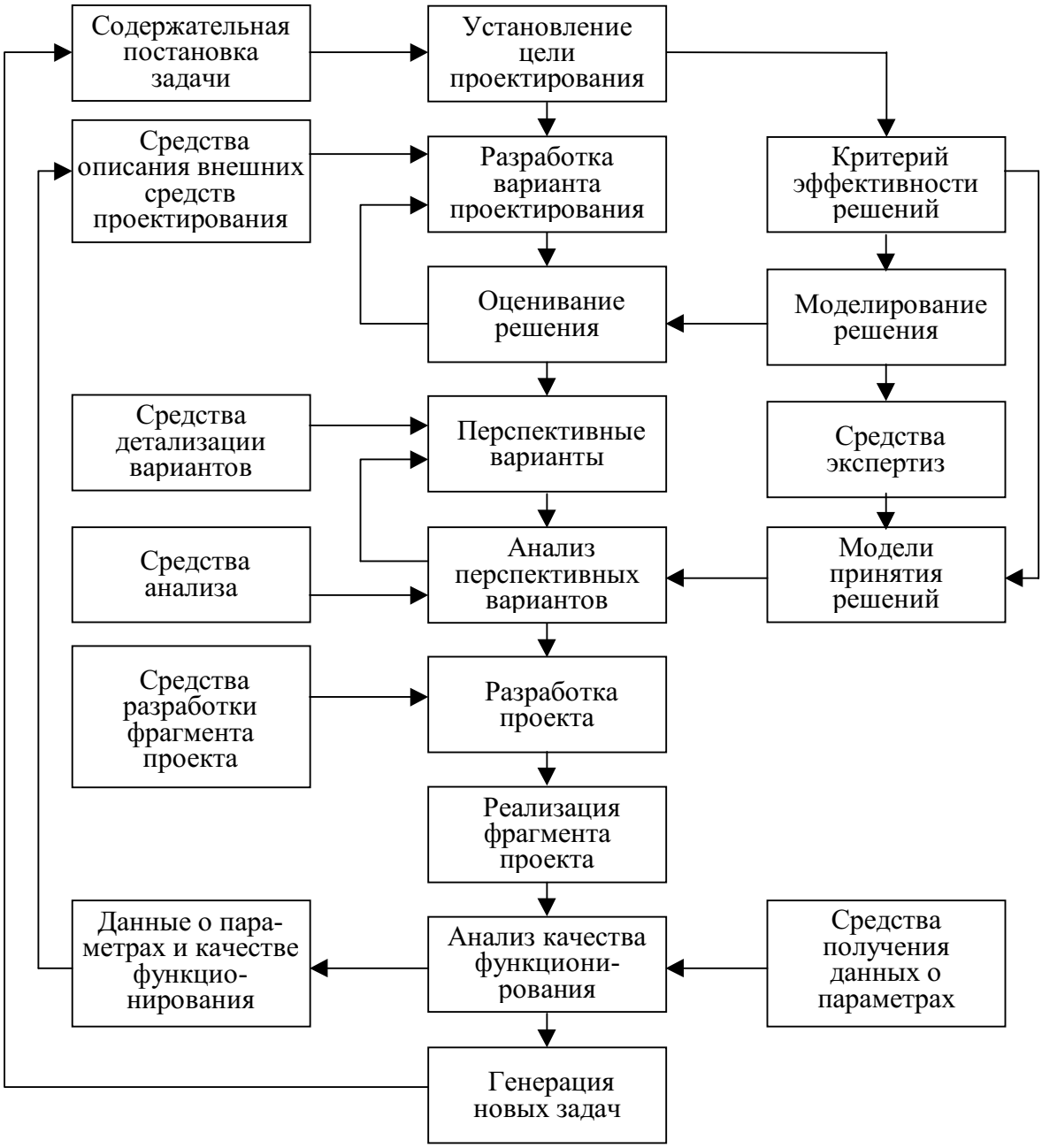


Рис. 2.1. Структурная схема жизненного цикла

Говоря о ЦСИО как о системе, прежде всего, имеют в виду ее структуру и архитектуру. Архитектура ЦСИО описывает ее внутреннее строение, алгоритмы работы, структуру и состав процедур доступа, обмена и управления, а структура характеризует внешнее строение ЦСИО, в частности, географическое размещение набора аппаратных средств и конфигурацию связей между ними.

Под *структурой* понимают вещественную основу организации сети (элементы и связи между ними). Структура ЦСИО является статистической характеристикой сети, так как не отражает способа доставки информации. Структура сети может рассматриваться в различных аспектах, отличающихся

степенью детализации данных и целевой направленностью задачи – абстрактном, географическом и физическом. Модель структуры ЦСИО – граф.

Структура (топология) сети с пакетной коммутацией зависит от большого числа переменных проектирования, а именно: расположения источников и получателей объема нагрузки, передаваемой между ними, требований к значениям задержек в сети и надежности, стоимости каналов и узлов и т.д. Поэтому разработка структуры сети является процессом оптимизации, в котором некоторые переменные или их функции принимаются как целевые, а остальные – как ограничения, и ставится задача определения таких параметров сети, как местонахождение узлов, трассы каналов между узлами, емкости каналов и потоки в каждом из них.

Лицо, принимающее решение (ЛПР) в части структурной и архитектурной организации перспективных ЦСИО, на первоначальных этапах проектирования интересует не единичное точное решение, заключающееся в поиске графа сети, а целый комплекс вопросов, из которых основными являются следующие:

– проверка технического задания (ТЗ) на непротиворечивость, оценка степени выполнимости ТЗ;

– обоснование выбора критериев оптимальности, систем ограничений и отдельных условий;

– оценка предельно допустимых значений надежностных, вероятностно-временных и стоимостных характеристик проектируемой ЦСИО;

– определение степени иерархичности, характера разветвленности, связности и других интегральных топологических характеристик;

– выбор наиболее предпочтительной концепции построения ЦСИО;

– расчет оптимального типажа технических средств (ТСС), в том числе технического обслуживания (ТО) и управления потоками, или определение требований к ним по надежности, быстродействию и т.п.

– исследование наиболее общих свойств, предлагаемого проекта сети, в частности, устойчивости к входным условиям и чувствительности интегральных показателей по отношению к внутренним параметрам;

– определение оптимальной этапности внедрения сети;

– выявление “узких” по тому или иному показателю звеньев сети и выработка предложений по их расширению.

ЦСИО является многофункциональной системой, в которой выделяются главная системовыделяющая функция (доставка информации) и набор составляющих ее подфункций [23, 24, 26]. К последним относятся: функции коммутации, маршрутизации, повышения достоверности, обеспечения надежности, устранения отклонений фактического состояния элементов от расчетного.

Поскольку в реальной сети эти процессы протекают параллельно и взаимосвязанно, ЦСИО следует рассматривать как некоторую кибернетическую систему, состоящую из управляемой и управляющей подсистем. Управляемой является подсистема доставки, параметры которой (пропускная способность,

верность, надежность) изменяются во времени. Управляющей системой в кибернетическом смысле является совокупность датчиков, средств обработки информации, контроля и регулирования работы управляемой подсистемы. Обе подсистемы связаны обратной связью (каналами “служебной” связи). Управление сетью определяют как реакцию на изменения характера входящей на обслуживание нагрузки и структуры сети, вызванные отказами (повреждениями) элементов, перегрузками и т.д.

Система управления сетью в соответствии с функциями может быть разделена на системы управления структурой сети, управления нагрузкой и управления потоками нагрузки.

Целью управления структурой сети при отказах (повреждениях) является обеспечение требуемого качества функционирования сети при неизменной внешней нагрузке путем изменения структуры перераспределения существующих средств связи и/или вводом резервных средств связи. При отсутствии функциональной и структурной избыточностей управление структурой сводится к вводу ее резервных средств.

Управление внешней нагрузкой заключается в поддержании уровня нагрузки по результатам ее контроля в пределах допустимых значений. Методом управления нагрузкой являются ограничение передачи информации по обходным путям и ограничение входной нагрузки.

Управление потоками нагрузки обеспечивает требуемое качество функционирования сети с учетом надежности элементов и локальных перегрузок. На основе контроля потоков нагрузки по заданной структуре сети и входящей нагрузке вырабатывается план распределения потоков нагрузки в сети, оптимальный с точки зрения выбранного критерия.

Таким образом, ЦСИО можно рассматривать [26] как совокупность управляемого объекта (подсистема “Доставка”), реализующего целевую функцию ЦСИО с требуемыми показателями назначения и управляющего объекта (подсистема “Эксплуатация”), обеспечивающего требуемые показатели надежности и реализующего функцию управления ЦСИО.

Рассматриваемые функции (целевая функция и функция управления) являются сложными составными функциям, которые могут быть декомпозированы.

В гл. 1 показана декомпозиция системы, которая представлена в виде многоуровневой архитектуры в соответствии с рекомендациями МККТТ серии “Эталонная модель взаимодействия открытых систем”.

Методы декомпозиции позволяют осуществлять последовательное расчленение системы на части, в свою очередь расчленяющиеся на составляющие их части. После чего может быть получено *математическое описание модели*.

Функционирование систем и сетей связи определяется как переход из одного состояния в другое, поэтому при математическом описании модели используются три метода математического моделирования: информационный, цепи Маркова и метод фазового пространства [29].

При использовании *информационного метода* на основании анализа информации контроля, как средства взаимосвязи объекта и субъекта, делается вывод о ценности указанной информации для субъекта как меры неопределенности (энтропии) объекта, величина которой возрастает с увеличением состояний системы. Задача контроля функционирования систем и сетей связи может быть представлена как задача процесса уменьшения неопределенности сведений о состоянии системы в требуемый момент времени. Вводя в рассмотрение меру априорных знаний о состоянии системы – среднюю априорную неопределенность и меру средней апостериорной неопределенности сведений о состоянии системы после контроля, можно определить среднее количество контролируемой информации между указанными величинами. Априорная неопределенность состояния сети связи в любой момент времени контроля определяется вероятностными свойствами этого состояния – законом распределения априорных вероятностей различных состояний. Неопределенность знаний о состоянии системы после контроля характеризуется апостериорными вероятностями, которые рассчитываются по формуле Байеса. Таким образом, находится мера неопределенности искомого состояния системы в момент времени контроля. В данной постановке задачи необходимо найти взаимосвязь апостериорных вероятностей с контролируемыми характеристиками объекта контроля.

При использовании *метода цепи Маркова* переходы между различными состояниями системы описываются как марковский процесс. Предполагая, что в любой момент времени контроля система находится в одном из состояний, процесс контроля функционирования представляется в виде вероятностной схемы известными двумя способами: 1) составлением матрицы вероятностей перехода; 2) составлением диаграммы переходов или графа вероятностей перехода системы из одного состояния в другое.

Применительно к определению вида ТС систем и сетей связи может быть использован *метод фазового пространства*. В этом случае состояние системы характеризуется векторами контролируемых величин и задающих воздействий, а процесс контроля функционирования определяется как процесс восприятия изменений управляемых величин, сбора, обработки, хранения и отображения информации о равносильности указанных векторов с целью принятия решения по выработке управляющих воздействий. Применение этого метода требует решить следующие проблемы:

– оценить точность отображения явления функционирования системы набором показателей;

– обосновать выбор признаков классификации показателей контроля;

– доказать необходимость и достаточность числа уровней иерархии;

– проверить их по критерию “существенности”;

– сформировать из полученной системы признаков системы показателей функционирования, для чего необходимо установить, “что, где, когда, как, в каком количестве и в каком объеме” необходимо контролировать;

– определить параметры, тесноту взаимосвязей и значимость показателей.

Считается, что математическая модель построена, если оформлен набор ограничений и выбраны целевые функции. Для определения характеристик математических моделей необходимо осуществить анализ параметров. Описание любой системы и условий ее функционирования характеризуется определенной совокупностью параметров, причем на разных этапах анализа и оптимизации требуются различные способы описания. Основой классификации являются группы параметров (параметрические базисы) [24]. Для произвольной системы выделяются базисы внешних и внутренних параметров. Внешние параметры, в свою очередь, разбиваются на два класса – входные и выходные.

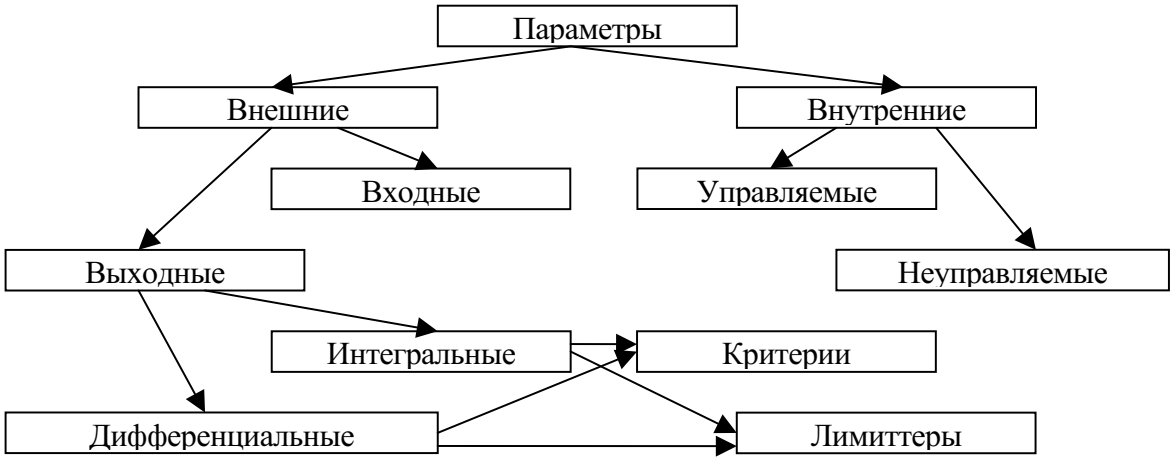


Рис. 2.2. Классификация сетевых параметров

В зависимости от степени комплексности и степени детализации выходные параметры подразделяются на интегральные и дифференциальные. Кроме этого, в зависимости от цели операции среди выходных могут быть выделены: переменные – критерии, максимизируемые или минимизируемые в процессе оптимизации, переменные – лимиттеры, на которые накладываются ограничения (рис. 2.2).

Внутренние переменные применительно к задаче оптимизации разбиваются на две группы: управляемые и неуправляемые. Первые – модельные параметры, непосредственно влияя на которые, алгоритм осуществляет оптимизацию, вторые – различные производные от управляемых, которые могут быть как контролируемыми, так и неконтролируемыми. На переменные этого базиса также могут накладываться ограничения (табл. 2.1).

Известны и более подробные системы классификации, различающие функциональные и стоимостные, доминирующие и недоминирующие, качественные и количественные, повышающие и понижающие, скалярные и векторные, непосредственные и системные.

Однако изменение постановки задачи сопровождается подчас переводом параметра в другой параметрический базис, и излишне подробная детализация свойств может мало что дать при построении гибких программных средств САПР многоцелевого назначения.

Таблица 2.1

Наиболее распространенные представители параметрических базисов используемой системы классификации применительно к подсистеме доставки сети связи

Параметр	Категория параметра
Число абонентов	Входной
Общественные затраты	Выходной, интегральный, минимизируемый
Среднесетевая задержка	Выходной, интегральный, лимиттер
Суммарная протяженность сети	Выходной, интегральный, неконтролируемый
Задержка сообщения в канале	Выходной, дифференциальный, неуправляемый
Сквозная задержка сообщения на маршруте	Выходной, дифференциальный, лимиттер
Пропускная способность канала	Внутренний, управляемый
Средняя длина маршрута	Внутренний, неуправляемый

Успех проектирования, внедрения и эксплуатации ЦСИО зависит не только от выбранных моделей функционирования, используемого математического аппарата, но и от выбранных критериев оценки эффективности системы. В качестве модели оценки эффективности воспользуемся моделью [1], включающей как систему, т.е. ЦСИО, так и метасистему, т.е. пользователей ЦСИО (уровни 5–7 ЭМВОС). При этом используемые критерии должны зависеть от системы привязки к реальным процессам, которые имеют место в ЦСИО. Кроме того, необходимо выделять взаимосвязанные процессы (подпроцессы) в едином процессе доставки информации в ЦСИО. Например, рассматривая с 1-го по 4-й уровни ЭМВОС, можно выделить процессы коммутации, маршрутизации и ограничения потоков. Обобщающим для всех перечисленных процессов является процесс доставки информации пользователям транспортной системы ЦСИО. В этом случае можно выделить следующую цепочку критериев:

1) функция ценности информации (для процесса доставки);
2) функция производительности сети (для процесса обмена информацией);
3) вероятностно-внешние характеристики (для процессов коммутации, маршрутизации и ограничения потоков).

Под производительностью ЦСИО понимают количество информации пользователей, содержащейся во всех сообщениях, обслуженных сетью полностью и с заданным качеством за единичный интервал времени ее функционирования.

Сообщение в режиме ВК (виртуального канала) считается обслуженным полностью и с заданным качеством, если установилось

соединение, обеспечивающее передачу всей необходимой пользователю информации с заданным качеством.

К вероятностно-временным характеристикам относится среднее время задержки $\overline{T}_c$ сообщения в сети как среднее время по всему множеству сообщений от момента первого бита сообщения от пользователя в УК-источник до передачи последнего бита сообщения из УК-получателя пользователю. $\overline{T}_c$ необходимо рассматривать отдельно для каждого вида информации, каждого приоритета и каждого реализованного режима коммутации.

Таблица 2.2

Критерии эффективности

Уровни модели ВОС				
Прикладной сеансовый представительный	Транспортный	Сетевой	Канальный	Физический
Защита Приоритет Темп остаточных ошибок Полоса пропускания Задержка передачи (для каждого направления) Оптимизация передачи Расширенное управление Задержка установления соединения Вероятность отказа от установленного соединения Вероятность ошибки передачи Задержка завершения соединения Вероятность ошибки завершения соединения Надежность соединения ПС	*С соединением* Защита Приоритет Фаза установления соединения: задержка установления; вероятность неустановления Фаза передачи данных: пропускная способность; транзитная задержка; КНО; надежность; вероятность отказа Фаза разъединения: задержка разъединения; вероятность неразъединения; *Без соединения* Транзитная задержка КНО Защита Приоритет ПС	*С соединением* Фаза передачи данных: пропускная способность; транзитная задержка; КНО; надежность; вероятность отказа; наибольшая сложность соединения Фаза становления соединения: задержка установления; вероятность установления Фаза разъединения: задержка разъединения; вероятность неразъединения *Без соединения* Транзитная задержка Защита ПС КНО Приоритет Возможность контроля нагрузок Вероятность сохранения последовательности Максимальное время существования сетевого сервисного блока данных	Пропускная способность Транзитная задержка Защита соединения КНО Надежность соединения ПС	Частота появления ошибок Доступность сервиса Скорость передачи Транзитная задержка ПС
Примечание. ПС – параметры стоимости; КНО – коэффициент необнаружения ошибок				

Кроме того, могут быть использованы такие оценки, как среднее по сети среднеквадратическое отклонение времени задержки сообщения для запроса на передачу информации определенного вида и приоритета при заданном режиме коммутации y_c или средняя дисперсия времени задержки $\overline{T}_c$. Величины y_c и являются важными показателями качества обслуживания, главным образом, для речи и оперативных данных.

В ряде случаев может быть использована такая оценка как верность передачи сообщения, в качестве меры которой может быть принято среднее число ошибок в кодовой комбинации или средняя вероятность искажения символа $\overline{D}_c$. Допустимые значения $\overline{D}_c$ могут находиться в весьма широких пределах для различных видов информации (от 10^{-12}–10^{-16} и менее для файлов и до 10^{-2} –10^{-1} для речи).

Существуют и другие виды критериев. В табл. 2.2 представлена многоуровневая система критериев эффективности интегральной сети [25], разработанная в соответствии с моделью взаимодействия открытых систем.

Синтез структуры сети связи

В прагматическом смысле интегральная цифровая сеть связи есть вторичная сеть связи, основная задача которой состоит в обеспечении обмена информацией между пользователями с заданным качеством.

Эта задача успешно решается лишь путем создания эффективной структуры системы доставки, системы эксплуатации (СЭ) и входящей в ее состав системы технического обслуживания (ТО).

Опыт эксплуатации зарубежных сетей [20] позволяет выделить ряд следующих принципиальных черт, характеризующих СЭ:

1) сети строятся как самоорганизующиеся и самовосстанавливающиеся, однако во всех сетях предусмотрена возможность вмешательства обслуживающего персонала;

2) обеспечивается большая степень автоматизации процессов диагностирования и изменения конфигурации сети и ее отдельных компонентов;

3) локальное диагностирование элементов сети позволяет обнаруживать отказы в момент их возникновения по назначению, а также при периодическом плановом диагностировании;

4) централизованное диагностирование предполагает наличие в сетях центров техобслуживания (ЦТО), выполняющих одновременно функции сбора и обработки статистических данных.

С позиции пользователя эксплуатация ЦТО (в широком смысле) – это процесс использования ресурсов сети в соответствии со своими потребностями в обмене данными, т.е. объектом эксплуатации в данном случае является сеть в целом. СЭ охватывает широкий круг вопросов и может декомпозироваться на подсистему общей эксплуатации (управление

состоянием внешней среды) и подсистему технического обслуживания (управление состоянием внутренней среды).

Воздействие внешней среды – входящий поток заявок; поток внешних воздействий; поток внешних поставок ЗИПа, КИПа, документации, материалов, энергии и т.д.; экологические и социальные воздействия.

Воздействием внутренней среды является поток отказов, вызванный несовершенством технологии изготовления, физической прочности (обрыв или короткое замыкание), конструктивными, алгоритмическими, программными, технологическими ошибками, ошибками обслуживающего персонала.

В основу системы ТО положена высокая надежность и автоматизация процессов восстановления работоспособности. Отказ отдельного устройства элемента сети, как правило, не оказывает значительного влияния на качество функционирования всей сети. Это объясняется различными видами избыточности, используемыми в ЦСИО.

Большие эксплутационные расходы, связанные с использованием большой численности обслуживающего персонала высокой квалификации могут быть снижены путем автоматизации процессов ТО и выбором его оптимальной системы. Очевидно, что полностью децентрализованная система ТО даже при высокой степени автоматизации процессов техобслуживания не будет оптимальной, так как требует присутствия технического персонала. С другой стороны, полностью децентрализованная система ТО также не решит поставленной задачи. Вот почему наряду с автоматизацией выдвигается проблема оптимальной структуры системы ТО, т.е. выбора такого количества ЦТО и такого их расположения, чтобы обеспечить минимум эксплутационных и капитальных вложений.

Синтез структуры сети связи с учетом системы ТО

Алгоритм поиска решения задачи оптимизации топологии сети основан на двух общих подходах: многократном построении решений и трансформации решений с целью улучшения некоторых начальных, заданных решений (см. гл. 1). Сначала задается некоторая исходная модель. Затем с помощью метода целенаправленного перебора структур исходная сеть оптимизируется путем включения или исключения отдельных ребер графа сети [1, 14]. На каждом этапе осуществляется расчет стоимостного критерия и ограничений, характеризующих различные показатели надежности, а также определяется направление траектории оптимизации. Структура полученного варианта сети зависит от структуры исходной сети, процедур изменения структуры и очередности их проведения [22].

Структура сети может быть задана географическим размещением своих элементов и связей между ними или получена специальным методом генерации решений, которые выполняются автоматически машинными алгоритмами поиска [20].

Задача совместной автоматизации структуры сети и системы ТО обеспечивает нахождение оптимума общей задачи, включая подсистемы доставки. Однако задача совместной оптимизации из-за большой размерности современных сетей представляет собой чрезвычайно сложную задачу, для которых нет методов поиска точного решения. Поэтому целесообразно рассмотреть задачу совместной оптимизации структуры сети и систем ТО для базовой системы передачи данных [21], которая достаточно просто может быть трансформирована для общей модели оптимизации иерархической сети.

Пусть заданы

1) координаты и УК

$$\{a_i\} = \{(x_i, y_i)\},\ i = 1, 2, \ldots, n;$$

2) матрица расстояний между узлами коммутации (УК)

$$\| l_{ij} \|,\ i, j = 1, 2, \ldots, n;$$

3) матрица трафика между УК

$$\| \lambda_{ij} \|,\ i, j = 1, 2, \ldots, n;$$

4) набор $P_{\text{л}}$ типов линий связи с соответствующими параметрами надежности и пропускной способности

$$\left\{k_{\text{г}i}^{\text{л}}, d_i^{\text{л}}, \mu_i^{\text{л}}\right\},\ i = 1, 2, \ldots, P_{\text{л}};$$

5) набор $P_{\text{у}}$ типов УК с характеристиками надежности и интенсивности обслуживания

$$\left\{k_{\text{г}i}^{\text{у}}, d_i^{\text{у}}, \mu_i^{\text{у}}\right\},\ i = 1, 2, \ldots, P_{\text{у}},$$

где $k_{\text{г}}$ – коэффициент готовности; d – интенсивность восстановления работоспособности; μ – интенсивность обслуживания;

6) набор $P_{\text{т}}$ типовых центров ТО (ЦТО) $g_i^{\text{т}}$, $i = 1, 2, \ldots, P_{\text{т}}$, различающихся составом средств тестового диагностирования и т.п.

Требуется определить:

1) топологию базовой сети A;
2) распределение потоков в линиях сети $\{\rho_{ij}\}$, $(i, j) \in A$;
3) распределение пропускных способностей линий

$$\left\{\tilde{\mu}_{ij}^{\text{л}}\right\},\ i, j \in A,\ \tilde{\mu}_{ij}^{\text{л}} \in \left\{\tilde{\mu}_i^{\text{л}}\right\};$$

4) производительность УК

$$\left\{\tilde{\mu}_j^{\text{у}}\right\},\ j = 1, 2, \ldots n,\ \tilde{\mu}_i^{\text{у}} \in \left\{\tilde{\mu}_i^{\text{у}}\right\};$$

5) число $n_{\text{у}}$ и места размещения ЦТО

$$b_1, b_2, \ldots, b_{n_{\text{у}}};\ \left\{b_j\right\} = \left\{x_j^{\text{т}}, y_j^{\text{т}}\right\};$$

6) типы ЦТО

$$\left\{\tilde{y}_j^{\text{т}}\right\},\ j=1,2,...,n_{\text{у}}\ ;\ \tilde{g}_j^{\text{т}}\in\left\{g_j^{\text{т}}\right\};$$

7) разбиение множеств УК на зоны ТО

$$\beta=\left\{B_1, B_2, \ldots, B_{n_{\text{у}}}\right\},$$

где $n_{\text{у}} - \bigcup_{i=1}^{n_{\text{у}}} B_i=\{1, 2, \ldots, n\};\ B_i\cap B_j\neq 0;\ i, j, = 1, 2, \ldots, n_{\text{у}},\ i\neq j$, с тем, чтобы минимизировать приведенные затраты на сеть в целом:

$$\text{П}=\text{П}_{\text{Б}}+\text{П}_{\text{ТО}},$$

где $\text{П}_{\text{Б}}$ – приведенные затраты на базовую сеть; $\text{П}_{\text{ТО}}$ – приведенные затраты на систему ТО.

Суммарные приведенные затраты [21] имеют вид:

$$\text{П}=\sum c_{ij}\left(l_{ij}, \tilde{\mu}_{ij}^{\text{л}}\right)+E_{\text{н}}\left[\sum_{i=1}^{n} k_i^{\text{у}}\left(\tilde{\mu}_i^{\text{у}}\right)+\sum_{j=1}^{n_{\text{у}}} k_i^{\text{т}}\left(\tilde{g}_j^{\text{у}}\right)\right]+\sum_{i=1}^{n} k_i \text{Э}_{\text{у}g_i}+\sum_{m=1}^{n_y}\sum_{i\in B_m} k_i \text{Э}_{\text{т}m_i}+\sum_{j=1}^{n_{\text{у}}} \text{Э}_{\text{з}g_i},$$

где $c_{ij}\left(l_{ij}, \tilde{\mu}_{ij}^{\text{л}}\right)$ – стоимость аренды линий длины l_{ij} пропускной способности $\tilde{\mu}_{ij}^{\text{л}}$; $k_i^{\text{у}}\left(\tilde{\mu}_i^{\text{у}}\right)$ – капитальные вложения на УК i-го с производительностью $\tilde{\mu}_i^{\text{у}}$.

В качестве ограничений выбираются ограничения на вероятностно-временные характеристики (ВВХ) сети, а также условие обеспечения заданной связности сети $\nu \geq \nu_3$ для структуры сети.

Для решения данной задачи используются субстантивные методы эвристического программирования.

Вначале проведем декомпозицию задачи на две подзадачи – оптимизацию топологии базовой сети и оптимизацию структуры системы ТО сети. Метод решения поставленной задачи включает итеративное решение поставленных подзадач с использованием информации, полученной на предыдущем шаге поиска: при оптимизации системы ТО топология сети учитывается посредством таких характеристик, как степень узлов связи (число линий, инцидентных УК) и производительность УК, которые определяют параметры $k_i, \text{Э}_{\text{у}g_i}, \text{Э}_{\text{т}m_i}$.

Структура системы ТО на этапе оптимизации топологии учитывается посредством повышения вероятности включения в топологию сети линий, инцидентных узлам, находящимся на возможно меньшем расстоянии от ЦТО.

Для оптимизации топологии сети используются так называемые MST-алгоритмы (см. гл. 1) или алгоритмы, построенные на основе MST-алгоритмов (например, метод размытых эвристик – МРЭ). При генерации сетей с высокой связностью, используется алгоритм, являющийся обобщением алгоритма Прима [20].

В качестве методов локальной оптимизации могут быть использованы наиболее распространенные и универсальные методы трансформации – методы замены линий.

Модификация, которая требуется в алгоритмах оптимизация структуры базовой сети для учета структуры системы ТО, полученной на некотором шаге, является весьма незначительной и состоит в следующем. При использовании алгоритмов генерации решений помимо “успешности” топологии сети, полученной на предыдущем шаге, учитывается такая структура системы ТО, а именно: места размещения и тип ЦТО.

Пусть ω_{ij} – исходный вес (значимость) линии (i, j) между УК$_i$ и УК$_j$, отражающий предпочтительность включения данной линии в окончательный вариант сети и используемый в алгоритмах генерации топологии для учета системы ТО. Тогда модернизацию алгоритмов можно свести к модификации весов $\{\omega_{ij}\}$; $i, j = 1, 2, \ldots, n$ в следующем виде:

$$\omega_{ij}^{'} = \omega_{ij} \frac{\left(M_{m_i} + M_{m_j}\right)}{\left(l_{im_i} + l_{jm_j}\right)}, \qquad (2.1)$$

где l_{im_i} $\left(l_{jm_j}\right)$ – расстояние от УК$_i$ (УК$_j$) до ЦТО, в зону обслуживания которого входит УК$_i$ (УК$_j$); M_{im_i} $\left(M_{jm_j}\right)$ – показатель, отражающий возможность зоны B_i (B_j), в которую входит УК$_i$ (УК$_j$).

В процессе генерации решений необходимо также ввести проверку дополнительно, нового для задач оптимизации структуры сети, ограничения. Оно состоит в недопустимости образования такой структуры сети, которая не может быть обслужена с заданным качеством данной системой ТО без ее изменения. Другими словами, на каждом этапе генерации решения частичные решения должны приниматься только тогда, когда не требуется изменения числа и состава зон ТО и типа ЦТО в зонах. Реализация такой проверки выполняется достаточно просто и по сложности линейно зависит от числа УК и числа ЦТО, несущественно влияя на общее время поиска решений.

Если для оптимизации структуры базовой сети используются алгоритмы локальной оптимизации, то для модификации весов линий, варьируемых в некоторой начальной (заданной или автоматически генерированной) структуре сети, надо использовать выражение (2.1). При этом для введения линий в сеть применяется (2.1), а для удаления линий – обратные к $\omega_{ij}^{'}$ значения.

В качестве весов ω_{ij} могут выступать веса вида

$$\omega_{ij} = \frac{k_{ij}}{d_{ij}},$$

где k_{ij} – путь длины k_{ij} между узлами сети i и j; d_{ij} – стоимость линии.

Подход, основанный на локальной оптимизации, содержит две возможности: 1) после каждой коррекции структуры сети оптимизировать структуру системы ТО; 2) проводить оптимизацию структуры системы ТО только после нахождения локального минимума в задаче оптимизации структуры сети.

Анализ, проведенный в работе [21], говорит в пользу первого подхода, так как оптимизация структуры при втором подходе не является асимптотически оптимальной и требует дополнительных вычислений.

2.2. Каналы связи

В результате анализа исследований эффективности и надежности средств связи установлено, что около 50% ошибок в принимаемом сообщении, а также повреждений происходят в каналах связи. Это говорит о том, что наибольшее число помех и разных мешающих факторов действует на передаваемое сообщение в канале связи, или другими словами, из всех элементов сети наиболее чувствителен к воздействию помех и других мешающих факторов канал связи. Поэтому одним из перспективных путей повышения достоверности и надежности сети передачи данных является разработка методов, позволяющих анализировать и предсказывать аварийные ситуации в каналах связи.

Модель канала

Согласно эталонной модели ЭМВОСЭ, канал передачи данных (ПД) представляет собой совокупность средств двух уровней: 1) физического; 2) канального. Структурная схема канала ПД показана на рис. 2.3. Состав средств приведен для случая, когда канал связи является непрерывным.

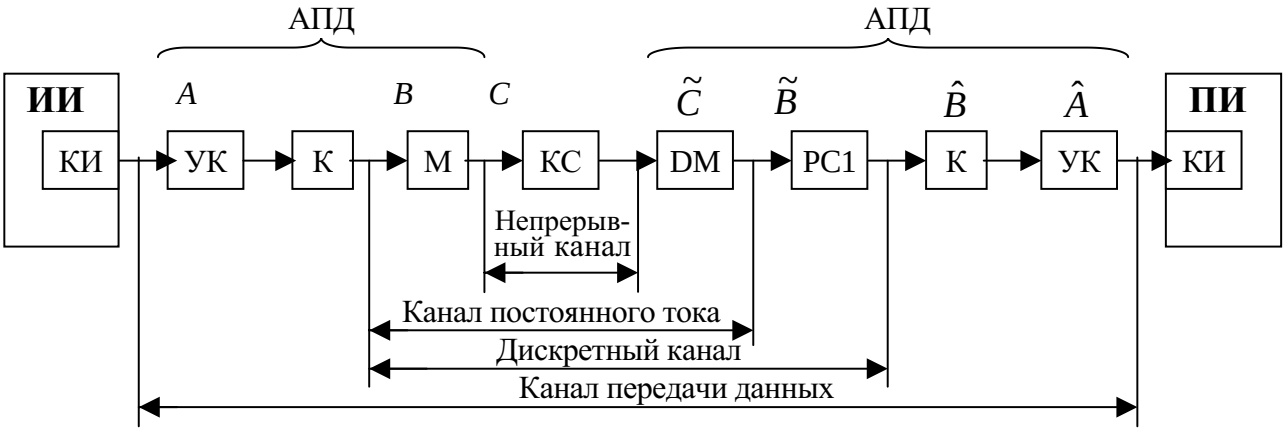


Рис. 2.3. Структура канала передачи данных

ИИ – источник информации; КИ – кодек (кодер-декодер) источника; АПД – аппаратура передачи данных; УК – средства управления каналами передачи данных; К – кодек канала (устройство, осуществляющее операции кодирования передаваемой информации и декодирования принимаемой); М – модулятор; КС – канал связи; ДМ – демодулятор; РС1 – первая решающая схема; ПИ – получатель информации; A, B, C, $\hat{A}$, $\hat{B}$, $\tilde{C}$, $\tilde{B}$ – обозначения сигнала на входе соответствующих устройств.

При использовании цифрового канала вместо модемов применяются контролеры, осуществляющие связь средств управления с каналом связи.

Стоимость структуры каналов ПД обусловлена в первую очередь тем, что для передачи по сети данные необходимо специальным образом отформатировать и согласовать темы их получения от источника со скоростью передачи канала. На приемной стороне нужно решить обратные задачи: восстановить исходную последовательность данных (с учетом возможных потерь и дублирований), согласовать скорость приема информации из канала с темпом ее выдачи получателю.

Физическому каналу на рис. 2.3 соответствует цепочка блоков: модулятор М – канал связи – демодулятор ДМ – первая решающая схема РС1. К функциям управления физическими каналами относятся следующие:

– установление и разъединение соединений;

– преобразование сигналов (изменение диапазона значений параметров переносчика информации или переход к новому переносчику с целью согласования частотных, временных и энергетических параметров сигнала и характеристики канала связи);

– реализация интерфейса (правила взаимодействия) со вторым (канальным) уровнем.

Канальный уровень представлен на рис. 2.3. средствами управления каналом передачи, на которые возлагаются следующие основные функции:

– форматирование данных (разбиение последовательности битов на блоки, добавление служебной информации, признаков канала и конца кадра – самостоятельно адресуемой единицы информации второго уровня);

– помехоустойчивое кодирование данных кадра на передающей стороне и декодирование на приемной стороне с целью обнаружения ошибок, возникающих в процессе передачи данных по каналу связи;

– организация запросов на повторную передачу кадров, принятых с ошибками;

– восстановление исходной последовательности передаваемых бит на приемной стороне (деформатирование данных или распаковка кадров);

– обеспечение прозрачности канала, т.е. кодовой независимости передаваемых данных от кодов управляющих команд (это дает возможность применять произвольный первичный командный код представления информации);

– согласование скорости передачи по каналу связи с темпом их приема и выдачи пользователю.

Совокупность правил взаимодействия смежных уровней (в одной системе), включающих регламент параметров сигнала, кода и целей обмена, называют *межуровневым интерфейсом*.

Характеристики канала

Распределение вероятностей, отдельные параметры и системные функции, отражающие случайные факторы, влияющие на качество передачи информации, составляют множество статистических характеристик канала.

Характеристики непрерывного канала связи называются *первичными*. Они отражают главным образом вызывающие искажения сигнала: нестабильность генераторов несущей и поднесущей частот, скачки и дрожания фазы, замирания (в радиоканалах) и изменения остаточного затухания (в кабельных и воздушных линиях), флуктуационные шумы, импульсные и гармонические помехи, перерывы и нелинейность преобразований, амплитудно- и фазочастотные характеристики. Неидеальность частотных и амплитудных характеристик возникает из-за регулярных искажений только регулярных сигналов. При передаче случайных (информационных) последовательностей искажения приобретают стохастический характер, так как их величина зависит от вида (формы) сигнала.

Характеристики сигнала на выходе канала постоянного тока называются *вторичными*. Они отражают степень искажения единичных элементов и включают краевые искажения, дробления, массу искажений.

Наиболее полной характеристикой качества дискретного канала является *статистика ошибок*, возникающих при передаче информации.

Перечень учитываемых характеристик в каждом конкретном случае зависит от специфики рассматриваемой задачи и типа используемого канала.
Важной характеристикой любого канала является его пропускная способность C, представляющая собой максимально возможную скорость передачи информации, т.е. максимальное количество информации, которое может быть передано по каналу за единицу времени (обычно C измеряется в двоичных единицах информации в сети). Скорость передачи информации по каналу называется отнесенное к единице времени количество взаимной информации между сигналами $A(t)$ и $\dot{A}(t)$, т.е.

$$I'\left(A, \hat{A}\right)= I'\left(A, \hat{A}^*\right)= I'\left(\hat{A}, A\right)= H'(A)- H'\left(A/\hat{A}\right)= H'\left(\hat{A}\right)- H'\left(\hat{A}/A\right),$$

где $H'(A)$, $H'\left(\hat{A}\right)$ – энтропия входного и выходного сигналов; $H'\left(A/\hat{A}\right)$, $H'\left(\hat{A}/A\right)$ – условные энтропии входного сигнала при известном выходном и наоборот, отнесенные к единице времени (секунде). Пропускная способность зависит только от свойств канала, так как представляет собой максимум значений $I'\left(A, \hat{A}\right)$, вычисленный по всем возможным статистикам сигналов, которые могут быть поданы на вход канала в соответствии с заданными ограничениями на передаваемые сигналы

$$C = \max_{\bar{P}(A)} I'\left(A, \hat{A}\right).$$

Характеристики ошибок в дискретном канале

На основе использующихся экспериментальных данных можно сделать следующие заключения о характере ошибок в реальных каналах [24].

Реальные дискретные каналы в общем случае неидеально синхронизированы, нестационарны, несимметричны и имеют память. Ошибки синхронизации (выпадения и вставки символов) связаны с нестабильностью генераторного оборудования и нарушением принудительной синхронизации в период сильного воздействия помехи.

Нестационарность обусловливается наличием детерминированной составляющей в процессах, влияющих на закономерность возникновения ошибок. Как правило, регулярные изменения статистических параметров дискретного канала происходят довольно медленно и в не слишком широких пределах.

Несимметричность реальных каналов обычно имеет сложный характер. Одной из вызывающих причин является инерционность решающих устройств, а также наличие прерываний в канале. В период сильного воздействия мультипликативной помехи при малой аддитивной помехе решающее устройство во многих каналах сохраняет состояние, соответствующее последнему решению, принятому перед прерыванием. При этом во время прерывания дискретный канал становится практически асимметричным. Другой причиной могут быть дискретные воздействия сильной аддитивной помехи одного знака, приводящие к выдаче решающим устройством символов одного вида. Важно отметить, что в обоих случаях несимметрия возникает в периоды, когда выходные символы не зависят от входных (являются искаженными).

Память в реальных дискретных каналах выражается в группировании ошибок. Она связана с тем, что длительность отдельных мешающих воздействий часто превышает длительности отдельных символов, и одно воздействие поражает сразу группу символов. Возникают относительно длинные серии пораженных символов, т.е. пакеты ошибок. Группирование ошибок во многих реальных каналах имеет весьма сложный характер (ошибки группируются в пакеты, пакеты – в более сложные структуры и т.д.).

Следует отметить, что ошибки синхронизации, нестационарность и асимметрия реальных дискретных каналов исследованы менее подробно, чем память. Большинство моделей построено в предположении, что канал идеально синхронизирован, стационарен и симметричен [16].

Требованию простоты и удобства использования наилучшим образом удовлетворяет модель стационарного симметричного двоичного дискретного канала без памяти и отсутствия стираний. Моделью потока ошибок в таком канале служит биномиальная модель, характеризуемая одним параметром – вероятностью неверного приема единичного элемента, основанной на предположении независимости возникновения ошибок.

Однако для большинства реальных каналов она оказывается непригодной. Модели, рассчитанные на отображении реальных дискретных каналов, должны

основываться на испытательных передачах символов по дискретным каналам, которые позволяют делать определенные качественные заключения о характере ошибок.

Для канала с идеальной синхронизацией вводится представление о некотором условном источнике ошибок и стираний. Этот источник выдает дискретный случайный процесс $\{E_i\}$, который назовем последовательностью ошибок. Каждая позиция $\{E_i\}$ складывается с соответствующей позицией процесса $\{B_i\}$ по определенному правилу. Реализация последовательности ошибок $\{E_i\}$ зависит от реализации помехи в непрерывном канале и реализации входного процесса $\{B_i\}$. В общем случае (несимметричный канал) статистика $\{E_i\}$, а следовательно, и верность передачи зависят от статистики помехи и статистики процесса $\{B_i\}$. При этом в стационарном канале при стационарной передаваемой последовательности $\{B_i\}$ последовательность ошибок $\{E_i\}$ также стационарна. В симметричном канале статистика последовательности ошибок $\{E_i\}$ не зависит от статистики входного процесса $\{B_i\}$ (несмотря на зависимость реализаций). Симметричный канал полностью определяется заданием статистики $\{E_i\}$, причем последняя зависит лишь от помехи в непрерывном канале и от построения дискретного канала.

Основные модели источника ошибок

1. *Описание источника ошибок на основе цепей Маркова (схема М)*

Сколь угодно хорошего согласия модели $\{E_i\}$ с экспериментальными данными можно достичь с помощью наиболее универсальных способов – через многомерные распределения или многомерные переходные вероятности, последовательные или интервальные. Однако трудности, связанные с их заданием и использованием, заставляют искать более удобные способы описания $\{E_i\}$ – по возможности через систему одномерных распределений или переходных вероятностей.

Один из таких способов состоит в представлении описанной двоичной последовательности $\{E_i\}$ функции простой цепи Маркова $\{c_i\}$ с k состояниями, которая определяется матрицей переходных вероятностей.

Пусть имеется k-ичный процесс состояний $\{c_i\}$, $c_i = 0, 1, \ldots, k-1$. Пусть эти состояния могут сменяться одно другим (или сохраняться) только в заранее фиксированные моменты времени $t_{-1}, t_0, t_1, \ldots, t_i, \ldots$, которые в дальнейшем будем обозначать их номерами $(-1, 0, 1, \ldots, i, \ldots)$ и называть позициями (шагами). Тогда, если вероятность того или иного состояния c_i на i-й позиции полностью определяется состоянием $c_{i-1}, \ldots, c_{i-n}$ на n предшествующих позициях и не меняется при получении информации о более ранних состояниях, то случайная последовательность указанных состояний $\{c_i\}$ $(c_i = c)$ называется n-связной k-ичной цепью Маркова. При $n = 1$ цепь Маркова называется *простой*.

Для описания простой цепи Маркова необходимо задать условные вероятности $P_{c_{i-1}c_i}$ того, что система на i-м шаге перейдет в состояние c_i при

условии, что на $(i-1)$-м шаге она находилась в состоянии c_{i-1}. Они определяют квадратную матрицу k-го порядка $\left\| p_{c_{i-1}c_i} \right\|$, которая называется *матрицей переходных вероятностей* или *матрицей переходов*.

В общем случае матрица переходов зависит номера шага i (неоднородная цепь Маркова). Если матрица переходов не зависит от i, то цепь Маркова называется *однородной*.

Для построения модели воспользуемся представлениями о последовательности двоичных состояний “0”, “1”, в которых ошибки независимы, но имеют произвольные условные вероятности ε. Первое состояние будем называть “хорошим” (полностью определяется переданными символами), второе – “плохим” (определяется помехой). Искаженные символы не зависят от переданных и совпадают с ними лишь случайно. Пораженный символ не содержит информации о переданном символе. При наличии памяти конечной величины N-последовательности двоичных состояний могут рассматриваться как N-связные цепи Маркова. При этом значения символов 0, 1 на каждой позиции рассматриваются как состояние цепи. В силу стационарности описываемых последовательностей отображающие их цепи Маркова являются однородными и определяются матрицей переходных вероятностей для произвольной позиции i. В общем случае для определения цепи Маркова необходимо также задавать начальные вероятности (вероятности символов на первой или первых позициях). Однако, если рассматривать последовательности двоичных состояний бесконечными и эргодическими, то начальные вероятности для них значения не имеют. Безусловные вероятности символов на любой позиции при этом называются финальными.

Наименьшая память системы имеет величину N. В этом случае вероятность того или иного символа данной позиции зависит лишь от символа непосредственно предшествующей позиции. При этом последовательность двоичных состояний является односвязной ($n = 1$) или простейшей цепью Маркова, которая полностью определяется матрицей одномерных переходных вероятностей

$$p = \left\| \begin{matrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{matrix} \right\|.$$

N-связная цепь Маркова с двумя состояниями “0” и “1” может быть определена при помощи односвязной цепи Маркова $\{c_i\}$ с $k = 2^n$ состояниями $c_i = 0, 1, \ldots, k-1$. При этом каждое из k-ичных состояний c_i на данной позиции $\{c_i\}$ соответствует вероятности того или иного из двух возможных значений c_i на данной позиции и $n-1$ предшествующих позиций. Таким образом, статистика $\{E_i\}$ полностью определяется матрицей переходных вероятностей p_{c-1,c_0} порядка k

$$p = \left\| p_{c-1,c_0} \right\| = \begin{Vmatrix} p_{00} & p_{01} & \cdots & p_{0,k-1} \\ p_{10} & p_{11} & \cdots & p_{1,k-1} \\ \vdots & \vdots & \cdots & \vdots \\ p_{k-1,0} & p_{k-1,1} & \cdots & p_{k-1,k-1} \end{Vmatrix} \qquad (2.2)$$

(сумма вероятностей в каждой строке равна 1) и значениями вероятности ошибки в каждом состоянии $\varepsilon_0, \ldots, \varepsilon_{k-1}$. В частности вероятность ошибки в канале записывается выражением

$$p_l = \sum_{c=0}^{k-1} P_c \varepsilon_c ,$$

где $P_c = p_{c_0}$ – финальные вероятности состояний, которые определяются по матрице (2.2) из системы уравнений

$$\sum_{c_0=0}^{k-1} p_{c_0} = 1; \quad \sum_{c_0=0}^{k-1} P_{c-1} p_{c-1,c_0} = p_{c_0}, \; c_0 = 0,\ldots,k-1 \; \left(P_{c-1} = p_{c_0}, \text{ если } c-1 = c_0\right).$$

Способ описания источника ошибок на основе простой цепи Маркова $\{c_i\}$ с k состояниями удобен лишь при небольших значениях k. Между тем для достижения удовлетворительного согласия модели с экспериментальными данными, особенно собранными за длительное время, значение k должно быть достаточно велико. Поэтому целесообразно использовать также и другие способы построения модели.

