Министерство общего и профессионального образования РФ
Санкт - Петербургский государственный институт точной механики и оптики
(технический университет)

А. Ю. ЩЕГЛОВ

МЕТОДЫ ДИСПЕТЧЕРИЗАЦИИ ЗАЯВОК НА ОБСЛУЖИВАНИЕ В ЛВС РЕАЛЬНОГО ВРЕМЕНИ

Учебное пособие

Санкт-Петербург 1997

Кафедра ВТ ИТМО (ТУ), 197101, Санкт-Петербург, Саблинская ул., 14

УДК 681.324

Щеглов А. Ю. Методы диспетчеризации заявок на обслуживание в ЛВС реального времени. Учебное пособие. - СПб. : ИТМО, 1997. - 77 с.

Рассматриваются проблемы диспетчеризации заявок на обслуживание в управляющих и информационных локальных вычислительных сетях (ЛВС) реального времени, в основе параллельной обработки которых находится принцип распараллеливания по функциям; ЛВС оперативной обработки и комбинированного обслуживания, в части синтеза дисциплин обслуживания и методов децентрализованного управления множественным доступом к общим ресурсам при многоточечном подключении абонентов к каналу связи. Иллюстрируется общность требований к обслуживанию заявок в реальном времени для современных управляющих и информационных ЛВС, для ЛВС комбинированного обслуживания и с интеграцией служб связи, рассматриваемых в рамках сетевой технологии АТМ.

Излагается современная концепция обслуживания с динамическими относительными приоритетами, изменяемыми по расписанию при каждом занятии общего связного ресурса, при кодовом управлении множественным доступом к ресурсам. Рассматриваются и исследуются принципы обслуживания по расписаниям - в реальном времени, с относительными, со смешанными и с многоуровневыми приоритетами, учитывающими в одной системе приоритеты заявок и абонентов системы в реальном времени и в режиме оперативной обработки данных. Излагаются принципы комбинирования и унификации альтернативных дисциплин обслуживания в рамках рассматриваемого подхода.

Рассматриваются вопросы повышения эффективности кодового управления множественным доступом, за счет оптимального кодирования относительных приоритетов. Вводится понятие и рассматриваются методы динамического кодирования относительных приоритетов, изменяемых по расписанию.

Рассматриваются альтернативные подходы к реализации передачи в системе прав на занятие ресурса при кодовом управлении доступом, излагаются принципы комбинирования и унификации альтернативных методов.

Рецензенты: докт.техн.наук, профессор Т.И.Алиев
докт.техн.наук, профессор А.Ю.Тропченко

Одобрено на заседании кафедры вычислительной техники 11 марта 1997 г. , протокол № 2.

ВВЕДЕНИЕ

До недавнего времени опережающее развитие в вопросах создания и использования распределенных вычислительных систем (ВС) имели сосредоточенные (или мультипроцессорные) ВС, отличающиеся высокой эффективностью межмодульных взаимодействий по параллельной системной шине, либо высокоскоростным коммутируемым каналам (транспьютерные системы). Эффективность функционирования сосредоточенных систем, в основе построения которых, как правило, находится принцип однородности структуры, определяется эффективностью языков параллельного программирования. Принципиально иные особенности имеют распределенные ВС, отличающиеся использованием последовательных каналов связи, подверженных воздействию помех и, как следствие, низкой эффективностью межмодульных взаимодействий. Такие системы называются вычислительными сетями, особое место среди которых занимают локальные вычислительные сети (ЛВС), отличающиеся небольшой территорией охвата и, как следствие, применением ориентированных на такую территорию средств и методов передачи данных. Низкая эффективность межмодульных взаимодействий в ЛВС накладывает ограничения на способы построения распределенных коллективов вычислителей. Здесь эффективен уже способ распараллеливания задач по функциям, где каждый вычислитель системы самостоятельно решает свою часть (функцию) общей задачи ВС, чем минимизируется число межмодульных взаимодействий при решении задачи системой. В качестве условия синтеза алгоритма параллельной обработки здесь рассматривается минимизация числа межмодульных взаимодействий абонентов системы.

Особое место среди ЛВС занимают проблемно-ориентированные ЛВС (ЛВСПО), в частности ЛВС, используемые в задачах управления (УВС) - в системах автоматического управления (САУ) и на нижних уровнях автоматизированных систем управления (АСУ). В соответствии с требованиями к работе ЛВС в реальном масштабе времени протекания управляемого процесса, функционирование ВС здесь характеризуется жесткими ограничениями, накладываемыми на время реакции системы на входные воздействия, значения которых могут быть достаточно малы (могут составлять миллисекунды и их доли). При возможности существенного увеличения мощности отдельных вычислителей ЛВС, реализующих отдельные функции общей задачи, решаемой ЛВС, например путем объединения их в сосредоточенные ВС (магистрально-модульные УВС [7]), остается ограниченной эффективность межмодульных взаимодействий ЛВС, что вызвано требованиями к высокой помехозащищенности передачи данных между вычислителями (в современных стандартах на интерфейсы распределенных УВС физическая скорость передачи данных ограничивается

единицами Мбит/с [7]). Отмеченные особенности ЛВС, с учетом того, что система может содержать десятки и сотни вычислителей, требующих высокой эффективности межмодульных взаимодействий, обусловливает, что связной ресурс является «узким местом» ЛВС, определяющим их производительность в целом, что требует эффективного решения задачи распределения прав на занятие ресурса между абонентами ЛВС, соответственно при аппаратурной реализации.