2. *Описание источника ошибок на основе процессов восстановления (схема В)*

Последовательность $\{E_i\}$ может быть разбита на отрезки – серии символов двух видов: пакеты ошибок и промежутки между ними. В каждом из отрезков возникают независимые ошибки с вероятностями ε_1 и ε_0,причем $\varepsilon_1 \geq \varepsilon_0$. Длины промежутков λ ($\lambda = 1, 2, \ldots$) и длины пакетов l ($l = 1, 2, \ldots$) независимы в совокупности. Поэтому статистика $\{E_i\}$ полностью определяется их одномерными распределениями $P(\lambda)$ и $P(l)$ и вероятностями ε_1 и ε_0. Это значит, что канал имеет два состояния: “хорошее” и “плохое” ($k = 2$), и последовательность состояний $\{c_i\}$ является процессом восстановления с конечным временем. Вероятность попадания символа в пакет ошибок равна

$$p = \frac{\bar{l}}{\bar{\lambda} + \bar{l}},$$

где $\bar{\lambda}$ и $\bar{l}$ – средние длины пакета и промежутка соответственно (в теории надежности отношение $\frac{\bar{\lambda}}{\bar{\lambda} + \bar{l}} = 1 - p$ называется коэффициентом готовности). Вероятность того, что данная позиция является началом пакета ошибок, и

равная ей вероятность того, что данный символ является началом промежутка между пакетами, равна

$$p_\lambda = \frac{1}{\bar{\lambda} + \bar{l}}.$$

Поэтому вероятность ошибки выражается как

$$p_l = \varepsilon_0(1-p) + \varepsilon_1 p = \frac{\varepsilon_0\bar{\lambda} + \varepsilon_1\bar{l}}{\bar{\lambda} + \bar{l}} = p_n\left(\varepsilon_0\bar{\lambda} + \varepsilon_1\bar{l}\right).$$

3. *Описание источника ошибок на основе процессов накопления (схема H)*

Данная модель источника ошибок учитывает допустимость перекрытия пакетов.

Любая позиция последовательности $\{E_i\}$ может стать началом пакета ошибок, причем длины интервалов между началами пакетов $\lambda_0^* = 0, 1, \ldots$ являются значениями независимых случайных величин Λ_{0j}^*. Последовательность состояний, где состояние “1” соответствует позиции, являющейся началами состояний пакетов, представляет собой процесс с мгновенным восстановлением, статистика которого полностью определяется распределением $p\left(\lambda_0^*\right)$ Длины пакетов $l^* = 1, 2, \ldots$ также являются значениями независимых случайных величин L_j^*, которые имеют распределения вероятностей $P(l^*)$. В пределах каждого отдельного пакета (не перекрывающегося с другими пакетам) ошибки независимы и имеют вероятность ε, или, что то же, позиции поражаются независимо с вероятностью 2ε. Таким образом, статистика $\{E_i\}$ по схеме H полностью определяется двумя одномерными распределениями – длин пакетов $P(l^*)$ и интервалов между началами пакетов $p\left(\lambda_0^*\right)$ (т.е. статистической последовательностью пар независимых чисел $\left\{\Lambda_{0j}^*, \lambda_0^*\right\}$) и вероятностью ошибки в такте ε.

Для рассматриваемой модели независимы не промежутки между пакетами, как в схеме B, а интервалы между началами пакетов. Это обстоятельство обусловливает возможность перекрытия и примыкания пакетов ошибок. Последовательность $\{E_i\}$ для описываемой модели может быть представлена последовательностью состояний $\{c_i\}$, в пределах которых ошибки имеют одинаковые вероятности. Число состояний при этом в общем случае уже не равно двум. Оно может быть сколь угодно большим, так как на участке, являющемся наложением нескольких пакетов, вероятность ошибки может превышать вероятность ошибки ε в каждом отдельном пакете.

Возможность перекрытия пакетов существенно усложняет расчет многих характеристик канала. В этом случае просто определяется лишь один параметр – вероятность того, что данный символ является началом пакета ошибок

$$p_p^* = \frac{1}{\overline{\lambda}_0^* + 1},$$

где $\overline{\lambda}_0^*$ – средняя длина интервала между началом пакетов.

2.3. Канал передачи данных

В большинстве современных систем передачи дискретной информации для помехоустойчивого кодирования используются блочные коды. При исследовании временных характеристик таких систем (средних значений и распределения числа переспросов, времени передачи и т.д.) исходным процессом является последовательность решений декодера – второй решающей схемы. Это обусловливает целесообразность двухуровнего описания канала: модель первого уровня, описывающая ошибки по символам, применяется для оценивания эффективности выбранного кода и расчета параметров модели второго уровня, аппроксимирующей поток искажений кодовых блоков.

На первом уровне может использоваться полная модель дискретного канала, а при отсутствии необходимых экспериментальных данных – частичная модель [17].

Если на первом уровне использовать полную модель дискретного канала, то структура второго уровня определяется однозначно. Например, если первый (символьный) уровень описан марковской моделью, характеризуемой числом состояний канала k, матрицей переходных вероятностей $B = \{B_{ij}\}$, $i = \overline{0, k-1}$, $j = \overline{0, k-1}$ и условными вероятностями ошибок в каждом состоянии ε_j, $i = \overline{0, k-1}$, то модель второго (блокового) уровня также марковская с тем же числом состояний k.

Матрица P переходных вероятностей второго уровня определяется соотношениями $P = B^m$, где m – число символов в блоке.

Условные вероятности q_i искажения блока вычисляются с помощью матриц:

$$B_1(0) = [(1 - \varepsilon_j)B_{ij}];\ B_1(1) = [\varepsilon_j B_{ij}],$$

представляющих собой матричные вероятности того, символ будет принят правильно и ошибочно соответственно. Для этого в начале определяются матричные вероятности приема кодового блока без ошибок и с ошибками:

$$P_1(0) = B_1^m(0);\ P_1(1) = B^m - B_1^m(0) = P - P_1(0).$$

Затем определяются искомые вероятности

$$q_j = E_{ij}(P_1(1)) \Big/ E_{ij}(P),\ j = \overline{0, k-1},$$

где i – число из множества $\{0, 1, \ldots, k-1\}$; E_{ij} – элемент матрицы, указанной в скобках.

2.4. Модель трафика

При буквальном переводе с латинского “трафик” (“tra-veho”) означает “перевести, переслать” [7]. Поэтому под понятием “трафик” можно подразумевать любые потоки информации в системах коммутации и распределения информации.

Основные виды информации, поступающие от пользователей в ЦСИО, включают в себя оперативные и диалоговые данные, речь, видеопотоки, фоновую информацию.

Оперативные данные представляют собой небольшие цифровые потоки, чувствительные к задержкам и шумам. Оперативные данные могут быть как пользовательскими, так и служебными, т.е. порождаться самой сетью для целей управления ею.

Диалоговые данные включают относительно короткие сообщения с достаточно большими допустимыми задержками, но критичные к шумам. Представление диалога в значительной степени зависит от случайного поведения пользователей при каждом конкретном комплекте оборудования терминалов.

Речь – это поток чередующихся интервалов активностей и пауз. Речевой сигнал может передаваться и цифровым способом. При исследовании пакетизированной передачи используется модель, отображающая распределения длин активностей и пауз по геометрическому (полигеометрическому) закону, причем математические ожидания их длительностей примерно одинаковы (около 10^{-3} с), а диапазон их изменения составляет около 10 μс – 2 с [1]. Пакетизированная речь принадлежит категории трафика реального времени, т.е. задает жесткие требования к временным характеристикам доставки.

Видеопотоки представляют собой большие (сотни мегабайт и более) потоки аналоговой по своей природе информации (промышленное телевидение, диспетчерская связь и др.), подобно речевым потокам, но требующие значительно большие полосы частот (скорости передачи).

Фоновая информация бывает двух типов: файлы данных и видеофайлы. Для передачи фоновой информации обычно используется режим коммутации каналов (КК) и коммутации сообщений (КС). Передача осуществляется сообщениями большей длины с обязательным подтверждением приема (рис. 2.4.).

Характеристики информационного трафика пользователей сети, как правило, менее устойчивы, чем характеристики телефонного трафика из-за широкого диапазона скоростей передачи, длин сообщений, частот их поступления. Доля каждого вида информации в общем ее объеме в ЦСИО и в отдельных ее участках может меняться в течение времени в широких пределах. Часто диалоговый трафик меняется скачкообразно (от периодов высокой к периодам низкой активности и наоборот). На рис. 2.4 показаны различные виды трафика и возможные скорости их передачи [1].

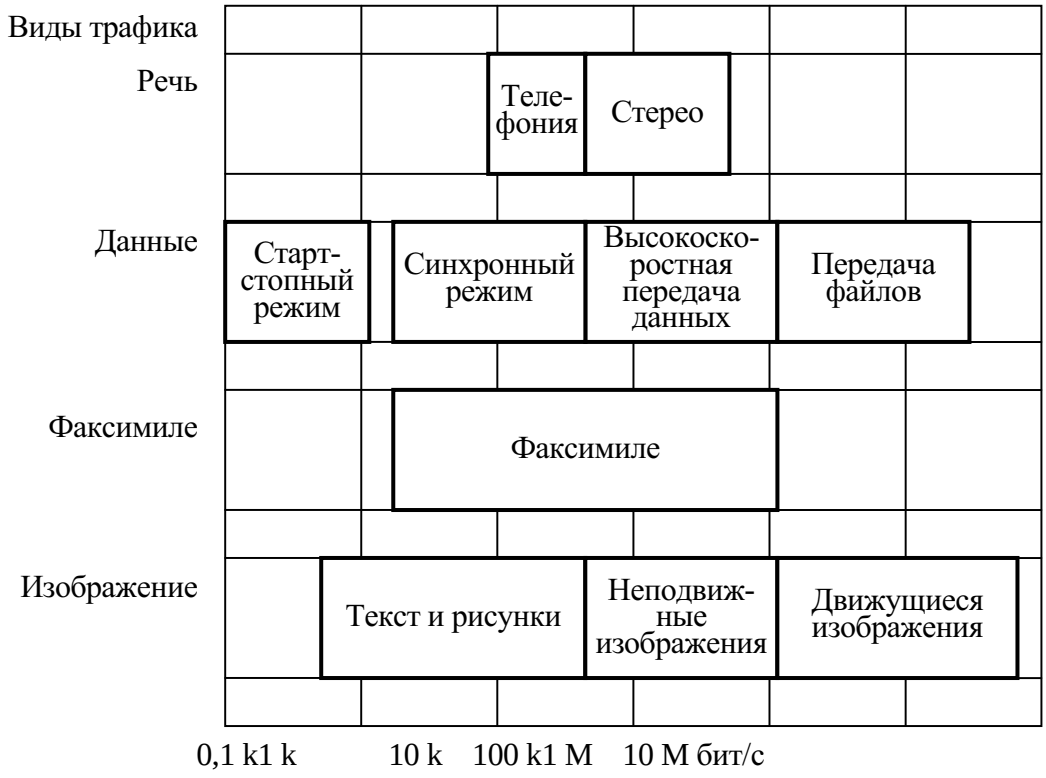


Рис. 2.4. Скорость передачи для различных видов трафиков

Модель трафика [1] определяется:

– законом распределения интервала времени между моментами поступлений информационных сообщений (пакетов);

– классом приоритетности, определяемым в общем случае скоростью доставки;

– законом распределения числа пакетов в информационном сообщении;

– ценностью информации;

– адресом источника и адресом получателя информации и т.д.

Основным назначением ЦСИО является обеспечение требуемого качества обслуживания пользователей, которое заключается в реализации требований пользователей на передачу их информации. Поэтому поток информации, поступающей в сеть, будет определять нагрузку сети. В общем случае входящий поток является случайным. При статистическом моделировании процессов обмена информации наиболее специфической моделью является модель потока [8].

2.5. Потоки вызовов

Под потоками вызовов понимают последовательность моментов поступления вызовов t_1, t_2, …, t_k, …, t_{n-1}, t_n. При этом длительность занятия приборов обслуживающих устройств не учитывается. Если вызовы поступают

через определенные промежутки времени, то поток называется *детерминированным*.

В реальных сетях вызовы поступают в случайные моменты времени, т.е. имеют случайный характер. Поэтому потоки задаются распределением вероятностей поступления вызовов за рассматриваемый интервал времени $P_k(t)$. Величина $P_k(t)$ означает вероятность поступления k вызовов за время t.

Потоки характеризуются следующими основными свойствами: стационарностью, ординарностью, отсутствием последствия.

Поток называется *стационарным*, если его вероятностные характеристики не меняются с течением времени, точнее – в стационарном потоке вероятность $P_k(t)$ зависит только от длины отрезка t, а не от момента его начала.

Ординарность означает практическую невозможность появления двух и более вызовов. Более строго можно сказать, что в ординарном потоке вероятность поступления двух и более вызовов в отрезке времени τ при $\tau \to 0$ является бесконечно малой величиной более высокого порядка, чем τ.

Отсутствие последствия означает, что прошлое в потоке никак не может воздействовать на будущее, т.е. вероятность поступления вызовов после момента t не зависит от вероятностного процесса до этого момента. При малом числе источников нагрузки ($N < 100$) вероятность нового вызова зависит от числа уже поступивших, т.е. имеется последствие.

Для описания реальных потоков чаще всего используются две теоретические модели: простейший и примитивный потоки вызовов.

Поток вызовов, обладающий одновременно стационарностью, ординарностью и отсутствием последствия, называется *простейшим*. Вероятность поступления точно k вызовов за время t $P_k(t)$ в простейшем потоке подчиняется распределению Пуассона

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

где λ – параметр потока, равный среднему числу вызовов за единицу времени.

Если рассматривается единичный интервал времени $t = 1$, то $P_k = \frac{\lambda^k}{k!} e^{-\lambda}$ практически поток считается простейшим при $N \geq 100$.

Если вероятность поступления вызовов зависит от состояния системы, то поток вызовов называется потоком с простым последствием.

Примитивный поток – это ординарный поток с простым последствием, у которого параметр изменяется пропорционально числу свободных источников нагрузки:

$$\lambda_i = (N - i)\alpha,$$

где N – число источников нагрузки; i – число занятых источников; α – коэффициент пропорциональности (равный параметру потока от одного свободного источника).

Примитивный поток создается конечным числом источников ($N < 100$). При увеличении N примитивный поток переходит в простейший. На практике при исследовании примитивного потока пользуются не параметром α, а так называемой интенсивностью нагрузки a, поступающей от одного источника, которая связана с α соотношением

$$\alpha = \frac{\alpha}{1+\alpha}.$$

Вероятность поступления вызовов примитивного потока подчиняется распределению Бернулли:

$$P_k(t) = c_N^k a^k (1-a)^{\lambda},$$

где k – число поступивших вызовов; t – период наблюдения.

Поток вызовов может также задаваться функциями распределения промежутков между вызовами (закон интервалов) или числом вызовов в фиксированные промежутки времени (закон реализаций). Поток, у которого интервалы между вызовами независимы и распределены одинаковым образом, называется *рекуррентным*. Простейший поток – это рекуррентный поток, у которого интервалы между вызовами распределены по показательному закону с функцией распределения

$$F(x) = P\{t_k < x\} = 1 - e^{-\lambda x},$$

где t_k – интервал между вызовами; λ – параметр экспоненциального распределения, численно совпадающий с интенсивностью потока – числом вызовов в единицу времени.

В теории трафика находят применение и другие модели потоков, которые с различной степенью точности отображают характеристики реальных потоков вызовов. Однако чаще всего рассматривают рекуррентные или простейшие (пуассоновские) потоки вызовов.

Так, например, при анализе многопролетных пакетных радиосетей (таких сетей, отдельные цепи которых осуществляют передачу с промежуточным хранением информации, поскольку не все узлы могут непосредственно связываться между собой) обычно предполагается [9], что в системах с непрерывными во времени протоколами поступающие сообщения распределены по пуассоновскому закону, а в сегментированных системах – по закону Бернулли.

Если задано распределение объемов сообщений (времени занятия) или их среднее значение, то поток сообщений на каком-либо участке сети может быть охарактеризован нагрузкой – общим временем занятия всех каналов на данном участке. Если характеристики потоков для всей сети можно считать одинаковыми, то все они могут быть охарактеризованы только одним параметром – нагрузками. В сети, имеющей N узлов, потребность связи между ними (тяготение) может быть задана совокупностью нагрузок для каждой пары узлов, представленной в виде матрицы или графа нагрузок [10].

При правильно спроектированной сети структура каналов должна соответствовать матрице (графу) нагрузок. Такое соответствие может продолжаться лишь некоторый период времени, так как нагрузки между узлами сети постоянно меняются. Подобные изменения нагрузок приводят к несоответствию (“перекосу”) сети нагрузкам, что вызывает ухудшение качества обслуживания на некоторых направлениях сети.

Тяготение зависит от вида информации, территориального распределения пользователей, их особенностей, а также от характера взаимосвязи.

Тяготение по своему характеру может быть детерминированным, т.е. иметь определенный объем, распределенный во времени для каждой пары или группы пунктов сети, или случайным с некоторым распределением по объемам, коэффициентам концентрации (отношение среднего тяготения в ЧНН – (часы наибольшей нагрузки) и коэффициентам тяготения (отношение тяготения между двумя заданными пунктами к суммарному тяготению между этим пунктом и остальными пунктами сети).

Будем рассматривать (в случае выбора в качестве моделей источников информации пуассоновских входных потоков) интенсивности потоков, входящих в ЦСИО, задаваемых матрицей тяготения в виде

$$\left\| \Lambda^{ij} \right\| = \left[\lambda_{sd}^{ij} \right],$$

где λ_{sd}^{ij} – численность потока информации для передачи информации *i*-го вида и *j*-го приоритета между *S*-м и *d*-м УК.

Так как с учетом подобия конкретных видов и приоритетов потоков информации для каждого режима коммутации

$$d_{sd} = \sum_{\substack{i \in I \\ j \in J}} \lambda_{sd}^{ij},$$

где *I* – множество видов информации; *J* – множество приоритетов, предусмотренных для передачи информации, то

$$\Lambda = \lambda_{sd}.$$

С потоками информации связана еще одна в общем случае случайная величина – длина сообщения в режиме КП. Под длиной сообщения будем понимать число содержащихся в них двоичных единиц информации пользователя. При аналитических расчетах будем считать, что длина сообщения для режима КП для информации каждого вида и приоритета распределена экспоненциально со средним числом $1/\mu^{ij}$ и одинаково для всех пар “источник-адресат”. Таким образом, усредненная длина сообщения в режиме КП будет равна

$$\frac{1}{\mu} = \sum_{\substack{i \in I \\ j \in J}} \frac{\lambda^{ij}}{\lambda} \frac{1}{\mu^{ij}}.$$

При определении нагрузок на отдельных участках сети необходимо учитывать, что кроме потоков сообщений, несущих информацию пользователей и информацию, связанную с доставкой каждого сообщения в сети, учитываются дополнительные нагрузки, определяемые самой сетью: повторные вызовы в сетях, задержки в процессе установления соединения в системах с ожиданием, повторные передачи и переспросы при обнаружении ошибок в процессе передачи, наличие служебной информации, связанной с управлением сетью.

Так, потоки сообщений – их последовательности во времени и в пространстве – характеризуются тремя основными распределениями:

1) моментов поступления отдельных сообщений или интервалов времени между этими моментами;

2) объемов, выраженных часо-занятиями, числом бит, знаков, слов, кадров и т.д.;

3) адресов (мест назначения).

По каждой из этих характеристик можно выделить детерминированные, стохастические и смешанные потоки. Обычно рассматривают случайные ординарные потоки вызовов в стационарном и нестационарном режимах без последствия и с последствием.

Нагрузка

Нагрузка определяется как суммарное время обслуживания сообщений.

Интенсивность нагрузки в общем случае характеризуется неравномерно. Наблюдением установлено, что наряду со случайными колебаниями интенсивности нагрузки по часам суток, дням недели и месяцам года существуют ее периодические, относительно регулярные колебания, поддающиеся расчету.

Под интенсивностью нагрузки понимается нагрузка за единицу времени, обычно за 1 ч. За единицу измерения интенсивности принят Эрланг (Эрл); 1 Эрл – это нагрузка в 1 ч.-зан. за 1 час.

За единицу измерения интенсивности нагрузки принято 1 часо-занятие (1 ч.-зан.), т.е. такая нагрузка, которая может быть обслужена одной двухполюсной сетью в течение 1 ч при непрерывном ее занятии. Для сетей ПДС приемлемо также измерение нагрузки в единицах измерения информации (байт, бит). Общепринято за максимальную брать нагрузку в ЧНН (час наибольшей нагрузки) – это непрерывный интервал времени, в течение которого средняя интенсивность нагрузки является наибольшей.

Основными параметрами нагрузки являются: N – число источников нагрузки; n_i – число источников i-го типа; c – среднее число вызовов (сообщений) от одного пользователя; Θ – среднее время занятия двухполосной сети ПДС для передачи одного сообщения.

При оценке среднего числа вызовов (сообщений) необходимо учитывать вид обмена (доставка сообщений или диалоговая связь) и способы коммутации. Для режима доставки число сообщений определяется однозначно. При этом оно не зависит от способа коммутации, кроме КП, для которого помимо общего числа сообщений необходимо определять число пакетов.

При диалоговой связи условия расчета сохраняются, если за сообщение принимается полностью завершенный переговор. При этом в случае КА необходимо знать число пакетов.

Различные виды пользователей могут классифицироваться как по пропускной способности, так и по интенсивности.

Пусть c_i – среднее число сообщений i-го типа, а n_i – число пользователей этого типа. Тогда

$$c = \frac{\sum_{i=1}^{k} c_i n_i}{\sum_{i=1}^{k} n_i},$$

где k – число всех классов сообщений.

Средняя длительность занятия двухполосной сети ПДС обычно оценивается через случайное время Θ. Для сетей с КП

$$\Theta = t_з + t_о + t_{ож} + t_{пер} + t_{подтв},$$

где $t_з$ – время запроса на передачу сообщения; $t_о$ – время ответа на разрешение передачи; $t_{ож}$ – время ожидания в очереди на передачу; $t_{пер}$ – время передачи сообщения; $t_{подтв}$ – время подтверждения времени сообщения. Величину $t_з + t_о$ можно рассматривать как время коммутации, а $t_{ож} + t_{пер} + t_{подтв}$ – время доставки сообщения.

Для сетей с КП необходимо учитывать дополнительно случайное время разбиения сообщений на пакеты и его сборки из пакетов.

Пусть каким-то образом при суммировании случайных величин будут получены некоторые средние значения перечисленных показателей. Тогда можно определить среднее время коммутации $T_к$, среднее время доставки $T_д$, включая время $T_{ож}$.

Средняя численность входящей нагрузки двухполюсной сети ПДС может быть определена по числу источников, среднему числу вызовов (сообщений) в ЧНН c_i для каждого источника и длительности занятия Θ_i также для каждого занятия:

$$\rho = \sum_{i=1}^{N/2} c_i \Theta_i.$$

Здесь значение $N/2$ принимается из тех соображений, что каждый вызывающий должен соединиться или доставить сообщение одному вызывающему, который в это время вызывать не может.

Необходимо отметить, что интенсивность входящей нагрузки можно также определить через интенсивности входящего потока λ и обслуживания μ:

$$\rho = \frac{\lambda}{\mu}.$$

Определим обслуженную и потерянную нагрузки:

$$\rho = \frac{\lambda_{со}}{\mu};\ \rho = \frac{\lambda_{п}}{N},$$

где $\lambda_{со}$ – интенсивность своевременного обслуживания потока сообщений; $\lambda_{п}$ – интенсивность потерянного потока сообщений. При этом $c = y + x$, $\lambda = \lambda_{со} + \lambda_{п}$.

2.6. Модели процесса обмена информацией в ЦСИО

Под задачей управления процессом обмена информацией в ЦСИО будем понимать задачу выбора оптимальных по отношению к некоторому критерию методов маршрутизации и ограничения интенсивности потоков в сети с заданной топологической структурой при соответствующих ограничениях. Управление процессом обмена информацией в ЦСИО тесно связано с используемыми в сети методами (режимами) коммутации. В дальнейшем будем считать, что в интегральной сети реализован метод КП, а остальные режимы коммутации будут рассматриваться как частные случаи.

Управление процессом обмена информацией в ЦСИО характеризуется рядом особенностей [1]:

1) сеть, по которой передается служебная информация об объекте управления, обладает теми же характеристиками, что и ЦСИО, в которой осуществляется обмен пользовательской информацией, так как в общем случае они совпадают;

2) элементы системы управления территориально удалены друг от друга.; это приводит к тому, что служебная информация о состоянии объекта управления всегда запаздывает и отражает прошлое состояние процесса обмена информацией, т.е. решения по управлению процессом обмена информацией всегда принимаются на основе информации о прошлом состоянии этого процесса;

3) пропускная способность ЦСИО, т.е. количество информации (бит/с), которое можно одновременно передавать между всеми УК за единицу времени, при некоторых условиях может быть меньше производительности источников информации, генерирующих сообщения;

4) условия работы ЦСИО изменяются, т.е. имеют место случайные изменения интенсивностей и направлений, входящих в сеть потоков сообщений, случайные воздействия ошибок сообщений в каналах связи (КС) на передаваемую цифровую информацию; случайные изменения топологической

структуры сети вследствие выхода из строя УК или КС (полностью или частично) и их последующего восстановления; эволюционные изменения топологической структуры ЦСИО, т.е. добавление новых УК, новых КС, удаление УК или КС и т.д.

Основные особенности обработки и передачи информации в ЦСИО могут рассматриваться на базе двух типов повторяющихся процессов: 1) процессов взаимодействия между парой пунктов ЦСИО; 2) процессов, происходящих в УК. Это возможно вследствие того, что процесс обработки и передачи сообщений состоит из повторяющихся циклов. Пакет (запрос на соединение) поступает в УК, к которому подключен отправитель, обрабатывается в нем и передается дальше через промежуточные УК к узлу, к которому подключен получатель. Цикл обработки аналогичен в каждом УК. Таким образом, управление процессом обмена информацией должно содержать в себе управление потоками по входу УК, внутри УК и по выходу из него. Управление по входу УК включает в себя процедуры управления интенсивностью передаваемых по сети потоков. Управление в УК представляет собой маршрутизацию потоков, а управление по выходу УК – совокупность управления структурой сети.

Модель процесса функционирования УК

Для исследования процесса функционирования УК воспользуемся схемами массового обслуживания (СМО). Система распределения информации обладает всеми признаками системы массового обслуживания, а именно:

– наличием потоков сообщений, которые характеризуются моментами поступления и упрощенно описываются этими моментами;

– наличием обслуживающей системы (собственно системы распределения информации);

– наличием дисциплины обслуживания, регламентирующей порядок обслуживания сообщений.

Будем использовать пятизначное символическое обозначение систем массового обслуживания в виде:

$$A/B/V/k/N,$$

где A – закон распределения промежутков времени между поступающими требованиями; B – закон распределения времени обслуживания; V – число обслуживаемых приборов; k – наибольшее число требований в системе (очередь S плюс обслуживание требования); N – число источников нагрузки.

Две последние позиции являются необязательными, пустое место в любой из них свидетельствует о том, что соответствующая величина равна ∞. Заметим, что среди конкретных законов распределения, которые указываются в пятизначном символическом представлении, вместо букв A и B чаще всего встречается показательное распределение (обозначаемое буквой M); возможен детерминированный процесс, обозначаемый буквой D.

Поступающие на вход УК вызовы могут немедленно получить возможность передачи информации по требуемому адресу, могут получить

отказ (снимаются с дальнейшего обслуживания) или могут быть поставлены в очередь для ожидания возможности предоставления соединения. На основании этого различают две дисциплины обслуживания: без потерь и с потерями. *Дисциплиной обслуживания без потерь* называется такая, при которой поступающему вызову немедленно предоставляется возможность соединения.

Реальные коммутационные системы обычно проектируются с допустимыми потерями. *Дисциплиной обслуживания с потерями* называется такая, при которой поступающий вызов может получить отказ при невозможности немедленного установления соединения либо обслуживание его задерживается на некоторое время.

Различают следующие виды потерь: явные, условные и комбинированные. При обслуживании с *явными потерями* поступающий вызов, получая отказ в соединении с требуемым абонентом, покидает систему и в дальнейшем не оказывает на нее никакого влияния. Система, обеспечивающая обслуживание с явными потерями, называется системой *с отказами*. Дисциплиной обслуживания *с условными потерями* называется такая, при которой в момент отсутствия соединительных путей вызов не получает отказ, а обслуживается с ожиданием.

На практике кроме систем с отказами и с ожиданиями встречаются различные их комбинации. Например, при ограничении числа мест на ожидание (длина очереди *r*) часть вызовов будет обслуживаться с ожиданием, а часть вызовов, поступающих в период, когда на ожидании уже находится r вызовов, обслуживается с отказами. Здесь говорят о дисциплине обслуживания *с комбинированными потерями*.

При обслуживании вызовов с потерями возникает необходимость предоставления определенных преимуществ (приоритетов) для абонентов, находящихся на верхних этажах иерархической структуры, или для срочных сообщений. В связи с этим вводится понятие дисциплины обслуживания *с приоритетом* и *без приоритетов*.

Различают два варианта схем коммутационной системы: 1) полнодоступные; 2) неполнодоступные [3, 4, 5].

В *полнодоступных схемах* любой вход может быть соединен с любым выходом, если последний находится в свободном состоянии (не занят обслуживанием какого-либо вызова).

В *неполнодоступных схемах* это условие не выполняется, все входы и выходы разбиваются на группы, и определенные группы входов имеют доступ (возможность соединения) к определенным группам входов.

Если для описания полнодоступной схемы требуется только знание числа входов и выходов, то для описания неполнодоступных схем указывается число групп входов и возможность доступа определенных групп входов к определенным группам выходов.

Для определения ВВХ, характеризующих качество обслуживания вызовов, необходимо определить вероятности состояний УК в зависимости от

характеристик поступающего потока вызовов, схемы коммутационной системы, дисциплины и длительности обслуживания вызовов.

В системах с ожиданием обслуживание вызовов осуществляется с неявными потерями, т.е. с потерями на время задержки в доставке сообщений. Такие потери наиболее характерны для систем с КС (КП).

В рассматриваемом случае постановка задачи конкретизируется следующим образом:

– на входы полнодоступной системы с V выходами $(1 \le V \le \infty)$ поступает простейший поток вызовов с параметром λ;

– при занятии всех ν выходов поступающий вызов ставится в очередь до освобождения одного из занятых выходов;

– длина очереди конечна ($k < \infty$);

– вызовы, находящиеся на ожидании, обслуживаются в порядке очереди;

– длительность обслуживания распределена по показательному закону с параметром μ.

Таким образом, при обслуживании и поступлении на вход системы одного требования модель системы будет M/M/1.

Пусть на один обслуживающий прибор поступает простейший поток вызовов с параметром λ, а время обслуживания одного вызова имеет экспоненциальное распределение с параметром м. Вызовы обслуживаются в порядке поступления. Тогда, если $y = \lambda/\mu < 1$, получим стационарный марковский процесс, т.е. процесс размножения и гибели, который может быть отображен графом, представленным на рис. 2.5.

Параметр $\lambda_k = \lambda$, т.к. поток простейший; параметр потока освобожденный $\mu = \mu$, так как в системе только один обслуживающий прибор.

Пользуясь формулами вероятности состояния для процессов размножения и гибели, получим

$$P_k = p_0 \prod_{i=0}^{k=1} \frac{\lambda_i}{\mu_i + 1} = p_0 y^k .$$

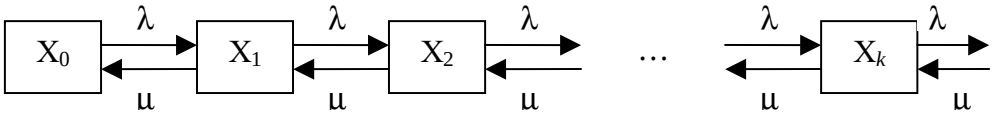


Рис. 2.5. Граф размножения и гибели для одноканального СМО с ожиданием

По условию нормировки $\sum_{i=0}^{\infty} y^i p_0 = 1$ или $p_0 \sum_{i=0}^{\infty} y^i = 1$. Так как $y < 1$, то ряд $\sum_{i=0}^{\infty} y^i$ сходится и имеет предел $\frac{1}{1-y}$. Отсюда $p_0 = 1 - y$. Следовательно

$$P_k = y^k(1-y). \qquad (2.3)$$

Полученное выражение представляет собой геометрическое распределение.

Пользуясь выражением (2.3), определим основные характеристики качества обслуживания для одноканальной системы:

1) среднее число вызовов, находящихся в системе,

$$\overline{n} = \sum_{i=0}^{\infty} i p_i = (1-y)\sum_{i=0}^{\infty} i y^i = \frac{y}{1-y};$$

2) среднее число вызовов, находящихся в системе

$$\overline{n}_0 = \sum_{i=0}^{\infty} (i-1) p_i = (\overline{n} - y) = \frac{y}{1-y} - y = \frac{y^2}{1-y};$$

3) дисперсия числа вызовов в системе

$$\sigma_n^2 = \sum_{i=0}^{\infty} (i - \overline{n})^2 p_i = \sum_{i=0}^{\infty} i^2 p_i \cdot \left(\frac{y}{1-y}\right)^2;$$

однако

$$\sum_{i=0}^{\infty} i^2 p_i = (1-y)\sum_{i=0}^{\infty} i^2 y^i = \frac{y}{1+y} + \frac{2y^2}{(1-y)^2},$$

поэтому

$$\sigma_n^2 = \frac{y}{1-y} + \frac{y^2}{(1-y)^2} = \overline{n} + \overline{n}^2,$$

или

$$\sigma_n^2 = \frac{y}{(1-y)^2};$$

4) для определения среднего времени пребывания в системе $t_{\text{пр}}$ заметим, что в среднем за единицу времени через систему проходит число вызовов

$$\lambda = (1 - p_0)\mu,$$

тогда

$$t_{\text{пр}} = \frac{\overline{n}}{\lambda} = \frac{1}{\lambda}\frac{y}{1-y} = \frac{1}{\mu}\frac{1}{1-y};$$

5) среднее время ожидания в системе

$$\overline{t}_{\text{ож}} = \frac{\overline{n}_0}{\lambda} = \frac{n_0}{(1-p_0)\mu} = \frac{y^2}{1-q}\frac{1}{\lambda} = \frac{1}{\mu}\frac{y}{1-y};$$

6) среднее время обслуживания

$$\overline{t}_{\text{пр}} - \overline{t}_{\text{ож}} = \frac{1}{\mu};$$

7) плотность распределения и функция распределения вероятностей времени ожидания

$$f_0(t) = (1-y)U_0(t) + \lambda(1-y)e^{-\mu(1-y)t},$$

где $U_0(t)$ – дельта-функция Дирака, $U_0(t) = \begin{cases} 0, & t > 0 \\ 1, & t = 0; \end{cases}$

$$F_0(t) = 1 - ye^{-\mu(1-y)t};$$

8) плотность распределения и функция распределения вероятностей времени нахождения в системе соответственно равны

$$f_0(t) = \mu(1-y)e^{-\mu(1-y)t};$$

$$F_0(t) = 1 - e^{-\mu(1-y)t}.$$

Отметим важное свойство одноканальной СМО. При приближении значения y к единице (снизу) среднее время ожидания и длина очереди растут неограниченно, т.е. с повышением использования обслуживающего прибора снижается качество обслуживания. Это справедливо, по существу, для любой СМО и показывает ту высокую цену, которую нужно платить за использование системы в режиме, близком к ее пропускной способности ($y = 1$) (рис.2.6).

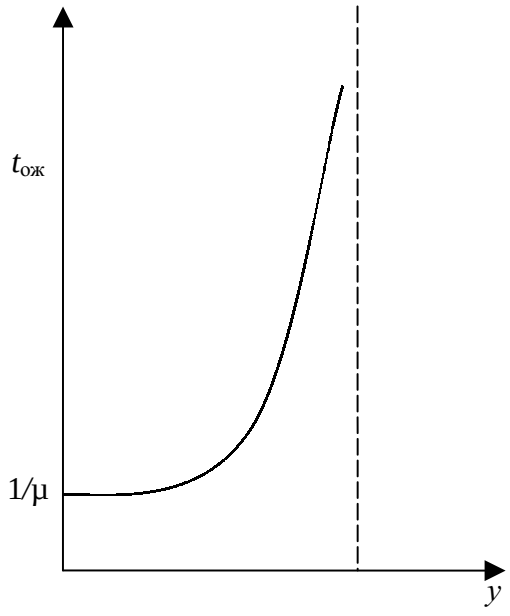


Рис. 2.6. Зависимость времени нахождения в системе от использования канала

В сетях передачи данных имеется много очередей на передачу, которые взаимодействуют друг с другом в том смысле, что поток, уходящий из одной очереди, поступает в одну или несколько очередей, возможно, после слияния с частями других очередей. С аналитической точки зрения это усложняет характер процессов поступления на

очереди, расположенные по течению потока. Трудность состоит в том, что, когда пакеты передаются за пределами первой по отношению к точке их входа в сеть очереди, интервалы между моментами поступления пакетов становятся сильно коррелированными с длинами пакетов. В результате невозможно выполнить точный и эффективный анализ, сравнимый с анализом, проведенным для таких СМО, как, например, M/1/1.

Для того чтобы разрешить эту трудность, делается предположение, что при объединении нескольких потоков пакетов в линии передачи сохраняется независимость между интервалами поступления и длинами пакетов. Было заключено, что при этом для каждой линии связи часто можно приближенно принять модель системы M/M/1 независимо от взаимодействия потока на этой линии с потоками на других линиях. Это известно как аппроксимация Клейнрока (или гипотеза о независимости); она дает довольно хорошее приближение при умеренных и больших нагрузках для сильно связанных сетей с пуассоновскими моментами поступления потоков во входные точки сети с длинами пакетов, которые распределены почти экспоненциально [6].

Рассмотренные подсети связи состоят из узлов, соединенных двухточечными (от точки к точке) линиями связи. Относительно двухточечных линий связи всегда неявно предполагается, что принятый из линии сигнал зависит только от переданного сигнала и шума этой линии. В спутниковых системах принятый сигнал является суммой претерпевших затухание сигналов, переданных множеством других узлов; эти сигналы приходят с задержкой, испытывая воздействие искажения и шума. Такие передающие среды называются средами с искусственным доступом.

В спутниковых каналах множество узлов совместно используют канал связи. Проблема состоит в том, чтобы упорядочить использование канала таким образом, чтобы каждый узел мог передавать в течение большей части времени.

В системе связи с геостационарными спутниками множество наземных станций может передавать сообщения общему приемнику спутника, и принятые им сообщения ретранслируются наземными станциями. Такие спутники часто имеют антенны с различными направлениями излучения для различных географических зон, что позволяет производить наземный прием и ретрансляцию между зонами. Кроме того, может использоваться ЧУ (или ВУ), что дает возможность производить независимый прием от различных наземных станций, находящихся в зоне действия одного и того же направления излучения антенны. Таким образом, спутниковый канал можно использовать как совокупность виртуальных двухточечных линий, при этом виртуальные линии создаются ими с помощью диаграммы направленности антенны или путем уплотнения.

Управление объемом передаваемых информационных потоков

В процессе передачи данных от источника к потребителю каждый пользователь занимает часть ресурсов (каналов, буферов, коммутационных процессоров). Если использование ресурсов сети не контролируется и не

ограничивается, то возможны падение эффективности, несправедливое распределение ресурсов и перегрузка.

Сеть считается перегруженной, если некоторое приращение предполагаемого (внешнего) трафика вызывает уменьшение эффективной производительности. Перегрузка является прямым следствием бесполезного расходования ресурсов сети. Поскольку такое расходование (в неконтролируемой сети) увеличивается пропорционально предполагаемому трафику, а не полезной производительности, должен существовать оптимальный уровень предполагаемой нагрузки, выше которого напрасный расход становится больше потенциального выигрыша в производительности. Вне такого оптимума сеть является перегруженной по двум причинам: из-за падения эффективности и возникновения блокировок.

Управление объемом передаваемых информационных потоков может осуществляться следующим образом:

1) между соседними УК путем изменения скорости передачи потоков по линии связи, соединяющей эти УК (протокол X25.2 – канальный уровень);

2) при доступе к сети (протокол X25.3 – сетевой уровень);

3) между исходящими УК и входящими УК (управление объемами потока “из конца в конец”) (протокол X25.3 – сетевой уровень).

Методы управления потоком на уровне доступа к сети связаны с ограничением входящей нагрузки, связанным с нагрузкой сети. Ограничение нагрузки является механизмом, с помощью которого сеть связи защищается от излишнего трафика путем регулирования процесса входа пакетов в сеть на основе локальных или глобальных измерений степени нагрузки. Ограничение нагрузки является результатом совместного действия протоколов нескольких уровней. Каждый из этих протоколов вносит вклад в управление потоком, в то же время выполняя другие функции.

Межузловое управление потоком (управление на уровне транзитного участка или управление на уровне передачи с промежуточным накоплением) выполняет функции предотвращения передачи трафика к перегруженному узлу.

Межузловые процедуры управления потоком отличаются друг от друга в зависимости от определения состояния нагрузки узла. Однако, предотвращая возникновение блокировок, они не исключают перегрузки сети, поскольку оказывают влияние только после того, как внутренние очереди становятся относительно большими.

Выбор межузловой процедуры может зависеть от того, используются ли механизмы виртуальных каналов или дейтограмм. В механизме виртуального канала буферы предварительно выделяются вдоль пути – таким образом, нет опасности возникновения прямых или косвенных блокировок передачи с промежуточным или косвенным накоплением, в то время как при дейтограммном обслуживании такие блокировки могут возникнуть.

Управление потоком по виртуальному каналу осуществляется только в сетях с виртуальными каналами. При управлении потоком по виртуальному каналу сеть может отказать в соединении, если имеющиеся свободные ресурсы

не могут удовлетворить требованиям, предъявляемым пользователем. После установления соединения выполняется избирательное управление по каждому соединению. В частности, если нагрузка превышает ресурс (например, из-за отказа), пути тех соединений, в которых расходуется этот ресурс, могут быть быстро отслежены в обратном направлении до своих источников и отрегулированы на уровне сетевого доступа, благодаря чему возникновение перегрузки не допускается еще до принятия специальных мер (как в случае дейтограммной сети).

Механизм управления виртуальным каналом удовлетворяет критерию предотвращения нагрузок, минимизирует несправедливое распределение ресурсов, но по критерию эффективности уступает дейтограммным сетям для соединений с пульсирующим трафиком, хотя более эффективен в случае стационарного потока (например, пересылка файлов, цифровая передача речи, факсимильная передача).

Методы управления потоком на уровне ввода-вывода основаны на понятии “окна” и сводятся к управлению пакетами, передаваемыми за период одного “окна”. Под шириной “окна” понимается допустимое число передаваемых пакетов с исходящего УК до получателя пакета, подтверждающего их прием в УК-адресата. Ширина, или размер “окна”, является основным параметром, характеризующим метод управления этого уровня.

Блокировка в сетях с КП

Основными факторами, приводящими в условиях перегрузки к резкому снижению пропускной способности сети, являются блокировки (запреты) на использование ресурсов сети (различных видов буферной памяти на узлах коммутации и у абонентов).

Блокировки возникают в тех случаях, когда несколько процессов транспортировки пакетов оказываются в состоянии неопределенного взаимного ожидания, и каждый из них не может продолжаться, поскольку некоторый другой процесс из этого множества либо не может освободить хотя бы один буфер, либо не может получить некоторое управляющее сообщение (квитанцию, отклик).

Главной причиной появления блокировки, наряду с несовершенством применяемых протоколов и алгоритмов управления, является ограничение на физически доступные ресурсы сети, в первую очередь – на объем буферной памяти УК. Потенциально возможны следующие виды блокировки с КП.

Прямая блокировка передачи возникает в двух смежных УК, когда буфер одного из них заполнен пакетами, предназначенными для другого. Эта блокировка предотвращается ограничением числа буферов, которые могут быть связаны с одной выходящей очередью.

Косвенная блокировка передачи возникает на участке сети, содержащем более двух УК и образующем логический цикл, в котором пакеты направляются узлом-адресатом, находящемся на расстоянии двух или более переприемов. Когда все буферы в УК заняты этими пакетами, передача полностью

блокируется. Очевидно, простое ограничение выходной очереди не предотвращает эту блокировку. В сетях с ВК косвенную блокировку можно предотвратить, если на этапе установки виртуального канала выделять буфера в узлах для каждого направления передачи на время сеанса связи. Для предотвращения косвенной блокировки предпочтительнее использовать метод разделения буферов на группы, доступные поступающим пакетам, в зависимости от числа уже проделанных переприемов в сети.

Блокировки сборки сообщения могут возникнуть, когда сообщение разбивается на пакеты в узле отправителя, а собирается в узле-адресате. Блокировка наступает, когда все буферы сборки узла-адресата заняты несобранными сообщениями, и новый пакет не может быть присоединен ни к одному из них. Такая ситуация предотвращается резервированием буферов для сборки сообщений.

Блокировка всевозможных квитанций возникает в узле отправителя, буфера которого заполнены копиями передаваемых пакетов, а пакеты с квитанциями, освобождающими эти буфера, не могут войти в узел. Для предотвращения этой блокировки следует избегать вкладывания квитанций в пакеты с данными и принимать отдельные служебные пакеты на отдельные буфера.

Блокировка, вызванная приоритетностью потоков, возникает, когда все буфера УК заняты низкоприоритетными пакетами, ожидающими другие низкоприоритетные пакеты для сборки сообщений или для извлечения из них квитанций и уничтожения копий пакетов. Но эти пакеты могут быть блокированы в соседних УК высокоприоритетными пакетами, ожидающими передачи в данный УК. Поэтому в сетях необходимо разделение буферной памяти в зависимости от приоритетности пакетов.

В целом блокировки, связанные с ожиданием свободных буферов, предотвращаются с помощью различных методов ранжирования буферов и организаций очередей пакетов к ресурсам. Ограничения, налагаемые на распределения буферов, позволяют локально регулировать нагрузку в сети и предотвращать некоторые типы блокировок, а организация очередей позволяет максимизировать степень использования сетевого ресурса.

Модели ограничения доступа в сеть

Основная задача динамического управления объемом потока информации при доступе к сети состоит в адаптации объема входящего потока к создавшимся условиям на сети, т.е. предохранение от перегрузок сети непосредственно.

Система управления доступом в сеть при наличии нагрузки должна обеспечить ограничение поступающей в сеть нагрузки, а при устранении перегрузки – частичное или полное снятие этого ограничения.

Наиболее просто задачи управления доступом решаются для сети КП с виртуальными каналами. На таких сетях объем поступающей в сеть нагрузки ограничивается числом виртуальных каналов, которые можно образовывать на

сети. Поступающей в сеть заявке на передачу дается отказ, если на исходящем УК отсутствуют свободные логические каналы или на сети невозможно установить виртуальный канал от исходящего УК к узлу назначения из-за отсутствия свободных логических каналов хотя бы на одном транзитном участке пути, по которому устанавливается этот виртуальный канал.

При дейтограммном режиме задачи управления доступом решаются значительно сложнее. Управление объемом поступающей в сеть нагрузки может основываться на учете локальной нагрузки (нагрузки БЗУ входящего УК) при глобальной перегрузке (перегрузке всей сети, в том числе переполнение всех БЗУ в УК сети). При этом в первом случае объем поступающей нагрузки может быть ограничен только на одном УК, где возникла перегрузка БЗУ, или наряду с ним еще на нескольких с ним УК. Во втором случае объем поступающей в сеть нагрузки будет ограничен на всех УК.

Кроме этих двух, возможны различные промежуточные случаи. Например, объем поступающей нагрузки может быть ограничен лишь для нескольких направлений (к определенным УК) из-за перегрузки БЗУ входящих или транзитных УК.

К наиболее характерным методам управления доступом относятся изоритмический метод, метод ограничения объема входного БЗУ и метод посылок.

Изоритмический метод основан на наличии циркуляции в сети так называемых пермитов (разрешений) [14, 19], которые наделяются как бы билетами передачи по сети пакетов. Пакет с исходящего УК будет передан по сети только в том случае, если на УК имеется хотя бы один пермит. Изоритмический метод, несмотря на его простоту, является довольно эффективным методом устранения перегрузок сети и возникновения тупиковых ситуаций, особенно в отсутствии системы управления потоком. Однако этот метод приводит к заметному снижению пропускной способности сети в условиях неоднородных тяготений между различными парами УК (т.е. в условиях перекосов нагрузки).

Метод ограничения объема входного БЗУ обеспечивает управление объемом поступающей в УК нагрузки в зависимости от загрузок его БЗУ. Данный метод в отличие от глобального изоритмического метода относится к методам локального управления объемом поступающей в сеть через УК нагрузки, так как он основывается только на нагрузке БЗУ одного данного УК. Однако перегрузка на каком-либо одном УК, как правило, является следствием общей перегрузки на сети или достаточно большом ее участке. При указанном методе определяется некоторый порог загрузки БЗУ, при достижении которого ограничивается объем поступающей через данный УК нагрузки в сеть, т.е. предусматривается определенное преимущество транзитной нагрузки перед поступающей, так как транзитный поток уже использовал определенные ресурсы сети.

Метод посылок обеспечивает ограничение поступающей в сеть нагрузки при перегрузке пути в оптимальном маршруте, выбранном системой

адаптивной маршрутизации. Линия связи, соединяющая два инцидентных УК, определяется как перегруженная, если ее использование превысило некоторый порог. В качестве порога взято использование линии на 80%. Весь путь считается перегруженным, если в ней перегружен хотя бы один из транзитных участков. Информация о перегрузке линии связи передается по сети вместе с маршрутной информацией о длине пути, определяемой числом транзитных участков, на основе которой определяется план распределения информации (ПРИ) (по методу рельефов). На каждом УК известно о загрузке кратчайшего пути к каждому входящему УК. Если на УК поступает входящий пакет, предназначенный для УК, кратчайший путь к которому перегружен, то этот пакет теряется (т.е. на исходящем УК он не принимается). Если в таких же условиях на УК поступает транзитный пакет, то он не теряется, а передается дальше. Однако при этом на данном УК формируется некоторый управляющий пакет, который передается на исходящий УК. Эта посылка информирует исходящий УК о перегрузке пути к входящему УК.

При получении на исходящем УК такой посылки прием в сеть пакетов, адресованных к входящему УК, прекращается. Снятие запрета передачи пакетов к входящему УК произойдет через некоторое время, если вместе с маршрутной информацией не будет больше поступать такого рода посылки о перегрузке.

Модели межузлового управления потоком

Процедуры управления потоками информации в сетях с КП в значительной степени определяются протоколами различных уровней сети.

Так, управление на уровне от УК до УК осуществляется децентролизованно в каждом узле путем ограничения нагрузки, поступающей в узел, когда превышается некоторый порог, ограничивающий максимальную длину очереди. Функция контроля длины очереди, сброса и повторной передачи пакетов выполняется протоколом управления каналом данных. Существуют несколько моделей этого уровня управления потоками. К ним относятся [11, 13] схемы ограничений очереди к каналу, схемы буферного класса и схемы подключения буферов, схемы, специфические для сетей с виртуальными каналами.

Таким образом, управление потоком внутри сети по выбранным маршрутам сводится к рациональной организации очередей пакетов при возникновении конкуренции за ресурс и распределение буферной памяти в узлах коммутации.

В первой группе моделей (*группа А*) накладываются ограничения на длину очереди к каждому исходящему каналу. Если очередь достигла предела, то пакеты сбрасываются. Схема ограничения очереди к каналу имеет несколько модификаций:

1. Модель полного разделения.

В этом случае емкость буфера равномерно делится между всеми исходящими линиями (классами потоков). Поэтому число пакетов n_i, стоящих в

очереди для передачи по *i*-ой исходящей линии, лежат в пределах $0 \le n_i \le B/N$, где B – размер буфера УК; N – число исходящих линий.

2. Модель разделения буфера с верхними пределами очередей.

В этом случае для *i*-й исходящей линии (потока *i*-го класса) отводится верхний предел очереди, равный $b_{\max} > B/N$. Число пакетов n_i в очереди по *i*-ой исходящей линии находится в пределах $0 \le n_i \le b_{\max}$ для $\forall i, \sum n_i \le B$.

3. Модель разделения буфера с гарантированной минимальной емкостью буфера для исходящей линии.

В данном случае для *i*-й исходящей линии гарантируется минимальная емкость буфера в размере $b_{\min}$, обычно $b_{\min} < B/N$. Оставшаяся емкость $B_{\text{ост}} = B - Nb_{\min}$ может распределятся между всеми исходящими линиями по мере необходимости. При этом очевидно, что $\sum \max(0; n_i - b_{\min}) \le B_{\text{ост}} = B - Nb_{\min}$.

4. Модель распределения по минимальному размещению и максимальной очереди (совмещение моделей 2 и 3).

Из этих четырех разновидностей модели первого вида не допускают динамического управления объемом потока на линии. Наибольшие возможности для динамического управления представляют модели четвертой модификации, в которых параметры $b_{\max}$ и $b_{\min}$ могут изменяться во времени и принимать различные значения для каждой исходящей линии (поток определенного класса) в зависимости от характера и интенсивностей поступающих в УК потоков и общей ситуации по сети.

Во второй группе моделей (*группа Б*) учитывается число пройденных потоков транзитных УК. Модели этой группы названы схемами буферного класса. В этом случае буферы распределяются в соответствии с длиной пути и числом ветвей, составляющих его, от исходящего узла до рассматриваемого.

Существуют схемы виртуального канала (*группа В*).

Другим параметром, характеризующим уровни управления от УК до УК, является схема подключения буферов. Существуют следующие способы подключения накопителей на УК: фиксированное разделение буферов между классами нагрузки, различающиеся по исходящим очередям, для которых они предназначены; распределение буферов пропорционально величине нагрузки каждого класса, динамическое регулирование буферных ограничений в соответствии с относительными флуктуациями нагрузки.

Управление потоком для сетей с виртуальными каналами

Принцип реализации метода управления потоком для сетей с виртуальными каналами основан на ограничении некоторого значения (ω) числа пакетов, передаваемых по одному логическому каналу на следующий УК. При этом принятый в качестве стандарта в протоколе X25.2 принцип ограничения передаваемых пакетов по логическому каналу (ЛК) основан на использовании “окна”. Размер “окна” равен двум пакетам. В случае

использования спутниковых линий из-за большого времени распространения необходимо выбрать большие размеры “окон” с целью беспрепятственной передачи в условиях малой нагрузки. Размер “окна” может быть также установлен равным 7 или 127. Управление потоком осуществляется в обоих направлениях. При реализации “оконной” стратегии каждый пакет должен содержать трехбитовый порядковый номер и трехбитовый прицепной номер. Если размер “окна” равен 127, то эти номера имеют длину 7 бит. Порядковый номер указывает на позицию пакета в “окне” отправителя, а прицепной номер равен номеру, который отправитель ожидает получить в качестве номера следующего пакета. Прицепной номер играет роль разрешения, позволяющего приемнику продвинуть соответствующее “окно”.

Управление потоком вдоль виртуальной цепи относиться к стратегии оконного управления потоком между каждой парой последовательных узлов. В этой стратегии имеется отдельное “окно” для каждого логического канала и пары смежных УК на пути виртуальной цепи. Основная идея узловой схемы [6] состоит в том, что из двух последовательных узлов на пути виртуальной цепи один из них (приемник) может избежать накопления большого числа пакетов в своей памяти путем уменьшения скорости, с которой он возвращает разрешение другому узлу (передатчику). В самой распространенной стратегии у приемника имеется буфер, в который можно записать ω пакетов для каждой виртуальной цепи, и приемник возвращает разрешения передатчику только тогда, когда в его ω-пакетном буфере имеется свободное место для записи еще одного пакета. Как только пакет покинет ω-канальный буфер, он либо будет отдан пользователю вне подсети, либо пойдет к следующему узлу по пути виртуальной сети.

Динамическое управление объемом потока, передаваемого по линии, состоит в адаптации размера “окна” для каждого ЛК в зависимости от поступающей на эту линию нагрузки. Простейшим методом динамического управления размером “окна” для каждого ЛК, а следовательно, и объемом потока на линии является следующий метод [12].

Пусть в данной линии связи возможно образование не более N логических каналов ЛК$_1$,…,ЛК$_N$. Введем n пороговых ограничений $\zeta_{\xi_1},...,\zeta_{\xi_N}, \zeta_{\xi_i} \in \{0,1\}$, таких, что при числе η ($0 \le \eta \le N$), занятых в линию ЛК, равном или больше ξ_i, но меньшем ξ_{i+1}, порог ζ_{ξ_i} принимает значение единицы, а в остальных случаях – нуля, т.е.

$$\zeta_{\xi_i} = \begin{cases} 1, \text{если } \xi_i \le \eta < \xi_{i+1} \\ 0 \text{ в остальных случаях.} \end{cases}$$

Таким образом, если число занятых в линии ЛК равно нулю или меньше ξ_1, то порог ζ_{ξ_0} равен единице, а остальное – нулю. При $\xi_1 \le \eta < \xi_2$ имеем $\zeta_{\xi_1} = 1$, а остальные пороги равны нулю и т.п. Максимальный порог ζ_{ξ_N} равен единице только в единственном случае – при $\eta = N$, т.е.

$$\zeta_{\xi_N} = \begin{cases} 1, \text{если } \xi_N = N; \\ 0 \text{ в остальных случаях,} \end{cases}$$

так как число ЛК в линии не может быть больше *N*. Сопоставляем пороги с размерами окна таким образом, что при увеличении $\eta = \xi_i$ (число занятых в линии ЛК) изменяется соответствующий порог ζ_{ξ_i} и уменьшается размер “окна” ω и наоборот. Тогда при изменении загрузки и изменении порогов, размер “окна” и, следовательно, объем потока в каждом ЛК будет динамически меняться.

Для устранения возникающего в данном случае непрерывного изменения размеров окна целесообразно ввести две разновидности порогов: величину ζ_{ξ_i} оставить в качестве порога при увеличении нагрузки, а при уменьшении нагрузки (т.е. при изменении числа занятых ЛК) ввести другое значение порога ζ'_{ξ_i}, отличающееся от ζ_{ξ_i} на $\Delta\xi_i$, т.е.

$$\zeta'_{\xi_i} = \zeta_{\xi_i} - \Delta\xi_i, \; 0 < \Delta\xi_i < \zeta_{\xi_{i-1}}.$$

Выбор значения порога и соответствующего ему “окна” непосредственно связан с распределениями буферов между входящими ЛК данной линии. Выбор емкости буфера, отводимого для каждого ЛК может осуществляется методом резервирования определенных емкостей буферов для каждой исходящей линии [12] (методы группы *А*). Использование ВК с соответствующими методами динамического управления потоками на каждом ЛК практически полностью исключает возможность возникновения тупиковых состояний, что, в свою очередь, исключает необходимость использования методов группы Б.

2.7. Алгоритмы маршрутизации в сетях КП

Под алгоритмами маршрутизации подразумевают протокол сетевого уровня, который управляет пакетами при их движении по подсети связи до требуемого места назначения.

В зависимости от принципа передачи сообщений от абонента-источника к абоненту-адресату различают несколько модификаций КП, основными из которых являются режим виртуальных сообщений и дейтограммный режим.

В *режиме виртуальных сообщений* абонент-источник перед тем, как послать сообщение абоненту-адресату, посылает специальный пакет виртуального вызова с информацией о величине посылаемого сообщения с целью резервирования ресурса (памяти) абонента-адресата для приема пакетов всего сообщения. Пакет виртуального вызова одновременно фиксирует маршрут передачи пакетов одного и того же сообщения (устанавливает виртуальный канал) и резервирует ресурсы УК для приема следующих друг за другом пакетов этого сообщения. По окончании сеанса связи виртуальный канал разрушается.

При *дейтограммном способе передачи* пакетов абонент-источник посылает пакеты сообщения без предварительного уведомления абонента-адресата, при этом пакеты одного и того же сообщения могут передаваться по различным маршрутам.

В соответствии с архитектурой цифровых сетей маршрутизация осуществляется на сетевом уровне. При этом процессы маршрутизации, в свою очередь, можно представить в виде трех уровней [11]:

1) передачи пакетов по выбранному маршруту (пути);

2) выбора пути передачи по маршрутным таблицам;

3) коррекции матриц маршрутов.

Методы коррекции матриц для сетей с КП часто называют методами адаптивной маршрутизации.

При дейтограммном режиме необходимость в выполнения процессов первого уровня отпадает, так как перед передачей информационных пакетов никакой маршрут не выбирается и не устанавливается определенный виртуальный канал между источником и потребителем информации. Процесс аналогичен передаче информации по установленному каналу.

Процессы второго уровня (выбор направления передачи по таблице маршрутов) осуществляется в процессе передачи информационных пакетов по сети в случае дейтограммного режима или с помощью специальных пакетов вызова при установлении виртуального канала. Процессы второго уровня аналогичны установлению соединения.

Процессы первого и второго уровней характерны для фиксированной маршрутизации в сетях с КП.

Процессы третьего уровня относятся к динамическому управлению распределением потока пакетов. Коррекция матриц маршрутов как при дейтограмном режиме, так и при коммутации пакетов с виртуальными каналами может выполняться аналогичными методами [12].

Первоначально принципы построения и функционирования систем адаптивного (динамического) распределения информационных потоков были сформулированы в 1964 г. В. Г. Лазаревым [19]. Несколько позднее на их основе был разработан метод распределенного управления выбором путей передачи информации, получивший название метода рельефов. Однако несмотря на тот факт, что научный приоритет идей адаптивного распределения принадлежит нашей стране, практическая реализация была осуществлена в США в сети ARPA (Adranced Research Project’s Agency) Управления перспективных научных исследований (DARPA) США в 1968 г.

Основным критерием оптимальной маршрутизации в сетях с КП является среднее время задержки пакета в сети, поэтому основной проблемой при адаптивной маршрутизации является оценка при передаче пакетов по различным маршрутам.

В сетях с КП в режиме виртуальных соединений на каждом УК маршрутизация осуществляется с помощью таблицы путевых номеров виртуальных каналов. Указанная таблица составляет номера линий, связанных с

данным УК, а номера логических каналов в этих линиях с номером виртуального канала. С помощью этих таблиц осуществляется выбор пути передачи пакетов в УК.

Выбор маршрута устанавливаемого виртуального канала может осуществляться централизованно или быть распределенным. При *централизованном способе* информация о состоянии сети (сведения о нагрузке, наличии свободных логических каналов в линиях связи, данных очередей в УК и др.) поступает в ЦУС или специальный центр маршрутизации. На основе этих данных ЦУС, используя алгоритм выбора кратчайших путей по критерию задержки пакета в сети, определяет по заявке от абонента-источника маршрут прохождения виртуального канала. Сведения о маршруте в виде управляющей информации передаются в УК, через которые проходит образованный Виртуальный канал. На основе этой информации заполняется таблица путевых номеров указанных УК.

Если виртуальный канал невозможно установить, то ЦУС посылает абоненту-источнику отказ в установлении канала, либо запрос ставит в очередь на ожидание. После окончания сеанса связи ЦУС осуществляет разъединение виртуального канала. Таким образом, выбор оптимального маршрута в случае централизованного управления в сетях с КП в режиме виртуального соединения осуществляется в ЦУС на основе методов потока кратчайшего пути, в частности, матричных методов. При этом каждой линии назначается определенный вес, зависящий от стоимости линии, ее длины, задержки во времени при передаче сигналов, нагрузки в линии, числа ошибок и др.

При *распределенном способе* выбора маршрутов виртуальных каналов на каждом УК имеется матрица маршрутов, определяющая порядок выбора исходящих линий для связи с УК, в который включен абонент-адресат. При поступлении пакета-вызова на определенный УК, в заголовке которого указан УК-адресат, по маршрутной матрице в порядке предпочтения отыскиваются свободные логические каналы в исходящих линиях, при этом номер линии и номер логического канала заносятся в таблицу путевых номеров. Таким образом, таблица путевых номеров УК при каждом запросе на установление виртуального канала записывается в соответствии с маршрутной матрицей и с учетом занятости определенных логических каналов.

В сетях с КП существуют также гибридные методы маршрутизации, сочетающие в себе элементы распределенной и централизованной маршрутизаций.

2.8. Классификация методов маршрутизации

Наиболее целесообразно в основу классификации методов маршрутизации положить одно из фундаментальных понятий теории адаптивного распределения информационных потоков – план распределения информации (ПРИ).

Если для УК$_i$ (i = 1, 2, …, n, где n – число узлов в сети) задан состав доступных исходящих направлений и порядок их выбора при установлении связи к любому из других узлов в сети, т.е. дана матрица маршрутов М$_i$, то говорят, что для УК$_i$ задан *план распределения информации*. Если такой план задан для каждого узла в сети, то считают, что он задан для всей сети [12, 19].

Методы маршрутизации, не производящие коррекции ПРИ в процессе функционирования сети, называются *статистическими*. Методы статистической маршрутизации делятся на методы без использования обходных направлений и методы с их использованием. В первом случае между двумя различными УК имеется возможность установления соединения только по одному заранее заданному пути. Во втором случае имеется возможность использовать не один, а несколько возможных направлений установления соединений, но порядок их выбора не меняется в процессе функционирования сети.

Методы маршрутизации *с использованием обходных путей* можно подразделить на *методы случайной маршрутизации* и методы *детерминированной статической маршрутизации*. В первом случае суть заключается в формировании поискового формата в случае поступления заявки на установление соединения и передачей его по одному из исходящих направлений, выбранных случайным образом. Во втором случае производится выбор направления установления соединения согласно упорядоченному множеству возможных направлений, определяющему порядок их просмотра.

Как методы случайной, так и детерминированной статической маршрутизации можно разделить на методы без перепоиска и с перепоиском направлений дальнейшего установления соединений. Суть метода статической маршрутизации с перепоиском заключается в следующем. Если в процессе установления соединения заявка достигла транзитного УК, из которого нет свободных (доступных) исходящих направлений с целью дальнейшего установления соединения, то допускается возврат данной заявки на предыдущее УК, где делается попытка установить данное соединение по альтернативному направлению.

Методы адаптивной маршрутизации наиболее эффективны в условиях изменяющихся тяготений между УК сети и (или) поражений элементов сети.

Методы адаптивной маршрутизации можно подразделить на следующие:

– детерминированной адаптивной маршрутизации;

– статистической маршрутизации;

– комбинированной маршрутизации.

К методам *детерминированной адаптивной маршрутизации* относятся такие методы, при которых план распределения информации (ПРИ) корректируется в соответствии с состоянием сети в данный момент времени. К методам статистической маршрутизации – такие, при которых ПРИ корректируется на основе предыстории об обслуживании предыдущих вызовов. К наиболее важным характерным методам детерминированной адаптивной маршрутизации относятся метод *лавинной маршрутизации* (в отечественной

литературе известен как волновой метод), метод *рельефов* и *матричный* метод. К основным статистическим методам маршрутизации принадлежат игровой метод, метод рельефов [карандаш], а также *вероятностно-игровой* [12, 19].

В зависимости от источника принятия решения о маршруте подлежащих передаче информационных потоков различаются:

– методы централизованной маршрутизации, в которых решение о маршрутах передачи информации принимается так называемым Главным администратором сети;

– методы локальной маршрутизации, в которых решение о маршрутах передачи информационных потоков принимается региональными службами;

– методы распределенной маршрутизации, где решение о маршрутах передаваемой информации принимается в каждом узле связи отдельно, по заранее известному алгоритму на основании межузлового обмена служебной информацией, реализуемой службой сигнализации;

– методы иерархической маршрутизации, суть которых заключается в зонировании сети связи, выделении узлов связи в каждой зоне, выполняющих функции региональных центров, и в организации межузловой сети связи;

– методы изолированной маршрутизации, в которых решение относительно выбора маршрута передачи информации или коррекции маршрутных таблиц основывается лишь на собственных измерениях каждого УК сети;

– методы гибридной маршрутизации, являющиеся композицией упомянутых выше методов (например, процедура дельта-маршрутизации [6,19].

Метод *дельта-маршрутизации* заключается в том, что решение о маршруте передачи сообщения принимается на каждом УК сети в процессе установления соединения на основании информации о маршрутах, получаемой от центра управления сетью (ЦУС), и локальной информации о длинах очередей по направлениям связи, включенным в данный узел связи. При этом предполагается, что ЦУС использует общесетевую информацию для формирования плана распределения информации, а дельта-фактор является степенью свободы выбора маршрута передачи, предоставляемого каждому УК.

Все указанные выше методы адаптивной маршрутизации могут быть разовыми или групповыми. К разовым относятся такие методы, в которых коррекция ПРИ осуществляется после поступления каждой заявки на соединение. Однако вряд ли можно предполагать, что в реальных сетях связи по истечении времени от возникновения одной заявки на установление соединения до другой произойдут существенные изменения. В связи с этим большее предпочтение отдается групповым методам, при которых ПРИ корректируется после поступления некоторой группы заявок.

По принципам изменения маршрутов установления соединений методы адаптивной маршрутизации могут быть разделены на синхронные и асинхронные. Метод *синхронной маршрутизации* позволяет корректировать маршруты передаваемых сообщений лишь в заданные моменты времени.

Методы *асинхронной маршрутизации* позволяют корректировать маршруты передачи информации в произвольные моменты времени.

2.9. Выбор алгоритма маршрутизации

Существенными характеристиками любого алгоритма адаптивной маршрутизации являются:

1) способ рассылки информации, используемой для построения маршрутных матриц УК;

2) период обновления маршрутных матриц.

Затраты на адаптацию складываются из расходов на сбор и рассылку служебной информации о состоянии сети и на вычислительные ресурсы для расчета маршрутных таблиц. Чем больше сеть подвержена резким колебаниям нагрузки и частым изменениям структуры, тем динамичнее должен быть алгоритм маршрутизации и тем чаще возникает необходимость в обмене служебной информацией о состоянии сети, что приводит к отвлечению значительных ресурсов сети.

Для применения оптимальных маршрутных решений на узлах необходимо располагать достоверной информацией о ситуации в сети. Эта информация может включать в себя загруженности узлов коммутации сети, длины очередей (например, в пакетах, заявках, блоках и т.п.) по направлениям связи в узлах, состояние каналов связи и т.д.; степень детализации информации зависит от конкретного алгоритма маршрутизации. Основная сложность, возникающая при этом, заключается в соизмеримости скорости изменения ситуации в сети со скоростью передачи информации об этих изменениях. На узлах при принятии маршрутных решений информация о состоянии сети оказывается устаревшей. Поэтому в большинстве случаев пользуются не мгновенными значениями контролируемых параметров, а их усредненными значениями за некоторый промежуток времени. Это связано с тем, что чрезмерно быстрая реакция на мгновенные колебания нагрузки приводит к неустойчивой работе алгоритма маршрутизации и большим издержкам на обмен служебной информацией [11, 13]. В общем можно сказать, что вопросы: какие параметры сети контролировать, как часто проводить обмен служебной информацией и что конкретно она должна включать в себя, сколько это потребует связных и вычислительных ресурсов сети, – являются ключевыми при разработке новых алгоритмов адаптивной маршрутизации.

Из качественного рассмотрения адаптивных детерминированных методов маршрутизации следует, что каждый из них ориентирован на решение определенного класса задач. Так основным достоинством адаптивных детерминированных методов (например, метода рельефа) является хорошая адаптация к структурным изменениям, но они малоэффективны при возникновении функциональных изменений (перекосов нагрузки и перегрузок). В то же время статистические методы (например, игровой), обладая достаточной эффектив-

ностью при отслеживании функциональных изменений, малоэффективны для адаптации к изменению топологии сети.

Для сетей, в равной степени характеризуемых как структурными, так и функциональными изменениями, задача совмещения в одном комбинированном методе достоинств детерминированных и статистических методов и нивелирование присущих им недостатков является актуальной. К числу таких методов относится *стохастическо-детерминированный* метод.

Несмотря на определяющую роль маршрутизации в деле обеспечения эффективной работы сети связи, в силу сложности процессов, протекающих в системах связи, ни одну из маршрутных стратегий нельзя назвать “наилучшей” вообще. Выбор той или иной стратегии маршрутизации необходимо проводить с учетом особенностей контроля и функционирования конкретной рассматриваемой системы связи, включая характеристики и виды передаваемого трафика, размерность сети и ее топологию, объем памяти и производительность управляющих вычислительных комплексов узлов сети, пропускную способность линии связи.

В сетях большой размерности (более 100 узлов) обычные стратегии маршрутизации оказываются неэффективными, так как возросший размер маршрутных таблиц (пропорционально числу узлов) обусловливает более высокие накладные расходы, связанные передачей по линии значительного объема служебной информации для коррекции маршрутных таблиц и большую загрузку памяти. Одним из путей решения данной проблемы является использование иерархической маршрутизации. При этом сеть разбивается на области, внутри которых маршруты вычисляются на основе региональных стратегий [1, 19]. Области объединяются посредством межрегиональной сети. В этом случае маршруты, соединяющие пользователей различных областей, представляют собой композицию трех локально оптимальных маршрутов (двух областных и одного межрегионального).

Литература

1. **Советов Б. Я., Яковлев С. А.** Построение сетей интегрального обслуживания. – Л.: Машиностроение, 1990. – 332 с.
2. **Мамонтов Н. Г., Шаль В. И., Белявская Г. Г., Сотников А. Д.** Применение ЭВМ для расчета систем распределения информации. – Л.: ЛЭИС, 1989. – 40 с.
3. **Ланко А. А., Див В. В. Журавин А. И.** Коммутация в сетях связи. – Изд-во МО СССР, 1988. – 373 с.
4. **Лаукс Г. Я., Осокина Н. Н.** Теория телетрафика. – Рига: РПИ им. А.Я. Пельше, 1983. – 123 с.
5. **Клейнрок Л.** Вычислительные системы с очередями. – М.: Мир, 1979. – 600 с.
6. **Бертсекас Д., Галлагер Р**. Сети передачи данных. – М.: Мир, 1989. – 544 с.

7. **Шнеис-Шнеллс М.А**. Численные методы теории телетрафика. – М.: Связь, 1974. – 232 с.
8. **Артюхин И. И., Буланов А. В.** Элементы теории телетрафика. – М: ВЗЭИС, 1979. – 51 с.
9. **Клейнрок Л., Сильвестр Дж.** Методы многократного использования пространства в многопролетных пакетных радиосетях // ТИИЭР. – Т. 75. – №1. – 1987. – С.187 – 200.
10. **Теория** сетей связи / Под ред. В. Н. Рогинского. – М.: Радио и связь, 1983. – 248 с.
11. **Агаян А. А., Захаренко Г. Д., Крутникова Н. П.** Алгоритмы функционирования интегральных цифровых сетей связи. – Л.: ЛО ин-та повыш. квалиф. руководящих работников и специалистов, 1986. – 59 с.
12. **Лазарев В. Г., Лазарев Ю. В.** Динамическое управление потоками информации в сетях связи. – М.: Радио и связь, 1983. – 216 с.
13. **Арипов М. Н., Присяжнюк С. П., Шарифов Р. А.** Контроль и управление в сетях передачи данных с коммутацией пакетов. – Ташкент: ФАН, 1988. – 160 с.
14. **Протоколы** и методы управления в сетях передачи данных / Под ред. Ф.Ф. Куо. – М.: Радио и связь, 1985. – 480 с.
15. **Присяжнюк С. П., Мигалин В. Н.** Овчинников Т.Р. Интегральные сети АСУВ. Системы коммутации пакетов. – Л.: ВИКИ им. А.Ф. Можайского, 1989. – 93 с.
16. **Блох Э. Л. и др.** Модели источника ошибок в каналах передачи цифровой информации. – М.: Связь, 1971. – 101 с.
17. **Коригиев Л. П., Королев В. Д.** Статистический контроль каналов связи. – М.: Радио и связь, 1989. – 240 с.
18. **Блох Э. Л.** Построение и анализ систем передачи информации. Сборник статей. – М.: Наука, 1980. – 140 с.
19. **Обельченко С. Е.** Разработка специальных систем связи. Методы адаптивного распределения информационных потоков во вторичных сетях связи с коммутацией каналов. – М.: Ин-т повыш. квалиф. руководящих работников и специалистов, 1988. – 91 с.
20. **Агаян А. А., Захаренко Г. П.** Оптимизация структур цифровых сетей связи и технического обслуживания. Ч.I. – М.: Ин-т повыш. квалиф. руководящих работников и специалистов, 1986. – 45 с.
21. **Агаян А. А., Захаренко Г. П.** Оптимизация структур цифровых сетей связи и технического обслуживания. Ч.II. – М.: Ин-т повыш. квалиф. руководящих работников и специалистов, 1987. – 39 с.
22. **Артыкова А. А., Лешевич В. К.** Использование алгоритма поиска кратчайшего пути при оптимизации топологии сети связи // Тез. докл. Респ. научно-технич. конф. “Автоматизированный контроль и повышение эффективности систем”, 3–5 июля 1985 г. Ч. I. – Ташкент, 1985. – 150 с.
23. **Рыбкин Л. В., Кобзарь Ю. В., Демин В. К.** Автоматизация проектирования систем управления сетями связи. – М.: Радио и связь, 1990. – 203 с.

24. **Лохмотко В. В., Пирогов К. И.** Анализ и оптимизация цифровых сетей интегрального обслуживания. – Минск: Навука і тэхніка, 1991. – 192 с.
25. **Печурин М. К.** Методы и модели синтеза и анализа структуры информационных сетей в период жизненного цикла. – Киев: КПИ, 1992. – 32 с.
26. **Захаренко Г. П., Иванов В. К.** Эксплуатация цифровых сетей связи. Ч. II. – М.: Ин-т повыш. квалиф. руководящих работников и специалистов, 1986. – 40 с.
27. **Иносэ Хироси** Интегральные цифровые сети связи. – М.: Радио и связь, 1982. – 320 с.
28. **Милер Б. М., Нильсон Д. Л., Тобачи Ф. А.** Проблемы проектирования пакетных радиосетей // ТИИЭР – Т. 75. – №1. – 1987. – С. 8 – 26.
29. **Борщ В. И.** Проблемы информационного обеспечения контроля функционирования систем и сетей электросвязи // Тез. докл. Респ. научно-технич. конф. “Автоматизированный контроль и повышение эффективности систем”, 3–5 июля 1985 г. Ч. II. – Ташкент, 1985. – 150 с.

ГЛАВА 3
ОЦЕНКА РАБОТОСПОСОБНОСТИ КАНАЛА

3.1. Определение условий работоспособности канала

Как отмечалось выше, характеристики канала разделяются на первичные и вторичные. В отношении параметров для оценивания состояния канала вторичные характеристики имеют ряд преимуществ по сравнению с первичными. Для них хорошо применим математический аппарат, они более быстро измеряются, их можно моделировать.

Первичные же характеристики определяют динамические явления. На их долю приходится наибольший процент всех ошибок, возникающих в передаваемом сообщении. Если первичные характеристики, представляющие интерес с точки зрения состояния канала, являются нестационарными, то в этом случае законы изменения характеристик соответствуют определенным отрезкам времени, в пределах которых их значения можно считать стационарными или медленно меняющимися (квазистационарными).

Пусть состояние канала характеризуется значениями контролируемых характеристик x_1, x_2, …, x_n. Изменение работоспособности канала можно представить как изменение целевой функции, которая имеет вид

$$S = f(x_1, x_2, \ldots, x_n).$$

Для прогнозирования состояния канала воспользуемся методом градиентного прогнозирования [1]. В этом случае функция работоспособности экстраполируется в градиентном направлении, т.е. в направлении вектора градиента функции работоспособности. Таким образом, вектор градиента определяет направление наибольшего изменения функции работоспособности. Характеристики канала меняют свои значения во времени, которые можно представить в виде $x_i = \varphi_i(t)$. В моменты времени $t_1, t_2, \ldots, t_m$, где $t_1 < t_2 < \ldots < t_m$, значение работоспособности S будет изменяться и принимать значения $x_1, x_2, \ldots, x_n$, т.е. имеем множество $\{S\}$, которое определяет пространство D.

$$\begin{aligned} S_1 &= f(x_{11}, \ldots, x_{1n}) \\ &\vdots \\ S_m &= f(x_{m1}, \ldots, x_{mn}) \end{aligned} \qquad (3.1)$$

Поскольку значение работоспособности S зависит от аргументов $x_1, x_2, \ldots, x_n$, то S можно рассматривать как вектор в многомерном пространстве. Конец многомерного вектора находится в пространстве, ограниченном гиперповерхностью. Положение гиперповерхности в пространстве определяется максимальными значениями выбранных характеристик, которые задаются в ТЗ, выбираются по экспериментальным данным или являются результатом исследования модели. Гиперповерхность разделяет пространство на две области: область допустимых исследуемых характеристик канала, которая

соответствует устойчивой работе канала и его пригодности для передачи дискретных сообщений; и область допустимых значений, которая представляет пригодность канала из-за низкой достоверности передачи сообщения.

Тенденция изменения работоспособности канала зависит от характера зависимости (3.1). Цель прогнозирования сводится к предсказанию по известным значениям характеристик x_1, x_2, …, x_n в моменты t_i значений этих характеристик в моменты t_{m+i}. Градиентное прогнозирование осуществляется в два этапа: на первом определяется направление градиента, а на втором осуществляется собственно прогнозирование. Направление вектора градиента определяется приращением значений отдельных характеристик в моменты t_m и t_{m+1} :

$$\frac{x_i(t_{m+1}) - x_i(t_m)}{t_{m+1} - t_m} \approx \frac{dx_i}{dt}.$$

Тогда вектор градиента функции состояния будет

$$\nabla \bar{S}(x) = \frac{dx_i}{dt} i_1 + \cdots + \frac{dx_n}{dt} i_n,$$

где i_1, …, n – единичные орты.

Положение конца вектора в пространстве определяется точкой. На основе математического описания гиперповерхности и используя усредненные значения положения конца вектора получим взаимное расстояние $\left|\bar{R}\right|$ в направлении вектора. Тогда время наступления отказа канала $t_{\text{ож}}$, т.е. время, когда вектор целевой функции состояния канала будет пересекать гиперповерхность, определяется выражением

$$t_{\text{ож}} = \frac{\left|\bar{R}\right|}{\nabla \bar{S}(x)}.$$

Характеристики канала имеют различную размерность, поэтому их необходимо привести к единой системе счисления, в которой они могут быть сравнимы. Такой системой является система безразмерного нормирования. Для пересчета в относительные единицы для каждой характеристики используется выражение

$$\hat{x}_{s(t)} = \frac{x_{s(t)}}{x_{s(\text{доп})}},$$

где $x_{s(t)}$ – текущее значение параметра, $x_{s(\text{доп})}$ – максимальное значение характеристики.

3.2. Контроль каналов

Параметры, характеризующие состояние дискретного канала, могут быть разделены на прямые и косвенные. *Прямыми* параметрами, определяющими непосредственно качество принимаемой информации, могут быть:

– средняя вероятность ошибочного приема двоичного элемента (бита);

– вероятность искажения кодовой комбинации (блока информации), характеризующая качество приема с учетом группирования ошибок.

К *косвенным* параметрам относятся:

– частота переспросов, искаженных ошибками блоков информации;

– средняя вероятность ошибочного приема символов;

– отношение “сигнал–помеха”;

– телеграфные искажения;

– параметры помех.

Отношение “сигнал/шум” позволяет достаточно точно характеризовать условия приема, но требует осуществления того или иного способа разделения и сигнала и помехи, что приводит к значительному усложнению устройства контроля. Однако в КВ- и УКВ-каналах этот принцип неприемлем вследствие значительных изменений (медленных и быстрых) уровня сигнала и характера помех.

Контроль состояния дискретных каналов с распределением ошибок, близким к независимому, основан на контроле одного параметра распределения, а именно частоты ошибок. Если контроль состояния производится по частоте потока ошибок двоичных элементов или искаженных комбинаций корректирующего кода (кодовых блоков), отпадает необходимость дополнительной избыточности для организации контроля и включения в аппаратуру дополнительных устройств обработки сигнала и помехи [2].

Следует отметить, что алгоритм функционирования и технические параметры устройства контроля существенно зависят от области применения, условий работы систем и характеристик используемых корректирующих кодов. При этом устройство должно быть элементом системы и должно рассматриваться как один из его функциональных узлов (подсистем) с учетом общих технических требований.

Реальные дискретные каналы характеризуются сложным групповым распределением ошибок. Поэтому естественно, что способы контроля, основанные на анализе статистики независимых испытаний, становятся неадекватными при их применении на реальных каналах. Это диктует необходимость создания устройств контроля, позволяющих в той или иной мере учесть групповой характер ошибок в каналах связи и, таким образом, повысить эффективность контроля.

При учете группового характера ошибок дискретный канал на втором уровне (уровне передачи блоков информации) описывается с помощью стационарных цепей Маркова с различным числом состояний.

Рассмотрим пример получения расчетных соотношений для построения статистических процедур, относящихся к конкретным простым Марковским цепям, а использование этих процедур – для контроля дискретных параметров.

Пусть имеется *однородная стационарная марковская цепь* с двумя состояниями и переходными вероятностями p_{ij} (i = 0, 1; j = 0, 1). Для канала с групповым образованием ошибок в качестве наблюдаемых случайных величин в задаче контроля удобно применить следующие характеристики канала:

k – общее число ошибочно принятых символов (число единиц);

v – число серий, ошибок (число серий единиц);

u – число серий без ошибок (число серий нулей) в последовательности n принятых символов (испытаний).

Если x_i – исход i-го испытания (i = 1, 2, …, n), то совокупность последовательных исходов $x_i = x_{i+1} = \ldots = x_{i+l} = 1$ образует серию единиц, если $x_{i-1} = x_{i+l+1} = 0$, аналогично образуется серия нулей.

Вероятность $W(n, k, v, u)$ появления конкретной выборки, содержащей k-единиц, $n - k$ нулей, v серий единиц, u серий нулей, т.е. функция правдоподобия может быть определена следующим образом:

$$W(n,k,v,u) = P(x_1) p_{01}^{n_{01}} \cdot p_{11}^{n_{11}} \cdot p_{10}^{n_{10}} \cdot p_{00}^{n_{i1}},$$

где $p(x_i)$ – вероятность исхода x_i в первом испытании; n_{ij} – число переходов из состояния i (i = 0, 1) в состояние j (j = 0, 1), связанное с параметрами n, k, v соотношениями:

$$\sum_{i=0}^{1}\sum_{j=0}^{1} n_{ij} = n - 1;$$

$$n_{ij} = k - v,\ n_{00} = n - k - u;$$

$$n_{01} = v,\ n_{10} = u - 1 \text{ при } x_1 = 0;$$

$$n_{01} = v - 1,\ n_{10} = u \text{ при } x_1 = 1.$$

Вероятность $P(x)$ для стационарной однородной Марковской цепи определяется выражениями:

$$P(0) = P_0 = p_{10}(p_{10} + p_{01})^{-1};$$

$$P(1) = P_1 = p_{01}(p_{10} + p_{01})^{-1}.$$

Тогда как при $x_1 = 0$, так и при $x_1 = 1$

$$W(n,k,v,u) = \frac{1}{p_{10} + p_{01}} \cdot p_{01}^{v} \cdot p_{11}^{k-v} \cdot p_{10}^{u} \cdot p_{00}^{n-k-u}. \qquad (3.2)$$

Поскольку состояние реального канала изменяется в широких пределах, переходные вероятности, по существу, являются функцией неизвестного параметра p, и, следовательно, последнее выражение можно записать в виде:

$$W\left(n,k,v,u/p\right)=\frac{1}{p_{10}(p)+p_{01}(p)}p_{01}^{v}(p)\,p_{11}^{k-v}(p)\,p_{10}^{u}(p)\,p_{00}^{n-k-u}(p).$$

Для нахождения вероятности $P\left(n,k,v,u/p\right)$ появления всех возможных последовательностей исходов, содержащих k единиц, $n-k$ нулей, v серий единиц и u серий нулей, воспользуемся выражением

$$P(n,k,v,u/p)=$$

$$=\frac{1{,}5+0{,}5(-1)^{v+u}}{p_{10}(p)+p_{01}(p)}\binom{k-1}{v-1}\binom{n-k-1}{u-1}p_{01}^{v}(p)\,p_{01}^{k-v}(p)\,p_{10}^{v}(p)\,p_{00}^{n-k-u}(p).$$

Будем рассматривать задачу оценки состояния дискретного канала по результатам испытаний как задачу проверки простой гипотезы H: $p = p_a$ (канал не пригоден для работы) против простой гипотезы $\overline{H}$: $p = p_0 < p_a$, $p_0 > 0$ (канал пригоден к работе). Для проверки этих гипотез воспользуемся критерием Неймана-Пирсона, который в случае Марковской цепи, является наиболее мощным из всех критериев с данным уровнем значимости. Составим отношение правдоподобия и сравним его с некоторым порогом

$$\Lambda=\frac{W\left(n,k,v,u/p_0\right)}{W\left(n,k,v,u/p_a\right)}\geq\Lambda^{*}. \qquad (3.3)$$

Это неравенство определяет некоторую критическую область G. Неравенство (3.3) с учетом (3.2) преобразуется к виду

$$\Lambda=\frac{p_{10}(p_a)+p_{01}(p_a)}{p_{10}(p_0)+p_{01}(p_0)}\left[\frac{p_{01}(p_0)}{p_{01}(p_a)}\right]^{v}\times$$

$$\times\left[\frac{p_{11}(p_0)}{p_{11}(p_{01})}\right]^{k-v}\left[\frac{p_{10}(p_0)}{p_{10}(p_a)}\right]^{u}\left[\frac{p_{00}(p_0)}{p_{00}(p_a)}\right]^{n-k-u}\geq\Lambda^{*}.$$

Неравенство, определяющее критическую область G, после преобразований будет иметь вид

$$k/K_a + v/K_b + u \le c, \qquad (3.4)$$

где

$$K_a = \log \frac{p_{00}(p_0)p_{10}(p_a)}{p_{00}(p_{01})p_{10}(p_0)} \left[\log \frac{p_{11}(p_a)p_{00}(p_0)}{p_{11}(p_0)p_{00}(p_a)} \right]^{-1};$$

$$K_b = \log \frac{p_{10}(p_a)p_{00}(p_0)}{p_{10}(p_0)p_{00}(p_a)} \left[\log \frac{p_{01}(p_a)p_{11}(p_0)}{p_{01}(p_0)p_{11}(p_a)} \right]^{-1};$$

c – константа, выбираемая из условия обеспечения заданной достоверности контроля; основание логарифма любое, больше единицы.

Необходимо также выполнение дополнительных соотношений, вытекающих из физического смысла n, k, v, u:

$n \ge 0, k \ge 0, v \ge 0$; n, k, v, u – целые числа;

$v \le k \le n$; $u \le n - k$,

если $k \le n/2$, то для $v > 1$

$u = [v - 1, v, v + 1]$, для $v = 1$ $u = 1, 2$;

если $k > n/2$, то для $u > 1$

$v = [u - 1, u, u + 1]$, для $u = 1$ $v = 1, 2$;

если $k = 0$, то $v = 0$ и $u = 1$;

если $k = n$, то для $v = 1$ и $u = 0$.

Вероятности ошибок первого и второго родов в общем виде определяются выражениями:

$$\alpha(p) = \sum_G P\left(n, k, v, u/p\right), \quad p \ge p_a;$$

$$\beta(p) = 1 - \sum_G P\left(n, k, v, u/p\right), \quad p \le p_0 \le p_a,$$

а функция мощности

$$M(p) = \sum_G P\left(n, k, v, u/p\right),$$

где $\sum_G$ – означает суммирование по всем точкам (k, v, u), принадлежащим критической области G, для определения которой необходимо найти константу c. Для этого можно воспользоваться геометрической интерпретацией критической области, которая в декартовой системе координат $0kvu$ представляет собой множество точек (k, v, u), удовлетворяющих ограничению (3.3) и попадающих внутрь или на грани треугольной пирамиды, ограниченной координатными плоскостями и наклонной плоскостью, описываемой уравнением

$$k/K_a + v/K_b + u = c.$$

При этом константа с может быть найдена из следующей системы неравенств:

$$\begin{cases} \sum\limits_{G_r} P\left(n, k, v, u/p_a\right) \le \alpha; \\ \sum\limits_{G_{r+1}} P\left(n, k, v, u/p_a\right) \ge \alpha, \end{cases}$$

где G – область, ограниченная наклонной плоскостью, проходящей через некоторую точку (k_r, v_r, u_r); G_{r+1} – область, ограниченная аналогичной плоскостью, проходящей через точку (k_{r+1}, v_{r+1}, u_{r+1}). Алгоритм поиска областей G_r и G_{r+1} описан в работе [3]. При этом константа c определяется при некотором фиксированном значении n. Окончательный выбор длины контрольной последовательности, а следовательно, и константы c производится исходя из условия обеспечения заданного значения ошибки второго рода в $\beta \ge 1 - M_0(p_0)$.

Процедура контроля на основе критерия отношения правдоподобия состоит в следующем: при заданных значениях α, β вычисляются параметры n, K_a, K_b, c; значения k, v, u, соответствующие контрольной последовательности длиной n, суммируются с вычисленными коэффициентами и проверяется соотношение (3.4). В случае выполнения неравенства канал считается пригодным для работы (принимается гипотеза $\overline{H}$), в противном случае канал бракуется (принимается гипотеза H). Эта процедура с теоретической точки зрения является наилучшей среди всех процедур с фиксированным числом испытаний в том смысле, что при заданных значениях α и n обеспечивает минимально возможную ошибку второго рода.

3.3. Контроль каналов связи

Каналы связи контролируются по таким параметрам, как уровень приема, длительность и интенсивность кратковременных перерывов связи, амплитуда и длительность импульсных помех, статистические характеристики помех, величина сдвига частоты, отклонения амплитудно-частотных и фазо-частотных характеристик, частоты и величины выбросов фазы сигнала. Одной из наиболее сложных задач контроля элементов сети является контроль качества дискретных каналов, формирующих сеть ПДС. Одним из основных методов является оценка состояний качества каналов, которые квалифицируются как работоспособное и неработоспособное состояние. Работоспособный дискретный канал обычно имеет стационарные и нестационарные состояния, а нестационарные состояния могут характеризоваться чередованием стационарных состояний или параметров во времени.

Качество дискретных каналов наряду с конструктивными данными косвенно оценивается качеством передачи информации по каналам:

– методом оценки через параметры помех;

– методом оценки через параметры сигналов;

– методом оценки через вторичные статистические характеристики сигналов (искажений элементов, импульсов дроблений, ошибок).

Результаты этих оценок используются как для установления технического состояния (диагностирования, прогнозирования) канала передачи данных, так и для повышения вероятности принимаемой последовательности сигналов.

Система технической диагностики должна состоять из *аппаратных* и *программных* средств, обеспечивающих оценку информативных диагностических признаков, по которым устанавливаются условия работоспособности систем и связи их с отказом в неисправном состоянии в контролируемых системах [3]. При этом должны быть установлены диагностические модели, позволяющие путем обработки диагностической информации выбранных контрольных точек с заданной вероятностью, глубиной и временем диагностирования распознать классы технического состояния контролируемых систем.