Практически аналогичные требования сегодня выдвигаются и к функционированию современных ЛВС общего назначения (ЛВСОН). Современные сетевые технологии, прежде всего АТМ (АТМ - Asynchronous Transfer Mode) [3] предполагают реализацию сетей связи, в том числе и ЛВС, с интеграцией служб связи (ЛВСИС). При этом в одной сети передаются сигналы реального времени (например, при передаче речи и подвижных изображений) и оперативной обработки (передача данных и текста) , причем сигналы реального времени для различных служб, прежде всего - передача речи и подвижных изображений, имеют существенно различающиеся требования ко времени обработки. Существенным отличием реализации ЛВС в этой концепции будет: требование к функционированию в реальном масштабе времени, обмен фиксированными короткими (53 байтными) пакетами данных (называемыми ячейками), эффективное аппаратурное решение задачи управления доступом абонентов к ресурсам ЛВС [3].

Качество использования общего ресурса в значительной степени определяется методом диспетчеризации [1] - в системах с распараллеливанием по функциям - дисциплиной обслуживания или выбора заявки на использование ресурса от одного из абонентов системы (дисциплина занесения заявки в очередь в ЛВС определяется месторасположением вычислителя с учетом функционального распараллеливания обработки в ЛВСПО, соответственно источника нагрузки в ЛВСОН).

С учетом требований к высокой надежности и модульности построения к реализации ЛВС сегодня, как правило, выдвигается требование распределенного или децентрализованного управления занятием ресурса, причем в большинстве приложений как ЛВСПО, так и ЛВСОН предполагается, что абоненты системы, равно как и общие ресурсы, подключаются к единому для системы каналу связи магистральной топологии. Тогда задача управления доступом абонентов к ресурсу должна быть в равной мере (симметрично) распределена между всеми абонентами ЛВС, т. е. должна решаться устройством децентрализованного управления, задача проектирования которого по существу в ЛВС сводится к проектированию дисциплины обслуживания и способа эффективной ее распределенной аппаратурной реализации.

В учебном пособии рассматриваются особенности и методы построения распределенных ВС, прежде всего в части диспетчеризации заявок на обслуживание методами децентрализованного управления

доступом абонентов к общим ресурсам системы, излагается наиболее перспективная на сегодняшний день концепция построения ЛВС реального времени, позволяющая отказаться от опроса очередей при обслуживании заявок по расписаниям, что существенно улучшает характеристики межмодульных взаимодействий в ЛВС, обсуждаются важнейшие вопросы проектирования и унификации алгоритмических и технических средств распределенных ВС реального времени.

При изучении материала необходимо знание теоретических основ и математических методов теории вычислительных систем и сетей в объеме монографии [1], желательно знакомство с принципами управления множественным доступом к каналу связи ЛВС, изложенными, например в [8, 10, 11], и с методами оптимального кодирования [4]. В своих материалах автор полностью следовал терминологии, введенной и использованной в [1, 2].

Учебное пособие содержит четыре раздела. В первом разделе рассматриваются классификация, принципы построения и модель ЛВС реального времени, формулируются задачи эффективного управления множественным доступом к ресурсам системы в реальном масштабе времени. Рассматриваются цели и критерии оптимальности приоритетного обслуживания заявок в ЛВС. Второй раздел посвящен методам диспетчеризации, применяемым в ЛВС, в частности эффективно распределяющим права на занятие ресурса по расписаниям. Исследуются цели и возможности обслуживания заявок на основе смешанных приоритетов. Раздел содержит описание концепции кодового управления доступом к ресурсам как эффективного унифицированного для ЛВС подхода к обслуживанию заявок в реальном времени и со смешанными приоритетами. Исследуется модель системы кодового управления. Вводится понятие и рассматриваются свойства канонического расписания. Третий раздел посвящен методам кодирования приоритетов абонентов ЛВС. Рассматриваются задачи и возможности повышения эффективности кодового управления доступом к каналу за счет оптимального кодирования приоритетов абонентов. Излагаются принципы динамического кодирования приоритетов абонентов, реализуемого в системе для передачи прав на занятие ресурса по расписанию. В четвертом разделе рассматриваются альтернативные механизмы кодового управления множественным доступом и исследуются возможности их унификации в рамках единой концепции кодового управления, обсуждаются пути повышения эффективности кодового управления, за счет эффективной передачи прав между абонентами на доступ к ресурсу.

Ограниченный объем пособия не позволяет изложить многие вопросы достаточно полно. Более подробно с ними можно ознакомиться с помощью дополнительной литературы [12 - 24].