3.4. Измерение вероятностных характеристик искажений элементов

Методика измерений искажений элементов состоит из следующих этапов: измерение, преобразование и регистрация результатов измерений, обработка данных и оценка распределений вероятностей. Измерение искажений производится двумя способами: 1) сравнением длительности измеряемого и эталонного элементов; 2) разверткой смещений фронтов элементов на градуированной шкале измерителей. Этап преобразования и регистрации измерений зависит от способа применяемой регистрации: индикаторных устройств (стробоскоп, осциллограф, табло, счетчики) или носителей (бумага, пленка, магнитная лента). Этап обработки данных и оценки распределений вероятностных характеристик зависит от выбора аппроксимационной модели распределения вероятностей.

Методы измерений вероятностных характеристик искажений могут быть классифицированы по четырем основным признакам:

1) по классу исследуемого случайного процесса − стационарный, нестационарный и т.д.;

2) по виду оцениваемой вероятностной характеристики − моментные функции, одномерные и многомерные распределения и т.д.;

3) по типу применяемого оператора усреднения − по времени, по совокупности и т.д.;

4) по варианту реализации методика измерения – аналоговый, дискретный, аппаратный, программный и .т.д.

При этом весьма важное значение имеют степень адекватности выбранной модели случайного процесса и правильный синтез измерительного алгоритма, которые оказывают существенное влияние на погрешности результатов измерений.

Приборы, позволяющие проводить статистические исследования характеристик искажений каждого элемента или группы элементов с соответствующим распределением искажений по величине или этапам действия с заданной точностью измерений, называются *анализаторами искажений*.

Предполагается не разделять искажения на краевые искажения и дробления и называть их “массой” искажений, поскольку при измерении не представляется возможным произвести такое разделение.

Программные варианты реализации анализаторов искажений позволяют:

– анализировать вероятностные характеристики одинарных искажений не для одиночного элемента, а для группы элементов;
– исследовать характер пакетообразования искажений;
– автоматизировать процессы аппроксимации законов распределения.

3.5. Организация контроля состояния каналов связи

Цель контроля. Совокупность технических средств и организованных мероприятий, обеспечивающих процесс передачи дискретной информации, представляет собой достаточно сложную систему. Качество ее функционирования определяется множеством факторов, среди которых наиболее существенными являются надежность аппаратных средств и состояние каналов связи. Изменение режимов работы и отказы элементов, помехи в канале вызывают снижение качественных показателей информационного обмена, а иногда приводят к полному прекращению процесса передачи информации.

Для обеспечения безусловной работоспособности в сложные системы закладывается определенный “запас прочности” в виде различного рода избыточности: сигнальной, информационной, аппаратурной, структурной, временной, энергетической, частотной, функциональной, эксплуатационной и т.д. Следует различать два направления использования избыточности: борьба с отказами и помехами. Дополнительные ресурсы, предназначенные для борьбы с отказами, должны обязательно находиться в состоянии резерва, поскольку для быстрого устранения неисправности необходимо наличие резервного комплекта оборудования. Задача обеспечения требуемой надежности системы возлагается на подсистему телеобслуживания оборудования, осуществляющую

оперативный контроль состояния, коммутацию и замещение функциональных узлов при их отказах, а также сигнализацию и отображение информации о произведенной замене. На службу планово-профилактического регламента возлагается задача поддержания постоянного уровня избыточности средств, направленных на борьбу с отказами.

В то же время постоянный уровень избыточности ресурсов, обеспечивающих необходимую помехозащищенность передаваемой информации, невыгоден с экономической и функциональной точек зрения, потому что помехи, как правило, не стационарны – их интенсивность изменяется во времени. При ухудшении помеховой обстановки происходит недоиспользование пропускной способности каналов, возникают неоправданные энергетические потери. Для обеспечения высокой эффективности функционирования системы ПД необходимо, чтобы в ее состав входила подсистема управления, направляющая дополнительные ресурсы либо на увеличение объемов передаваемой информации, либо на повышение помехозащищенности высокоприоритетных сообщений. Указанная подсистема осуществляет адаптацию к помеховой обстановке, обменивая скорость на верность передачи при возрастании интенсивности помех и повышая скорость без снижения верности передачи при устойчивом низком уровне возмущений.

Выбор управляющего воздействия в подсистемах телеобслуживания и адаптации в каждой конкретной ситуации производится на основании оценки внутреннего состояния системы и внешних условий ее функционирования. Сбор и обобщение осведомительной информации осуществляется подсистемой контроля, на которую возлагается:

– диагностирование (оценка технического состояния, локализация и предупреждение отказовых и предотказовых состояний оборудования);

– испытания (оценка качества передачи информации в естественных и искусственно создаваемых условиях);

– исследование (накопление и обобщение статистических данных о распределениях и корреляции ошибок в каналах связи, об эффективности применяемых процедур передачи и защиты данных и т.п.);

– эксплуатационные измерения частотных параметров каналов связи, основного и резервного оборудования;

– оценка текущей помеховой обстановки по первичным и вторичным статистическим характеристикам;

– оценка текущей системной функции или отдельных параметров канала (идентификация);

– оценка качества обмена информацией (вычисление значения критерия).

Классификация [2]. Различные виды контроля можно классифицировать по целям, времени подключения контролирующих устройств, способу получения и источнику осведомительной информации, форме представления выходной информации, отношению к процессу передачи информации, способу реализации процедур.

По целевому признаку различают следующие виды контроля:

– *предупредительный*, проводимый для определения характеристик различного рода возмущающих факторов (потока ошибок в канале, интенсивности отказов в оборудовании); необходим для разработки эффективных алгоритмов обмена и управления ресурсами; отличается наличием большого парка контрольной аппаратуры; осуществляется на свободных каналах, выделяемых для исследования на длительное время (сотни и тысячи часов);

– *функциональный*, проводимый для определения степени соответствия частных параметров канала и всего оборудования заданным нормам и для регулировки аппаратуры в случае выявления несоответствия каналов; каналы, поставленные на функциональный контроль, не используются в течение всего времени контроля, длительность которого составляет от десятков минут до нескольких часов;

– *оперативный*, организуемый для достоверной оценки текущего состояния и помеховой обстановки в каналах и трактах, занятых под передачу, а также для оценки показателей качества передачи за время, соизмеримое с длительностью цикла обмена информацией; необходим для оперативной проверки готовности аппаратуры и организации оперативного управления ресурсами системы с целью адаптации к изменениям условий функционирования; отличается многообразием форм сбора и представления осведомительной информации, зависящих от типа контролируемого оборудования (показатели качества) и критерия, на основании которого осуществляется выбор управляющих воздействий.

По *времени подключения* различают следующие виды контроля:

– *непрерывный*, когда устройства контроля постоянно подключены к контролируемому оборудованию;

– *периодический*, когда одно и то же устройство контроля применяется для определения состояния нескольких функциональных устройств.

Наиболее характерен для предупредительного и функционального контроля.

По *способу получения* и *источнику осведомительной информации* можно выделить следующие виды контроля:

– *тестовый*, если на вход оборудования (канала, тракта) подается определенная последовательность символов, и по результату приема оцениваются внутреннее состояние и помеховая обстановка; обычно применяется для проверки работоспособности всей системы передачи дискретной информации в начале ее работы и при возникновении “пауз” в информационном обмене;

– *косвенный*, если источником осведомительной информации являются искажения параметров сигнала, потоки решений различного рода детекторов качества – первой решающей схемы при решении со стираниями;

– *кодовый*, если источником осведомительной информации служит последовательность решений декодера – второй решающей схемы;

– *прямое измерение*, если отдельные параметры оборудования, уровни помех, отношение “сигнал/шум” и т.д. могут непосредственно измеряться специальными приборами; широко используется при функциональном контроле и в радиосистемах с частотной адаптацией.

Косвенный и кодовый методы используются главным образом в оперативном контроле.

3.6. Организация ограничения доступа в сеть

В сети с коммутацией пакетов резкое и глубокое падение пропускной способности вызывает лавинообразное распространение явления блокировки по всем узлам коммутации, что в считанные секунды может привести к блокировке всех элементов сети. Поэтому система управления должна обеспечить оперативное и надежное ограничение доступа потоков в сеть путем выставления порогов $\{K\}$ и рассылки необходимой служебной информации до всех абонентов (интерфейсов).

С этой целью используется алгоритм, сущность которого заключается в образовании замкнутых ненаправленных путей, покрывающих все вершины графа, отображающего сеть, и проходящих через центральную вершину. Этот алгоритм нашел широкое распространение для контроля сетей [3].

Служебное сообщение, несущее информацию о всех порога $\{K\}$, с центра управления сетью лавиной передается на все смежные узлы коммутации. На каждом узле коммутации запоминается ветвь, из которой пришло первым служебное сообщение. Из сообщения считывается предназначенная для данного узла информация о порогах $\{K\}$, после чего производится смена старых порогов на новые $\{K\}$. В само же сообщение заносится квитанция о доведенной команде (пороге). Дополнительное служебное сообщение направляется во все смежные с данным узлом узлы коммутации.

На каждом узле запоминаются ветви, по которым поступили служебные сообщения по времени вторыми, третьими и т.д. Эти сообщения без изменений возвращаются назад по пути прихода первого по времени сообщения. При возвращении служебного сообщения с квитанциями в центр управления в случае выхода из строя одной из ветвей, входящей в кратчайший путь возврата, сообщения направляются по пути перехода второго по времени служебного сообщения, а в случае отказа и этого пути – по пути третьего и т.д. Тем самым обеспечивается как высокая надежность и оперативность доведения служебных сообщений, так и высокая надежность и оперативность сбора квитанций о выставленных порогах, что исключает повторную передачу служебного сообщения из центра управления сетью.

Следует отметить, что даже при достаточно оперативных алгоритмах ограничения нагрузки возможность блокировок полностью не исключается, что требует дополнительных мер по борьбе с блокировками. Такие меры

обеспечиваются алгоритмами управления внутренними потоками, т.е. теми потоками, которые уже допущены в сеть.

3.7. Методы выбора кратчайших путей

Распределение потоков информации производится с учетом длин дуг. Для оценки длин пути используются различные критерии, например, число транзитных узлов, протяженность пути, качество тракта передачи, надежность передачи информации и т.п.

Кратчайшим путем называется путь, для которого показатель длины пути имеет наименьшее значение по сравнению с его значениями для других возможных путей.

Все методы выбора кратчайших путей, развитые в теории потоков, основаны на достаточно очевидном утверждении о том, что если кратчайший путь μ_{ij} от произвольного УК$_i$ и УК$_j$ проходит через промежуточные УК$_{i_1}$, …, УК$_{i_k}$, то кратчайшие $\mu_{i_1 j}$, …, $\mu_{i_k j}$ от УК$_{i_1}$, …, УК$_{i_k}$ к УК$_j$ соответственно являются частями кратчайшего пути μ_{ij} от УК$_i$ к УК$_j$.

Легко понять, что для нахождения кратчайшего пути от некоторого узла к узлу j необходимо просмотреть все возможные пути и выбрать тот, у которого наименьшая длина.

В настоящее время существует ряд методов, позволяющих упорядочить процедуру определения длин кратчайшего пути. Условно эти методы можно разделить на две группы: 1) методы нумерации узлов и ветвей; 2) матричные методы.

Наиболее характерный для 1-й группы метод заключается в выполнении следующих операций:

1) выделенному УК$_j$ приписывается вес $\omega_j = 0$, а каждому из остальных УК сети – вес $\omega_i = \infty$;
2) выбирается произвольный УК$_j$, $i = 1, \ldots, N$, $i \neq j$, и проверяется неравенство

$$\omega_i > (l_{i,\xi} + \omega_\xi), \qquad (3.5)$$

где ω_ξ – вес последнего УК$_o$. При этом, если неравенство (3.5) выполняется, то прежний вес ω_i УК$_i$ заменяется на вес $\omega_i' = l_{i,\xi} + \omega_\xi$, в противном случае вес УК остается без изменений.

Указанная процедура пересчета узлов производится до тех пор, пока хотя бы для одного узла выполняется неравенство (3.5). Веса ω_i после пересчета будут равны длине кратчайшего пути от УК$_i$ к УК$_j$. Повторяя эту процедуру для каждого из N узлов сети, можно найти длины кратчайших путей между всеми узлами.

Матричный метод позволяет определить длины кратчайших путей между всеми узлами сети одновременно. Он основывается на применении операций над матрицами.

Структуру сети связи с указанием длин ее ветвей можно считать в виде матрицы расстояний (длин) непосредственных связей $L' = \left\| l'_{i,j} \right\|$, где $l'_{i,j}$ – длина ветви $\beta_{i,j}$.

В матрице расстояний непосредственных связей элементы главной диагонали всегда равны нулю, так как расстояние внутри узла принимается равным нулю. Если между парой узлов графа связи отсутствует ребро, то соответствующий элемент матрицы принимается равным ∞. Если же между узлами графа i и j имеется дуга $\beta_{i,j}$, то элемент матрицы $l'_{i,j}$ также принимается равным бесконечности. При анализе сетей передачи данных длину ветви удобно трактовать как задержку, которую вносит ветвь при передаче через нее информации.

Матрица расстояний непосредственных связей неориентированной сети всегда симметрична относительно своей главной диагонали, для ориентированной сети связи она может быть несимметрична.

Возведем матрицу L' в квадрат: $L^2 = L'L'$. Тогда

$$l^2_{i,j} = \sum_{k=1}^{N} l'_{i,k} l'_{k,j} = l'_{i,1} l'_{1,j} + l'_{i,2} l'_{2,j} + \ldots + l'_{i,N} l'_{N,j}. \qquad (3.6)$$

Интерпретируя умножение как последовательное, а сложение как параллельное соединение ветвей, легко понять (рис.3.1), что произведение соответствует двухтранзитному пути (т.е. пути, проходящему через две транзитные ветви сети) от узла i к узлу j через узел k, а сумма трех произведений $l'_{i,i} l'_{i,j} + l'_{i,j} l'_{j,j} + l'_{i,k} l'_{k,j}$ – трем двухтранзитным путям.

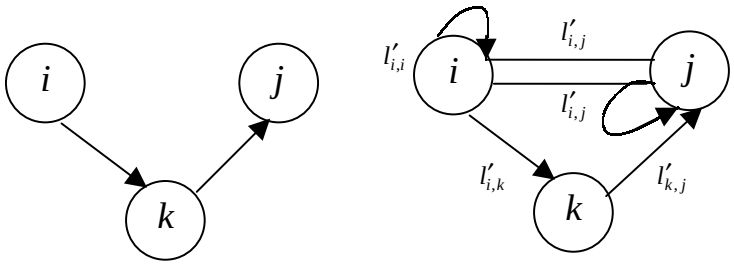


Рис.3.1. Интерпретация операций умножения и суммирования трех произведений

При этом произведения $l'_{i,i} l'_{i,j}$ и $l'_{i,j} l'_{j,j}$ фактически соответствуют однотранзитным путям (т.е. путям, включающим только одну ветвь), поскольку длина пути (задержка) внутри УК (т.е. $l'_{i,i}$ и $l'_{j,j}$) не принимается во внимание. Для подсчета длины каждого из таких транзитных путей необходимо операцию умножения заменить операцией сложения, т.е. вместо $l'_{i,k} l'_{i,k}$ будем иметь $l'_{i,k} + l'_{i,k}$.

При наличии нескольких параллельных одно- и двухтранзитных путей для определения длины между узлами следует операцию сложения заменить операцией выбора из всех длин минимальной длины, т.е. вместо (3.6) будем иметь

$$l_{j,j}^{2} = \min\left(l'_{i,k} + l'_{k,j}\right) = \min\left[\left(l'_{i,1} + l'_{1,j}\right); \left(l'_{i,2} + l'_{2,j}\right); \ldots \left(l'_{i,N} + l'_{N,j}\right)\right].$$

Таким образом, элемент $l_{i,j}^{2}$ матрицы L^2 равен длине кратчайшего пути от УК$_i$ к УК$_j$ среди всех одно- и двухтранзитных путей.

При возведении матрицы L' в r-ю степень с использованием указанных выше операций получим матрицу $L^r = L^{r-1}L'$, элемент которой

$$l_{i,j}^{r} = \min\left(l_{i,k}^{r-1} + l'_{k,j}\right) = \min\left[\left(l_{i,1}^{r-1} + l'_{1,j}\right); \left(l_{i,2}^{r-1} + l'_{2,j}\right); \ldots \left(l_{i,N}^{r-1} + l'_{N,j}\right)\right]$$

будет равен длине кратчайшего пути среди всех одно-, двух- и т.д. r-транзитных путей. Легко показать, что при наличии на сети N узлов коммутации число транзитных ветвей в пути без петель не может быть больше $N - 1$. Следовательно, при определении среди возмущенных путей кратчайших для всех узлов сети может потребоваться вычисление матрицы L^r, у которой $r \leq N - 1$. Для конкретной сети может оказаться, что при $r < N - 1$

$$L^r = L^{r-1}. \qquad (3.7)$$

Так как всегда выполнится равенство $L^r = L^{r-1}$, вычисление матрицы более высокой степени прекращается, если в процессе вычисления матрицы встретится равенство (3.7).

Равенство (3.7) означает, что кратчайшие пути между каждой парой УК находятся среди одно-, двух- и т.д. $(r - 1)$-транзитных путей, а любой r, (r + 1) и т.д. (N–1)-транзитный путь имеет длину (задержку), большую, чем кратчайший путь.

Матрица L^{r-1} при выполнении условия (3.7) называется *дистанционной матрицей* и обозначается

$$D = L^{r-1} = L^{r} = \left\| d_{i,j} \right\|.$$

Таким образом, элементы дистанционной матрицы равны длинам кратчайших путей между соответствующими узлами сети связи. Поэтому дистанционная матрица называется *матрицей расстояний* (длин, задержек) *кратчайших путей.*

Метод Флойда

Определение самого кратчайшего пути связано с дополнительной процедурой. Так, если для определения длин кратчайшего пути применяется способ нумерации узлов, то при выполнении дополнительной процедуры учитывается свойство веса УК$_i$. Это свойство заключается в том, что существует УК$_j$, $i \neq j$, для которого выполняется равенство

$$\omega_i = l_{i,j} + \omega_j.$$

Отсюда следует, что

$$\omega_i - \omega_j = l_{i,j}\,. \qquad (3.8)$$

Поэтому, если выполняется условие (3.8), то кратчайший путь проходит по ветви $\beta_{i,j}$. Переходя к УК$_i$, находим следующую ветвь, для которой выполняется условие (3.8) и которая также входит в кратчайший путь. Так, шаг за шагом, можно определить все ветви, образующие кратчайший путь. Исключив затем кратчайший путь из рассмотрения, аналогично определяются и другие пути от исходящего УК$_i$ к входящему УК$_j$. Данный метод выбора кратчайших путей называется *методом Флойда* [12].

При матричном методе определения кратчайшего пути дополнительно к дистанционной матрице на основе матрицы длин непосредственных связей составляется так называемая *модернизированная матрица длин непосредственных связей.*

Метрика

Определим метрику, в соответствие с которой выбираются кратчайшие маршруты между узлами сети [13]. В этом случае необходимо учесть тот факт, что сеть осуществляет обмен данными между абонентами как в диалоговом режиме, требующем большой оперативности, так и в режиме передачи файлов данных, не требующем большой оперативности. С целью обеспечения большой пропускной способности сети выбор путей для каждого из видов трафика необходимо осуществлять дифференциально. Данные диалогового режима целесообразно обслуживать с приоритетом, а в качестве критерия выбора путей передачи данных использовать минимум доведения использования пакетов. Передача файлов ведется со вторым приоритетом, а в качестве критерия выбора пути целесообразно использовать максимум пропускной способности пути или другую метрику, приводящую к увеличению пропускной способности сети. В качестве веса ветви можно использовать коэффициент недоиспользованной пропускной способности ветви (i, j):

$$\beta_{ij} = -\ln(1 - c_{ij}),$$

где $\beta_{ij} = \lambda_{ij}/c_{ij}$; c_{ij} – пропускная способность ветви; λ_{ij} – интенсивность поступления пакетов данных на вход ветви (i, j). Предложенный показатель позволяет эффективно бороться с локальными перегрузками в сети с одновременным увеличением ее пропускной способности.

3.8. Критерии выбора оптимальных путей

Существуют общие и частные критерии оптимизации. При статистической маршрутизации, как правило, используются общие критерии, и производится системная оптимизация, тогда как при адаптивной маршрутизации наиболее часто принимаются частные критерии и происходит выбор оптимального пути с точки зрения пользователя. Естественно,

пользовательская оптимизация не гарантирует системной оптимизации, однако в некоторых случаях они могут давать практически одинаковые результаты [14].

За редким исключением алгоритмы маршрутизации, использующие частные критерии, основаны на алгоритмах выбора пути в графе, т.е. осуществляется выбор минимального по “весу” пути. За “вес” пути (оптимизируемый критерий) принимается определенный параметр сети, который необходимо минимизировать по заданному алгоритму. Параметрами могут служить длина линии, число транзитных участков в пути, суммарная задержка при передаче по данному пути и т.д.

Например, в известном методе рельефов [10, 12] (см. Приложение) в качестве критерия оптимальности берется число транзитных центров коммутации в пути. Путь, где число транзитных центров коммутации пакетов (ЦКП) наименьшее, выбирается за оптимальный. Главный недостаток этого метода – нечувствительность к задержкам в очередях.

При формировании веса пути решающее значение имеет доступная в данный момент динамическая информация о состоянии сети. Если имеется возможность вычислить или измерить загрузку линий, входящих в маршрут, то возможно применение нескольких стратегий выбора пути. Например, если маршрут выбирается по максимальной остаточной пропускной способности, то используется следующее правило выбора маршрута

$$\max_{\gamma}\{\min(c_i(1-\rho_i))\},$$

где γ – суммарная выходная нагрузка; c_i – пропускная способность линии связи; ρ_i – вероятность использования линии.

Если маршрут выбирается по минимальной задержке передачи пакетов, то можно применить следующее правило выбора маршрута:

$$\min_{\{\gamma\}}\left\{\sum_{g_i\in\{\gamma\}}\frac{1}{\mu c_i(1-\rho_i)}+t_i\right\},$$

где t_i – задержка распространения.

В известном алгоритме “отклонения потоков” [5, 14] применяется критерий

$$\min_{\{\gamma\}}\left\{\sum_{g_i\in\{\gamma\}}\frac{1}{\mu c_i(1-\rho_i)^2}+t_i\right\}.$$

Если в сети используется приоритетная дисциплина, то маршрут может выбираться следующим образом:

$$\min_{\{\gamma\}}\left\{\sum_{g_i\in\{\gamma\}}(T_{ik}+t_i)\right\},$$

где T_{ik} – задержка пактов k-го приоритета.

В сетях с приоритетной дисциплиной обслуживания возможно применение отдельных, отличающихся друг от друга, правил выбора маршрута для

каждого приоритетного потока. Например, для потоков высшего приоритета маршрут может выбираться по минимальной задержке, а для потоков низкого приоритета – по максимальной остаточной пропускной способности.

Объем служебной информации при использовании в ЦСИО того или иного метода адаптивной маршрутизации целесообразно оценивать по степени ухудшения качества обслуживания запросов пользователей, вызванного наличием служебной информации.

Важным показателем, характеризующим эффективность метода маршрутизации, являются среднее время доведения служебной информации и ее объем. Под средним временем T_c будем понимать статистически усредненное значение интервала времени между моментами возникновения служебного сообщения и моментом изменения маршрутной таблицы (МТ) во всех УК сети, вызванного этим служебным сообщением.

Наличие служебной информации вызывает ухудшение качества обслуживания запросов пользователей, поэтому при том или ином методе адаптивной маршрутизации целесообразно оценивать объем служебной информации по степени ухудшения качества обслуживания.

Будем оценивать объем служебной информации по увеличению среднего времени задержки сообщений ΔT_n :

$$\Delta T_n = T_{\text{а}} - \Delta T_0,$$

где $T_{\text{а}}$ – среднее время задержки сообщений в режиме КП при адаптивной маршрутизации при наличии служебных данных; T_0 – среднее время задержки сообщений в режиме КП при фиксированной маршрутизации, т.е. когда служебные сообщения не передаются. Значение T_0 может быть определено по модели УК, описываемой моделью ТМО типа $M/M/1$ (см. п.2.6). Аналитическая оценка значения $T_{\text{а}}$ зависит от используемых методов адаптивной маршрутизации.

Для оценки эффективности для сетей с КП могут использоваться и другие подходы [14, 19]. В этом случае в качестве меры оценки эффективности могут выступать производительность, объем передаваемых сообщений, накладные расходы и т.п. Однако вследствие трудности аналитического решения большинство оценок и сравнений адаптивной маршрутизации в настоящее время традиционно базируются на имитационном моделировании и результатах измерений.

3.9. Оценка вероятностно-временных характеристик

К основным вероятностно-временным характеристикам (ВВХ) относятся время доставки информации между корреспондирующими абонентами и вероятность доставки.

Оценка этих характеристик осуществляется с помощью прикладных математических моделей. Одной из отличительных особенностей рассматрива-

емых систем связи является динамика изменения как структуры и параметров, так и информационной нагрузки. Учет динамики накладывает определенные ограничения на используемые математические модели. Выбор математической модели для оценки ВВХ зависит от соотношения интенсивностей информационных процессов в системе и интенсивностей изменений, происходящих в ней. Если они соизмеримы, т.е. соотношение интенсивностей не превышает $10–10^2$, то модели должны учитывать динамику изменений в сети. В противном случае (квазистационарный режим) для оценки ВВХ можно воспользоваться "стационарными" моделями и выполнять расчеты для некоторого числа фиксированных вариантов (субвариантов) системы.

Рассмотрим "стационарные" модели, которые могут быть использованы для расчета таких ВВХ, как время доставки и вероятность доставки. Описываемые модели основываются на декомпозиции сетей и представлении информационных процессов в них полумарковским процессом с конечным числом состояний (конечным полумарковским процессом). Таким образом, информационный процесс представляется как процесс перемещения информации от абонента-источника к абоненту-получателю последовательно через выделенные в результате декомпозиции элементы сети, на которых осуществляется либо переприем (в случае коммутации сообщений или пакетов), либо транзит (в случае коммутации каналов). Формально это может быть описано следующим образом [5].

Пусть $J = \{1\}$ – множество элементов декомпозиции, а $S = \{s\}$ – множество типов информации, циркулирующей в системе (принадлежность информации одному и тому же типу определяется однотипным обслуживанием на элементах декомпозиции). Пусть также $M(J)$ и $M(S)$– мощности множеств J и S соответственно.

Введем в рассмотрение конечную цепь Маркова, число состояний которой оценивается сверху величиной

$$N = M(J) \cdot M(S) + 2. \qquad (3.9)$$

Процесс передачи информации от источника к получателю описывается как бы "цепочкой" перемещений из состояний в состояние, в последнем из которых происходит "поглощение". В случае нормального доведения "поглощение" происходит в состоянии, соответствующем нормальному доведению, а в случае потери – соответствующем "потере". Эти поглощающие состояния – дополнительные два состояния в правой части (3.9). Если предположить, что тип информации в процессе ее доведения не изменяется (как правило, на практике это бывает в подавляющем большинстве случаев), то цепочка доведения информации от источника через ряд элементов сети (узлов) может быть условно представлена в следующем виде:

$$(j_0, s) \rightarrow (j_1, s) \rightarrow (j_2, \mathrm{s}) \rightarrow \ldots \rightarrow (j_k, s),$$

где k – длина пройденного пути.

Переходы между состояниями описываются матрицей переходных вероятностей **P**, элементы которой задают структуру вероятности переходов между состояниями. Матрица **P** имеет структуру

$$\begin{matrix} \mathbf{I} & 0 \\ \mathbf{R} & \mathbf{Q} \end{matrix}$$

Здесь **Q** – вероятности переходов между невозвратными состояниями (квадратная матрица); **I** – единичная диагональная матрица; **R** – матрица переходов из невозвратных состояний в поглощающие.

В общем случае элементы **Q** (задают вероятности переходов между элементами сети) являются функциями интенсивностей информационных потоков. Тогда, если обозначить **V** вектор интенсивностей потоков, поступающих на входы системы, а **Λ** – вектор интенсивностей потоков, обслуживаемых на элементах декомпозиции, то распределение информационных потоков по элементам системы описывается в общем случае системой нелинейных уравнений относительно **Λ**, имеющей в матричном представлении следующий вид:

$$\mathbf{\Lambda} = \mathbf{V}(1 - \mathbf{Q}(\mathbf{\Lambda}))^{-1}. \qquad (3.10)$$

Решение этой системы позволяет определить потоки **Λ** и переходные вероятности – основы для оценки ВВХ.

Так, расчет среднего времени доставки производится по формуле

$$\mathbf{T} = (1 - \mathbf{Q})^{-1}\mathbf{Q}(\mathbf{\Lambda}),$$

где **Q**(**Λ**) – вектор времени обслуживания информации по элементам декомпозиции.

Вектор второго начального времени доведения информации определяется по формуле

$$\mathbf{W} = (1 - \mathbf{Q})^{-1}(\mathbf{M}(\xi^2) + 2\mathbf{C}\mathrm{T}), \qquad (3.11)$$

где $\mathbf{M}(\xi^2)$ – вектор вторых моментов времени обслуживания информации на элементах декомпозиции; **C** – матрица, элементы которой являются произведениями вероятностей переходов между состояниями на соответствующие средние времена обслуживания.

Оценка второго момента времени доведения обеспечивает получение оценок своевременности доставки в предположении, что время доведения является нормально распределенной случайной величиной.

Элементы σ_i – среднеквадратические отклонения времени доведения получаются из (3.11) как

$$\sigma_i = \sqrt{W_i - T_i^{\,2}}.$$

При этом оценка своевременности $P_{\text{св}}$ доведения (как вероятность превышения критического времени $T_{\text{кр}}$ доведения) выражается как

$$P_{\text{св}} = \max\{P_{\text{св}} : T_{\text{кр}} = T_i + k(P_{\text{св}\,i})\sigma_i\},$$

где $k(P_{\text{св}\,i})$ – коэффициент, являющийся квантилем нормированного нормального распределения для вероятности $P_{\text{св}\,i}$.

Для оценки надежности доставки может быть использован аналогичный математический аппарат. Вероятность $\mathbf{P}_{\text{д}}$ доставки информации получается из выражения

$$\mathbf{P}_{\text{д}} = (\mathbf{I} - \mathbf{Q}^{-1})\mathbf{b},$$

где $\mathbf{b}$ – вектор вероятностей переходов в поглощающее состояние, соответствующее нормальному доведению.

Рассмотрим теперь модели, которые могут быть использованы для получения характеристик системы связи, учитывающие ее динамику.

Общее представление информационного процесса аналогично описанному выше для статистических моделей. Не ограничивая общности, будем полагать, что в сети отсутствуют потери информации. Изменения входных потоков на элементах декомпозиции, описываемых моделями СМО (моделями очередей), можно представить в виде совокупности случайных скачков. Тогда изменение интенсивности потока на входе очереди для одного скачка определяется в виде интеграла Дюамеля

$$\lambda_{ij}(t) = \int_0^t \varphi_{ij}\left(t - \tau, \lambda_i^0(\tau)\right)d\tau,$$

где λ_{ij} – интенсивность потока на входе элемента i в направлении элемента j; φ_{ij} – плотность распределения времени пребывания сообщения в очереди ij для интенсивности потока на входе λ_i^0 в нулевой (начальный) момент времени; λ_i^0 – интенсивность потока на входе i-й очереди.

Уравнение для потоков в сети в целом имеет сложный вид, аналогичный (3.10), и в алгебраическом виде представляется как

$$\lambda_i\left(t, \lambda_i^0\right) = \sum P_{ij}\left(\lambda_i^0\right)\int_0^t \varphi_{ij}\left(t - \tau, \lambda_i^0(\tau)\right)d\tau + V_j(t_0), \qquad (3.12)$$

где $P_{ij}\left(\lambda_i^0\right)$ – вероятности переходов, рассчитанные для значения интенсивностей потоков λ_i^0 в стационарном случае из уравнения (3.10); $V_j(t)$ – поток извне на входе j-й очереди. Полученная система является системой интегральных уравнений Вольтера второго рода.

Анализ динамики включает следующие шаги:

1) расчет стационарных характеристик сети с использованием модели;

2) составление системы интегральных уравнений, описывающих динамику сети в окрестности точки равновесия;

3) решение системы интегральных уравнений;

4) расчет стационарных характеристик в новой точке;

5) решение системы интегральных уравнений как продолжение предыдущего решения с неявными значениями переменных из (3.12) и т.д. до тех пор, пока не будет проведен расчет для всего интервала времени наблюдения за поведением системы.

Для оценки ВВХ сетей с учетом динамики, используя получаемые значения потоков, и в предположении, что интенсивности их изменений по сравнению с интенсивностями доставки информации (обратными значениями к средним временам доставки) невелики, можно воспользоваться выражением для **T,** а значения параметров **Q**(**Λ**) должны выясняться с использованием полученных из (3.12) интенсивностей потоков для определенных моментов времени интервала наблюдения.

3.10. Методы измерения нагрузки и показателей качества обслуживания

Измерение нагрузки и показателей качества обслуживания может производиться для целей:

- технической эксплуатации, прогнозирования нагрузки;
- изучения потоков нагрузки и качества, накопления статистических данных.

В зависимости от цели выбираются методы измерения, измерительные приборы и измеряемые показатели (объем и периодичность измерения, время их проведения и т.д.), т.е. к системе измерений в целом предъявляются различные требования. Эти требования также зависят от применяемого на сети метода коммутации. Кроме того, система измерений должна позволять вести контроль за такими факторами, которые в совокупности могут значительно увеличить нагрузку, изменить характер ее поступления и значительно снизить качество обслуживания.

Необходимо подчеркнуть, что измеряется только обслуживаемая нагрузка, например, ее интенсивность *y*. Интенсивности входящей и потерянной *x* нагрузок не могут быть измерены, так как не может быть измерена длительность необслуживаемого вызова или объем непереданного сообщения. Поэтому значения с и *x* определяются расчетом через *y*, если измерены показатели качества обслуживания, а именно вероятность своевременной доставки сообщений **Q** или величина потерь.

При измерении *нагрузки* в основном используется непрерывный метод измерения и метод сканирования. При непрерывном методе измерения за отрезок времени от 0 до *T* нагрузка *H* равна сумме нагрузок H_i по интервалам *n* обслуживающего устройства, на котором производится измерение:

$$H = \sum_{i=1}^{n} H_i .$$

Если полученную таким образом нагрузку *H* разделить на общее время измерений в часах, т.е. отнести к единице времени, то искомая величина будет являться интенсивностью нагрузки y_i:

$$y = \frac{H}{T} = \frac{\sum_{i=1}^{n} H_i}{\sum_{i=1}^{n} t_i}.$$

Метод сканирования заключается в подсчете числа занятых приборов в отдельные моменты времени. Известно, что среднее значение нагрузки за период измерения равно среднему числу одновременно занятых приборов. В этом случае

$$H = \frac{1}{N} \sum_{i=1}^{n} k_i,$$

где k – число одновременно занятых устройств при i-м сканировании; N – общее число сканирований.

Так как за нагрузкой не ведется непосредственного наблюдения, то метод сканирования вносит погрешность и оказывается менее точным по сравнению с непрерывным методом. Средняя ошибка при определении нагрузки H определяется по формуле Пальма:

$$dH = H \sqrt{\frac{1}{n} \cdot \frac{1 + e^{\alpha}}{1 - e^{-\alpha}} (\alpha - 2)},$$

где n – число занятий; α – отношение интервала Δt между двумя моментами сканирования (интервалами сканирования) к среднему времени занятия Θ.

Для определения интенсивности обслуженной нагрузки необходимо полученное значение нагрузки H разделить на общий интервал времени измерений.

В сетях с КС (КП) при измерении различных *показателей качества* обслуживания (числа сообщений, пакетов, ожидающих обслуживания, времени ожидания начала обслуживания и т.д.), как правило, используются методы прямого отсчета числа сообщений. Они основаны приеме импульсов для каждого события и накопления их в пакеты (например, в счетчиках). В современных УК нагрузка и показатели качества обслуживания измеряются программными методами [4].

3.11. Контроль эффективности входного потока. Ограничение нагрузки

Организация контроля за текущим значением интенсивности входного потока осуществляется следующим образом [13, 15]. На каждом узле I имеется текущий счетчик разрешений $\Pi_i(t)$. В начальный момент времени $\Pi_i(0) = K$. В дальнейшем значение K не может быть превышено. С интенсивностью $\lambda_i(t)$ на каждом узле генерируется новое разрешение, увеличивающее значение

счетчика $\Pi_i(t)$ на единицу, т.е. $\Pi_i(t) = \Pi_i(t) + 1$ если $\Pi_i(t) < K$, в противном случае счетчик сохраняет свое значение.

Прием сообщений от абонента возможен лишь при наличии свободного разрешения в узле коммутации $\Pi_i(t) > 0$ с одновременным изменением значения счетчика $\Pi_i(t)$. Начальное значение счетчика фактически определяет максимально допустимое число пакетов, которое может принять узел от абонента одновременно даже при нагруженной сети ПД. Основная трудность в организации ограничения внешних потоков как раз и состоит в нахождении значения $\{K\}$.

Для эффективного управления доступом необходимо определить все $\{K_{sl} \ (sl \in Q)\}$, где Q – множество всех процессов, взаимодействующих через сеть, и изменять значения K во время функционирования системы с учетом применяемой адаптивной маршрутизации и управления потоками на транспортном уровне.

Изменения осуществляются в целях лучшего согласования ограниченных возможностей сети с потребностями в обслуживании внешних потоков сообщений. Для осуществления управления доступом внешних потоков необходимо вначале получить оценку пропускной способности сети. Воспользуемся сведениями о топологической структуре сети, модели трафика, модели канала, модели коммутирующего устройства, а также выбранном алгоритме маршрутизации.

Пусть модель сети задана в виде взвешенного неориентированного графа без петель $\Gamma(N, L)$, где N – множество узлов коммутации, а L – множество ветвей (трактов передачи данных). На входе сети поступает пуассоновский поток интенсивностью $\Lambda = \sum_{r=1}^{R} \lambda_r$, где r – номер приоритета потока, $r = \overline{1, R}$. Задана матрица тяготения потоков. Плотность распределения интенсивности обслуживания подчинена экспоненциальному закону. Среднее число пакетов в сообщениях S_r, $r = \overline{1, R}$ для каждого приоритетного потока. Заданы также требования по качеству обслуживания.

Задача оценки пропускной способности сети определяется в два этапа. На первом этапе вычисляются суммарные потоки на входе ветви с учетом функционирования системы маршрутизации, тяготений и приоритетного характера обслуживания потоков. На втором – определяют количество сообщений, обслуженных сетью с заданным качеством.

На первом этапе решается оптимизационная задача. Для каждого из приоритетных потоков λ_r необходимо найти такие λ_r^+, при которых достигается максимум произведения коэффициентов недоиспользования пропускных способностей ветвей связности:

$$\lambda_r^+ = \arg\max_{\lambda_{ijr}^{kl} \in \lambda_r} \prod_{i \in N} \prod_{j \in N} \left(1 - \frac{c_{ij} + \lambda_{ij(r-1)} + \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_{ijr}^{kl}}{K_{rij}} \right),$$

при ограничениях

$$\lambda_{kir} = \sum_{j=1}^{N} \lambda_{kjr}^{kl}, \quad k, l = \overline{1, N}, \ r = \overline{1, R},$$

$$\sum_{\substack{m=1 \\ i \neq k}}^{N} \lambda_{imr}^{kl} = \sum_{\substack{m=1 \\ i \neq k}}^{N} \lambda_{mir}^{kl}, \quad k, l = \overline{1, N}, \ r = \overline{1, R},$$

$$\lambda_{ijr}^{kl} \geq 0, \ \mu_{ij} > 0, \ k, l, i, j = \overline{1, N}, \ r = \overline{1, R}, \ \mu_{ij} \in \mathrm{M},$$

$$\lambda_{ij(r-1)} + c_{ij} + \sum_{k=1}^{N} \sum_{l=1}^{N} \lambda_{ijr}^{kl} \leq \rho_r k_{rij} \mu_{ij}, \ i, j = \overline{1, N}, \ r = \overline{1, R},$$

$$\omega_{kl} \sum_{k=1}^{N} \sum_{l=1}^{N} \lambda_{klr} = \Lambda_{klr}, \quad k, l = \overline{1, N}, \ r = \overline{1, R}, \ \omega_{kl} \in W_l,$$

$$\sum_{k=1}^{N} \sum_{l=1}^{N} \omega_{kl} = 1, \quad 0 \leq \omega_{kl} \leq 1,$$

где $c_{ij} = a_{ij} + b_{ij} + d_{ij}$, $a_{ij} \in \mathbf{A}$, $b_{ij} \in \mathbf{B}$, $d_{ij} \in \mathbf{D}$, $r_{rij} \in K_r$, $\mu_{ij} \in \mathbf{M}$,

$\Lambda_{ij(r-1)} \sum_{i=1}^{r-1} \sum_{k=1}^{N} \sum_{i=1}^{N} \lambda_{ij\varepsilon}^{kl}$ – сумма всех потоков с первого до $(r-1)$-го приоритета, передающихся по ветви (i, j); c_{ij} – поток служебной информации в ветви (i, j); $\rho_{rij} = \sum_{k=1}^{N} \sum_{i=1}^{N} \lambda_{ijr}^{kl} \Big/ \mu_{ij}$ – допустимый коэффициент искажения ветви (i, j) потоком r-го приоритета; ω_{kl} – доля пропускной способности сети, выделенная для потока между узлами k и l. Значение $\rho_r = f(\tau_r)$ (предполагается заданным), где ε_r – допустимое время доведения пакетов в ветви для потока с r-м приоритетом; **B** – матрица потока служебной информации; **M** – матрица пропускных способностей трактов; **W** – матрица важности потоков; **D** – матрица потока квитанций.

Предположим, что в сети реализован алгоритм адаптивной маршрутизации, осуществляющий выбор пути передачи по критерию максимума пропускной способности сети.

Для решения сформулированной задачи предлагается следующий адаптивный алгоритм нахождения субаптимального решения, который заключается в том, что сначала фиксируется поток с наивысшим приоритетом, а затем выполняются следующие действия:

1) производится ранжирование потоков выбранного приоритета по критерию максимума тяготения;

2) выбирается первый из ранжированных потоков;

3) для выбранного потока с помощью алгоритма Флойда находятся все пути доведения информации и производится их ранжирование по критерию максимума пропускной способности; в случае равенства пропускных способностей путей предпочтение отдается обходному пути с меньшим числом транзитных участков; при равенстве и этого показателя выбор пути осуществляется произвольно;

4) весь поток направляется по пути с максимальной пропускной способностью;

5) из сети исключается часть пропускной способности, задействованной под передачу потока;

6) выбирается следующий по рангу поток и для него выполняются операции, начиная с п.3; если распределены все потоки, то выполняется следующий пункт;

7) вычисляются значения целевой функции $F(1)$;

8) выбирается поток, имеющий минимальный ранг по критерию максимума тяготения;

9) некоторая доля выбранного потока $\Delta\lambda^{kl}$ направляется по первому обходному пути;

10) если нагрузка в этом пути превышает пороговое значение для данного приоритета, то путь исключается из рассмотрения; переходим к выполнению п.12, в противном случае выполняется п.11;

11) вычисляются значения целевой функции $F(2)$ из выражения ( ); если $F(2) > F(1)$, то $F(1)$ присваивается значение $F(2)$, и распределение потока принимается, иначе значение $F(1)$ не изменяется и распределение не принимается;

12) выбираем следующий обходной путь, по нему направляется $\Delta\lambda^{kl}$; переход к п.10, повторение пп. 11–12 осуществляется до тех пор, пока не будут просмотрены все обходные пути;

13) из оставшегося нераспределенного потока берется новая доля $\Delta\lambda^{kl}$ и выполняется п.9; п. 12 повторяется до тех пор, пока не будет путей, для которых $F(2) > F(1)$, либо не исчерпается поток;

14) выбирается следующий поток по п.8 и для него повторяются операции, начиная с п.9, только с учетом, что некоторая доля ресурса сети уже использована ранее распределенными потоками.

15) после того, как будут распределены все потоки старого приоритета, вычисляются новые значения $\rho(\tau_r)$ для потоков с меньшим приоритетом $(r+1)$ и для них выполняются операции, начиная с п.1.

Таким образом, в результате решения получили суммарные потоки на входе ветви с учетом функционирования системы маршрутизации, управления потоками, тяготений и приоритетного характера обслуживания потоков.

На втором этапе определяются потери пакетов при передаче в сети из-за старения и возможной блокировки узлов коммутации. Для этого используют модель многоканального тракта передачи данных с различными каналами, что

характерно для систем, использующих кабельные сети и радиоканалы [5]. Согласно этой модели, в первую очередь определяется вероятность своевременной передачи пакетов *r*-го приоритета через ветвь $\beta - P_{\beta r}$. Ветвь в представляется в виде *b*-канальной системы массового обслуживания с очередью ограниченной длины Q_r и надежными обслуживающими приборами. На вход системы поступают потоки пакетов с относительными приоритетами. Потоки приоритетных пакетов представляются как потоки отказов, приводящие к блокированию каналов на время обслуживания этих каналов. В этом случае $P_{\beta r}$ можно оценить по формуле полной вероятности:

$$P_{\beta r} = \sum_{u=0}^{b} P_r(H_u) \cdot P_r\left(A/H_u\right), \qquad (3.13)$$

где $P_r(H_u)$– вероятность гипотезы H_u, заключающейся в безотказной работе U каналов ветви; u =1, b ($b \geq 1$); $P_r\left(A/H_u\right)$ – условная вероятность события *A*, состоящего в обслуживании пакета при выполнении гипотезы H_u.

Для определения $P_r(H_u)$ и $P_r\left(A/H_u\right)$ используются методы теории массового обслуживания. Вероятность $P_r(H_u)$ для случая одинаковой надежности каналов определяется по формуле Бернулли

$$P_r(H_u) = C_p^u P_r^u (1 - P_r)^{b-u},$$

где P_r – вероятность безотказной работы канала при передаче пакетов *r*-го приоритета.

Значение P_r определяется по формуле

$$P_r = \frac{\bar{\mu} + \kappa}{\mu + \kappa + \alpha + \lambda_{r-1}/b} \cdot \exp\left(-\alpha - \frac{\lambda_{r-1}}{b\bar{\mu}}\right),$$

где $\bar{\mu}$ – интенсивность обслуживания в канале; к – интенсивность восстановления работоспособности канала; α – интенсивность отказа канала; λ_{r-1} – суммарная интенсивность всех потоков до (r – 1)-го приоритета.

Вероятность $P_r(H_u)$ для случая неоднородных каналов определяется по схеме Бернулли

$$P_r(H_u) = \sum_{k=0}^{C_b^U} \prod P_{ri} \prod_{i \in (b-k)} (1 - P_{rj}), \quad i \neq j. \qquad (3.14)$$

Условная вероятность обслуживания пакетов *r*-го приоритета $P_r\left(A/H_u\right)$ представляет собой вероятность обслуживания потоков в *u*-канальной полиодоступной системе массового обслуживания с очередью ограниченной длины Q и ограниченным временем ожидания в очереди.

Обозначим $\rho_r^l \Big/ \prod_{i=1}^{l}(U + i\delta_r)$ через E. Тогда

$$P_{ru} = P_r\left(A \Big/ H_u\right) = 1 - \frac{\frac{\delta_r \rho_r}{U!}\sum_{i=1}^{Q_r} i \cdot E}{1 + \sum_{k=0}^{U}\frac{\rho_r^k}{k!} + \frac{\rho_r^U}{U!}\sum_{i=1}^{Q_r}\rho_r^i\left(\prod_{i=1}^{l} E\right)^{-1}}, \qquad (3.15)$$

где $\delta_r = \frac{\tau_r}{\mu}$, μ – интенсивность старения информации r-го приоритета; U – число каналов ветви.

Подставляя (3.14) и (3.15) в (3.13), получаем вероятность для определения $P_{\beta r}$:

$$P_{\beta r} = \sum_{u=1}^{b}\left[P_{ru}\sum_{k=0}^{C_b^U}\prod_{i\in k} P_{ri}\prod_{j\in(b-k)}(1 - P_{rj})\right].$$

Теперь, зная $P_{\beta r}$, можно определить вероятность доведения пакетов по пути

$$P_{\gamma r} = \prod_{\beta=1}^{\overline{\omega}} P_{\beta r},$$

где $\overline{\omega}$ – число ветвей в пути γ.

При передаче потока по нескольким путям вероятность доведения пакетов определяется по формуле

$$P_{klr} = 1 - \prod_{j=1}^{\pi}(1 - P_{jr}),$$

где π – число нулей передачи заданного потока.

Теперь можно вычислить обслуженный поток:

$$\lambda_{klr}^{\text{обсл}} = \lambda_{klr} \cdot P_{klr}, \quad \lambda_{klr} \in \Lambda_r.$$

Число сообщений, передаваемых по сети в единицу времени определяется по формуле

$$\Lambda_{\text{обсл}} = \sum_{r=1}^{R}\sum_{k=1}^{N}\sum_{l=1}^{N}\frac{\lambda_{klr}^{\text{обсл}}}{S_r}.$$

Величина $\Lambda_{\text{обсл}}$ характеризует пропускную способность сети.

Для вычисления значения порогов необходимо найти дополнительные потоки пакетов $\lambda_{r\,\text{доп}}^{sl}$, $r = \overline{1,R}$, генерируемых транспортным уровнем при восстановлении пакетов (повторная передача). Вычисление осуществляется для

каждой корреспондирующей пары (*sl*) и для каждого приоритетного потока $r=\overline{1,R},$ по формуле

$$\lambda^{sl}_{r\,\text{доп}}=\lambda^{sl}_{r\,\text{доп}}\cdot P_r^{sl}\left(1+\left(1-P_r^{sl}\right)+\ldots+\left(1-P_r^{sl}\right)^k\right),\ \forall(sl)\in Q,\ r=\overline{1,R},\ \lambda^{sl}_{r\,\text{пот}}\in\lambda_{r\,\text{пот}}.,$$

где P_r^{sl} – вероятность своевременного обслуживания пакетов *r*-го при передаче их от процесса *s* к процессу *b* через сеть; *k* – число повторных передач с транспортного уровня при их потерях в сети.

Для фиксирования момента времени t^1 значения порогов вычисляются по формуле

$$k_r^{sl}(t)=\lambda_r^{sl}-\lambda^{sl}_{r\,\text{доп}}$$

для $\forall(sl)\in Q$ и $r=\overline{1,R}.$

3.12. Контроль и сбор служебной информации в сети ПД

Принципы организации системы контроля и сбора служебной информации о состоянии сети в основном определяются:

- структурой сети ПД;
- характером потоков информации, циркулирующих в сети;
- требованием к качеству обслуживания потоков информации;
- степенью влияния внешней среды;
- организацией систем управления сетью;
- видом технических средств и каналов связи, используемых в сети;
- организацией систем технического обслуживания.

Процесс функционирования подсистем контроля и сбора служебной информации состоит из следующих этапов:

- непосредственного контроля за состоянием элементов сети и непрохождением информации;
- сбора служебной информации;
- анализа служебной информации.

Определяющим моментом при выборе алгоритма управления сетью с различной структурой является объем служебной информации, циркулирующей в сети, и время ее передачи, поскольку оперативность управления обеспечивается именно быстродействием передачи служебной информации. Адекватность и полнота служебной информации должна соответствовать требованиям к качеству управления. Служебная информация, поступающая от других узлов сети, может быть как быстроменяющейся (информация об уровне очередей на различных узлах), так и медленно меняющейся (информация об отказах каналов связи), поэтому важно определить объемы и темпы обновления служебной информации.

Одним из наиболее сложных вопросов являются вопросы учета влияния очередей на УК. В сетях с напряженным трафиком информация о состоянии очередей быстро стареет, поэтому приходится прибегать к прогнозированию очередей или пользоваться средними характеристиками длин очередей.

Одним из важнейших требований к передаче служебной информации является обеспечение заданной достоверности ее доставки. Эта задача успешно решается выделением методов временной и аппаратурной избыточностей. Однако при принятой скорости передачи методов повышения достоверности ведут к увеличению времени запаздывания в управлении и дополнительной загрузке сети служебной информацией. В этой связи перспективным является использование для передачи служебной информации шумоподобных сигналов, передаваемых на фоне основной информации [3, 7].

Сбор служебной информации в реальных сетях является очень важной проблемой, от правильного решения которой в значительной степени зависит эффективность использования сети, ее стоимость, вероятностно-временные характеристики, надежностные характеристики обмена информацией.

Для сбора служебной информации в сети ПД применяются следующие методы [6]: служебных сообщений, сопровождающей информации, зондирования.

При использовании *методов служебных сообщений* для передачи служебной информации между узлами сети применяются специальные сообщения, с помощью которых организуется обмен сведениями о состоянии элементов сети, о загрузке и уровнях очередей на узлах. В этом случае сообщения содержат сведения только о собственных очередях на узле, загрузке инцидентных трактов (включая оповещение о переполнении буферных накопителей) и о состоянии технических средств.

Для передачи служебной информации при использовании *методов сопровождающей информац*ии используются сообщения, несущие полезную информацию (см. гл. 2). Скорость передачи служебной информации при этом зависит от категории срочности того сообщения, с которым оно передается, что является недостатком этого метода. Метод находит широкое применение в сетях, характеризующихся преимущественно передачей высококатегорийных сообщений сравнительно небольшой длины.

Методы служебных сообщений и сопровождающей информации предполагают передачу служебной информации по мере ее возникновения с последующей принудительной рассылкой узлом сети, а не по мере необходимости. Алгоритм рассылки может быть произвольным.

При использовании *метода зондирования* вначале по инцидентным узлу-источнику трактам рассылается короткий прямой зонд-сигнал, содержащий адрес получателя. Затем этот сигнал транспортируется по всем инцидентным трактам на других узлах. На каждом узле запоминается входящий тракт, по которому поступал прямой зонд-сигнал. На узле, к которому непосредственно подключен абонент-получатель, формируется отраженный зонд-сигнал, который передается по запомненному ранее на узлах тракту и поступает на

узел-источник. Узел-источник осуществляет выдачу сообщения в тот тракт, из которого на него поступил отраженный зонд-сигнал. По мере прохождения сообщения запомненный путь “разрушается”. Этот метод целесообразно использовать на сетях с небольшим трафиком, поскольку при возрастании трафика резко растет поток прямых и отраженных зонд-сигналов, что приводит к уменьшению полезной производительности сети.

Литература

1. **Сокол Ш.** Прогнозирование состояний дискретного канала – Л.: ЛЭИС, 1985. – 17 с.
2. **Коричнев Л. П., Королев К. Д.** Статистический контроль каналов связи. – М.: Радио и связь, 1983. – 240 с.
3. **Арипов М. Н., Присяжнюк С. П., Шарифов Р. А.** Контроль и управление в сетях передачи данных с коммутацией пакетов. – Ташкент: ФАН, 1988. –160 с.
4. **Захаров Г. П., Архипов М. Н.** Проектирование и техническая эксплуатация сетей передачи данных. – М.: Радио и связь, 1989. – 360 с.
5. **Райцис Я. Н., Соколов В. А.** Специальные системы связи. Введение в системотехническое проектирование: Учебное пособие. – М.: МИС, 1991. – 81 с.
6. **Журавин А. И., Родионов А. В.** Управление сетями связи: Учебное пособие. – Л.: ВИКИ им. А.Ф. Можайского, 1989. – 50 с.
7. **Шумоподобные** сигналы в системах передачи информации. – М.: Сов. радио, 1973. – 424 с.

ГЛАВА 4
МОДЕЛИРОВАНИЕ ПРОЦЕССОВ В СЕТЯХ СВЯЗИ

4.1. Анализ методов моделирования трафика

Модели трафика условно можно разделить на две группы: традиционные и нетрадиционные. К традиционным моделям относят модели, основанные на состоянии, модели, основанные на временных рядах; к нетрадиционным- модели, основанные на данных наблюдения.

Модели, основанные на состоянии. Наиболее общим методом моделирования источника АТМ является предположение о существовании некоей машины с конечным числом состояний (FSM - finite state machine), определяющей поведение источника. Существуют два различных метода в моделировании, основанных на состоянии: либо процесс генерации ячеек модулируется FSM, либо процесс генерации ячеек моделируется непосредственно.

Процессы с модуляцией. Основная идея метода состоит в том, что FSM модулирует ожидаемую скорость (среднюю величину) основного процесса. Обычно предполагается, что FSM обладает марковскими или полумарковскими свойствами. Типовые модели источников этого класса различаются в следующих аспектах.

- По типу распределения времени пребывания. В большинстве моделей принимается, что время пребывания в определенном состоянии подчиняется отрицательному экспоненциальному распределению. Однако, время пребывания не всегда успешно можно моделировать отрицательным экспоненциальным распределением. Кроме того, оно не всегда может быть описано, как независимое от предыдущего состояния. В этом случае активности источника моделируются группами состояний, которые могут иметь распределение Кокса, эрланговское или гиперэрланговское.
- По типу модулируемого процесса. Модулируемый процесс часто (из соображений легкости математической обработки) выбирается пуассоновским. При этом ячейки генерируются независимо от всех предыдущих ячеек при постоянной скорости, определяемой модулятором. Если время пребывания выбрано в соответствии с отрицательным экспоненциальным распределением, этот вид стохастического процесса рассматривается как пуассоновский процесс, модулированный марковским (MMPP- Markov Modulated Poisson Process).

  Детерминированный процесс. Этот тип модели менее легко обрабатывается математически, чем пуассоновский, однако, он во многих случаях лучше отражает более низкие уровни модели источника АТМ (например, уровень ячейки или пачки).

  Флюидный поток (Fluid Flow). В этом виде модели "поток" ячеек источника аппроксимируется непрерывным флюидом, который "втекает" в

систему, то есть в модели не изучаются отдельные ячейки. Этот тип модели нацелен, в основном, на математический анализ системы.

- По структурным аспектам модели. Последним свойством, которое отличает различные типовые модели источника, является размер и структура пространства состояния. Часто вводят ограничения, чтобы уменьшить число параметров и увеличить легкость математической обработки. Одной из популярных моделей является модель с двумя состояниями типа ON-OFF (ВКЛ.-ВЫКЛ.), в которой ячейки формируются в период ON и не формируются в период OFF. Активность источника часто моделируется как цепь, то есть одномерная структура, где каждое следующее состояние зависит от предыдущего, но не более раннего состояния. Модель этого типа носит название мульти- мини- источника (multi-mini-source model). Такой источник может быть рассмотрен как объединение M независимых ON-OFF источников со средними ON- и OFF- периодами $1/\alpha$ и $1/\beta$ соответственно и со скоростью λ в ON-периоде. Несмотря на ее простоту, эта модель хорошо отражает некоторые из характеристик, например, источника с VBR (variate bit rate). Однако, она не способна отразить периодичности (например, кадровую), свойственные источнику с VBR. Для таких случаев разработаны процедуры, названные циркулянтными цепями.

Моделирование процесса генерации ячеек непосредственно. В этом методе уровень ячейки моделируется прямо. Это наиболее просто делается с помощью марковской модели дискретного времени с дискретными состояниями. Один временной шаг равен одному периоду ячейки. Процесс Бернулли является простейшим из них. В процессе Бернулли ячейка поступает в ячеечный слот с постоянной вероятностью p. Прибытие /генерация ячейки не зависит от предыдущего прибытия/генерации. Это соответствует геометрически идентично и независимо распределенным периодам между прибытиями ячейки со средним 1/p. Эта модель популярна благодаря легкости математическкой обработки и широко используется, несмотря на то, что не существует физически обоснованной мотивации, почему течение ячеек должно иметь именно эту характеристику.

Специальным классом моделей источника, который должен быть упомянут, является детерминированная модель источника, в которой ячейки источника всегда имеют одни и те же интервалы между прибытиями. Этот класс используется, чтобы моделировать источники с CBR (constant bit rate) . Фазы между прибытиями ячеек из различных источников определяют задержки буферизации и возможные потери ячеек в системе.

Модели, основанные на временных рядах. Эти модели получили наибольшее внимание в моделировании видео источников с VBR. Информационный поток источника может быть промоделирован как временной ряд, где, например, ожидаемое число бит в виде фреймов определяется числом бит в предыдущем фрейме. Другой простой моделью, ориентированной на временные ряды, которая является связующей с моделью, ориентированной на состояние, является дискретная модель

авторегрессии (DAR). В вышеупомянутых моделях предполагается, что ячейки из одного фрейма равномерно распределены на периоде фрейма, и что все фреймы закодированы подобно.

Пакетизированные данные из- за их пачечного характера и нелинейной природы особенно трудны для моделирования и предсказания при использовании традиционных моделей. Измерения трафика, собранные за большой период во многих странах мира, были использованы для получения статистических характеристик, необходимых для выработки стратегий управления сетью. В результате проведенных исследований появилась возможность сравнивать ранее используемые аналитические модели и подходы теории очередей с моделями, основанными на измерениях реального трафика, в силу чего были признаны непригодными ранее используемые модели в части общности, адаптивности и устойчивости. Для преодоления этих проблем многие исследователи повернулись лицом к альтернативным (“нетрадиционным”) техникам моделирования, таким как: калмановский фильтр, байесовский метод, аппарат нейронных сетей и нечеткой логики, фрактальный анализ .

Байесовский метод. Для целей управления важным параметром является пиковая и средняя интенсивность трафика. На практике вероятность получения среднего различных типов вызовов при их поступлении мала. Упомянутая проблема и проблема учета нелинейной природы трафика являются фундаментальными при построении генеральной модели потока трафика. Баейесовская теория обеспечивает метод для преодоления этих проблем без необходимых предположений о характере процесса прибытия (то есть, пуассоновский процесс или процесс Бернулли). Нелинейности моделируются в АТМ- потоке хорошо известным бета - распределением В (α, β), где распределение меняет форму, когда меняются переменные α и β. Решение основывается на функции, которая указывает, является ли вызов признанным или отвергнутым. Это решение может быть задано в виде зависимости

$$S_n = \sum_i \frac{N_i}{a_i},$$

где N_i- ширина полосы на вызов, a_i- пиковая скорость ячейки на вызов. N_i и a_i – содержат априорную информацию о источнике трафика в некоторое время и определяют ожидаемую степень точности. Недостатком байесовского метода является то, что оценки обычно, высоко "настраиваемые" и оптимизированы для некоторого источника трафика и, таким образом, не вносят большой устойчивости решения. Из-за этой относительно высокой степени оптимизации байесовский метод позволяет создавать очень точные модели потока трафика, до тех пор, пока источник существенным образом не изменит свои характеристики, и является мощным инструментом для описания нелинейного или пачечного трафика. Это положение обеспечивает успех упомянутого метода по сравнении с обычным методом , подобным MMPP.

Фильтр Калмана долгое время рассматривался, в основном, как метод для моделирования и рекурсивного предсказания поведения динамических систем. В последнее время делались попытки использовать фильтр Калмана для предсказания интенсивности голосового трафика на один шаг вперед по времени, основанного на предшествующей интенсивности трафика и среднем наблюдаемого трафика. Проблемой в использовании фильтра Калмана является выбор метода моделирования нестационарной и нелинейной системных динамик. Эта проблема может быть решена на основе известного метода Box'a и Jenkins'a с использованием модифицированной версии Sage-Huza. Результаты, полученные применением этого метода на French Network, для голосового трафика в реальном масштабе времени дали удовлетворительные результаты. Основным недостатком этого метода является недостаток устойчивости фильтров Калмана.

Нейронные сети и самообучающиеся системы с нечеткой логикой могут служить примерами алгоритмов, которые приспособлены к адаптивным моделям с высокой нелинейностью процессов при минимуме априорных предположений. Метод нечеткой логики имеет два преимущества: устойчивость к шумам и способность к самообучению. Ряд авторов основывали фазу обучения своего алгоритма на методе построения нечетких отношений посредством адаптивной кластеризации. Нечеткие отношения рассматривались при этом в качестве аналога функции передачи системы. После фазы обучения использовались нечеткие правила вместе с наблюдаемыми величинами для предсказания будущих величин трафика. Метод использовался для демонстрации эффективности моделирования пачечного видео трафика в пакетизированной сети. Однако, несмотря на достаточно высокую степень точности оценивания, все вышеописанные нетрадиционные методы, все- таки требуют некоторого количества априорных предположений. Вместе с тем использование моделей пакетного трафика большей частью игнорирует вопросы, относящиеся к физической основе, на которой они имеют значение. В результате сложные модели трафика требуют большого числа параметров, но обеспечивают малое проникновение в динамику трафика, наблюдаемого на реальных сетях. Поэтому ввиду наблюдаемой самоподобной природы измеренного сетевого трафика и его поведения, радикально отличного от поведения трафика, предсказанного используемыми моделями, возобновился интерес в обеспечении физического базиса для предлагаемых моделей с целью идентификации некоторых существенных характеристик высокосложной структуры "живого" сетевого трафика. К таким характеристикам относятся долговременная зависимость (LRD), медленно спадающая дисперсия, распределения с утяжеленными хвостами, фрактальные характеристики. В частности в ряде работ был развит подход, обеспечивающий простое толкование для наблюдаемого самоподобия пакетного трафика в терминах природы трафика, генерируемого парой источник- получатель. Суперпозиция многих источников типа ON-OFF, известных также как "packet train models", каждый из которых демонстрирует

феномен, называемый "эффект Ноя" (синоним "бесконечной дисперсии"), проявляется в самоподобном агрегированном трафике. Структура ON-OFF-моделей источников устанавливает тождество эффекта Ноя как существенную точку отклонения от традиционного трафика к самоподобному. Эффект Ноя для а ON-OFF-моделей отдельных источников приводит к высокопеременным периодам ON-OFF, которые с некоторой вероятностью могут быть очень большими. Другими словами, эффект Ноя гарантирует, что каждый ON-OFF – источник отдельно показывает характеристики, которые перекрывают широкий диапазон временных шкал. Математически для расчета эффекта Ноя использовались распределения с утяжеленными хвостами, то есть некоторый Парето- тип. Параметр α, описывающий утяжеление хвоста такого распределения, дает измерение интенсивности эффекта Ноя. Они также обеспечивают простое соотношение между α и Херст-параметром H, который был предложен как мера степени самоподобия (или эффект Иосифа) трафика. Наоборот, традиционное моделирование существенно ограничивают активности ON-OFF- источников и, как следствие, много таких источников ведет себя подобно белому шуму, в том смысле, что агрегированное течение трафика лишено любой значительной корреляции, исключая возможно, короткие периоды. Статистический анализ трасс трафика на сетях Ethernet (для 100- 1000 пар активных источников) показал, что реальный трафик совпадает с ON-OFF- моделями и распределение времени пребывания в ON-OFF-состояниях может быть точно описано с использованием распределений типа Парето. Очевидно, что эти данные помогают объяснению наблюдаемой устойчивости характеристик фрактального трафика. Кроме того, это также устанавливает потенциальные подходы к моделированию самоподобного трафика и его анализу.

Одна из первых моделей для моделирования самоподобного трафика основана на хаотических отображениях. Хаос – это явление, которое описывается детерминированным процессом, причем такое описание возникает при анализе даже достаточно простых нелинейных динамических систем. При описании системы задаются ее начальное состояние и динамические законы, описывающие ее работу, т.е. процесс изменения состояния во времени.

Хаос (стохастическое по внешнему виду поведение) возникает вследствие чувствительной зависимости траектории изменений состояний системы от начальных условий. Если f(x) – хаотическое отображение и существуют две траектории с почти одинаковыми начальными условиями x_0 и $x_0 + \varepsilon$, то чувствительная зависимость от начальных условий может быть определена в виде

$$|f^N(x_0+\varepsilon)-f^N(x_0)|=\varepsilon e^{N\lambda(x_0)},$$

где N- номер состояния системы. Иначе говоря, траектории, начинающиеся от произвольного близких начальных условий, тем не менее, могут расходиться с экспоненциальной интенсивностью. Параметр $\lambda(x_0)$, описывающий экспоненциальную расходимость, называется показателем Ляпунова, причем

для того, чтобы отображение было хаотическим, он должен быть положительным "почти для всех" x_0. Основная предпосылка классического анализа динамических систем состоит в том, что если известны начальные условия, дальнейшее поведение системы может быть вычислено для всех моментов времени. На практике же начальные условия могут быть заданы лишь с некоторой конечной точностью. Такого рода неопределенность в начальных условиях растет по экспоненциальному закону, что и делает непредсказуемыми долгосрочные характеристики подобных систем. Моделирование нагрузки систем цифровой передачи может осуществляться с помощью кусочно-линейного и прерывистого отображения. Даже, если пакетная нагрузка является очень нерегулярной и пачечной, предоставляется возможность построения простых нелинейных моделей первого или второго порядка, которые позволяют преодолеть многие трудности. Моделирование может осуществляться при условии, что источник генерирует пачку пакетов при пиковой скорости (соответствующей состоянию ON), когда переменная состояния выше порога , и не генерирует никаких пакетов, когда она ниже порога (состояние OFF). Используя подходящий выбор $f(\cdot)$ можно моделировать распределение времени пребывания либо с "легким хвостом", либо с "тяжелым хвостом" с бесконечной дисперсией в ON- и OFF- состояниях.

Другим методом является моделирование процесса частичного броуновского движения (fractional Brownian motion process- fbm) с Херст-параметром $H \in [½, 1]$. Это гауссовский процесс с нулевым средним и стационарными инкрементами и ковариационной структурой:

$$Cov[Z(t), Z(s)] = \sigma^2/2(t^{2H}+s^{2H}-|t-s|^{2H}).$$

В особом случае, когда $H=½$, $Z(t)$ есть стандартное броуновское движение. Самоподобные свойства $Z(t)$ основываются на факте, что $Z(\alpha t)$ идентичен по распределению с $\alpha^H Z(t)$. Инкрементный процесс $X(k)=Z(k+1)-Z(k)$, $k\geq 0$ называется частичным гауссовым шумом (fractional Gaussian noise- fGn) и является стационарным (дискретно- временным) гауссовым процессом с автокорреляционной функцией

$$r(k)=½(|k+1|^{2H}- 2|k|^{2H}+ |k-1|^{2H}),\ k\geq 1.$$

Асимптотически $r(k)\approx H(2H)|k|^{-2H-2}$, то есть демонстрирует долговременную зависимость (LRD). Экспериментальным анализом трасс было показано, что fBm- модель является разумным представлением “строго самоподобного трафика”, например, трафика данных, который имел ту же самую корреляционную структуру в сетях Ethernet в целом ряде диапазонов временных шкал (~10 мс через ~1000 сек). Наоборот, как показано в ряде работ трафик VBR должен иметь коротковременную зависимость (SRD), а корреляционная структура долговременной зависимости (LRD) наблюдается только на временных шкалах после нескольких секунд (“асимптотическое самоподобие”).

Таким образом, анализ методов моделирования трафика показал, что методы моделирования могут быть условно разделены на две группы:

традиционные и нетрадиционные. К традиционным моделям относятся модели, основанные на состоянии и временных рядах, к нетрадиционным – модели, основанные на данных наблюдения. Модели, основанные на состоянии, классифицируются как модели процессов с непосредственной генерацией ячеек и модели процессов с модуляцией. Типовые модели источников этого класса различаются в следующих аспектах: по типу распределения времени пребывания в определенном состоянии (например, отрицательное экспоненциальное распределение), по типу модулируемого процесса (например, пуассоновский) по структурным аспектам модели, то есть размеру и структуре пространства состояний (например, модель с двумя состояниями типа ON/OFF). Наиболее часто используемой моделью для пакетизированного трафика является модель пуассоновского процесса, модулированного марковским (Markov Modulation Poisson Process- MMPP). Модели, основанные на временных рядах, получили наибольшее внимание в моделировании видео - источников с VBR.

где ожидаемая скорость (средняя величина) основного процесса, модулируется каким- либо процессом, например, пуассоновским, а время пребывания в определенном состоянии выбрано в соответствии с отрицательным экспоненциальным распределением, этот вид стохастического процесса рассматривается как пуассоновский процесс, модулированный марковским (MMPP- Markov Modulated Poisson Process)

К нетрадиционным моделям относятся модели, основанные на данных наблюдения. Измерения трафика, собранные за большой период времени во многих странах мира были использованы для получения статистических характеристик, необходимых для выбора стратегий управления сетью. В результате было установлено, что традиционные модели непригодны в части общности, адекватности и устойчивости. Для преодоления этих проблем были разработаны альтернативные техники моделирования, такие как калмановский фильтр, байесовсий метод, аппарат нейронных сетей и нечеткой логики. Несмотря на достаточно высокую степень точности оценивания, все вышеизложенные нетрадиционные модели требуют некоторого количества априорной информации, кроме того, они мало соответствуют динамике высокосложной структуры “живого” трафика. Поэтому в связи с наблюдаемой самоподобной природой трафика, который показал такие характеристики, как LRD, медленно спадающую дисперсию, распределения с утяжеленными хвостами, в качестве моделей стали использоваться хаотические отображения и частичное броуновское движение. Анализ системы очередей, приводимых в движение потоками фрактального трафика, показал, что модель fBm воспроизводит поведение очереди, наблюдаемой в реальных трассах Ethernet.

4.2. Фрактальные свойства трафика современных сетей связи

Постановка задачи и ее решение обусловлены тем, что существующие и используемые модели процессов в высокоскоростных сетях не соответствуют реальным характеристикам потоков информации. Использование теории самоподобных процессов позволило разрешить эту проблему и создать математические модели, учитывающие характерные особенности высокоскоростных сетей связи.

Развитие сетей связи, а также успехи новых сетевых технологий (таких как SMDS, Frame Relay, ATM) и приложений (таких как видео по требованию и т.д.) возобновили интерес к изучению трафика современных сетей, генерируемого реальными службами и приложениями. Главными побудительными причинами этого интереса послужило желание проверить выводы, сделанные на основе расчетов с использованием традиционных моделей трафика с характеристиками реальных потоков трафика. Расчеты, основанные на традиционных представлениях о том, что мультиплексирование большого числа независимых потоков цифровой передачи приводит к пуассоновскому процессу, явилось причиной грубых ошибок при проектировании коммутаторов ATM первого поколения [1]. Когда такие коммутаторы с небольшими накопителями (10 – 100 ячеек) были пущены в эксплуатацию, потери ячеек оказались недопустимо большими, что заставило конструкторов внести необходимые изменения. Высококачественные измерения трафика с высоким разрешением обнаружили, что фактическая нагрузка в исследованных сетях существенно отличается как от классических представлений (телефонный трафик), так и от новых моделей (пакетный трафик), рассматриваемых в литературе. Отличительная особенность нагрузки быстродействующих цифровых сетей - ее пачечный характер, причем пачки (скученности) появляются в разных масштабах времени, и это затрудняет определение длин пачек: в разных шкалах времени длительность пачки может изменяться в пределах от миллисекунд до минут и часов в зависимости от разрешающей способности измерительной аппаратуры. Трафик, который является пачечным на многих или всех масштабах времени может быть описан статистически, используя понятие самоподобия. Самоподобие- это свойство фрактала- объекта, чье проявление не изменяется, несмотря на масштаб, при котором он наблюдался.

Математическая теория фракталов восходит к концу 1920-х годов к работам таких известных математиков как Хаусдорф, Безикович, Урысон Однако, широкую известность теория фракталов получила после исследований Мандельброта, который показал, что фракталы могут быть использованы, чтобы описать модель естественного явления широкой изменчивости. До сих пор строгого и полного определения фрактала не существует. Обычно под фракталом понимают самоподобные объекты,

инвариантные относительно локальных дилотаций, то есть объекты, которые подобны на различных пространственных масштабах рассмотрения.

Термин "фрактал" был введен Бенуа Мандельбротом в 1975 году. Исследуя фигуры произвольной сложности и неупорядоченности, Мандельброт использовал размерность Хаусдорфа. В результате в 1977 году Мандельброт провозгласил существование множеств с дробной хаусдорфовой размерность. Таким образом, появилось определение: фракталом называется множество, размерность Хаусдорфа- Безиковича которого строго больше его топологической размерности. Это определение при всей его правильности в точности слишком ограничительно. Оно не могло охватить многие "пограничные" фракталы, встречающиеся в физике. До сих пор понятие размерности среди всех обсуждений фракталов остается центральным. Позднее Мандельброт сузил свое предварительное определение, предложив заменить его следующим: фракталом называется структура, состоящая из частей, которые в каком-то смысле подобны целому. Второе определение включает существенно отличительный признак: наблюдаемый в эксперименте фрактал выглядит одинаково, в каком бы масштабе его не наблюдали. В основе этого понятия содержится важная идеализация действительности. Нет ни одной реальной структуры, которую можно было бы последовательно увеличивать бесконечное число раз и которая выглядела бы при этом неизменной, но тем не менее принцип самоподобия в приближенном виде имеется в природе: в линиях берегов морей и рек, в очертаниях облаков и деревьев, в турбулентном потоке жидкости и иерархической организации живых клеток. Хотя такая идеализация и может оказаться слишком большим упрощением действительности, она на порядок увеличивает глубину математического описания природы.

С момента появления работы Мандельброта началось стремительное проникновение идей фрактальной геометрии в различные области современного естествознания. Интерес к фракталам стал проявляться в большей степени после того, как было установлено большое число явлений и задач, где фрактальная структура (размерность) служит основой характеристики системы. Фрактальная геометрия, сформировавшая некую фундаментальную основу, позволила идти дальше и создать фрактальный анализ. Появился прорыв в понимании сложных явлений ранее не поддававшихся математическому описанию. Фракталы обнаруживаются и в структуре твердых тел, и в турбулентных потоках, и на фазовых пространствах динамических систем. В случае стохастического объекта, подобного временным рядам, самоподобие используется в статистическом смысле: статистические характеристики пакетной нагрузки имеют структурное сходство при ее измерении в разных масштабах времени. Формально в рамках временных рядов и сопровождающих их статистических процедур самоподобие может быть описано следующим образом. Пусть $X=(X_t, t=1,2,3,\ldots)$- стационарный случайный процесс с нулевым среднем и функцией автокорреляции $r(k)$, $k>0$. Для каждого

m=1, 2, 3… может быть определена новая стационарная последовательность случайных величин

$$X^{(m)}=(X_k^{(m)}, k=1,2,3,\ldots),$$

которая получается путем усреднения первоначальной последовательности X по непересекающимся блокам размера m. Иначе говоря, для каждого m (m=1, 2, 3…) случайная величина $X^{(m)}$ задается в виде

$X_k^{(m)}=(1/m)\,(X_{km-m+1}+\ldots+X_{km}),\ k\geq 1.$

Процесс называется самоподобным с параметром H (H- самоподобным), если для всех положительных m $X^{(m)}$ имеет то же самое распределение, как X с изменением масштаба в m^H, т.е.

$$X_t \overset{d}{=} m^{-H} \sum_{i=(t-1)m+1}^{t_m} X_i \quad \text{для всех } m\in N, \tag{4.1}$$

где $\overset{d}{=}$ означает равенство по распределению.

Если X – H- самоподобный, он имеет ту же самую автокорреляционную функцию

$$r(k)=E[(X_t-\mu)(X_{t+k}-\mu)]/\sigma^2,$$

как $X^{(m)}$ для всех m. Это означает, что агрегированные последовательности дистрибутивно самоподобны: распределение агрегированных последовательностей то же самое (за исключением изменения в масштабе), что и распределение первоначальной последовательности.

Как результат, самоподобные процессы могут показывать долговременную зависимость (Long- range dependence - LRD). Процесс с LRD имеет автокорреляционную функцию $r(k)\sim k^{-\beta}$, когда $k\to\infty$, где $0<\beta<1$. Таким образом, автокорреляционная функция такого процесса следует степенному закону в отличие от экспоненциального спада, показываемого традиционными моделями трафика. Спад по степенному закону медленнее, чем экспоненциальный спад, а так как $\beta<1$, ряд, образованный последовательными значениями коэффициента автокорреляции расходится, то есть $\sum_{\kappa=1}^{\infty} r(k)=\infty$ (признак долговременной зависимости Кокса). Из этой ситуации следует ряд выводов:

1. Дисперсия среднего значения выборок из таких рядов не уменьшается с увеличением объема выборки по закону обратной пропорциональности от этого объема, что типично для традиционных стационарных случайных процессов. Для самоподобных процессов характерно более медленное уменьшение дисперсии по закону $D[X^{(m)}]\cong c_1 m^{-\beta}$ при $m\to\infty$, $0<\beta<1$, c_1 и c_2- некоторые константы. Этот признак называется медленным убыванием дисперсии.
2. Эквивалентная формулировка долговременной зависимости в частотной области (по теореме Винера- Хинчина) может быть описана в виде

соответствующей степенной функции. Спектральная плотность самоподобного процесса в окрестности начала координат имеет вид $f(\lambda) \cong c_2\lambda^{-\gamma}$ при $\lambda \to \infty$, причем $0<\gamma<1$, где $\lambda=1-\beta$. Этот признак получил название спектра типа 1/f.

Одной из привлекательных характеристик использования самоподобных моделей для временных рядов, является характеристика степени самоподобия, которая выражается с использованием только единственного параметра. По историческим причинам используемый параметр называется Херст-параметром. Открытие Х.Э. Херстом (Hurst) (1951 г.) нового статистического метода (метода нормированного размаха, или метода R/S) послужило важным толчком для фундаментальных исследований, направленных на изучение самоподобных процессов. Как обнаружил Херст, для многих естественных процессов нормированный размах R/S очень хорошо описывается эмпирическими соотношениями для больших N:

$$R/S=(N/2)^H,$$

где R- разность между максимумом и минимумом; S- стандартное отклонение, то есть корень квадратный из дисперсии; N- дискретное время; H- параметр Херста. Параметр Херста лежит в пределах $0,5 \leq H \leq 1$, причем H=0,5 соответствует случаю отсутствия самоподобия, а H=1- случаю, означающему детерминированный характер процесса. Таким образом, самоподобный процесс характеризуется значениями параметра Херста, ограниченными строгим неравенством $0,5< H < 1$. Параметр Херста более или менее симметрично распределен вокруг среднего значения 0,73 со стандартным отклонением, равным ≈ 0,09 [3]. Построив график зависимости R/S от N в логарифмическом масштабе по обеим шкалам, Херст нашел, что график R/S в зависимости от N имеет наклон, который является оценкой H. Именно такой метод и положен в основу современного анализа статистических данных и определения параметра Херста. Таким образом, для самоподобных рядов с LRD, $0,5<H<1$. Когда $H \to 1$, степень самоподобия и LRD возрастает.

Мандельброт в своей работе заметил, что много негладких структур природы, которые привлекают внимание, во многих случаях трудны для документирования, однако, Библия предлагает два исключения. Трудно не увидеть историю Ноя [4], как притчу о неравномерном выпадении осадков на Среднем Востоке, и историю Иосифа как притчу о склонности сырых и сухих лет группироваться в мокрые периоды и периоды засухи. Мандельброт Б.Б. дал этим историям термины «эффект Ноя (Noah Effect)» и «эффект Иосифа (Joseph Effect)». Таким образом, в качестве меры степени самоподобия часто используется термин «эффект Иосифа».

В своей основе понятие "долговременная зависимость" и "самоподобие" не эквивалентны. Понятие долговременной зависимости включает поведение хвоста автокорреляционной функции стационарных временных рядов, в то время как самоподобие относится к поведению

масштабированных процессов с конечномерными распределениями непрерывного или дискретного типа.

Однако Кокс (Cox) ввел термин "строгое самоподобие в широком смысле" (exactly second- order self-similar) для стационарных рядов, чьи агрегированные процессы обладают той же самой невырожденной автокорреляционной функцией, как и исходный процесс.

Процесс X называется строго самоподобным в широком смысле случайным процессом с Херст- параметром H=1- (β/2), если для всех

$$m \in N=\{2,3,\ldots\}$$

$$\text{var } X^{(m)} = \sigma^2 m^{-\beta}, \tag{4.2}$$

$$r^{(m)}(k)=r(k), \quad k \in N=\{0,1,2,\ldots\}. \tag{4.3}$$

Согласно [2] такое определение самоподобия не вполне подходит для расширения определения на асимптотически самоподобные процессы в широком смысле. Было показано, что X удовлетворяет (4.2), если и только если его автокорреляционная функция имеет вид

$$r(k)=1/2[(k+1)^{2-\beta}- 2k^{2-\beta}+(k-1)^{2-\beta}] \stackrel{\Delta}{=} g(k),\ 0<\beta<1,\ k \in N=\{0,1,2,\ldots\}. \tag{4.4}$$

Его спектральная плотность

$$f(\lambda) = c\left|e^{2\pi i\lambda} - 1\right|^2 \sum_{t=-\infty}^{\infty} \frac{1}{|\lambda + l|^{3-\beta}}, \quad -\frac{1}{2} \le \lambda \le \frac{1}{2}, \tag{4.5}$$

где c- постоянная, заданная нормализацией $\int_{-1/2}^{1/2} f(\lambda)d\lambda = \sigma^2$.

Как следует из (4.5) функция f(λ) имеет сингулярность типа $f(\lambda) \sim \text{const}\,|\lambda|^{\beta-1}$ при λ=0.

Таким образом, процесс X - строго самоподобный в широком смысле с параметром H=1-β/2, 0<β<1, если и только если:

1. его спектральная плотность имеет форму (4.5), или если и только если
2. он удовлетворяет условию (4.2).

Гауссовская последовательность с нулевым средним H- самоподобна, если и только если ее автокорреляционная функция равна g(k).

Асимптотическое самоподобие в широком смысле было определено соответствующим образом. Процесс X называется асимптотически самоподобным в широком смысле с параметром H=1-β/2, 0<β<1, если все $k \in N$

$$\lim_{m\to\infty} r^{(m)}(k) = \frac{1}{2}[(k+1)^{2-\beta} - 2k^{2-\beta} + (k-1)^{2-\beta}] \stackrel{\Delta}{=} g(k). \tag{4.6}$$

Строго самоподобный процесс с параметром H=1-β/2, 0<β<1, должен иметь $r(k) \sim H(2H-1)k^{-\beta}$, в то время как процесс, являющийся асимптотически самоподобным с параметром H=1-β/2, 0<β<1, имеет $r(k) \sim ck^{-\beta}$ с некоторой постоянной c, которая не является необходимо равной H(2H-1). Смысл

определения (6) состоит в том, что X- является асимптотически самоподобным процессом в широком смысле, если после усреднения по блокам длины m и $m\to\infty$ он сходится к строго самоподобному в широком смысле процессу. Здесь сходимость уже не к исходному доусредненному процессу X, а к строго самоподобному в широком смысле процессу.

Следующие утверждения являются эквивалентными:

а) X есть процесс асимптотически самоподобный в широком смысле, то есть справедливо (4.6);

б) $(V_{km} / V_m)\sim k^{-\beta}$, целое $m\to\infty$, $k\in N$;

c) $r(k)\sim H(2H-1)L(k)k^{-\beta}$, целое $k\to\infty$ влечет асимптотическое равенство

d) $V_m \sim\sigma^2 L(m)m^{-\beta}$, целое $m\to\infty$,

где L(k)- функция, медленно меняющаяся на бесконечности и каждое из (c) и (d) влечет (a) и (b) (то есть измеримая функция $f(x)>0$ называется медленно меняющейся на бесконечности, если при каждом $u>0$ $f(ux)\to 1$ при $x\to\infty$).

С точки зрения введенных определений термины " долговременная зависимость" и "самоподобие строгое или асимптотическое в широком смысле" могут использоваться во взаимозаменяемой форме, потому что оба отсылают к поведению хвоста автокорреляционной функции и существенно эквивалентны.

Например, что касается гауссовых процессов, то в случае фрактального броуновского движения (fBm), также как и процесса приращения (т.е. фрактального гауссовского шума- fGn) они рассматриваются как самоподобные. В то же время, в первом случае самоподобие относится к поведению процессов масштабированных с конечномерными распределениями непрерывного времени, а во втором случае, оно понимается как самоподобие строгое в широком смысле и является синонимом долговременной зависимости.

Связь между строго самоподобным в широком смысле и самоподобным в узком смысле процессом, введенным Колмогоровым, Мандельбротом и др., аналогична связи между процессами, стационарными в широком и узком смысле.

Необходимо отметить, что если имеется два процесса X' и X'' такие, что $r(k)\sim c_1k^{-\beta}{}_1$, $k\to\infty$ для X' и $r(k)\sim c_2k^{-\beta}{}_2$, $k\to\infty$ для X'', где c_i и β_i, i=1,2- постоянные, $0<c_i<\infty$, $0<\beta_i<1$, тогда $(X'+X'''')$- процесс асимптотически самоподобный в широком смысле с параметром $H=1-\beta/2$, где $\beta=\min(\beta_1,\beta_2)$. При этом оба процесса X' и X'''' асимптотически самоподобны, X' с $H_1=1-\beta_1/2$, $0<\beta_1<1$, а X" с $H_2=1-\beta_2/2$, $0<\beta_2<1$. Таким образом, объединение потоков, асимптотически самоподобных в широком смысле, производит асимптотически самоподобный поток.

Пусть процессы X' и X"- строго самоподобные в широком смысле, X' с H_1, а X" с H_2. Если $H_1=H_2=H$, тогда $(X'+X'')$- строго самоподобный с параметром H. Если $H_1\neq H$, тогда $(X'+X'')$- процесс не строго самоподобный в широком смысле, но асимптотически самоподобный в широком смысле с

$H=\max(H_1, H_2)$. Эти результаты важны для анализа применений АТМ-мультиплексоров и коммутаторов, так как они дают условия, при которых самоподобные в широком смысле течения объединяются в строго или асимптотически самоподобные в широком смысле течения.

Литература

1. **Нейман В.И**. Новое направление в теории телетрафика// Электросвязь, 1998, №7, с.27-30.

2.**Tsybakov B., Georganas N.D.//** IEEE Trans. on Information Theory, 1998, v.44, №5, p.1713-1725

ГЛАВА 5
КАЧЕСТВО ОБСЛУЖИВАНИЯ

5.1. Модель качества обслуживания в среде B-ISDN

Технология пакетной коммутации была введена в 1970-х годах. При этом в процессе стандартизации была обсуждена и определена концепция качества обслуживания (Quality of Service- QoS), однако, она не получила достаточного развития и много пунктов, относящихся к QoS, были оставлены с формулировками "для дальнейшего изучения", или "оставить провайдеру сетевой службы", или не упоминались вовсе . До сих пор разработка стандартов была сосредоточена на протоколах взаимодействия, а не на обмене информации, поэтому аспекты QoS выпадали из контекста стандартизации (т.е. были служебно- и протокольно- ориентированными, а не QoS-ориентированными). В настоящее время недостатки протоколов с отсутствием QoS отчетливо прояились в работе Internet в ситуациях с высокой нагрузкой, когда Internet периодически так перегружалась, что практически не могла использоваться для плановой профессиональной работы. Теперь, когда технологии ATM предназначается роль стержня будущей информационной суперскоростной магистрали, концепции QoS уделяется гораздо больше внимания в стандартах ATM. Кроме того, в последнее время наметился рост числа приложений с хорошо определенными требованиями к QoS, например, развертывание ряда приложений, так называемой, "мультимедийной культуры". Не исключено, что в ближайшем будущем могут появиться совершенно новые виды приложений, для которых определяющим будет выполнение жестких требований к QoS.

Однако, важной стороной этого вопроса является трудность обработки информации с учетом QoS. Соответствующая обработка требует общего вида параметров поведения пользователя и понимания различных сторон функционирования системы. Хорошо известно, что ATM более уязвима с точки зрения плохих измерений и неоптимального функционирования, чем сеть с пакетной коммутацией. Способность системы обеспечивать некоторое QoS зависит от архитектуры системы и, таким образом, от принципов реализации функций системы. Оптимальная архитектура ориентирована на получение оптимального трафика. Таким образом, получение оптимального трафика существенно зависит от оптимальной архитектуры QoS.

Архитектура QoS основана на ряде рекомендаций и документов различных международных организаций, занимающихся разработкой стандартов, или организаций, способствующих ускорению развития и размещению ATM-продуктов через спецификации взаимодействия. Как известно, стандарты ATM разрабатываются, по крайней мере, двумя организациями: ITU (International Telecommunication Union) и Форумом ATM (ATM Forum). Вследствие этого, одни и те же функции определяются зачастую

несколькими параллельными стандартами. В рекомендации ITU E.800 (Quality of Service and Dependability Vocabulary) "качество обслуживания- QoS" определено как "суммарный эффект характеристик службы, который определяет степень удовлетворения пользователя службы". Это очень общее определение и, поэтому необходимо более точное определение QoS, которое было бы основано на современной версии "качества обслуживания", опирающейся на ряд документов, относящихся к концепции QoS. В первую очередь это основная эталонная модель взаимодействия открытых систем (ЭМВОС-OSI), определенная в рекомендациях X.200 и ISO/IEC 7498-1, в которых описана модель и действия, необходимые для взаимодействия систем, использующих коммуникационную среду. В ЭМВОС дана концепция QoS. Дополнением к описанию QoS в ISO/IEC "Quality of Service Framework" содержится определение модели QoS и определение семантики параметров QoS. Некоторые положения, касающиеся аспектов QoS, в рекомендациях I.350 из серии рекомендаций для N-ISDN (I.350: General Aspects of QoS and Network Performance in Digital Networks, Including ISDN) также применимы к B-ISDN. Кроме того, в основу современной концепции QoS входит ряд рекомендаций: I.356 (B-ISDN ATM Layer Cell Transfer Performance); I.321 (B-ISDN Protocol Reference Model); I.371 (Traffic Control and Congestion Control in B-ISDN); Q93B (User-Network and Network-Network Signaling); ATM Forum: UNI 3.0 – и ряд других.

В сфере ВОС (OSI) четко различаются 3 степени абстрагирования: архитектура, спецификация услуг и спецификация протоколов.

Архитектура ВОС представляет собой высшую степень абстрагирования в схеме ВОС. Архитектура ВОС определяет типы объектов, используемые для описания открытой системы, общие соотношения между этими типами и общие условия, налагаемые на них и на их соотношения, а также семиуровневую модель взаимодействия между процессами, сконструированную на основе этих объектов, соотношений и условий (ЭМВОС). Для обозначения произвольного уровня модели используется условное алфавитно- числовое обозначение как (N)- уровня, а смежных с ним нижнего и верхнего уровней как (N-1)- уровня и (N+1)- уровня соответственно. Функциональные возможности (N)- го уровня, которые предоставляются в распоряжение (N+1)- компоненты, называются услугами. К понятию услуг относятся не все функции, выполняемые внутри (N)- уровня, а только те из них, которые могут использоваться смежным верхним уровнем. (N)- услуги предоставляются (N+1)- компоненте в (N)- точках доступа (N- ТДУ) . Служба определяется через набор услуг, которые она предоставляет. В такой интерпретации служба предоставляет услуги тем частям системы, которые находятся над границей службы и в совокупности называются пользователями службы. Части системы, находящиеся ниже границы службы, в совокупности называются исполнителем (провайдером службы).

"Качество обслуживания" в документе МОС/ВОС (ISO/ OSI) определяется как "ряд качеств, отнесенных к обеспечению (N)- службы, которые воспринимаются пользователями (N)- службы.

ITU определяет телеслужбы (teleservice) и опорные службы (bearer service). Телеслужба- это служба, которую пользователь получает из пользовательского терминала, в то время как опорная служба- это служба, предоставляемая на некотором интерфейсе между пользователем и сетью. В дополнение к концепции службы ITU в рекомендации I.350 ввел концепцию сетевой характеристики (network performance- NP), определенную как "способность сети или части сети предоставлять функции, связанные с коммуникацией между пользователями". QoS ориентировано на пользователя, а NP- на провайдера. NP-измеряется в терминах параметров, которые указываются сетевыми провайдерами и используется для целей проектирования систем, конфигурации, работы и управления. NP не зависит от действий пользователя и терминала.

QoS измеряется в терминах параметров, которые могут непосредственно наблюдаться и измеряться в точке, в которой служба доступна для пользователя. На рис.5.1 представлено QoS и NP в среде ISDN. В связи с QoS или вместо QoS часто используется концепция характеристики трафика. QoS прямо связывается с использованием общих ресурсов трафика. Примерами ресурсов трафика являются: узлы, емкость передачи, передающие каналы, маршрутизаторы, логические каналы, буферы, окна, а также ресурсы обработки и схемы интерфейсов в узлах и оконечных системах. Таким образом, количественная мера QoS прямо относится к использованию ресурсов, вовлеченных в обеспечение службы, т.е. трафику на эти ресурсы. Поэтому характеристика трафика и QoS являются двумя строго родственными концепциями.

В то время как служба включает в себя некоторые общие функциональные возможности, QoS- служба включает только те аспекты службы, которые имеют значение для определения QoS. QoS- служба определяет природу QoS-параметров, переносимых на служебных примитивах, и сохраняет те же самые отношения, что служба, между близлежащими уровнями.

Функциональная архитектура коммуникационной службы определяется как общий набор функциональных элементов и динамических отношений между этими функциональными элементами. Эта архитектура имеет операционную и управляющую (менеджментную) части. Операционная архитектура определяет основное назначение, относящееся к обработке в реальном времени вызова, в то время как архитектура менеджмента определяет дополнительное назначение, необходимое для администрирования этой операционной части.

Функциональная архитектура состоит из плоскости пользователя для передачи данных, плоскости управления для управления вызовом и соединением и плоскости менеджмента для менеджмента (рис 5.2).

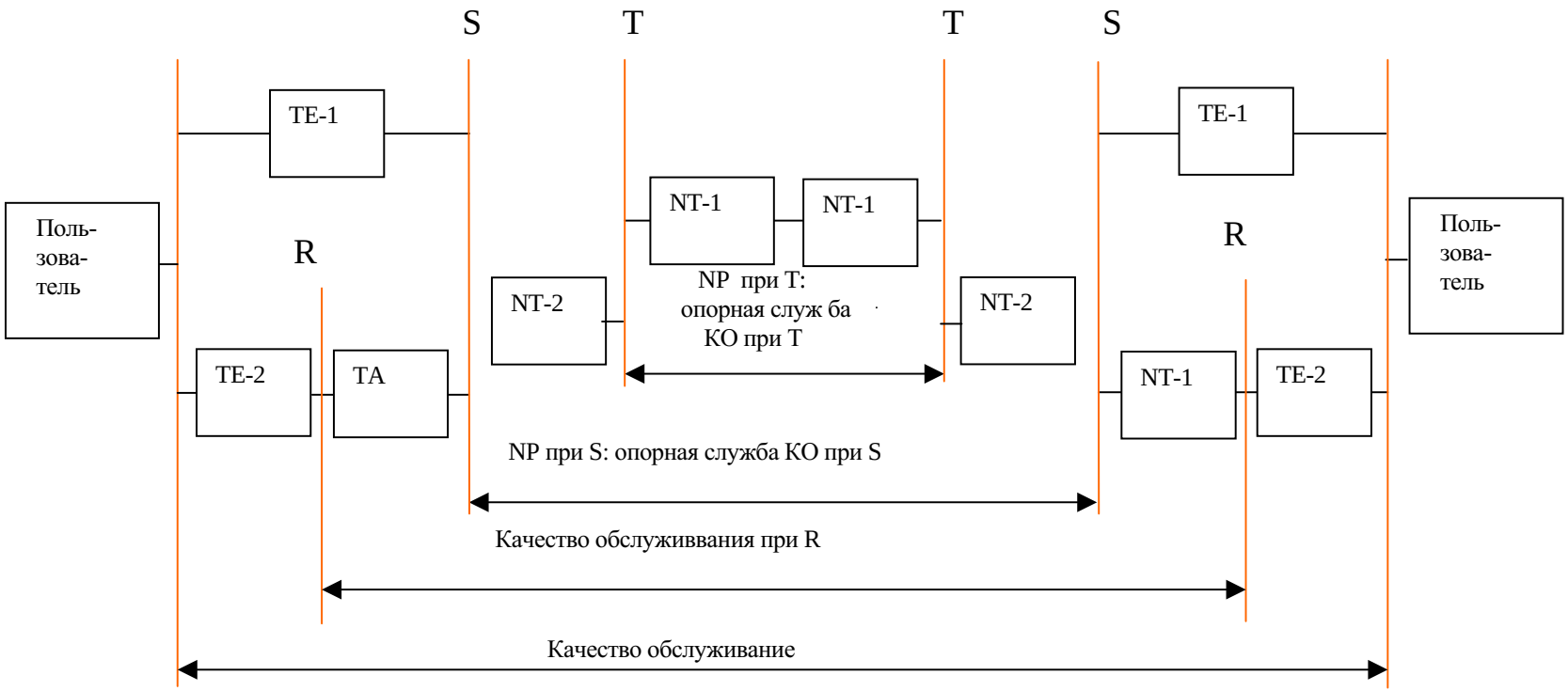


Рис. 5.1. Качество обслуживания и сетевая характеристика
в телекоммуникационных системах

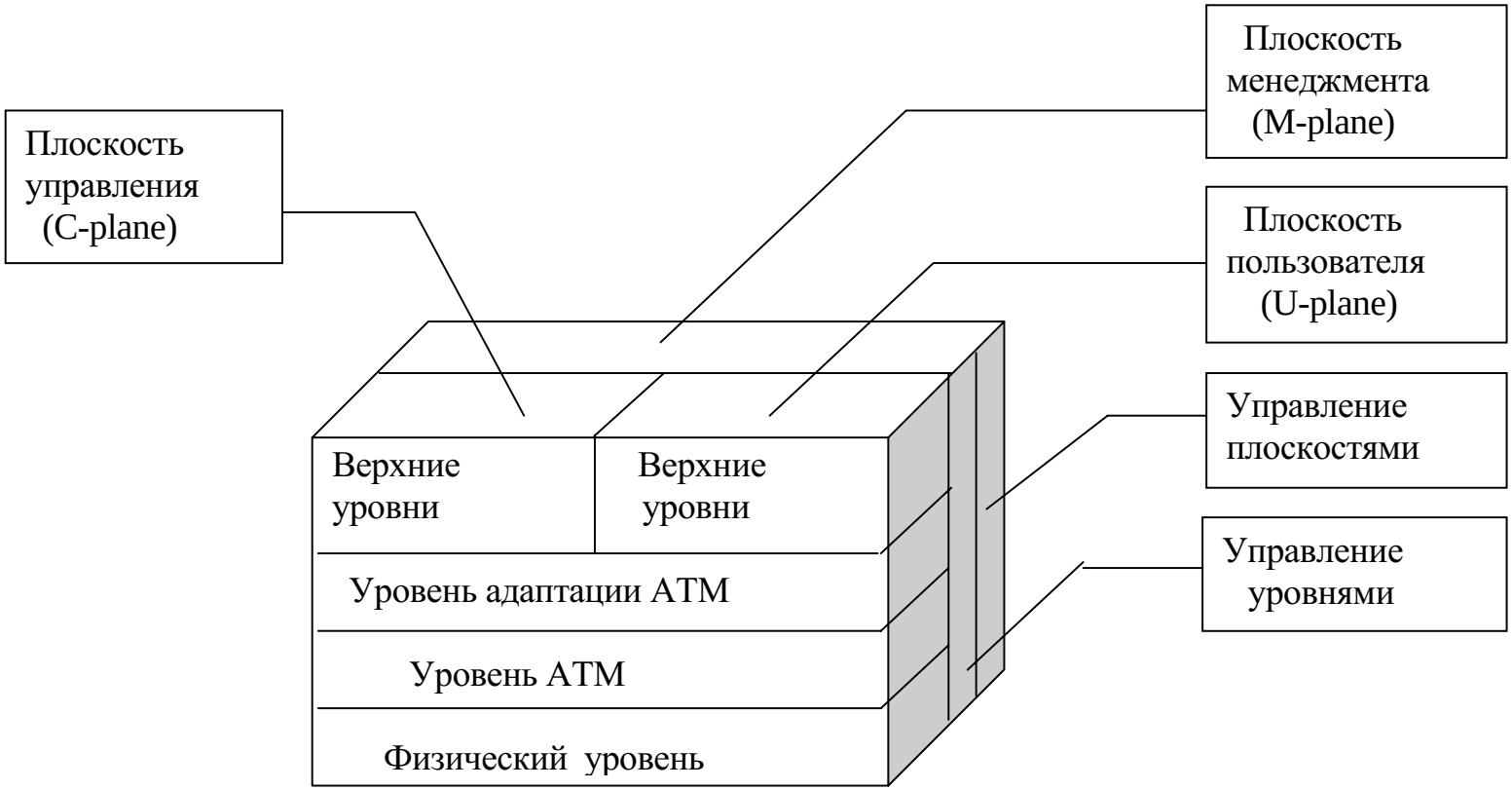


Рис. 5.2. Эталонная модель протоколов Ш-ЦСИО

Архитектура QoS представляет собой вид функциональной архитектуры, рассматривающей аспекты, относящиеся к ресурсам и трафику, и функции для администрирования этих ресурсов. Функции обработки неисправностей будут частью архитектуры QoS ввиду тесной связи между ресурсами трафика и обработкой неисправностей. Таким образом, QoS- архитектура будет иметь операционную и управляющую QoS - архитектуры соответственно. QoS-архитектура в основном сосредотачивается на том, как моделируются функции QoS. В QoS- архитектуре QoS является целью, а средством достижения цели является QoS- служба. При этом служба в функциональной архитектуре заменяется на QoS- службу в QoS- архитектуре. Параметры службы замещаются QoS- параметрами, а определение оптимального трафика относится к существованию оптимальной QoS- архитектуры.

Функции по распределению, администрированию и перераспределению ресурсов, несущих трафик, внутри операционной архитектуры QoS обозначаются как функции управления трафиком. Целью проектирования телекоммуникационных служб является предоставление определенного ряда служб с определенными QoS при минимальной стоимости. Архитектура QoS и функции управления трафиком являются инструментом для достижения этой цели. Существует следующая классификация для B-ISDN- функций управления трафиком:

- управление (control) трафиком: ряд действий, производимых сетью, чтобы избежать условий перегрузки;
- управление перегрузкой (congestion): ряд действий, производимых сетью для минимизации интенсивности, распространения и продолжительности перегрузки.

В сетях N-ISDN функциями управления трафиком были: коммутация, маршрутизация, доступ, управление приоритетом, управление потоком, основанное на подтверждении. Традиционные механизмы, основанные на подтверждении (например, "стой- и- жди", "скользящее окно") для сетей большой емкости в общем случае недостаточны. В сети с большой емкостью пришли такие концепции как:

- управление вхождения в соединение (Connection Admission Control- CAC);
- управление эксплуатационным/сетевым параметром (Usage/Network Parameter Control- UPC/NPC);
- формирование трафика (Traffic Shaping).

CAC- это ряд действий, имеющих место в течение организации вызова, для установления трафик- контракта и соединения. UPC/NPC- это ряд действий, производимых сетью для наблюдения и управления трафиком. Формирование трафика представляет собой модификацию характеристик трафика. Управление сетевыми ресурсами- это распределение ресурсов с целью выделения потоков согласно характеристикам службы. Управление с обратной связью- это ряд действий, производимых сетью и пользователями для регулирования трафика, передаваемого по ATM- соединению.

Установленными целями для трафика уровня ATM и управления перегрузкой являются:

- поддержка ряда QoS- классов ATM- уровня, достаточных для всех служб ATM, которые можно предвидеть;
- не основывать работу на протоколах AAL, которые являются специфическими службами B-ISDN, или на протоколах более высоких уровней, которые являются специфическими приложениями, но позволять протокольным уровням выше ATM использовать информацию, обеспечиваемую ATM- уровнем, для улучшения использования этих протоколов, которая может происходить от сети;
- при проектировании оптимального ряда управлений трафиком ATM- уровня и управлений перегрузкой, минимизировать сложность сети и оконечной системы, в то же время максимизировать использование сети.

Модель QoS в терминах BOC (OSI) представляется двумя типами организации: уровневой организацией QoS и системной организацией QoS. Уровневая организация QoS связана с работой отдельных (N)- подсистем, а системная организация- с работой полной системы. Уровневая организация QoS содержит: пользователя (N)-службы, (N)-функцию контроля управления ((N)-policy- control- function (N)-PCF), (N)- функцию управления QoS ((N)-QoS-control-function- (N)- QCF), (N)- протокольная организация ((N)-protocol entity-(N)-PE) и провайдера (N-1)- службы ((N-1)-service-provider). На рис.5.3 показано прохождение исходящего потока QoS- требований. (N)-PCF принимает (N)-требования QoS, предоставленные пользователем (N)-службы и применяет специальное управление.. Функция (N)-QCF будет решать, могут ли требования QoS быть удовлетворены действиями существующей (N)-PE. Если могут, то такая (N)-PE выбирается, в противном случае вызов отвергается. Для осуществления своих функций (N)-QCF необходим допуск информации, который обеспечивается системной организацией QoS. В системную организацию входят: функции контроля управления системой (SPCF), функция управления QoS системы (SQCF), агент управления системой (SMA), менеджер управления системой (SMM), менеджер ресурса (RM). Эта системная организация касается, главным образом, функций менеджмента, а ряд системных и уровневых организаций QoS устанавливают функциональную декомпозицию QoS для целей описания менеджмента QoS. На рис.5.4 представлено отношение между уровневой и системной организацией. SMA представляет функции менеджмента для обработки с помощью агента, которая позволяет, используя протоколы менеджмента системы OSI, управлять ресурсами на расстоянии. Она обеспечивает ряд функций системного менеджмента, которые будут зависеть от конфигурации отдельных систем и используют пакет OSI, включая CMIP (Common Managment Int. Protocol). RM представляет управление ресурсами оконечной системы. SQCF комбинирует две функции:

- системная функция для настройки протокольных функций, которые находятся в работе;

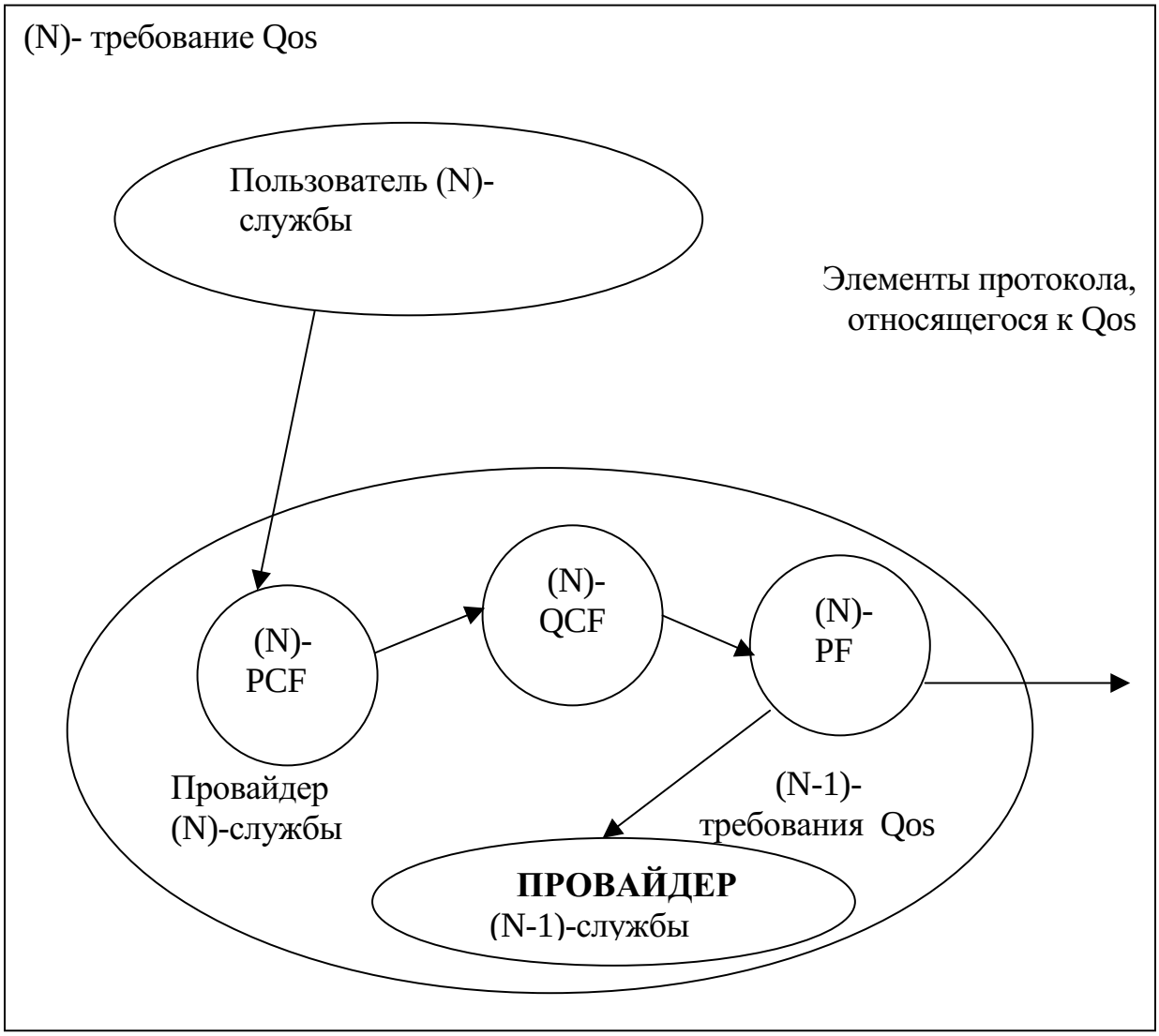


Рис.5.3. Уровневая организация Qos

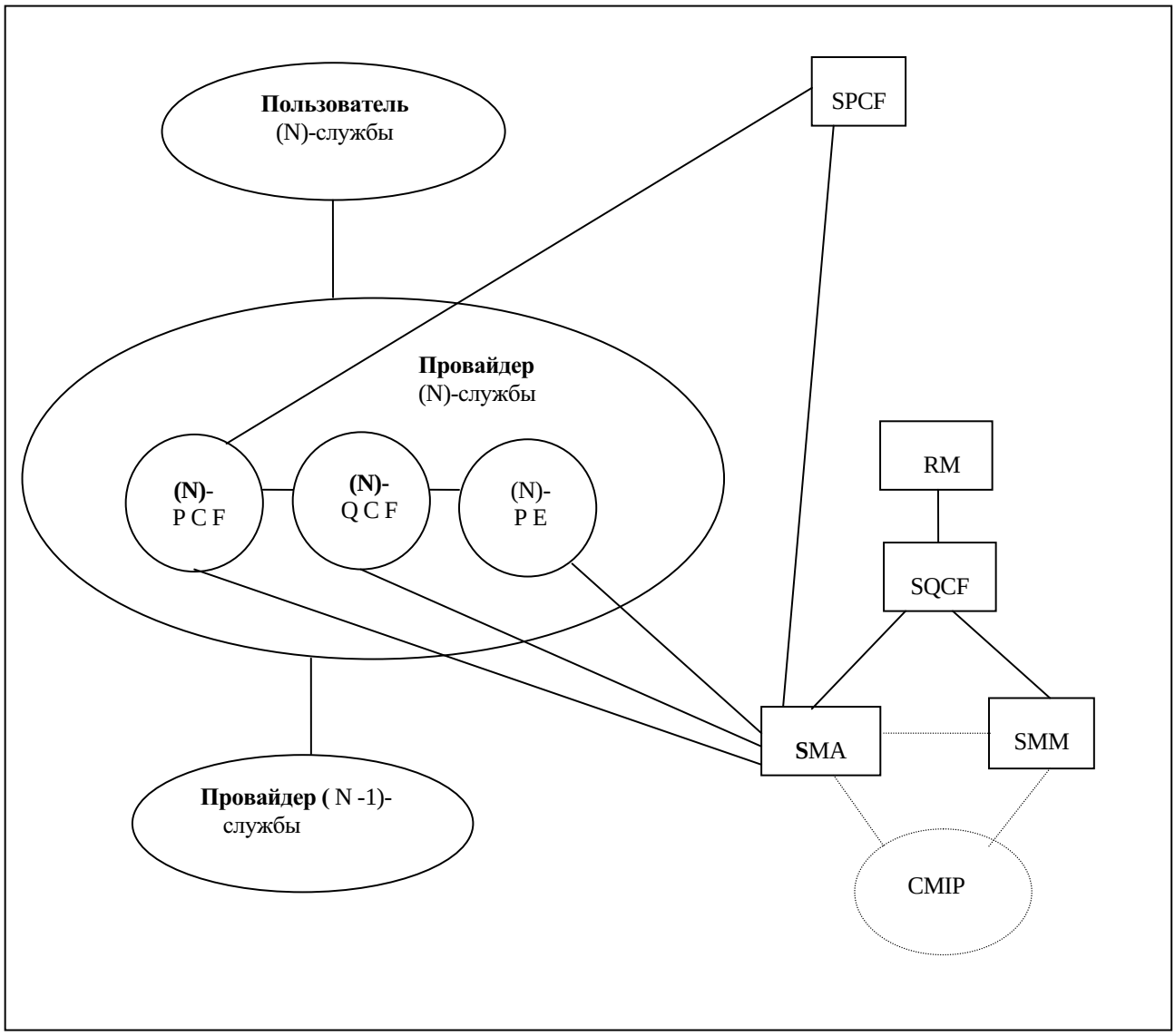


Рис. 5.4. Отношение между уровневой и системной организацией QoS

- обеспечение координации любых требований по модификации поведения любых удаленных систем, посредством менеджмента OSI.

SMM представляет работу открытой системы в роли менеджера, как определено в ISO/IEC 10040. Ее включение в модель делает возможным описание функций менеджмента QoS, которая необходима системе, чтобы управлять или получать информацию из других систем, используя протоколы менеджмента систем OSI. Роль SPC, по- существу, подобна роли (N)-PCF на каждом уровне. Включение SPCF показывает, что любое управление, выполненное на любом отдельном уровне должно зависеть от управления, которое было установлено для всей открытой системы.

Внутри эталонной модели ISO/OSI (MOC/BOC) концепция QoS определяется на различных уровнях:

- в части, поддерживающей приложения, состоящей из уровней 5-7;
- транспортном уровне;
- сетевом уровне и канальном уровне.

Прикладной уровень, как наивысший в архитектуре BOC, не обеспечивает услуг другим уровням. Главной его задачей является обработка семантики (смыслового содержания) прикладных процессов. К его компетенции относятся те его функции, которые связаны с организацией межпроцессных взаимодействий (называемые компонентами прикладного уровня). В каждой открытой системе прикладной уровень содержит набор конкретных сервисных элементов, каждый из которых принимает и обрабатывает запросы на предоставление той или иной услуги, предусматриваемой эталонной моделью или управляющего взаимодействия со стороны эталонной модели. Конкретный поднабор таких сервисных элементов составляет уникальный тип прикладной компоненты. Структура прикладной компоненты предусматривает две категории сервисных элементов: специальные прикладные сервисные элементы – SASE (Specific Application Service Elements) и стандартные прикладные сервисные элементы - CASE (Common Application Service Elements). SASE предусматривают возможности пересылки информации (например, обмен файлами, доступ к базе данных, пересылку заданий) или возможности, требуемые для выполнения функций конкретных прикладных процессов (например, для реализации банковских операций и т.п.). К SASE относятся : FTAM (File Transfer Access and Management), VT (Virtual Terminal), JTM (Job Transfer and Manipulation), RRC (Remote Procedure Call), а также INAP (Intelligent Network Application Protocol), EDI (Electronic Data Interchange), ELCOM (Electric Power Communication Protocol), SGML (Standard Graphical Markup Language). CASE предусматривают возможности, требуемые для пересылки информации между прикладными процессами, которые не зависят от природы приложения (например, установление связи между прикладными процессами, разрыв связи между прикладными процессами). К CASE относятся : ACSE (Associate Control Service Element), ROSE (Remote Operation Service Element), CCR (Concurrency,

Commitment and Recovery). Каждая прикладная компонента может содержать сервисные элементы обеих категорий. QoS- параметры не определены на службах, относящихся к SASE. Сервисным элементом приложения, который имеет ясно определенные QoS- параметры, является только CASE- элемент ACSE. Поле QoS- параметра определяется на служебных примитивах A.ASSOCIATE.

Основное назначение уровня представления состоит в том, чтобы обеспечить независимость прикладных процессов от различий в форме представления данных, то есть от синтаксиса данных. Стандарты уровня представления определяют протокол, который связан с синтаксисом (формой представления) этих данных и позволяют определять синтаксис через имена примитивов или описательные имена. Конкретная реализация подразумевает, что система будет применять протоколы уровня представления для установления соединения на всех более низких уровнях и перед началом соединения на прикладном уровне. Поведение системы будет таким, как если бы была подключена служба «ASSOCIATE» (установить связь) прикладного уровня, за которой последовал бы запрос P. CONNECT службы уровня представления, а затем на организацию сессии и т.д. Примитив A. ASSOCIATE отображается в служебный примитив P. CONNECT уровня представления. Не существует QoS- параметров, переносимых на ACSE или PDU уровня представления.

Основное назначение уровня сессий- обеспечение механизмов организации и формирования структуры взаимодействия между прикладными процессами. По- существу уровень сессий обеспечивает структуру управления взаимодействия. Уровень представления отображает QoS-параметры примитива P.CONNECT уровня представления в QoS- поле примитива S.CONNECT уровня сессий. QoS- параметры службы сессий передаются далее к транспортному уровню. QoS- параметры примитива S.CONNECT подобны QoS- параметрам транспортной службы, ориентированной на соединение, определенным в таблице 5.1. Таким образом, не существует QoS-ориентированных протокольных функций, на уровнях ВОС 5-7. Параметры QoS примитива A.ASSOCIATE службы ACSE только посылаются далее к транспортному уровню.

Транспортная служба обеспечивает прозрачную передачу данных между пользователями транспортных служб. Службы, могут быть ориентированными на соединение, или без соединения, но протокол - только на соединение. Транспортный уровень должен оптимизировать использование доступных коммуникационных ресурсов для обеспечения затребованных QoS при соединении транспортных служб пользователей с минимальной стоимостью. QoS определяется через выбор величин для параметров QoS, представляющих такие характеристики, как пропускная способность, транзитная задержка, темп остаточных ошибок, вероятность отказа (см. табл.5.1)

Таблица 5.1.

Уровни модели ВОС				
Прикладной Представления Сеансовый	Транспортный	Сетевой	Канальный	Физический
Защита Приоритет Темп остаточных ошибок Полоса пропускания Задержка передачи (для каждого направления) Оптимизация передачи Расширенное управление Задержка установления соединения Вероятность отказа от установления соединения Вероятность ошибки передачи Задержка завершения соединения Вероятность ошибки завершения соединения Надежность соединения Параметры стоимости	**С соединением** Защита Приоритет Фаза установления соединения: -задержка установления, -вероятность неустановления. Фаза передачи данных: - пропускная способность, - транзитная задержка, - коэффициент необнаруженных ошибок(КНО), - надежность, - вероятность отказа. Фаза разъединения: - задержка разъединения, - вероятность неразъединения. **Без соединения** Транзитная задержка КНО Защита Приоритет Параметры стоимости	**С соединением** Фаза передачи данных: - пропускная способность, - транзитная задержка, - КНО, - надежность, - вероятность отказа, - наибольшая стоимость соединения. Фаза установления соединения: - задержка установления, - вероятность установления. Фаза разъединения: - задержка разъединения, - вероятность неразъединения. **Без соединения** Транзитная задержка Защита Параметры стоимости КНО Приоритет Возможность контроля нагрузок Вероятность сохранения последовательности Максимальное время существования сетевого сервисного блока данных	Пропускная способность Транзитная задержка Защита соединения КНО Надежность соединения Параметры стоимости	Частота пользования Доступность сервиса Скорость передачи Транзитная задержка Пропускная способность

QoS обговаривается при установлении соединения. Пропускная способность и транзитная задержка основаны на среднем размере T- SDU (Transport service data unit- сервисный блок данных транспортного уровня). Для каждого направления определены величины максимальных и средних величин. QoS-параметры для службы без соединения являются подмножеством параметров службы, ориентированной на соединение (см. табл. 5.1). В рекомендациях ISO/OSI для транспортного уровня рассматриваются 3 аспекта обработки QoS:

- QoS- соглашения с сетевым уровнем;

  QoS- параметры сетевых служб, ориентированные на/без соединения, подобные QoS- параметрам соответствующих транспортных служб, так что требование QoS транспортной службы могут быть посланы на сетевой уровень.

- QoS- соглашения с равноуровневым транспортным уровнем;

  Определены 5 классов транспортных протоколов, ориентированных на соединение, в соответствии с типом сети и мультиплексирования, а также возможностью исправления ошибок. Класс 0: Простой класс. Класс 1: Базовый класс с исправлением ошибок. Класс 2: Класс с мультиплексированием. Класс 3: Класс с исправлением ошибок и мультиплексированием. Класс 4: Класс с обнаружением и исправлением ошибок. За исключением класса 0 для всех классов T-PDU требования соединения и подтверждения соединения имеют следующие поля для QoS-параметров: максимальная и средняя пропускная способность в обоих направлениях, определенная как “целевая” и “минимально доступная”, транзитная задержка и темп остаточной ошибки в обоих направлениях, определенные как “целевая” и “максимально доступная”, и приоритет. Классы 2-4 имеют механизмы управления потоком, основанные на подтверждении. Отображение требования QoS к размеру окна должно быть функцией транспортного уровня.

- обмен статусами и мониторинг QoS.

  В существующую версию определения транспортного протокола не включен обмен статусами и мониторинг QoS.

В современном определении транспортного уровня обработка QoS основывается на “максимально доступном качестве” (best effort). Семантика параметров QoS рассмотрена в .

Таким образом, параметры QoS сетевой и транспортной служб аналогичны. Функции QoS протоколов сетевого уровня очень малы, за исключением механизмов управления потоком, основанным на подтверждении. Протокол X.25 «Essential Optional Packet Switched User Facilities» определяет необязательные возможности, относящиеся к QoS: соглашение по классу пропускной способности, быстрый выбор и селекция транзитной задержки и идентификация. Класс пропускной способности является концепцией без определения точности. Он дает возможность пользователю выбрать из числа классов, с различными пропускными способностями. В протоколе не оговорены функции обработки трафика,

такие как отображение QoS- требований в размер окна и управление признанием вызова. Протокол IP имеет необязательную часть в заголовке, включающую: транзитную задержку, защиту от несанкционированного доступа, определители стоимости, вероятность остаточной ошибки и приоритет, соответствующие параметрам QoS сетевой службы без соединения.

Для традиционных канальных уровней, основанных на HDLC (High Data Level Control- высокоуровневое управление каналом передачи данных) не существует никакой определенной службы канального уровня, и HDLC не имеет полей PDU (протокольный блок данных), относящихся к QoS, исключая поля, относящиеся к управлению потоком, основанному на подтверждении, и управлении ошибкой. Если уровень звена данных разделен на 2 подуровня: управление доступом к среде (MAC) и управление логическим звеном данных (LLC), - то для канальных уровней, основанных на LLC/MAC, LLC- служебные примитивы имеют поле QoS- параметров для указания приоритета. Нет параметра QoS в LLC PDU-блоках, за исключением полей управления потоком и ошибкой. В служебных примитивах MAC также существует поле служебного класса, используемого для приоритета. Использование этого поля зависит от соответствующих механизмов в специальных протоколах.

Аспекты параметров QoS, относящихся к служебным примитивам и PDU проиллюстрированы на рис.5.5.

QoS- служба BOC может быть охарактеризована как ориентированная на приложения, поскольку (N+1)- уровень определяет свои параметры требования QoS, основанные на концепциях (N+1)- уровня и осуществляет нисходящее отражение требований QoS, определенных приложениями.

Модель операционных функций B-ISDN включает: пользователя службы, функции пользователя, функции управления, функции управления вхождения в соединение (CAC), пользовательский сетевой интерфейс (UNI) (рис.5.6). Пользователь службы состоит из дух частей: приложений и транспортной системы. Приложения содержат функции уровня 5-7, в то время как транспортная система состоит из функций уровня 4 или скомбинированных функций уровня 3-4. Функции плоскости пользователя состоят из уровня адаптации ATM (AAL), уровня передачи ячеек ATM (уровня ATM) и физического уровня ATM.

QoS - характеристики определяются как "количественный аспект QoS, который определен независимо, посредством чего он представлен или управляем".

QoS- параметр определяется как "переменная по отношению одному или более QoS- характеристикам, величины которых передаются между объектами как часть QoS- механизма".

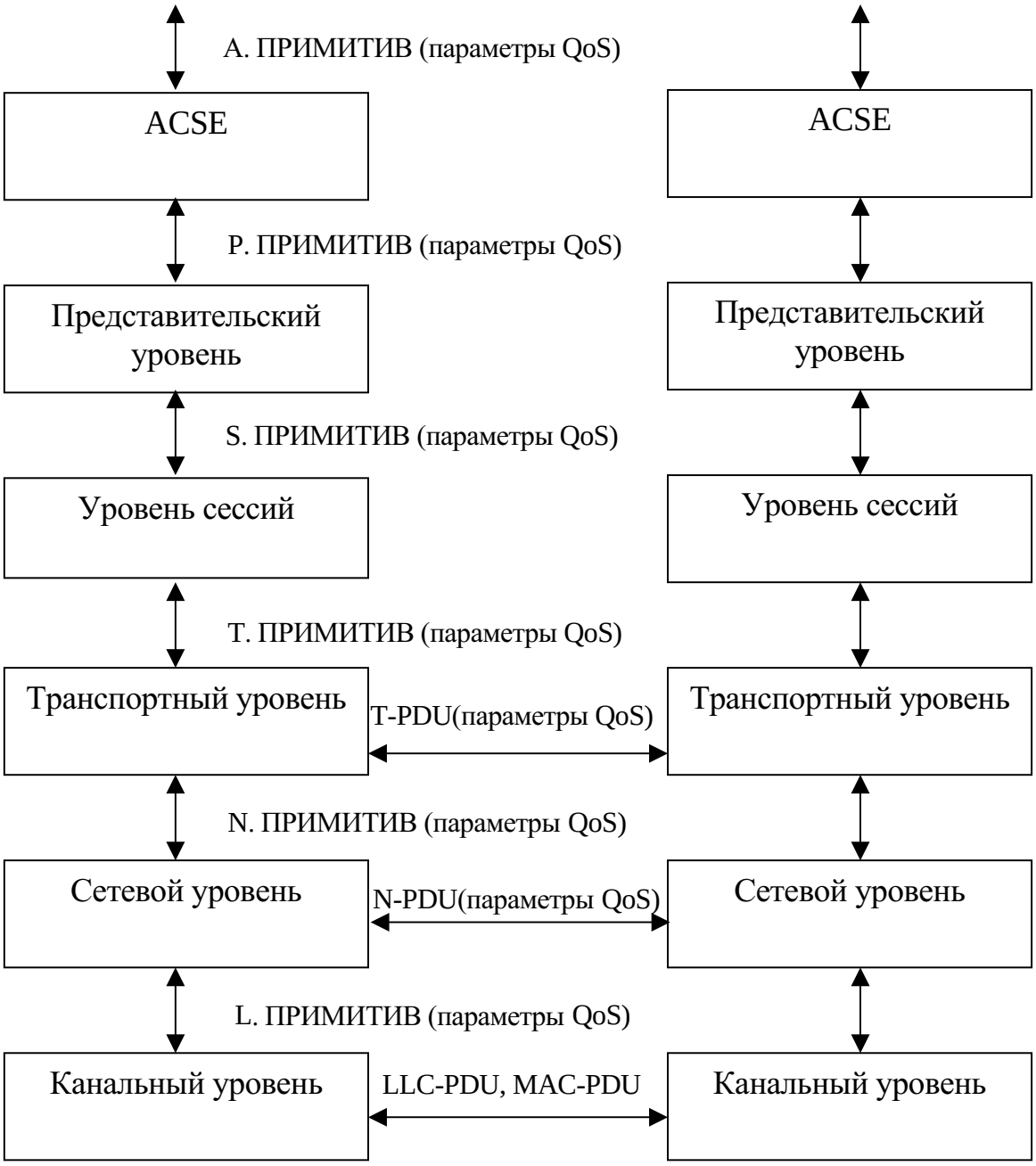


Рис. 5.5. QoS- параметры в среде LAN

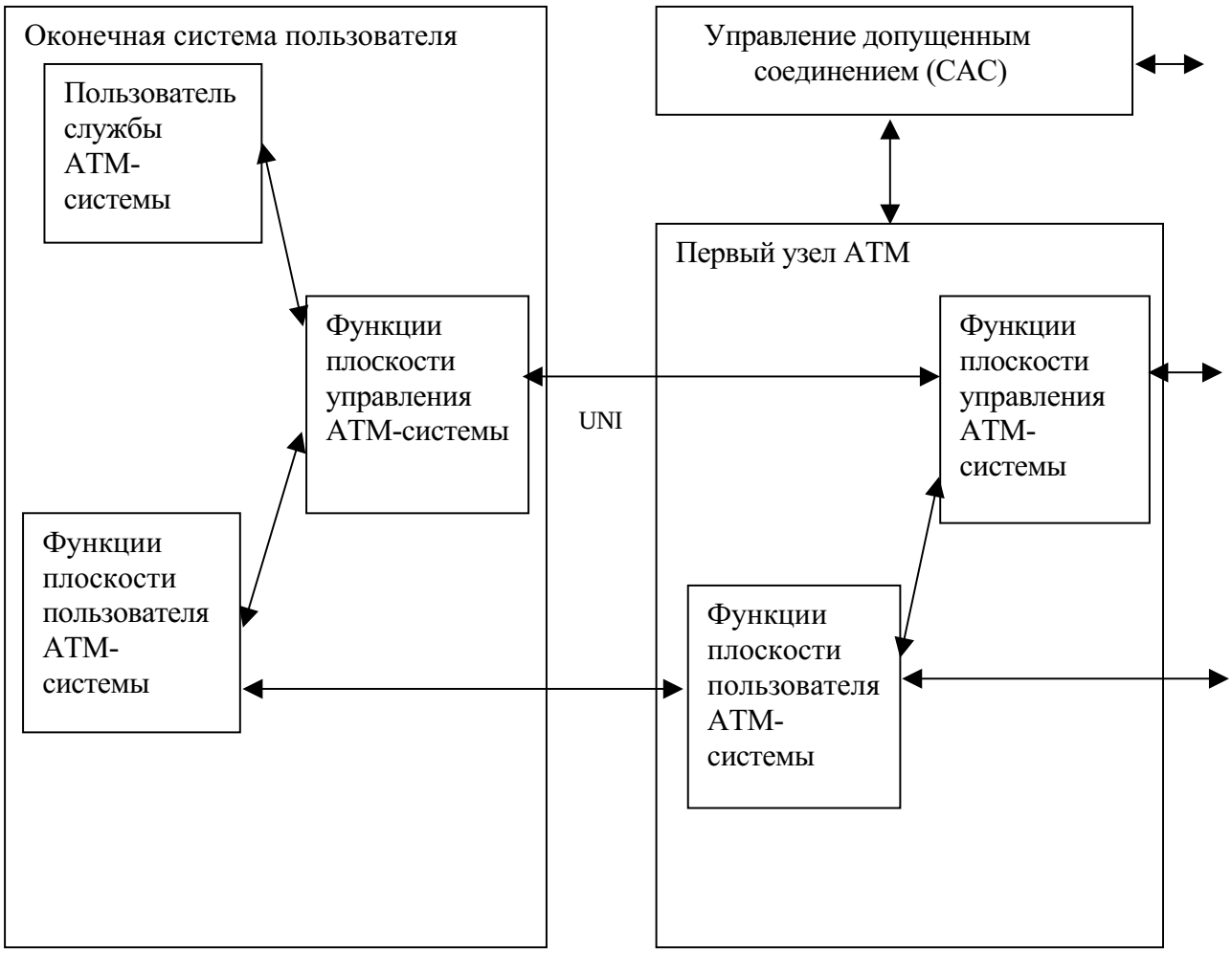


Рис.5.6. Операционная архитектура B-ISDN

В рекомендации ITU –T I.350 определен матричный метод для идентификации параметров, который должен браться в расчете NP и QoS. Каждая строка матрицы представляет одну из основных связных функций. Каждый столбец представляет один из трех возможных критериев. Эта матрица представлена в таблице 5.2.

Таблица 5.2

Функция	Критерий		
	Скорость	Точность	Надежность
Доступ			
Передача пользовательской информации			
Разъединение			

Тремя основными связными функциями является доступ, передача пользовательской информации, разъединение. Они определяются следующим образом.

Доступ начинается с требования доступа при интерфейсе сети и заканчивается, когда первый бит информации входит в сеть. Это измерение соответствует физическому установлению соединения (виртуального или нет).

Передача пользовательской информации начинается с комплекса функции доступа и окончания, когда требование испускается. Критерии определяются следующим образом.

Скорость является критерием, который описывает временной интервал, используемый для осуществления функции, или скорость, при которой функция осуществляется.

Точность- критерий, который описывает степень корректности, с которой функция осуществляется .

Надежность- критерий, который описывает степень уверенности, с которой функция осуществляется, независимо от скорости или точности. Эти девять параметров трудны для использования и, кроме того, не могут использоваться на фазе оценки QoS.

В АТМ- сети было предложено в качестве основы стандартизации для использования три следующих параметра :

- время задержки при передачи ячеек (cell-transfer delay, CTD) ;
- непостоянство времени задержки (cell- delay variation, CDV);
- процент потерянных ячеек (cell- loss ratio, CLR).

CTD- это максимальное время передачи ячейки от одной конечной станции до другой. Оно зависит от задержек передачи и времени нахождения ячеек в очередях коммутирующих устройств.

CDV-представляет собой разницу между максимальным и минимальным временем передачи ячеек между оконечным оборудованием. Оно определяется числом виртуальных каналов, мультиплексируемых в рамках одного физического соединения, и непостоянством времени задержки ячеек в очередях АТМ-коммутаторов.

Значение CLR связано с уровнем ошибок в заданном физическом соединении и алгоритмом, предусмотренном в АТМ- коммутаторе для обработки перегрузок. Именно этот алгоритм, а также метод обслуживания очередей играют решающую роль в достижении высоких характеристик, которые отличают сеть АТМ.

Основными параметрами трафика, которые определяют тип трафика, являются:

- пиковая скорость передачи (peak cell rate, PCR)- максимальное количество ячеек, которое источнику разрешено передавать за единицу времени;
- максимальный размер залпового выброса (maximum burst size, MBS)- количество ячеек, которое источник имеет право отправить с пиковой скоростью (PCR);
- нормальная скорость передачи (sustainable cell rate, SCR)- среднее количество ячеек, которое источнику разрешено передавать за единицу времени;
- минимальная скорость передачи (minimum cell rate, MCR)- минимальное количество ячеек, которое источник должен отправить за единицу времени.

Параметры обратной связи - совокупность параметров, относящихся к сервису с доступной скоростью передачи (available bit rate, ABR) и позволяющих источнику установить количество доступных сетевых ресурсов. Основные механизмы обратной связи - явная индикация перегрузки при прямой передачи (explicit forward congestion indication, EFCI) и явная индикация скорости (explicit rate, ER).

Пользователь и сеть должны согласовывать между собой характеристики трафика и требуемый режим обслуживания. Такое соглашение, именуемое трафик-контрактом, состоит из 3-х частей:

- дескриптер исходного трафика, использующий 4 атрибута для описания трафика пользователя, а именно: PCR, гарантированную скорость передачи ячеек (SCR), наибольшее число ячеек, переданных максимальной скоростью (MBS), минимальную скорость передачи ячеек, поддерживаемую отправителем (MCR);
- определение согласованности и допустимости отклонения во времени задержки (CDVT). В определении согласованности указывается, какой трафик будет приемлемым для сети, т.е. с какой скоростью и с каким дроблением допускается отправка трафика пользователем. Такое определение вносится управляющей функцией. Допустимое отклонение во времени задержки представляет собой некий запас надежности, который берет на себя разброс по времени задержки в оборудовании на удаленном конце;
- набор параметров качества обслуживания. Требуемое QoS должно гарантироваться сетью. Примерами параметров QoS являются время задержки при передаче ячек (CTD), допустимое отклонение во времени задержки (CDV), коэффициент потерянных ячеек (CLR).

Все пункты контракта (PCR, SCR и др.) основаны на алгоритме GCRA (Generic Cell Rate Algoritm) (рис.5.7) и эталонной модели оконечных систем. В модели QoS - службы (B-ISDN) различают 5 различных служб и протокольных концепций. Это:

- служебные классы;
- протокольные типы AAL– уровня;
- типы протоколов сигнализации;
- классы с определенным и неопределенным QoS.

Существуют 4 основных служебных классов ATM- системы: классы A, B, C, D. Недавно был определен класс X. В качестве квалификационных признаков были выбраны следующие основные характеристики :

- синхронизация устройств между конечными точками передачи (требуется или нет);
- скорость передачи битов информации (постоянная или изменяющаяся);
- режим соединения (с установкой соединения или без).

Первая основная характеристика означает требование синхронизации конечных устройств в ATM- сети. Часто бывает необходимо, например, чтобы каждое устройство в соединении получало сетевую синхронизацию. Теоретически все эти устройства могут иметь различные ее источники с одинаковыми тактовыми частотами, но с различными сдвигами фаз. Вторая основная характеристика ATM- скорость передачи информации. Поддерживает как постоянную, так и переменную скорость передачи. Третья основная характеристика- это требование к установлению соединения между точками передачи. Установление соединения требуется, когда передающая станция хочет удостовериться в досягаемости станции назначения и ее готовности принять информацию.

Все вышеперечисленные характеристики можно сгруппировать так, чтобы получить разные классы сервиса (таблица 5.3).

Таблица 5.3

	Класс A	Класс B	Класс C	Класс D
Синхронизация	ДА		НЕТ	
Постоянная скорость	ДА	НЕТ		
Установление соединения	ДА			НЕТ

Отношения между этими классами и протоколами показаны на рис. 5.8. Вместе с тем, следует учитывать тот факт, что с течением времени число классов служб (а, следовательно, и протоколов) может быть расширено.

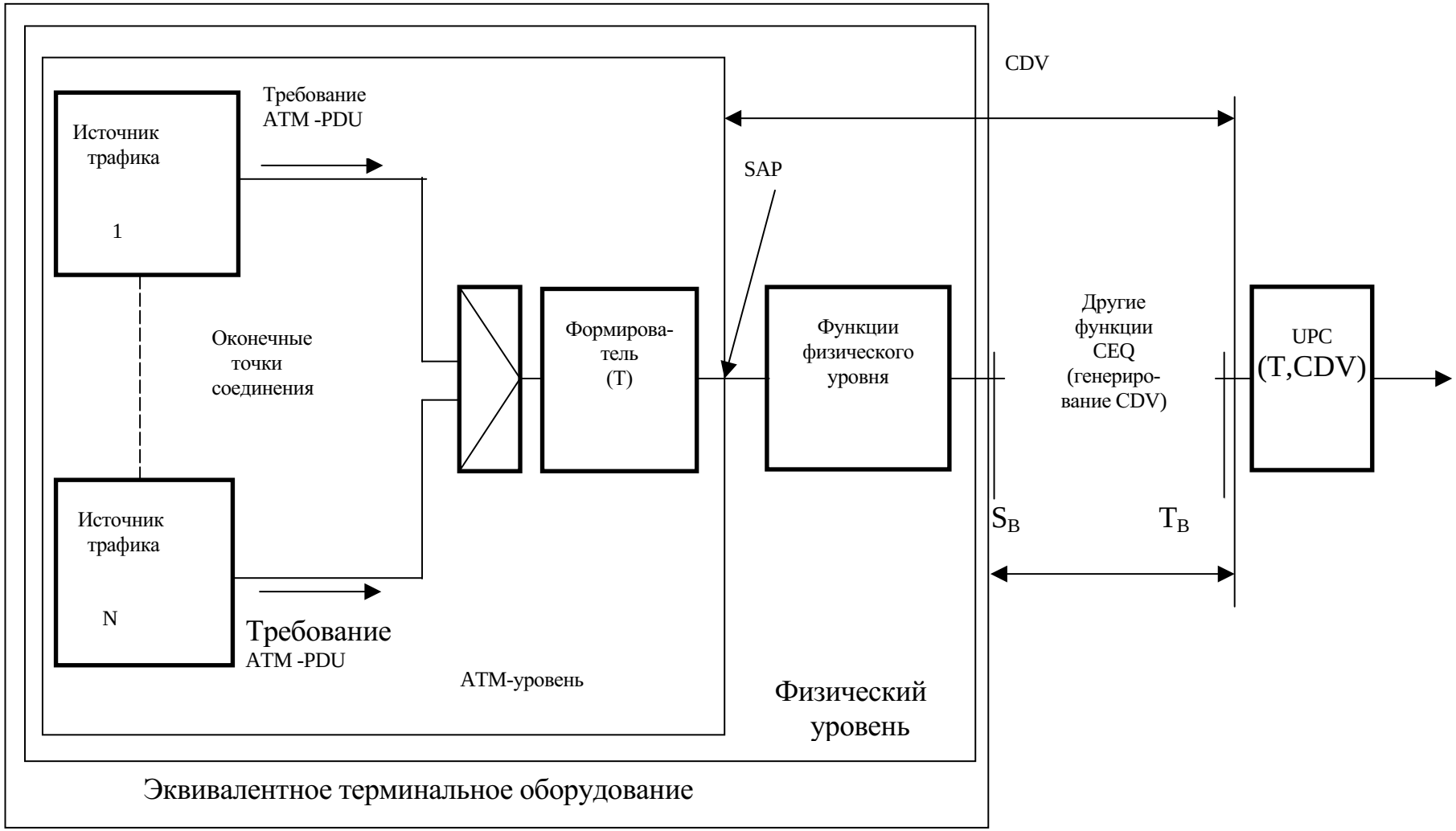


Рис. 5.7. Алгоритм GCRA (Generic Cell Rate Algorithm)

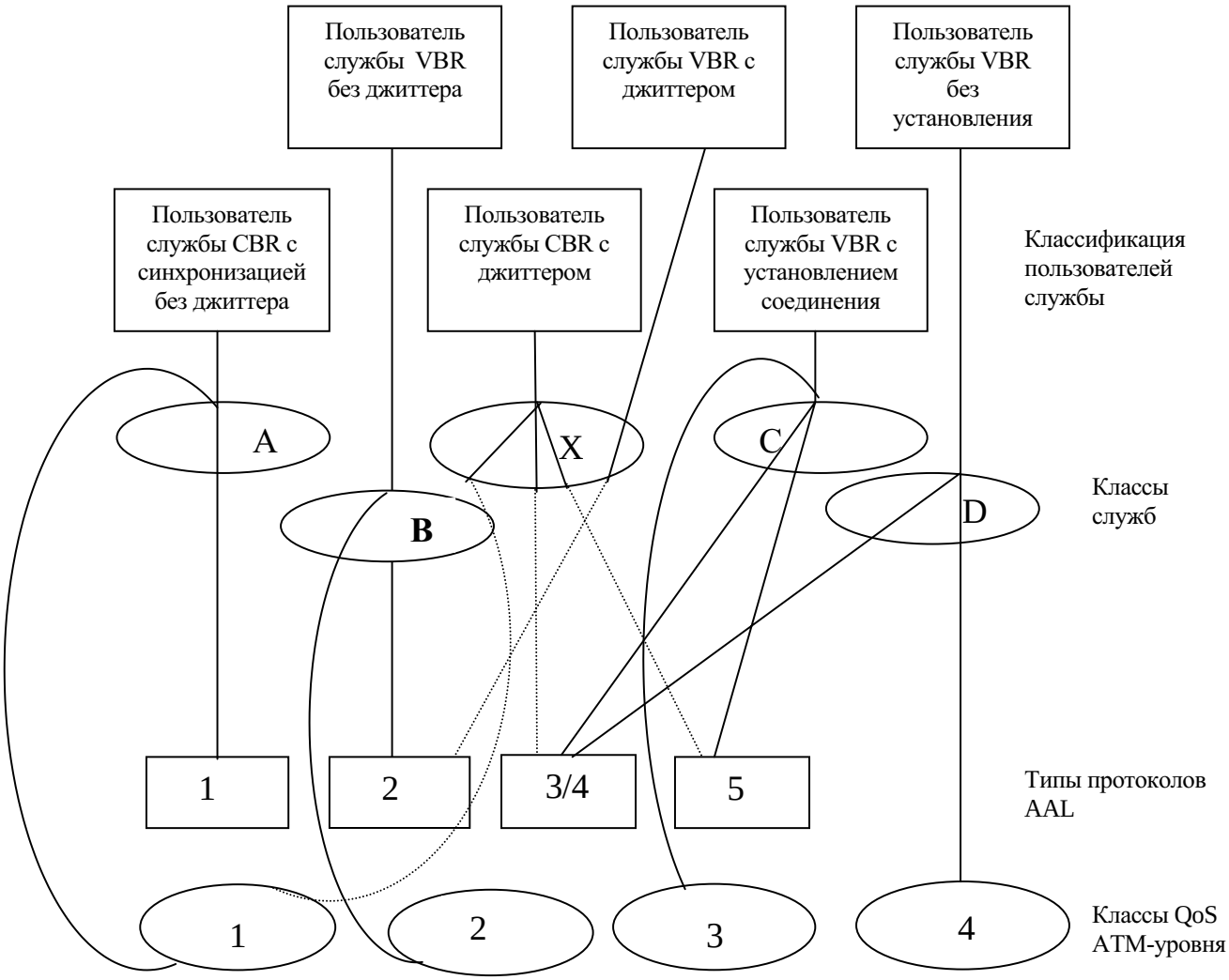


Рис. 5.8. Соотношение между классами обслуживания

Класс А - соответствует постоянной битовой скорости передачи (Constant Bit-Rate - CBR) служб, ориентированных на соединение с синхронизацией между источником и назначением. Класс CBR используется для восприимчивого к задержкам трафика, такого как аудио- и видео- информация, при котором данные передаются с постоянной скоростью и требуют малого времени ожидания. CBR гарантирует самый высокий уровень качества сервиса, но использование полосы пропускания неэффективно. Чтобы защитить трафик CBR от влияния других передач, CBR всегда резервирует для соединения определенную часть полосы пропускания, даже, если в данный момент в канале не происходит никакой передачи. Таким образом, резервирование полосы пропускания является особенно большой проблемой при работе по каналам широкомасштабных сетей, когда абоненту приходится платить за каждый мегабит полосы пропускания независимо от того, используется ли виртуальный канал.

Величина ширины полосы определяется пиковой скоростью ячейки (Peak Cell Rate- PCR). PCR определяется с допуском, известным как CDV (Cell Delay Variation). Точное определение дается посредством алгоритма, называемого GCRA (Generic Cell Rate Algorithm), где PCR и CDV являются параметрами алгоритма (рис. 5.7). В CBR трафик, предоставленный пользователем, согласовывается с определенным PCR, сетевой оператор связывает его с QoS, включает определенный целевой коэффициент потери ячейки и целевую дисперсию задержки ячейки из конца в конец.

В рекомендации I.363 приведены структуры форматов протокольных блоков данных (PDU) для всех 4-х классов служб. Формат пакета служб класса А содержит номер пакета (НП) и защиту номера пакета (ЗНП)- всего 8 бит и информационный поток (ИП) длиной 47 байт.

Класс В- соответствует переменой битовой скорости (Variate Bit-Rate- VBR) служб, ориентированных на соединение с синхронизацией между источником и назначением. Различают 2 вида VBR, которые используются для различных видов трафика:

- VBR- реального времен (Real-time VBR - rt VBR);
- VBR нереального времени (Non-real-time VBR- nrtVBR).

rt VBR требует жесткой синхронизации между ячейками и поддерживает восприимчивый к задержкам трафик, такой, как уплотненная речь и видео. nrt VBR не нуждается в жесткой синхронизации между ячейками и поддерживает допускающий задержки трафик, такой как трансляция кадров (frame relay).

Поскольку VBR не резервирует полосу пропускания, она используется более эффективно, чем в случае с CBR. Однако, в отличие от CBR, VBR не может гарантировать качество сервиса. Оба варианта VBR характеризуются двумя скоростями передачи:

- пиковой скоростью (PCR), с которой разрешено передавать ограниченное число ячеек (не более заданной величины MBS) и

- нормальной (SCR – Sustainable Cell Rate) (SCR всегда меньше, чем PCR), поддерживаемую неограниченно долго. При том передача данных регулируется таким образом, что ее средняя скорость не превышала допустимую. Обратная связь не используется. Единственное различие режимов rtVBR и nrtVBR состоит в том, что в первом должны быть заданы параметры качества обслуживания.

Формат пакета служб класса В содержит НП (номер пакета), тип информации и начало сообщения, продолжение или его конец, идентификатор длины (НП) и поле защиты от ошибок информации пользователя (ПЗО).

Класс С. Это служба с переменной скоростью, без синхронизации и с установлением соединения. Категории обслуживания ABR уделено основное внимание в спецификации Traffic Managment 4.0, принятой недавно ATM Forum. Ее основная особенность- наличие механизма обратной связи, который позволяет источнику определить количество сетевых ресурсов, доступных в данный момент времени. Поскольку АТМ- соединения работают с очень высокими скоростями, обратная связь должна осуществляться на аппаратном уровне. АТМ- коммутаторы способны поддерживать три типа обратной связи: EFCI, ER и метод виртуальных источников (virtual source/virtual destination, VS/VD). При использовании механизмов ABR необходимо периодически включать в поток данных служебные ячейки для управления ресурсами (RM). Как правило, на 32 ячейки данных требуется две управляющие ячейки, которые доставляют информацию о степени загруженности сети обратно к передающей станции. Ячейки, передаваемые вместе с потоком данных, называются прямыми (forward resource managment, FRM), а отсылаемые в противоположном направлении- обратными (backforward resource managment, BRM).

Формат пакета служб класса С содержит НП, резервное поле, поля ИП, ИД, ПЗО.

В то время как Форум АТМ предложил категорию услуг ABR ITU-T ввела альтернативную категорию для передачи данных АВТ. Категория АВТ позволяет передавать полные блоки с использованием RM (Resourse Managment) ячеек, одна перед первой ячейкой блока, а другая- после последней ячейки блока. В этом случае АВТ формируется протоколами более высокого уровня как АТМ- блок, хотя использование АТМ- блоков не ограничивается этим случаем. В АВТ при установлении соединения ширина полосы не распределяется до начала передачи пользовательских ячеек. Вместо этого ширина полосы резервируется не однократно при установлении соединения, а периодически по мере необходимости. ITU -T определяет 2 типа АВТ. В АВТ с задержанной передачей (ABT/DT) источник посылает RM- ячейку, чтобы затребовать скорость, при которой нужно передать блок, и затем источник ждет ответа от сети для RM- ячейки до посылки блока. В АВТ с немедленной передачей (ABT/IT) пользователь, желающий передать АТМ- блок, посылает RM- ячейку, а затем немедленно

оставшуюся часть АТМ- блока. Если узел сети не может обеспечить затребованную скорость, ячейки блока в случае ABT/IT могут быть отброшены.

Класс D. Оставшуюся после резервирования для категорий услуг CBR и VBR пропускную способность делят между собой все оставшиеся приложения. В рамках Форума АТМ им выделена собственная категория услуг- UBR. Приложениям, использующим категорию услуг UBR, не гарантируется качество сервиса или величина полосы пропускания и отсутствие потерь ячеек при возникновении перегрузок в сети. Сеть оставляет за собой право изъять данные пользователей по своему усмотрению, без уведомления. Для передачи с заранее определенной скоростью необходим протокол более высокого уровня, например, TCP, позволяющий обнаруживать и обрабатывать ошибки. Именно он должен регулировать скорость передач, исходя из количества потерянных пакетов. Этот класс соответствует службе без установления соединения, без синхронизации. Формат пакета класса D отличается от предыдущего тем, что вместо резервного поля вставляется индикатор мультиплексирования. В целом длина пакета без заголовка АТМ сохраняется равной 48 байтам.

Класс X. В дополнении UNI3.0 определен класс службы X, которая позволяет использовать собственное AAL в терминальном оборудовании, которое поддерживает частные AAL, определенные продавцом сети.

В рекомендации I.363 рассматриваются особенности функций каждого класса. Анализ содержания рекомендаций позволяет выделить несколько общих для классов А и В функций в том числе:

а) сегментация и сборка информации пользователя;
б) контроль и регулировка задержки пакетов АТМ;
в) контроль прохождения пакетов АТМ через сеть;
г) контроль в информационных полях пакетов AAL;
д) контроль тактовой синхронизации на отдельных участках сети.

Основными функциями для классов С и D являются сегментация и сборка информации пользователя и обнаружение ошибок в пользовательских данных. Для класса D в состав функций включены также процедуры, требуемые для поддержания режима без соединения. Эти функции определены в самом общем виде, однако указывается на их связь с функциями адресации и маршрутизации сетевого уровня.

Классы служб невидимы ни в каком определенном служебном примитиве, ни в каком информационном элементе сигнализации. Служебный класс есть результирующая последовательность выбранных классов QoS уровня АТМ и протокольных типов AAL.

Протокольными типами AAL являются типы 1,2, 3/ 4 и 5 (рис.5.9). Тип 1 поддерживает службу класса А, тип 2 поддерживает службу класса В. Типы 3/ 4 и 5 поддерживают служебные классы С и X, в то время как тип 3/ 4 поддерживает службу класса D. Плоскость управления использует протокол AAL типа 5. Классы QoS предоставляют возможности QoS, обеспеченные

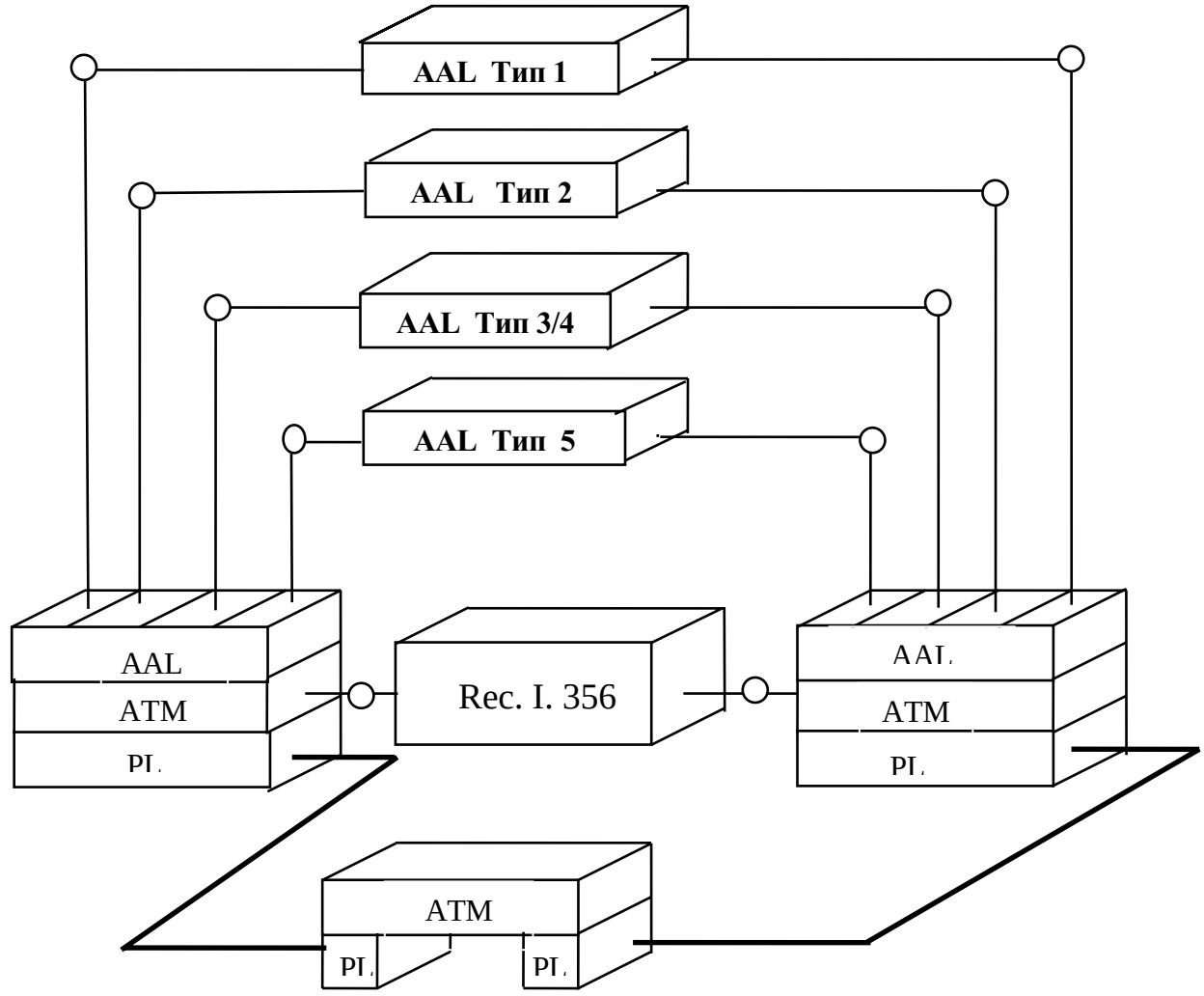


Рис.5.9. Протокольные типы уровня AAL

уровнем ATM. Классы QoS могут иметь специфицированные или неспецифицированные параметры исполнения (performance parameters), т.е. параметры, определенные на основе прямых наблюдений событий в точке доступа к службе или границам элементов соединений . Эти случаи обеспечены как специфицированные или неспецифицированные классы QoS соответственно. Форум ATM определил 4 неспецифицированных классов QoS, один для каждого из служебных классов A, B, C и D.

Классы QoS отличаются схемами буферизации в том или иномм сетевом элементе. Такие схемы имеют различную степень сложности- от очередей с внутренними приоритетами, работающих по алгоритму FIFO, до очередей на соединении с усовершенствованным механизмом диспетчеризации. Наиболее широко известны следующие механизмы:

- "Первым вошел, первым вышел" (First- In First- Out, FIFO)- обслуживание в порядке поступления пакетов- наиболее простой для реализации подход. Однако при использовании данного механизма пакет с высоким приоритетом может долго ждать своей очереди.
- "Строгий учет приоритетов" (Strict priority scheduling)- обслуживание пакетов определенного класса производится лишь в том случае, когда отсутствуют очереди пакетов более высокого класса. Механизм прост для реализации, но может возникнуть проблема, связанная с задержкой пакетов всех классов, кроме одного.
- "Честное формирование очередей" (Fair Queuing, Round Robin (RR))- реализация механизма выбора из множества очередей. Позволяет эффективно распределять полосу пропускания между различными очередями. Одна из основных проблем данного механизма заключается в том, что потоки с длинными пакетами могут захватывать значительную часть доступной полосы пропускания.
- "Взвешенное честное формирование очередей" (Weighted- Fair Queuing, WFQ)- усовершенствованный механизм честного формирования очередей, соответствующих различным классам трафика. Возможно применение различных методов обслуживания или планирования очередей.

"Формирование очередей на основе иерархии классов" (Hierarchical Class Based Queuing, CBQ)- трафик разделяется на классы, каждый класс, в свою очередь, может иметь подклассы. такая иерархия хорошо описывается с помощью деревьев. Если подклассу требуется больше выделенной ему полосы пропускания, то он сначала пробует заимствовать дополнительную полосу у своих дочерних подклассов. Такая схема может использоваться для обработки различных типов трафика на множестве иерархических уровней.

В сети ATM существует возможность измерения качества предоставляемого пользователям обслуживания и обнаружения каких-либо ухудшений. Это требует наличия путей обнаружения любых отказавших элементов, которые позволяли бы выполнять необходимую конфигурацию. Точное указание места отказа особенно ценно в сложной сети.

Как и уровень ATM, управляющий соединениями, образуемыми виртуальными путями и виртуальными каналами, нижележащая система передачи состоит из нескольких компонентов: среды передачи, участков регенерации, участков мультиплексирования и трактов передачи.

Потоки информации для обслуживания определены в рекомендации МККТТ I.610. Они необходимы для реализации следующих функций:

- административного управления характеристиками, включающего проверку на четность BIP (Bit Interleaved Parity- четность перемежающихся битов) и сбор результатов FEBE (Far End Block Error- ошибка в блоке на дальнем конце) для оценки коэффициентов ошибок;
- административного управления устранением неисправностей, использующего постоянные проверки и механизмы для сигнализации о событиях AIS (Alarm Indication Signal-сигнал индикации аварии) и возвращаемую назад индикацию неисправностей FERF (Far Receive Failure-неисправность при приеме на дальнем конце).

Имеется пять потоков информации для обслуживания, как показано на рис.5.10.

Потоки информации F1, F2 и F3 переносятся по каналам, предоставляемым физическим уровнем в зависимости от типа поддержки (непрерывный поток ячеек или разбитый на кадры); потоки F4 и F5 используют виртуальные соединения (тракты или каналы), предоставляемые уровнем ATM.

Потоки информации для обслуживания физического уровня. Потоки информации для обслуживания F1, F2 и F3 ответственны соответственно за контроль регенерационного участка (называемого также цифровым участком) и тракта передачи с помощью, главным образом, средств систем передачи.

Для физического уровня SDH потоки информации обслуживания используют вспомогательную информацию участка синхронной иерархии (SOH) и тракта (POH). Результаты измерения параметров переносятся в блоках байтов, размер которых точно равен полезной нагрузке виртуальных контейнеров (2340 байт для цикла STN-1 при скорости 155,520 Мбит/с, 9360 байт для цикла STN-4 при скорости 622,080 Мбит-с).

Поэтому производится проверка на четность всех байтов ячеек (включая заголовки), передаваемых в контейнерах. В потоках F1 и F2 измеренные параметры передаются байт за байтом(BIP=8), тогда как в потоке F2 они передаются 3- байтовыми словами (BIP=24) со скоростью 155,520 Мбит/с или 6- байтовыми словами (BIP=96) со скоростью 622,080 Мбит/с.

Аналогичным образом физический уровень PDH использует некоторые двоичные элементы вспомогательной информации систем со скоростями 34,368 и 139,264 Мбит/с для размещения потоков информации обслуживания. Здесь же производят проверку всей полезной нагрузки, которая может содержать 530 или 2160 байт.

Физический уровень систем, основанных на передаче ячеек, не обеспечивает априори каких-либо специальных средств для передачи потоков информации обслуживания. В этом случае в поток ячеек постоянно вводятся

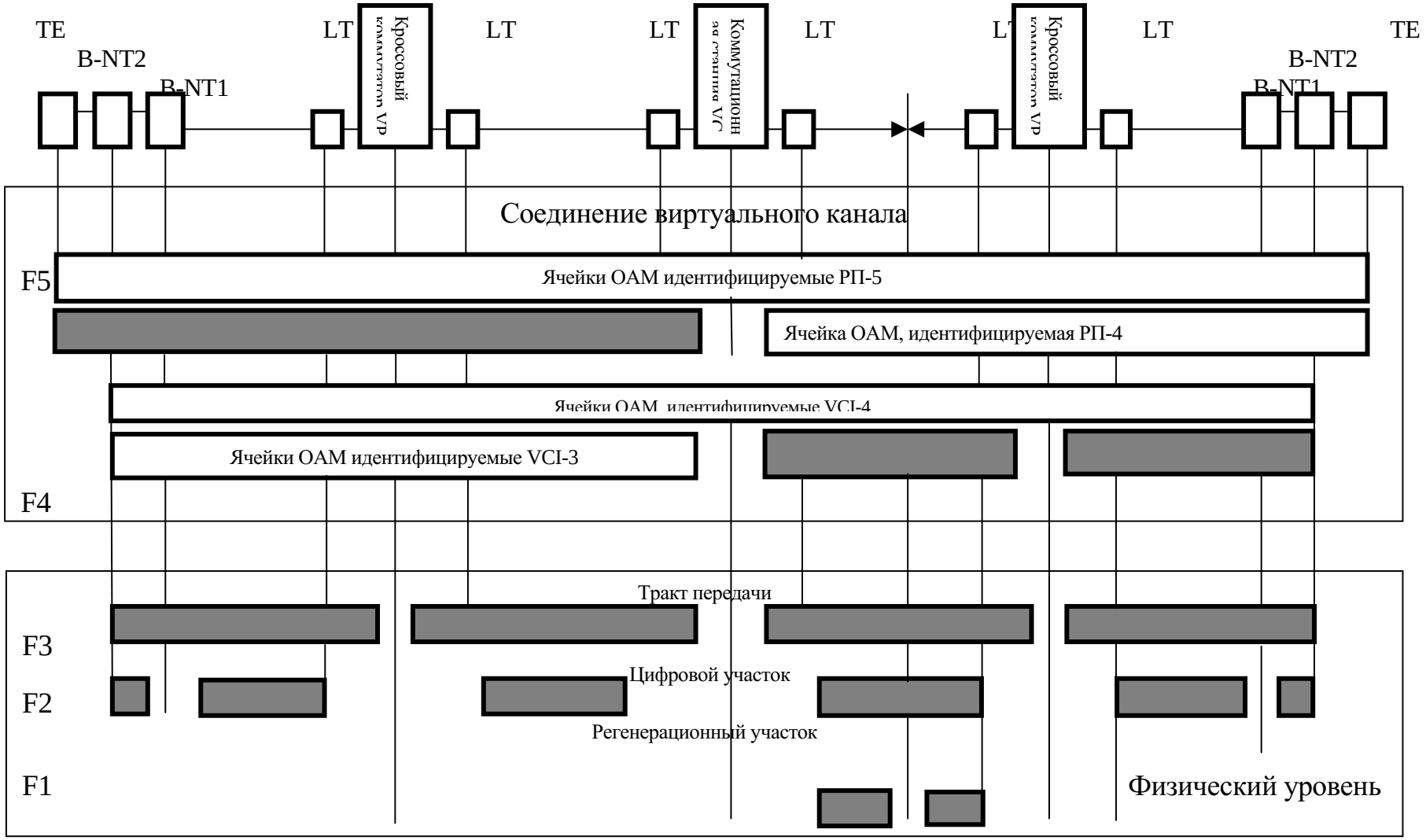


Рис. 5.10. Потоки информации для обслуживания сети ATM

специальные ячейки обслуживания OAM (Operation Administration and Maintenance- эксплуатация, управление и техническое обслуживание). Они определяются по заголовку, который указывает также, относится ли этот поток к типу F1 или типу F3. Измеренные параметры передаются фиксированным числом ячеек. Ячейки OAM могут содержать информацию проверки на четность (BIP=8), результаты (число ошибок, обнаруженных при проверке четности) или указания на события AIS и FERE. Их содержимое защищается 10 проверочными разрядами циклического кода (полином $x^{10}+x^9+x^5+x^4+x+1$).

Потоки информации для обслуживания уровня ATM. Потоки информации физического уровня могут быть доступны только оператору сети, потоки информации обслуживания F4 и F5 могут быть использованы пользователем. Как правило, они формируются по запросу. Это потоки, идущие из конца в конец, но существуют также сегментные потоки (потоки подсетей).

Виртуальный путь проверяется (поток F4) передачей ячеек OAM по резервному виртуальному каналу (VCI=4 для идущего из конца в конец потока F4, VCI=3 для потока F4 подсети). Потоки информации обслуживания, относящиеся к заданному виртуальному каналу (поток F5), используют тот же путь, что и рабочие ячейки: их можно отличить по особому кодированию поля PTI в их заголовках (PTI =5, если поток F5 идет из конца в конец, PTI=4, если он относится к подсети).

Потоки F4 и F5 используют тот же самый механизм измерения параметров. При этом применяются блоки номинальных размеров (N=128,256, 512 или 1024). Ячейка прверки на четность (BIP=16) OAM вводится только после N ячеек (в случае отсутствия активности передающего абонента) таким образом, чтобы это действие не могло привести к джиттеру в рабочем потоке. Ввод ускоряется, если активность отсутствует после прохождения 3N/2 ячеек; следующий ввод остается фиксированным на протяжении передачи 2N ячеек, и, таким образом, размер защищаемого блока уменьшается до N/2 ячеек.

Для проверки нахождения соединения в активном состоянии могут передаваться контрольные ячейки, если в течение заданного периода не передаются рабочие ячейки и нет сигнала об отказе.

В заключение отметим, что рекомендации ITU E.800 "качество обслуживания" определяется как "суммарный эффект характеристик службы, которые определяют степень удовлетворения пользователя службы". Согласно определению ISO/OSI (MOC/BOC) "качество обслуживания- это ряд качеств, отнесенных к обеспечению (N)-службы, которое воспринимаются пользователем (N)-службы". Оба определения слишком общи и не дают четкого представления о качестве обслуживания (QoS), поэтому необходима модель, раскрывающая содержание QoS. QoS прямо связывается с использованием общих ресурсов трафика, к которым относятся: узлы, передающие каналы, буферы, окна, а также ресурсы обработки и схемы интерфейсов в узлах и оконечных системах. Таким образом, количественная мера QoS прямо относится к использованию ресурсов, вовлеченных в обеспечении службы, т.е.

трафику на эти ресурсы. Поэтому концепции характеристики трафика и «качества обслуживания» являются родственными.

Функциональная архитектура коммуникационной службы определяется как общий набор функциональных элементов и динамических отношений между этими функциональными элементами. Она состоит из плоскости пользователя для передачи данных, плоскости управления для управления вызовом и соединением и плоскости менеджмента для администрирования.

Архитектура QoS представляет собой вид функциональной архитектуры, рассматривающей аспекты, относящиеся к ресурсам и трафику, и функции для администрирования этих ресурсов. Таким образом, QoS-архитектура имеет операционную и управляющую части. Внутри операционной архитектуры QoS функции по распределению, администрированию и перераспределению ресурсов, несущих трафик, обозначается как функции управления трафиком. Целями управления трафиком и перегрузкой является минимизация сложности сети и поддержка ряда QoS- классов уровня ATM. Отдельные QoS- классы уровня ATM и AAL- протокольные классы образуют служебные классы QoS. Существуют 4 служебных класса (A, B, C, D), которые классифицируются по следующим характеристикам: временная связь, скорость передачи, режим соединения. Таким образом, QoS- служба может быть охарактеризована как ориентированная на передачу. QoS- служба определяет природу QoS-параметров.

В рекомендации ITU –T I.350 определен матричный метод для идентификации параметров, который должен браться в расчете NP и QoS. Каждая строка матрицы представляет одну из основных связных функций (доступ, передача пользовательской информации, разъединение). Каждый столбец представляет один из трех возможных критериев (скорость, точность, надежность). Ввиду трудности для использования этих параметров, особенно на этапе оценки QoS, в ATM- сети было предложено в качестве основы стандартизации для использования три параметра: время задержки при передачи ячеек (cell-transfer delay, CTD) ; непостоянство времени задержки (cell- delay variation, CDV); процент потерянных ячеек (cell- loss ratio, CLR). Этими рамками задается трафик- контракт, который устанавливается между ATM- пользователем и сетью через UNI на фазе вызова и является основой для механизмов управления трафиком. Таким образом, QoS-модель операционной части OSI и ISDN содержит в себе статическую структуру для выполнения функций обработки трафика, параметры, относящиеся к QoS, и службы QoS.

Таким образом, была разработана модель качества обслуживания, раскрывающая содержание понятия качества обслуживания как "суммарного эффекта характеристик службы, который определяет степень удовлетворения пользователя службы".

Разработанная модель позволила

- раскрыть структуру качества обслуживания, состоящую из статической структуры для функций обработки трафика, параметров, относящихся к качеству обслуживания, и службы качества обслуживания;
- показать, что способом достижения цели качества обслуживания являются классы обслуживания и функции управления трафиком, которые зависят от архитектуры конкретной сети, при этом показана родственная связь понятий "качество обслуживание" и "характеристики трафика";
- рассматривать управление перегрузкой как часть общей проблемы управления трафиком, при этом оказываются неопределенными главные желаемые свойства алгоритмов для внутренних сетевых элементов,

Исследования последних лет, связанные с изучением трафика, который должен быть эффективно приспособлен к появлению новых сетей и новых служб, обнаружили, что реальный характер нагрузки сетей с АТМ-технологией носит фрактальный характер. Фрактальный характер нагрузки влияет на все аспекты управления трафиком и, следовательно, на количество (то есть производительность) и качество передаваемой информации.

Решение задачи доставки информации с заданным качеством обслуживания в сети с АТМ-технологией требует фрактальной формализации сетевого трафика, а также разработки метода управления сетевым элементом на уровне АТМ эталонной модели взаимодействия открытых систем.

5.2. Методы управления трафиком и перегрузкой

В рамках концептуальной модели качества обслуживания телекоммуникационной системы с АТМ- технологией, как было определено ранее, способом достижения цели качества обслуживания являются функции управления трафиком.

Обеспечение параметров качества обслуживания для различных типов трафика является сложной задачей, решение которой требует применения специальных методов управления трафиком в условиях изменяющихся требований к сетевым ресурсам. Это особенно важно для типов трафика с низкой предсказуемостью, не допускающего превентивного выделения сетевых ресурсов.

В основе метода АТМ лежит принцип статистического уплотнения, предполагающий отсутствие процедуры предварительного (статического) выделения ресурсов. АТМ- технология является гибкой технологией, которая поддерживает различные виды трафика и различные характеристики (потеря ячейки, задержка ячейки и непостоянство времени задержки), обеспечивая затребованную ширину полосы. Однако именно из-за гибкости возникает

опасность риска переполнения, поэтому управление переполнением есть критическая проблема, которая требует обеспечения согласованного КО для установленных соединений.

Большинство обычных сетей с коммутацией пакетов передают только трафик не реального времени, который может быть подвергнут управлению в случае перегрузки сети. Поскольку каждый пользователь виртуального канала передает всплесковый трафик, каждому каналу/ узлу позволяется более высокая пиковая нагрузка виртуального канала, чем они способны разместить, так, чтобы использовать статистические флуктуации каждого виртуального канала и, таким образом сохранить ширину полосы. В этом случае, если несколько виртуальных каналов передают информацию одновременно при пиковых скоростях, то канал/ узел становится перегруженным и должны быть привлечены алгоритмы управления перегрузкой для управления входными потоками каждого VC.

Характерные схемы управления перегрузкой, используемые в обычных сетях с коммутацией пакетов, включают оконное управление потоком и “обратное давление с запиранием”. Существуют ряд факторов, которые делают трудным управление перегрузкой в среде АТМ. К ним относятся широкий диапазон приложений, требующий большого диапазона ширины полосы, большое разнообразие различных структур трафика, которые требуют различных служб (например, службы чувствительные к задержке или службы низко чувствительные к задержке для передачи данных). Кроме того, очень высокие скорости при коммутировании и передаче делают сети более изменчивыми для целей управления трафиком. Так в схемах с обратным давлением как только перегрузка обнаруживается при узле в сети, узел посылает эту информацию к другим узлам так, что узлы, ответственные за перегрузку могут управлять этим трафиком. Такие схемы не подходят для управления в среде АТМ, так как большая часть трафика не является управляемой (то есть источники видео трафика и голоса не могут остановить генерирование ячейки данных, когда сеть перегружается).. Кроме того, из-за больших скоростей обратная связь становится неэффективной. Следовательно, требуется новая концепция для управления перегрузкой в среде АТМ.

Архитектура управления перегрузкой включает управление в оконечных терминалах, точках сетевого доступа и во внутренних элементах сети. Два последних понятия подразумевают использование термина “управление сетью”. Управление сетью действует на уровне соединения и уровне АТМ- ячейки, таким образом, предусматривая эффективное управление перегрузкой, встречающееся в различных временных масштабах. Целью управления сетью является обеспечение защиты сетевых ресурсов и каждого АТМ- соединения от других соединений, которые могут быть конкурирующими для ресурсов. Целью управления оконечным терминалом является действие, которое улучшает использование и характеристики, полученные от АТМ- соединения. Объединение сетевого управления и управления оконечным терминалом, примененное к данному АТМ- соединению, является “служебным контрактом”

(трафик-контрактом), включающим компоненты, согласованные с КО и трафиком.. Оконечные терминалы играют важную роль в архитектуре управления. Тем не менее архитектура не должна зависеть от специфических действий управления оконечным терминалом.

Рассмотрим влияние на качество обслуживания объектов в сетях с АТМ-технологией методов управления соединениями, глобального сетевого управления и внутреннего управления сетевыми элементами. Управление уровня соединения обеспечивает канал между служебным контрактом терминал-сеть и областью сетевого управления. Трафик-контракт устанавливается между пользователем и сетью до установления соединения. Он определяет переговорные характеристики соединения АТМ-уровня при интерфейсе пользователя с сетью (рис. 5.6). Трафик- контракт состоит из трех частей:

1. дескриптера исходного трафика (TD). Использующего четыре атрибута для описания трафика пользователя:
   - пиковая скорость ячейки;
   - гарантированная скорость ячейки;
   - наибольшее число ячеек, переданных с максимальной скоростью;
   - минимальная скорость передачи ячеек, поддерживаемая отправителем;
2. определения соглашения и допуска отклонения задержки во времени, который представляет собой некоторый запас надежности, который берет на себя разброс по времени задержки в оборудовании на удаленном конце. В определении соглашения указывается, какой трафик будет приемлемым для сети, то есть, с какой скоростью и с каким дроблением допускается отправка трафика пользователями. Такое определение вносится управляющей функцией. Соглашение применяется к ячейкам, когда они проходят интерфейс пользователя с сетью;
3. набор параметров качества обслуживания;
   - время задержки при передаче ячеек;
   - непостоянство времени задержки;
   - процент потерянных ячеек.

Требуемое КО должно гарантироваться сетью. Кроме того, трафик-контракт может содержать экспериментальные характеристики, которые могут включаться в сообщения сигнализации. Это особенно важно при испытаниях новых сетевых структур в том числе перспективных НОСС с АТМ-технологией. Все пункты контракта основаны на Общем алгоритме скорости ячейки. Сеть должна идентифицировать трафик при превышении служебных контрактов и предупреждать неблагоприятное влияние избыточного трафика не передачу неизбыточного трафика на других соединениях. Эти факторы относятся к управлению вхождения в соединение, управлению уровня соединения для разрешения или отказа соединения или переговоров для установки служебных параметров, которые могут быть поддержаны сетью.

Решение на принятие вызова базируется на характеристиках трафика и требованиях QoS, предварительно согласованных с сетью.

Глобальное сетевое управление на АТМ- уровне, обеспечивает глобальную стратегию управления, но в значительной степени включает действия распределенного управления при отдельных сетевых элементах.

Кроме служебных параметров, установленных на уровне соединения, необходим ряд механизмов для наблюдения за этими служебными соглашениями во время переноса АТМ- ячеек. Это одна из функций глобального управления АТМ- уровня. Другой функцией является снабжение информацией о состоянии перегрузки оконечных терминалов. Возможности управления в этой области могут быть обеспечены функционированием заголовка АТМ- ячейки. И состоит в следующем:

1. Структура управления имеет возможность для выборочного формирования нагрузки при условия перегрузки и, таким образом, обеспечить “упругость” для неизвестного трафика. Это выполняется посредством отдельного индикатора в заголовке ячейки АТМ, называемого указателем “приоритета потери ячейки” (CLP) . Если установлен указатель {CLP=1}, это означает, что ячейка может быть (выборочно) отброшена в любом сетевом элементе в соединении, если ячейка встретит в этом сетевом элементе локальную перегрузку выше порога. Этот CLP- указатель служит двойной цели:
    - Установление {CLP=1} пославшим терминалом должно означать, что ячейка несет несущественную информацию (и, таким образом, ячейка селективно отбрасывается при условии перегрузки);
    - Установление указателя {CLP=1} во время доступа в сеть, если сеть находит, что ячейка не в согласии с ограничениями трафика, обговоренными в служебном контракте. Установление сетью{CLP=1} для ячеек избыточного трафика называется “меткой избыточного трафика” для описанных причин. Однако, после установления {CLP=1} обработка ячейки в дальнейшем производится независимо от причин этого установления.
2. Существует возможность для передачи условий встретившихся перегрузок вперед вдоль виртуального канала или виртуального пути к АТМ- терминалу назначения, осуществляемая посредством индикатора EFCI (явная индикация перегрузки при прямой передаче) В заголовке АТМ- ячеек, и установлении {EFCI=1} любым элементом перегруженной сети, когда перегрузка превысит некоторый определенный порог. В зависимости от того, какой указатель установлен {EFCI=1} или {EFCI=0} при прибытии к АТМ– назначению EFCI означает наличие условий значительных перегрузок в некоторой точке ВК/ВП в первом случае, или отсутствие значительных перегрузок – во втором случае. В соответствие с подходящими алгоритмами терминал назначения может передавать обратно (например, периодически) терминалу источнику (возможно внутри ряда возможных AAL) информацию, позволяющему

источнику производить действия (например, формирование трафика и/ или контроль ошибок) для получения улучшенного использования соединения из конца в конец.

3. Для управления эксплуатационным параметром (динамического отслеживания параметров заявки пользователя) были определены в стандартах ATM механизмы мониторинга трафика/ установления излишнего трафика для UPC. Они могут рассматриваться в качестве фильтра (TBF). Необходимость фильтрации исходящего трафика может определяться требованиями как защиты, так и предотвращения блокировки линии низкоприоритетным трафиком. Например, стратегия фильтрации может быть использована для ограничения объемов некоторого трафика, чтобы критически важный трафик получил приоритет над менее значимым. Фильтрация может также потребоваться для обеспечения безопасности и блокировки несанкционированного трафика. Основными целями мониторинга трафика являются:
   - Оказание влияния на поток трафика с {CLP=0}- и {CLP=1}- ячейками, входящими в сеть по ВК/ВП;
   - Извлечение пользы терминалом– адресатом из возможности терминала источника посылать ячейки с {CLP=1};
   - Получение возможности предсказания терминалом источником определения мониторингом трафика избыточности его трафика; уменьшение возможности принятия ошибочного принятия решения о некотором количестве избыточного трафика, посылаемого терминалом- источником в соответствии с соглашением. Путем введения минимального джиттера в точке мониторинга.

   Структура мониторинга трафика представлена на рис. 5.11.

Глобальное сетевое управление уровнем ATM зависит от действий внутри элементов сети, то есть от внутреннего управления сетевыми элементами, включая мониторинг трафика в точке доступа в сеть, выборочного отбрасывания ячейки (SCD) и установления EFCI. В достижении целей управления трафиком и перегрузкой наиболее существенным является способность сетевого элемента осуществить формирование служб и распределение ресурса, соответствующее классам служб, поддерживаемых этим элементом. Она может включать распределение буфера между классами служб и распределение в реальном времени ширины полосы передающего канала так, чтобы обеспечить, по крайней мере. Минимальную ширину полосы для каждого служебного класса и, таким образом, предотвратить “выключение” из передачи одного служебного класса другим. Это устанавливается системой “относительных приоритетов” между служебными классами.

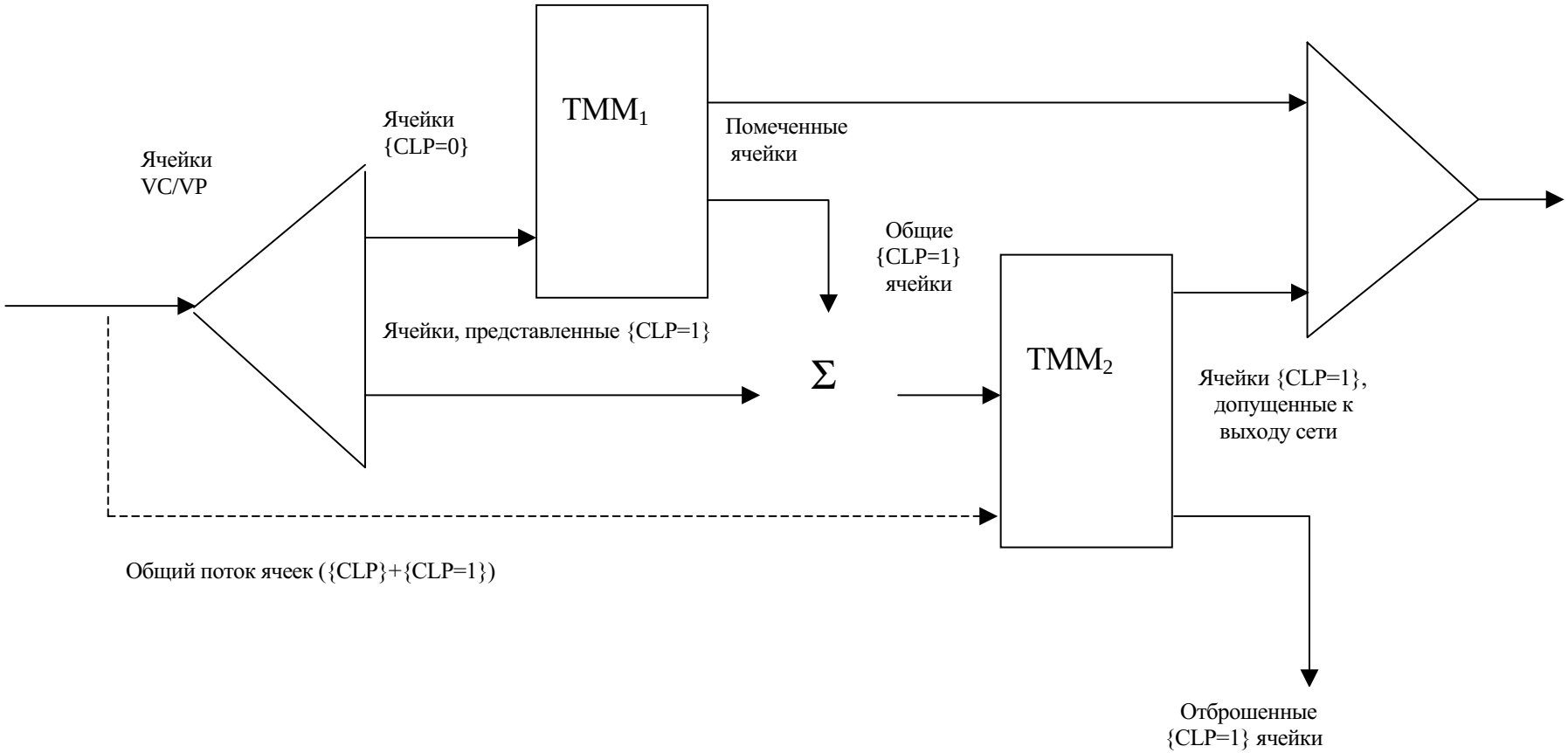


Рис. 5.11. Структура мониторинга трафика

5.3. Вопросы прогнозирования трафика в высокоскоростных сетях связи

В настоящее время вопросы прогнозирования трафика, в основном, решены применительно к низкоскоростным сетям связи. В статье эта проблема исследуется с учетом того, что в телекоммуникационной сети обмен информацией происходит с большой скоростью. Важный прикладной вывод из излагаемого материала — это формулировка рациональных аспектов управления перегрузкой.

Переход к новым сетям типа Ш-ЦСИО определяет необходимость расширения известных подходов в управлении сетевыми ресурсами, в том числе, к управлению потоками. Предполагая, что в сети применяются методы маршрутизации, обеспечивающие выбор оптимальных по заданному критерию маршрутов доставки сообщений, управление потоками определяется как совокупность механизмов, обеспечивающих доступ к ресурсам сети – системам передачи, хранения и обработки информации. Введение указанных механизмов в состав алгоритмов системы доставки информации обеспечивает устранение перегрузок, возникающих в отдельных точках сети, и разрешение тупиковых ситуаций. Под перегрузкой понимается такое состояние сети, при котором основные показатели качества обслуживания начинают быстро ухудшаться. В зависимости от типа сети эти ухудшения могут выражаться в увеличении числа отказов при установлении соединения, в увеличении числа потерянных пакетов и в увеличении средних значений и джиттера задержек. Перегрузки могут возникать как на отдельных участках сети (локальная перегрузка), так и распространяться на всю сеть в целом (глобальная перегрузка), что приводит к явлению полной блокировки всех участков сети, когда межконцевые задержки стремятся к бесконечности, а производительность – к нулю. Важность и сложность решения задач проектирования алгоритмов управления потоками и борьбы с перегрузкам определяются, в первую очередь, высокой стоимостью потерь информации при перегрузках, вызванных наличием больших объемов информации в виртуальном пути, что особенно важно в системах военного назначения.

Существуют три аспекта управления перегрузкой [1]: а) предупреждение, б) предотвращение, в) восстановление. Предупреждение перегрузки включает оптимально выбранные компоненты, хорошо спроектированный алгоритм маршрутизации, механизм принуждения для гарантии того, что пользователь не превысит установленную скорость трафика, управление очередями, которые защищают критические классы трафика (трафик управления сетью, сообщения маршрутизации). Предотвращение перегрузки - это действия, предпринимаемые сетью для избежания возможности перегрузки. Примером может служить изменение таблиц маршрутизации, чтобы направить трафик в обход тяжело нагруженной сетевой компоненты. Восстановление после перегрузки- действия, предпринимаемые

сетью после того, как перегрузка обнаружена, для ограничения влияния перегрузки. Примером является отбрасывание низкоприоритетных пакетов, когда буферы переполнены. Для выбора стратегии управления очень важным вопросом является возможность предсказания с разумной точностью, когда появятся пиковые требования и каков должен быть размер сети, чтобы можно было работать без существенных перегрузок. В такой ситуации предупреждение перегрузки есть наиболее существенная сторона управления перегрузкой. Однако анализ высококачественных измерений трафика, проведенный на нескольких локальных сетях типа Ethernet в Морристаунском центре института Bellcore (Bellcore Morristown Reserch and Engineering Centre) в штате Нью-Джерси, США [2] и др. показал, что нагрузка в исследованных сетях существенно отличается как от классических представлений (телефонный трафик), так и от новых моделей (пакетный трафик), рассматриваемых в литературе. Отличительная особенность нагрузки быстродействующих цифровых сетей - ее пачечный характер, причем пачки (скученности) появляются в разных масштабах времени, и это затрудняет определение длин пачек: в разных шкалах времени длительность пачки может изменяться в пределах от миллисекунд до минут и часов в зависимости от разрешающей способности измерительной аппаратуры. Трафик, который является пачечным на многих или всех масштабах времени может быть описан статистически, используя понятие самоподобия. Периоды наибольшей нагрузки в условиях самоподобного ("фрактального") трафика не поддаются предсказанию. В связи с этим возникает вопрос, если предсказание пика эксплуатации так низко, то будет ли оно иметь влияние на проектирование сети. Другим вопросом технической важности является вопрос, возможно ли предупредить перегрузку увеличением емкости буферов. Механизмы наблюдения трафика в сети позволяют понять, каким образом в будущих сетях можно будет эффективно избежать перегрузки. Решение этого вопроса зависит от того, существуют ли такие структуры в самоподобном трафике, которые могут быть использованы для предсказания перегрузки. Кроме того, когда встречается перегрузка, предпринимаются действия, чтобы помочь сети выйти из этого состояния. Перегрузка может иметь как незначительные, так и серьезные последствия. Поэтому для того чтобы, могли быть сделаны любые реальные предложения относительно восстановления сети, необходимо знать, как долго перегрузка продолжается и каковы структуры потерь, которые появляются во время перегрузки. Для разрешения этих вопросов необходима модель трафика, которая позволила бы рассмотреть поведение алгоритмов управления перегрузкой не только в режиме реального времени, но и использовать прогнозные значения для предотвращения перегрузок. В качестве модели самоподобного трафика рассмотрим частичное броуновское движение (fBm) [3]. Процесс fBm Z_t^H с индексом $H \in (½, 1)$ является гауссовым процессом с нулевым средним значением и корреляционной функцией

$$\Gamma^H(t,s) = ½ (t^{2H} + s^{2H} - |t-s|^{2H}) \, Var(Z_t^H), \qquad (5.1)$$

где H- Херст- параметр.

Для простоты обозначений будем предполагать, что

$$\mathrm{Var}\,(Z_t^{H}) = \tfrac{1}{2}H + \int_0^{\infty} (1+s)^{H-\frac{1}{2}} - s^{H-\frac{1}{2}})^2\, ds \qquad (5.2)$$

Необходимо отметить, что Z_t^{1} – процесс ограниченной вариации, $Z_t^{1/2}$ – стандартный винеровский процесс. Распределение Z_t^{H} сходится к винеровскому при $H \to \tfrac{1}{2}$.

Очевидно, что процесс Z_t^{H} имеет стационарные приращения и обладает свойствами автомодельности распределений приращений. Из определения процесса Z_t^{H}

$$Z_t^H = \int_{-\infty}^{0} (|t-r|^{H-\frac{1}{2}} - |r|^{H-\frac{1}{2}})\, dw_r + \int_0^t |t-r|^{H-\frac{1}{2}}\, dw_r, \qquad (5.3)$$

где W_t -стандартный винеровский процесс на ∇.

Такое же представление справедливо в случае, когда Z_t^{H} – d-мерный fBm с индексом $H \in (\tfrac{1}{2},1)$, т.е. когда все его компоненты суть независимые fBm с одним и тем же индексом H. В этом случае W_t- стандартный d-мерный винеровский процесс на ∇ с независимыми компонентами. Отметим также характерное свойство траекторий процесса Z_t^H:

$$\lim_{\substack{|\Delta t_i| \to 0 \\ N \to \infty}} \sum_{i=1}^{N} |Z_{t_{i+1}}^H - Z_{t_i}^H|^{\frac{1}{H}} < \infty \text{ P- п.н.},$$

т.е. траектории процесса Z_t^H имеют ограниченную 1/H -вариацию .

Определим интеграл по процессу Z_t^H. Этот процесс не является семимартингалом [см., например, 4]. Поскольку fBM – не семимартингал, можно ожидать, что стохастические интегралы не являются непрерывными в отношении подынтегрального выражения. Для простого процесса

$$Y_t = \sum_{j-1}^{\kappa} X_j 1_{(T_{j-1}, T_j)}(t), \text{ интеграл } \int_{-\infty}^{\infty} Y_t dZ_t \overset{def}{=} \sum_{j=1}^{k} X_j (Z_{T_j} - Z_{T_{j-1}}), \quad (5.4)$$

где $\overset{def}{=}$ равенство по определению. Для процесса Y с локально ограниченной дисперсией всегда возможно использовать определение формулы интегрирования по частям.

$$\int_a^b Y_t dZ_t \overset{def}{=} Y_b Z_b - Y_a Z_a - \int_a^b Z_t dY_t. \qquad (5.5)$$

Если Y- детерминированный, интеграл, определенный в [4], получается как предел в L^2 сумм Римана типа (5.1) с детерминированным T_j. Определение может быть расширено к более широкому классу детерминированных функций, используя пределы в L^2. Если Y- случайный процесс, то обозначая

$$p_\beta(x) = \mathrm{sing}\,(x)|\,x|^{\beta},\ x \in \nabla,$$

$$\text{и } U_{n,k} = p_\beta(\otimes[\, Z_{k/n} - Z_{(k-1)/n} \mid \Phi_{(k-1)/n}\,]),$$

где $\Phi_t=\sigma\{Z_s : s\le t\}$, а $\beta > 0$, определим

$$Y_n(t) = \sum_{k=1}^{n} U_{n,k}\, l_{\{[(k-1)/n,k/n]\}}(t).$$

Можно увидеть, что $Y_n(t)\to 0$ для каждого t, когда $n\to\infty$. Принимая во внимание самоподобие Z имеем

$$\int_0^1 Y_n(t)dZ_t =$$

$$= \sum_{k=1}^{n} U_{n,k}(Z_{k/n} - Z_{(k-1)/n}) \underset{D}{=} \sum_{k=1}^{n} n^{-\beta H} p_\beta(E[(Z_k - Z_{k-1})|\Phi_{k-1}])n^{-H}(Z_k - Z_{k-1}) =$$

$$= n^{1-H-\beta H}\,\frac{1}{n}\sum_{k=1}^{n}(p_\beta(E[Z_1|\ \Phi_0])\, Z_1 \circ T_t^{\,k-1},$$

где $\underset{D}{=}$ равенство по распределению, а T_1 – оператор сдвига. $Z_t(\omega) = Z_{t-1}(T_1\omega)$.

Поскольку процесс Z_t^H имеет стационарные приращения, стационарная последоваательность $Z_{n+1} - Z_n$ (часто называемая частичным гауссовым шумом) является эргодической [5]. Эргодическая теорема Биркгофа дает

$$\lim_{n\to\infty}\int_0^1 Y_n(t)dZ_t = (\lim_{n\to\infty} n^{1-H-\beta H})\cdot E(p_\beta[Z_t|\ \Phi_0]).$$

Математическое ожидание справа положительно, так как p_β (*) - возрастающая нечетная функция. Таким образом, предел стремится к бесконечности, когда $\beta < (1-H)/H$. Подынтегральное выражение не является непрерывным (как было показано ранее) и имеет следствием то, что определение Римановой суммы (1) не может быть расширено к основному определению с пределами в L^2 подобно классическому интегралу Ито. Однако, это возможно для детерминированного подынтегрального выражения. Для простоты ограничимся случаем с детерминированным подынтегральным выражением. Для $f\in L^2(\nabla;\nabla)\cap L^1(\nabla;\nabla)$ определим

$$\int_R f(t)dZ_t = c_H(H-\tfrac{1}{2})\int_R\left(\int_t^\infty (t-\tau)^{H-\frac{3}{2}} f(t)dt\right)dZ_\tau,$$

где $c_H = \sqrt{2H\Gamma(\frac{3}{2}-H)/\Gamma(2-2H)}$, Г- гамма- функция.

Чтобы это определение имело смысл, необходимо, чтобы f было таким, что функция $\tau \to \int_\tau^\infty (t-\tau)^{H-\frac{3}{2}} f(t)dt$ была квадратично интегрируема, и условие $f\in L^2(\nabla;\nabla)\cap L^1(\nabla;\nabla)$ является достаточным. Следующие положения являются основными в интегрировании по отношению к Z. Это обеспечивается изоморфизмом Гильбертова пространства, которое является центральным в

теории интегрирования гауссовых процессов. В качестве доказательства, рассмотрим f, g$\in$ $L^2(\nabla;\nabla) \cap L^1(\nabla;\nabla)$. Тогда имеем

$$E(\int_R f(s)dZ_s \int_R g(t)dZ_l) = H(2H-1)\iint_{R^2} f(s)g(t)|s-t|^{2H-2}dtds. \qquad (5.6)$$

В соответствии с доказательством, приведенным в [6],

$$\int_{-\infty}^{min(s,t)} (s-\tau)^{H-1/2}(t-\tau)^{H-\frac{3}{2}}d\tau = \int_0^{\infty} (|t-s|+\tau)^{H-\frac{3}{2}}\tau^{H-\frac{3}{2}}d\tau =$$

$$= |t-s|^{2H-2}\int_0^{\infty}(1+\tau)^{H-\frac{3}{2}}\tau^{H-\frac{3}{2}}d\tau = |t-s|^{2H-2}\frac{\Gamma(H-\frac{1}{2})\Gamma(2-2H)}{\Gamma(\frac{3}{2}-H)},$$

результат получается прямым вычислением

$$(E(\int_R f(s)dZ_s \int_R g(t)dZ_t) =$$

$$= c_H^2(H-\tfrac{1}{2})^2 \iint_{R^2} dtdsf(s)g(t) \int_{-\infty}^{min(s,t)} d\tau(s-t)^{H-3/2}(t-\tau)^{H-3/2} =$$

$$= H(2H-1)\iint_{R^2} f(s)g(t)|s-t|^{2H-2}dtds. \qquad (5.7)$$

Уравнение (7) может рассматриваться как

$$Cov(dZ_s, dZ_t) = H(2H-1)|s-t|^{2H-2}dsdt.$$

Таким образом, задача предсказания Z_a (a>0) на основе величин Z_t, полученных на интервале (-T,0), из-за стационарности приращений Z эквивалентна проблеме предсказания ($Z_{t+a} - Z_t$) при любых t на основе ($Z_s - Z_t$), s$\in$(t-T,t).

Учитывая вышесказанное, можно вычислить предиктор в виде интеграла от fBM следующим образом. Пусть Z – это fBM с параметром Херста H$\in$(½,1). Тогда для каждого a>0 и T$\in$(0,∞] выражение для предиктора

$$\overset{\cap}{Z}_{a,T} = E[Z_a|Z_s,\ s\in(-T, a)].$$

может быть представлен как интеграл $\int_{-T}^{0} g_t(a,t)dZ_t$, где для T<$\infty$, t$\in$(0,T)

$$g_t(a,-t) = \frac{Sin(\pi(H-1/2))}{\pi}t^{-H+1/2}\int_0^a \frac{\sigma^{H-1/2}(\sigma+T)^{H-1/2}}{\sigma+t}d\sigma, \qquad (5.8)$$

и для T=∞, t>0

$$g_\infty(a,-t) = \frac{Sin(\pi(H-1/2))}{\pi}t^{-H+1/2}\int_0^a \frac{\sigma^{H-1/2}}{\sigma+t}d\sigma =$$

$$= \frac{Sin(\pi(H-1/2))}{\pi}\left(\frac{1}{H-1/2}\left(\frac{t}{a}\right)^{-H+1/2} - B_{a/(t+a)}(H-1/2, 3/2-H)\right), \quad (5.9)$$

где B.(∗,∗)- неполная бета- функция.

Функция g_T(a, ∗)- решение интегрального уравнения

$$(2H-1)\int_0^T g_T(a,-t)|t-s|^{2H-2}\,dt = (a+s)^{2H-1} - s^{2H-1}, s \in (0,T), \quad (5.10)$$

и имеет свойства масштабирования:

$$g_T(a,t) = g_{T/a}(1, t/a). \quad (5.11)$$

Так как Z-гауссов процесс, $\overset{\cap}{Z}_{a,T}$ - линейный функционал (Z_s; s∈(-T,0)). Т.о., пытаемся найти гладкую функцию g_T (a, ∗): (-T,0) →∇, удовлетворяющую условию ортогональности

$$E((Z_a - \int_{-T}^{0} g_T(a,t)dZ_t(a,t)dZ_t)(Z_u - Z_v)) = 0, \quad -T < v < u \le 0. \quad (5.12)$$

Формула для решения этого уравнения может быть найдена в [6]. Т.о., получаем

$$h_{T,a}(t) = -c(\alpha)t^{-\alpha/2}\frac{d}{dt}\int ds s^{\alpha}(s-t)^{-\alpha/2}\frac{d}{ds}\int du u^{-\alpha/2}(s-u)^{-\alpha/2} f_{T,a}(u), \quad (5.13)$$

где $c(\alpha) = (\Gamma(1-\frac{1}{2}\alpha)^2\Gamma(\alpha)2Cos(\frac{1}{2}\pi\alpha))^{-1}$.

Заменой переменной получаем

$$\int_0^s u^{-\alpha/2}(s-u)^{-\alpha/2}(\sigma+u)^{\alpha-1}du = \frac{\Gamma(1-\alpha/2)^2}{\Gamma(2-\alpha)}\left(\frac{s}{\sigma}\right)^{1-\alpha} F\left(1-\alpha, 1-\frac{\alpha}{2}, 2-\alpha, -\frac{s}{\sigma}\right),$$

где F-гипергеометрическая функция. Учитывая (8, 9), имеем

$$\lim_{T\to\tau} g_T(a,-t) = \frac{Sin(\pi(H-\frac{1}{2}))}{\pi} t^{-H+\frac{1}{2}} \int_0^a \frac{\sigma^{H-\frac{1}{2}}}{\sigma+t}\,d\sigma =$$

$$= \frac{Sin(\pi(H-\frac{1}{2}))}{\pi}\left(\frac{1}{H-\frac{1}{2}}\left(\frac{t}{a}\right)^{-H+\frac{1}{2}} - \int_{\frac{t}{a}}^{\infty}\frac{\sigma^{-H+\frac{1}{2}}}{1+\sigma}\,d\sigma\right) =$$

$$= \frac{Sin(\pi(H-\frac{1}{2}))}{\pi}\left(\frac{1}{H-\frac{1}{2}}\left(\frac{t}{a}\right)^{-H+\frac{1}{2}} - B_{\frac{a}{(t+a)}}(H-\tfrac{1}{2}, \tfrac{3}{2}-H)\right). \quad (5.14)$$

Легко проверить, что т.к. H∈ (1/2, 1), то мы имеем g_T(a,∗) ∈ $L^1 \bigcap L^2$ как для конечного, так и бесконечного T. Для малых T отметим также, что

$$\lim_{T\to 0}\min_{-T<t<0} g_T(a,t)=\infty \qquad \lim_{T\to 0}\int_{-T}^{0} g_T(a,t)dt=0.$$

Интересно отметить, что весовая функция стремится к бесконечности при истинных значениях T, а также при – T, когда T конечно. В частности, в последнем случае весовая функция не монотонная . Интуитивно это может быть понято так, что "ближайшие свидетельства" о наблюденном прошлом имеют специальный вес.

Дисперсия предиктора $E[Z_a|Z_s,\ s\in(-T,0)]$ при $a>0$, $T\in(0,\infty)$ имеет вид

$$D^2(E[Z_a|Z_s,\ s\in(-T,0)])= D^2(Z_a)H\int_0^a g_{T/a}(1,-s)((1+s)^{2H-1}-s^{2H})ds.$$

Используя (7), получаем

$$D^2\Big(\int_{-T}^{0} g_T(a,t)dZ_t\Big)=H(2H-1)\int_0^T\int_0^T g_T(a,-s)g_T(a,-t)|s-t|^{2H-2}\,dsdt=$$

$$D^2(Z_a)H\int_0^{\frac{T}{a}} g_{\frac{T}{a}}(1,-t)((1+t)^{2H-1}-t^{2H-1})ds.$$

В случае $T=\infty$, можно, интегрируя по частям, получить

$$H\int_0^\infty g_\infty(1,-t)((1+t)^{2H-1}-t^{2H-1})dt=\frac{Sin(\pi(H-\frac{1}{2}))}{2\pi}\int_0^\infty\frac{(1+t)^{2H}-t^{2H}-1}{(1+t)t^{H+\frac{1}{2}}}dt$$

Для $-1/2<H<0$, последний интеграл может быть вычислен, используя формулы из [ ], что дает

$$\int_0^\infty\frac{(1+t)^{2H}-t^{2H}-1}{(1+t)t^{H+\frac{1}{2}}}dt=\frac{\Gamma(\frac{1}{2}-H)\Gamma(\frac{1}{2}-H)}{\Gamma(1-2H)}-2\Gamma(\frac{1}{2}+H)\Gamma(\frac{1}{2}-H)=$$

$$=\frac{2\Gamma(\frac{1}{2}-H)^2}{(H-\frac{1}{2})\Gamma(2-2H)}+\frac{2\pi}{Sin(\pi(H-\frac{1}{2}))}.$$

Теперь аналитическим продолжением этот результат может быть расширен для случая $½<H<1$. Относительная дисперсия ошибки как функция H

$$\frac{D^2(Z_a-\hat{Z}_{a,\infty})}{D^2(Z_a)}=\frac{Sin(\pi(H-\frac{1}{2}))}{\pi(H-\frac{1}{2})}\frac{\Gamma(\frac{3}{2}-H)^2}{\Gamma(2-2H)}$$

независима от a в результате самоподобия.

Анализ относительной дисперсии ошибки $\frac{D^2(Z_a-\hat{Z}_{a,\infty})}{D^2(Z_a)}$ – как функции T показал, что предсказание Z_a дает очень малое отличие от того, знаем ли мы Z на (-a, 0) или на (-∞,0). Однако, это маленькое отличие станет существенным,

если попытаться моделировать fBM шаг за шагом вперед. Таким образом, получаем простое правило: использовать только самую последнюю секунду, чтобы предсказать следующую секунду, самую последнюю минуту для предсказания следующей минуты и т. д.

Литература

1. Ефимушкин В.А., Ледовских Т.В. Методы управления перегрузками в сетях ATM// LV научная сессия, посвященная дню радио «Радиотехника, электроника и связь на рубеже тысячелетия». Труды, М.: 2000, С.41- 43.
2. Gershi A., Lee K.J. A Congestion Control for ATM Networks// IEEE Journal on Select Areas in Communication, 1991,v.9, №7, p. 1119-1130.

ГЛАВА 6
ОЦЕНИВАНИЕ ПАРАМЕТРОВ ТРАФИКА

Существенным аспектом техники трафика пакетных сетей является определение возможного набора рабочих измерений, которые могли бы быть собраны сетевыми элементами на уровнях мониторинга трафика. После появления в 85- 90- х годах высокоточных измерений LAN Ethernet трасс, записанных в Bellcore при различных условиях, и установления факта масштабно- зависимых свойств трафика были сделаны заключения о том, что для описания трафика требуется только три параметра (скорость, параметр Херста (Hurst- parameter) и параметр пиковости), чтобы решить такие технические проблемы, как, например, определение размера буфера и другие проблемы. Поэтому Херст- параметр держит центральное место в описании самоподобного трафика.

Как известно [1], статистически самоподобные процессы (Statistically Self-Similar - SSS) имеют общую характеристику, такую как $1/f^{\gamma}$, в частности, в диапазоне $0<\gamma< 2$. Другой общей характеристикой является долговременная зависимость (Long- Range Dependence- LRD) как в самом процессе, так и в его инкрементах.

Входной трафик со скоростью, зависящей от времени, моделируется стационарным стохастическим процессом. Основными характеристиками этого процесса являются его среднее $\mu_x=E[x]$, дисперсия $\sigma_x^2 =E[(x-\mu_x)^2]$ и корреляционная функция $\gamma_x(k)=E[(x(t+k)-\mu_x)(x(t)-\mu_x)]$. В этом контексте самоподобные свойства трафика проявляются сами в особой форме $\gamma_x(k)$, а именно, она уменьшается с увеличением k так медленно, что сумма всех корреляций на любом данном промежутке времени всегда ощутима, даже, если каждые в отдельности корреляции очень малы. Поэтому прошлое оказывает долговременное влияние на будущее, преувеличивая влияние изменчивости трафика и делая статистическое оценивание проблематичным. Этот феномен известен как долговременная зависимость (LRD) и обычно определяется

$$\gamma_x(k)\sim c_\gamma|k|^{-(1-\alpha)},\ \alpha\in(0,1), \qquad (6.1)$$

где c_γ- положительная константа. Эквивалентное утверждение для спектра имеет вид для частот, близких к нулю

$$f_x(v)\sim c_f|v|^{-\alpha},\ |\,v\,|\to 0, \qquad (6.2)$$

где $f_x(v)$ в случае дискретно- временного процесса удовлетворяет

$$\gamma_x(0)=\sigma_x^2\int_{-1/2}^{1/2} f_x(v)dv\,,$$

где σ_x^2 - дисперсия (или мощность) x_t.

Каждое из этих определений включает два параметра: (α, c_γ) или (α, c_f) соответственно, которые эквивалентны, когда

$$c_f = 2(2\pi)^{-\alpha} c_\gamma \Gamma(\alpha) \sin\left(\frac{(1-\alpha)\pi}{2}\right),$$

где Г- функция Эйлера.

В каждой паре параметров α - наиболее важный, так как он определяет существование самого феномена и управляет поведением масштабной характеристики также, как статистиками, произведенными из него.

Параметр Херста описывает (на практике асимптотически) самоподобие кумулятивного трафикового процесса $\int_0^t x(s)ds$, в то время как α описывает LRD скоростного процесса x(t). Существует не менее общая практика рассматривать H по отношению к LRD посредством выражения $H=(1+\alpha)/2$, которое будет использоваться в дальнейшем.

Херст- параметр является мерой самоподобия или статистической инерции процесса. Оценки Херст- параметра (H) основываются на идее измерения наклона линейного приближения на графике log- log. Примером такой оценки является, так называемая вариограмма или R/S- оценка [3]. График зависимости R/S от N (дискретное время) в логарифмическом масштабе по обеим шкалам использует тот факт, что для самоподобной последовательности данных диапазон изменения масштаба или R/S-статистика растет согласно степенного закона с экспонентой H как функция числа включенных точек (N). Таким образом, график R/S в зависимости от N на графике log-log имеет наклон, который является оценкой H. Такие оценки имеют бедные статистические характеристики, а именно: большое смещение и субоптимальную дисперсию. Во временной области также существует оценка, известная как Allan variance (дисперсия Аллана) [1,3], которая состоит в измерении ожидания квадрата разности средних значений внутри окон длины T

$$V_a(T) = \frac{1}{K}\sum_{k=1}^{K}\left(\int_{t_k-T}^{t_k} x(u)du - \int_{t_k}^{t_k+T} x(u)du\right)^2,$$

где K- число сегментов данных размера T .

Эта величина ведет себя в присутствии долговременной зависимости также как степенной закон T и уже допускает несмещенную оценку H.

Для оценки H можно использовать оценки в частотной области. Стандартная спектральная оценка состоит в усреднении сглаженных периодограмм, вычисленных на различных отрезках данных

$$\hat{\Gamma}_2(\nu) = \sum_{k=1}^{P}\left|\int x(t-kL)w_1(t)\exp(i2\pi\nu t)dt\right|^2,$$

где P- число отрезков данных, L- их длина, w_L- взвешенное окно. Было показано [1], что применяемые к $1/|f|^\alpha$- сигналам такие спектральные оценки результируются в оценку H, основанную на линейном приближении на графике $\log(\nu)$ от $\log(\hat{\Gamma}_2(\nu))$, которая сильно смещена.

Для анализа LRD и родственной оценки Херст- параметра H для стационарных и стационарно- инкрементных данных Patrice Abry (Франция) и Darryl Veitch (Австралия) в 1996 году предложили оценку (AV- оценку), которая вводит инструмент оценивания, основанный на вейвлетах. AV- оценка может быть применена как к сигналам непрерывного, так и дискретного времени. Может быть использована в режиме реального времени (on- line) с малыми вычислительными затратами и требованиями памяти, а также для накопленных данных для анализа не в режиме реального времени (off- line), при этом радикально уменьшается объем данных, которые необходимо сохранять для анализа off- line. AV- оценка является несмещенной и эффективной при гауссовских предположениях.

Она использовалась при анализе трафика LAN Ethernet и продемонстрировала превосходное согласие с теоретическими данными [4].

В частности, Abry и Veitch показали, что среднее $| d_{j,k} |^2$ (где $d_{j,k}$ – вейвлетные коэффициенты) на каждой шкале j- полезная спектральная оценка. Действительно, если E_j обозначает среднее $|d_{j,k}|^2$ на каждой шкале

$$E_j = \frac{1}{N_j} \sum_k \left| d_{j,k} \right|^2 ,$$

где N_j- число вейвлетных коэффициентов на каждой шкале j, тогда E_j- мера энергии, которая лежит внутри данной ширины полосы 2^{-j} около частоты $2^{-j}\lambda_0$, λ_0 - произвольная эталонная частота, полученная выбором ψ_0, и, таким образом может быть рассмотрена как статистическая оценка для спектра $\Gamma_x(\lambda)$ от x. Действительно, может быть показано, что когда x- процесс стационарный в широком смысле, ожидание E_j задается

$$E[E_j] = \int f(\lambda) 2^j \left| \hat{\psi}(2^j \lambda) \right|^2 d\lambda = c_f \left| 2^{-j} \lambda \right|^{1-2H} \int |\lambda|^{1-2H} \left| \hat{\psi}(\lambda) \right|^2 d\lambda ,$$

где $\hat{\psi}(\lambda)$ - есть Фурье- преобразование анализирующей вейвлеты $\psi(t)$.

Если построить график $\log_2 E_j$ в зависимости от шкалы j, то получаим несмещенный масштабный анализ X. Масштабный анализ сигнала, который является асимптотически самоподобным, для больших шкал показывает линейное отношение между $\log_2 E_j$ и шкалой j. Если сигнал точно самоподобный, график $\log_2 E_j$ от j будет демонстрировать линейное отношение для всех шкал.

AV- оценка может быть описана в виде последовательности 4- х шагов: вейвлетная декомпозиция, оценка дисперсии деталей, анализ с использованием диаграммы с логарифмической шкалой, оценивание параметров.

Вейвлетная декомпозиция. Вейвлетная декомпозиция может быть осуществлена на основе дискретного вейвлетного преобразования (Discrete Wavelet Transform), используя алгоритм кратномасштабного анализа (Multiresolution Analysis).

Кратномасштабный анализ (КМА) является инструментом для построения вейвлетных базисов, и в настоящее время нет других сколько-нибудь универсальных способов его построения.

Кратномасштабный анализ- это последовательность $\{V_j\}_{j\in Z}$ вложенных друг в друга замкнутых подпространств $L^2(R)$, удовлетворяющих следующим свойствам [1, 2]:

1. $\bigcap_{j\in Z} V_j = \{0\}, \bigcup_{j\in Z} V_j$ - компактно в $L^2(R)$.
2. $V_j \subset V_{j-1}$.
3. $x(t) \in V_j \Leftrightarrow x(2^j t) \in V_0$.
4. Существует функция $\phi_0(t)$ в V_0, называемая масштабирующей функцией такая, что последовательность $\{\phi_0(t-k), k\in Z\}$ образует базис Рисса в V_0.

Аналогично, функции сдвига и масштаба $\{\phi_{j,k}(t)=2^{-j/2}\phi_0(2^{-j} t-k), k\in Z\}$ устанавливают базис Рисса для пространства V_j.

Применение КМА сигнала x означает нахождение его проекций в каждом из аппроксимационных подпространств V_j

$$approx_j(t) = (\Pr oj_{V_j} x)(t) = \sum_k a_x(j,k)\phi_{j,k}(t).$$

Поскольку $V_j \in V_{j-1}$, $approx_j$ есть более грубая аппроксимация x, чем $approx_{j-1}$, то ключевой идеей КМА является рассмотрение потери информации, то есть детали, при переходе из одной аппроксимации к другой, более грубой аппроксимации $detail_j(t)= approx_{j-1}(t) - approx_j(t)$.

Кратномасштабный анализ показывает что сигналы $detail_j$ могут быть прямо получены из проекций x на последовательность подпространств W_j, называемых подпространством вейвлет. Кроме того, КМА- теория показывает, что существует функция ψ_0, называемая материнской вейвлетой, которая должна быть произведена из ϕ_0, так что ее образы

$$\{\psi_{j,k}(t)=2^{-j/2}\psi_0(2^{-j} t-k), k\in Z\}$$

устанавливают базис Рисса для W_j

$$\det ail_j(t) = (\Pr oj_{W_j} x)(t) = \sum_k d_x(j,k)\psi_j.$$

По-существу, КМА состоит в переписывании информации x в виде последовательности деталей различных разрешений и аппроксимации с низким разрешением

$$x(t) = appox_j(t) + \sum_{j=1}^{j=J} \det ail_j(t) = \sum_k a_x(J,k)\phi_{j,k}(t) + \sum_{j=1}^{J} d_x(j,k)\psi_{j,k}(t).$$

$Approx_j$ является существенно грубой, а грубая аппроксимация x означает, что ϕ_0 необходимо должна быть функцией нижних частот. $Detail_j$, будучи информационным "дифференциалом", указывают, что ψ_0 скорее есть функция полосы пропускания и, таким образом, маленькой волной, волночкой (wavelet' ой).

Более того, КМА показывает, что материнская вейвлета должна удовлетворять $\int \psi_0(t)dt = 0$ и что ее Фурье- преобразование подчиняется

$$|\psi_0(\nu)| \sim \nu^N, \nu\to 0,$$

где N- положительное целое, называемое числом исчезающих моментов вейвлеты. При заданной масштабирующей функции ϕ_0 и материнской вейвлете ψ_0 дискретное вейвлетное преобразование есть отображение $L^2(R) \rightarrow l^2(Z)$, задаваемое

$$x(t) \rightarrow \{\{a_x(J,k),\ k\in Z\},\ \{d_x(j,k),\ j=1,...,J,\ k\in Z\}\}.$$

Реконструкция производится используя функцию $\tilde{\psi}$, называемую синтезированной вейвлетой такую, что два семейства $\{\psi_{l,n}\}_{(j,n)\in Z^2}$ и $\{\tilde{\psi}_{j,n}\}_{(j,n)\in Z^2}$ являются биортогональным базисом Рисса $L^2(R)$, где

$$\psi_{j,n}(t) \stackrel{\Delta}{=} \frac{1}{\sqrt{2^j}}\psi\left(\frac{t-2^j n}{2^j}\right)$$

и подобна для $\tilde{\psi}_{j,n}$. Синтезированная вейвлета $\tilde{\psi}$ получается из функции $\tilde{\phi}$, которая является двойственной (дуальной) к ϕ, то есть которая удовлетворяет $\left\langle \phi(t), \tilde{\phi}(t-n) \right\rangle = \delta(n)$, где $\langle . \rangle$ - стандартное внутреннее произведение в $L^2(R)$, а $\delta(\cdot)$-функция Дирака. Масштабирующие функции ϕ и $\tilde{\phi}$ должны удовлетворять

$$\phi(t) = \sqrt{2}\sum_n h(n)\phi(2t-n)$$

и

$$\tilde{\phi}(t) = \sqrt{2}\sum_n \tilde{h}(n)\tilde{\phi}(2t-n),$$

где h и $\tilde{h}$ - дискретные фильтры, удовлетворяющие условию биортогональности в $l^2(Z)$

$$\sum_k h(k)\tilde{h}(k-2n) = \sum_k \tilde{h}(k)h(k-2n) = \delta(n).$$

ψ и $\tilde{\psi}$ задаются

$$\psi(t) = \sqrt{2}\sum_n g(n)\phi(2t-n),$$

$$\tilde{\psi}(t) = \sqrt{2}\sum_n \tilde{g}(n)\tilde{\phi}(2t-n),$$

где

$$\begin{cases} g(n) = (-1)^{1-n}\tilde{h}(1-n) \\ \tilde{g}(n) = (-1)^{1=n}h(1-n). \end{cases}$$

Дискретные фильтры h, g, $\tilde{h}$ и $\tilde{g}$ должны удовлетворять условию реконструкции, которое может быть найдено в [2].

Быстрое вейвлетное преобразование вычисляет вейвлетные коэффициенты дискретного сигнала $a_J(\cdot)$. Алгоритм быстрой вейвлетной декомпозиции есть

$$\begin{cases} a_j(n) = \sum_p h(p-2n)a_{j+1}(p) \\ d_j(n) = \sum_p g(p-2n)a_{j+1}(p). \end{cases}$$

Реконструкционный алгоритм имеет вид

$$a_{j+1}(n) = \sum_p \tilde{h}(n-2p)a_j(p) + \sum_p \tilde{g}(n-2p)d_j(p).$$

Коэффициенты $a_j(n)$ и $d_j(n)$ называются соответственно масштабирующими и детальными коэффициентами при j-ой шкале и n- ом сдвиге.

Практически они могут быть вычислены быстрым рекурсивным пирамидальным алгоритмом, основанном на банке фильтров, чья вычислительная стоимость экстремально низкая.

Оценка дисперсии деталей. При анализе феномена LRD следующие две характеристики (A1, A2) играют ключевые роли.

A1: Базис строится из оператора дилатации (изменения шкалы):

$\psi_{j,0}(t) = 2^{-j/2}\psi_0(2^{-j}t)$. Это означает, что анализирующее семейство проявляет масштабно- инвариантное свойство. Феномен LRD может быть понят как отсутствие любой характеристической частоты (и, следовательно, шкалы) в диапазоне частот, близких к нулю. Таким образом, свойство LRD может быть интерпретировано как характеристика, инвариантная к масштабу, которая эффективно анализируется вейвлетами.

A2: ψ_0 имеет число нулевых, или исчезающих моментов, которое может быть произвольно выбрано при условии $N \geq 1$. По определению это означает, что $\int t^k \psi_0(t)dt \equiv 0$, k=0,..., N-1 (но не для k $\geq$ N), или равносильно, Фурье- преобразование ψ_0 удовлетворяет $|\Psi_0(\nu)|=O(|\nu|^N)$ в нуле. Это свойство может быть использовано, что- бы управлять отклонениями, появляющимися в процессах, имею- щих спектр по степенному закону в нуле.

Для процессов , таких как LRD- процесс, эти характеристики порождают следующие ключевые свойства вейвлетных коэффициентов $d_x(j, k)$ на диапазоне шкал 2^j, $j=j_1...j_2$, где имеет силу масштабирование по степенному закону [3].

B1: Благодаря A1, масштабная инвариантность (поведение степенного закона) точно захватывается

$$Ed_x(j, \cdot)^2 = 2^{j\alpha} c_f C, \qquad (6.4)$$

где

$$C = \int |\nu|^{-\alpha} |\Psi_0(\nu)|^2 \, d\nu. \qquad (6.5)$$

Это точное выделение степенного закона получается прямо из оператора дилатации, лежащего в основе проектирования вейвлетного базиса. (Время- частотные или периодограммные оценки не показывают таких характеристик).

B2: Благодаря A1 и A2, $d_x(j, k)$ есть последовательность случайных переменных, которые квазидекоррелированы. В частности, LRD,

присутствующая в области временного представления, полностью отсутствует в плоскости вейвлетных коэффициентов {j, k}.

Свойство B2 заслуживает тщательной разработки. Было показано [3], что корреляции в плоскости время- масштаб спадают, по крайней мере, гиперболически во всех направлениях с экспонентами, управляемыми числом исчезающих моментов и соответствующими LRD. Поскольку, по определению, октава $j=\log_2$ (шкала), это означает наличие экспоненциального спада в октаве j.

Начальной точкой для анализа является выражение (4), которое может быть записано как

$$\log_2(Ed_x(j,\cdot)^2) = j\alpha + \log_2(c_f C)$$

и которое обеспечивает линейный регрессионный подход для оценивания (α, c_f), где наклон регрессии должен оценить α, а пересечение должно быть отнесено к c_f. Эта идея использования графика log- log является очевидной и общей для многих случаев, когда экспонента является объектом изучения.

Однако, на пути ее реализации должны быть взяты в расчет неминуемые усложнения.

Первым существенным усложнением является, конечно, что $Ed_x(j,\cdot)^2$- величина второго порядка. которая может быть отнесена к неизвестному спектру x, который должен быть оценен. В настоящем контексте это принципиальная трудность, так как хорошо известно, что оценка величин второго порядка (или других) в присутствии LRD- задача деликатная. Однако, свойство B2, квзидекорреляции $d_x(j, k)$ позволяет эффективно использовать простое "временное усреднение"

$$\mu_j = \frac{1}{n_j}\sum_{k=1}^{n_j} d_x^2(j,k), \qquad (6.6)$$

где n_j- число коэффициентов, доступных для анализа, при октаве j. Эта величина является несмещенной и эффективной оценкой $Ed_x(j,\cdot)^2$.

Вторым усложнением является нелинейность, вводимая $\log_2$, которая смещает оценку. Эту проблему также можно будет обойти при разумных гипотезах.

Слегка упрощая предмет, можно утверждать, что фундаментальный подход, лежащий в основе оценки, есть линейная регрессия $\log_2(\mu)$ на $\log_2(2^j)=j$. Взвешенная линейная регрессия будет использоваться как дисперсия $\log_2(\mu)$, которая меняется с j.

Анализ с использованием диаграммы с логарифмической шкалой. При рассмотрении линейной регрессии вида $Ey_j=b_j+a$ в качестве переменной y_j можно рассматривать $\log2(\mu_j)$. Такое допущение не является точно справедливым, так как $E\log_2(\mu_j) \neq \log_2(E\mu) = j\alpha + \log_2(c_f C)$. Поэтому допускаются следующие дополнительные идеализации.

C1: Процесс x, и следовательно, $d_x(j,\cdot)$ - гауссовский.

C2: Для фиксированного j процесс $d_x(j,\cdot)$ независимый и идентично распределенный.

C3: Процессы $d_x(j,\cdot)$ и $d_x(j',\cdot)$, $j \neq j'$ - независимые.

Идеализация C1 подтверждается численным доказательствам, которое показывает, что метод очень нечувствителен к форме маргинальных распределений x. Идеализация C2 и C3 хорошо подтверждаются свойством B2. Эти дополнительные условия, хотя и кажутся очень ограничительными на первый взгляд, являются очень разумными на практике, как подтверждено в моделированиях.

В [1] было показано, что при предположении гауссовости и квазидекорреляции вейвлетных коэффициентов, выражение для дисперсии $\log2(\mu_j)$ в асимптотическом пределе имеет вид

$$Var(\log_2(\mu_j)) \approx \frac{2}{n_j \ln^2 2}, \tag{6.7}$$

где $n_j \approx n2^{-j}$.

Это выражение важно при выборе доверительных интервалов при численом моделировании.

Учитывая вышесказанное, можно показать, что

$$Ey_j = j\alpha + \log_2 c_f C,$$

$$Var(y_j) = \varsigma(2, n_j/2)\ln^2 2,$$

и, таким образом, удовлетворяются требования для взвешенной линейной регрессии. Затем производится оценка $(\hat{b}, \hat{\delta})$ взвешенной линейной регресси y_j по $j=x_j$ с $\sigma_j^2 = Var(y_j)$.

Оценивание параметров. Объединенная несмещенная оценка $(\hat{\alpha}, \hat{c_f C})$ задана $\hat{\alpha} = \hat{b}$, $\hat{c_f C} = p \cdot 2^{\hat{a}}$, где p- коэффициент, корректирующий смещение. Выражение для p может быть найдено в [3]. Оценка $\hat{C}$ интеграла может быть определена как

$$C(\alpha, \psi_0) = \int |v|^{-\alpha} |\psi_0(v)|^2 dv,$$

когда

$$\begin{cases} C(0,\psi_o), \hat{\alpha} \le 0 \\ C(\hat{\alpha},\psi_0), 0 < \hat{\alpha} < 1 \\ C(1,\psi_0), \hat{\alpha} \ge 1. \end{cases}$$

Таким образом, можно определить $(\hat{\alpha}, \hat{c}_f)$ как

$$\hat{\alpha} = \hat{b}, \quad \hat{c}_f = \hat{c_f C} / \hat{C}.$$

В заключении необходимо отметить, что если параметр α важен, так как он определяет существование самого феномена и управляет поведением масштабной характеристики, то параметр c_f имеет силу в любом LRD-контексте, где он играет важную роль в проблеме оценки среднего. Для процессов с LRD классическое асимптотическое выражение σ_x^2 / n для дисперсии выборочного среднего с размером выборки n замещается $(2c, n^{\alpha}/(1+\alpha)\alpha) \cdot (1/n)$. Так как дисперсия пропорциональна c_f, следует

немедленно, что доверительные интервалы по оценкам выборочного среднего, по- существу, пропорциональны $\sqrt{c_f}$. Следовательно, величина c_f - критическая даже для вопросов простейшего практического оценивания в случаях процессов с LRD. Кроме того, в такой области применения, как телекоммуникации, этот параметр играет важную роль, поскольку, как показано в ряде исследований, его увеличение увеличивает задержку очереди, что, в свою очередь, подчеркивает соответствие α как меры размера влияний LRD.

Литература

1.**Abry P., Veith D.** Wavelet Analysis of Long- Range- Dependent Traffic// IEEE Transactions on Information Theory, 1998, v.44, ¹1, P. 2- 15.

2. **Новиков И.Я., Стечкин С.Б.** Основы теории всплесков// Успехи математических наук, 1998, т. 53, вып.6 (324), С.53-128.

3. **Veith D., Abry P. A** Wavelet- Based Joint Estimation of Parameters of Long Range Dependent// IEEE Transactions on Information Theory, 1999, v.45, ¹3, P. 878-897.

4. **Roughan M., Veith D., Abry P.** Real- Time Estimation of the Parameters of Lonf-Range Dependence// IEEE/ACM Transaction on Networking, 2000, v.8, ¹4, P.467-477.

Приложение

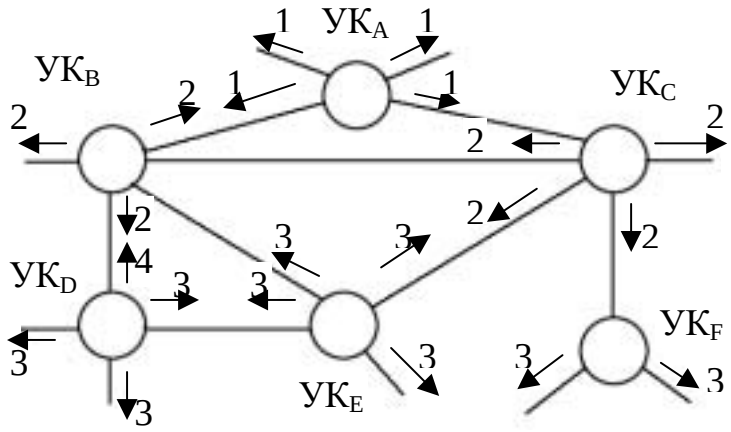
Метод рельефов

Метод рельефов (МР) относится к групповым распределенным методам. Критерием выбора пути в этом случае является минимальная длина пути, выраженная числом транзитных участков. Допустимо использование других критериев (времени установления связи).

На сети связи, где реализуется МР (рис.П.1), выполняются следующие операции: формирование рельефа, коррекция рельефа. Первая выполняется при начальном запуске сети, а также в процессе развития сети (роста числа УК сети). Вторая – периодически в процессе функционирования сети или в момент возникновения повреждений либо перегрузок на сети.

В момент пуска сети формирование ее рельефа начинается с некоторого узла $УК_\alpha$ (α = 1, 2, ..., N), являющегося инициатором, т.е. начинается формирование α-рельефа. В ЗУ каждого $УК_i$ отводится некоторый объем памяти $N \times M_i$ (где M_i – число исходящих направлений из $УК_i$), в который будет записываться матрица рельефов $\mathbf{R}_i$.

При формировании рельефа из УК инициатора по всем инцидентным направлениям передается цифра “1”. Эта единица на инцидентных узлах заносится в матрицу $\mathbf{R}_i$ по координатам (n, m^1), где n – номер УК инициатора административного сеанса, m^1 – номер ветви, по которой



поступила “1”.

Рис.П.1. Реализация метода рельефов на сети связи

Рассмотрим в качестве примера сеть следующей конфигурации. Пусть узел-инициатор – $УК_A$. В этом случае “1” будет записана в $\mathbf{R}_B$ и $\mathbf{R}_C$. Все УК, получившие “1”, передают по всем исходящим направлениям (за исключением того направления, по которому получено от $УК_A$ – “1”) цифру

“2”. Эта цифра во всех УК, в которые она поступила, заносится в матрицу **R** по координатам (n, m^2). Для нашего примера цифра “2” будет записана в матрицы $\mathbf{R}_B$, $\mathbf{R}_C$, $\mathbf{R}_E$, $\mathbf{R}_F$ и т.д.

При этом должны соблюдаться следующие правила.

1. Если в УК поступили одинаковые цифры с двух и более направлений, данный УК инициирует передачу цифры на единицу больше по всем без исключения исходящим направлениям. Например, В $УК_E$ цифра “2” поступает от $УК_B$ и $УК_C$. В этом случае $УК_E$ передает “3” по всем исходящим направлениям.

2. Если в УК поступает цифра с одного направления, на данном УК происходит инициация для передачи цифры на единицу больше той, которая поступила по всем направлениям, за исключением того направления, по которому была получена эта цифра. Передача цифры по этому направлению возможна лишь при поступлении в данный УК следующей цифры. Передаваемая по этому направлению цифра должна быть на единицу больше поступившей второй по порядку цифры. Например, в $УК_D$ цифра “2” поступает по одному направлению от $УК_B$. Тогда цифра “3” с $УК_D$ должна передаваться по всем направлениям, за исключением направления к $УК_B$ – по нему будет передана цифра “4”, т.к. следующая поступившая по порядку цифра – “3”.

3. Инициация передачи цифр по всем направлениям на каждом УК происходит один раз после поступления первой цифры по порядку.

Таким образом, мы сравнивали α-рельеф. Аналогичным образом строятся рельефы для всех остальных узлов сети. Считается, что рельеф сети сформирован, если построены все α-рельефы ($\alpha = 1, 2, ..., N$). Поиск оптимального пути установления соединения от $УК_i$ к $УК_j$ состоит в отыскании на $УК_i$ и каждом промежуточном УК ветви, которой соответствует минимальное число в строке матрицы рельефов для $УК_j$.

Пусть требуется установить соединение от $УК_D$ к $УК_A$. В этом случае на $УК_D$ производится обращение к строке матрицы рельефов $\mathbf{R}_D$, соответствующей $УК_A$. При этом соединение устанавливается по ветви, которой соответствует минимальное число в этой строке.