РАЗДЕЛ 1. ПРИНЦИПЫ ПОСТРОЕНИЯ И МОДЕЛЬ ЛВС РЕАЛЬНОГО ВРЕМЕНИ

1.1. Классификация ЛВС. Основные понятия

В соответствии с общей классификацией ВС [2], рассматривая ЛВС как вычислительную сеть, представляющую собою совокупность ВС, объединенных при помощи средств связи, включающих каналы связи и устройства сопряжения ВС с каналами связи, расположенную на небольшой территории, и использующую ориентированные на эту территорию средства и методы передачи данных; получаем классификацию ЛВС, представленную на рис. 1.1. Все многообразие ЛВС можно разделить на две большие группы - проблемно ориентированные (ЛВСПО) и оперативной обработки (ЛВСОН), в основу такого деления положено наличие (отсутствие) распараллеливания решаемой ЛВС задачи. Под **ЛВСПО** будем понимать ЛВС, ориентированные на решение ими в процессе всего времени работы ограниченного числа задач, связанных между собою по данным или требующих взаимодействий при их параллельной обработке, что позволяет синтезировать алгоритм функционирования ВС при проектировании ЛВС. Под **ЛВСОН** будем понимать ЛВС, круг решаемых задач которых не может быть ограничен, либо решаемые ЛВС задачи независимы, т. е. не связаны между собою по данным, что не позволяет синтезировать алгоритм функционирования ВС при проектировании ЛВС.

По виду **заявки на обслуживание**, под которой здесь будем понимать причину инициирования передачи данных по каналу связи ЛВС, будем делить ЛВС на ЛВС реального времени (ЛВСРВ) и ЛВС оперативной обработки (ЛВСОО). Под **ЛВСРВ** будем понимать ЛВС, функционирующие в реальном масштабе времени, характеризуемом нарушением законов функционирования системы или потерей заявок на обслуживание при превышении предельных ограничений на время реакции системы на входное воздействие. Под **временем реакции системы** понимается задержка во времени между моментом поступления в систему данных и моментом окончания их обработки системой. Под **ЛВСОО** будем понимать ЛВС, функционирование которых характеризуется снижением качества обслуживания без нарушения законов функционирования системы или отсутствием потерь заявок на обслуживание при превышении предельных ограничений на время реакции системы на входное воздействие. ЛВС, сочетающие в себе принципы обслуживания в реальном времени и оперативной обработки будем называть ЛВС комбинированного обслуживания (**ЛВСКО**).

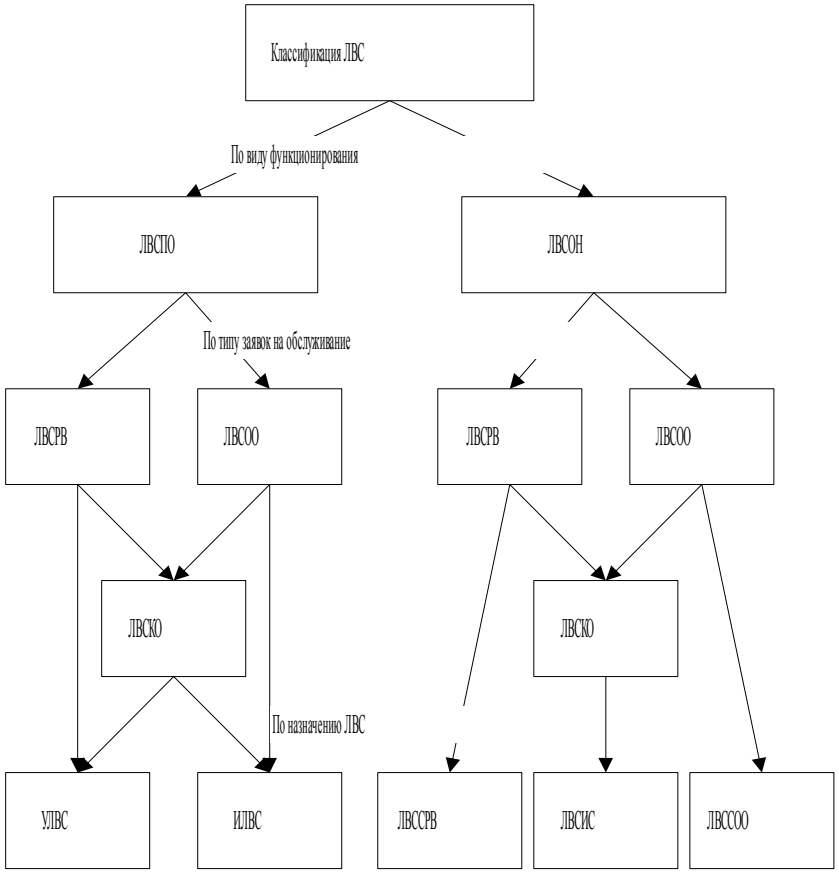


Рис. 1.1

По виду решаемых ЛВСПО задач (назначение), будем их делить на управляющие и информационные ЛВС. Под **управляющими ЛВС** (УЛВС) будем понимать ЛВС, работающие совместно с объектом управления, непрерывно функционирующим во времени. Соответственно, по наличию (отсутствию) оператора, принимающего решение в контуре управления, УЛВС могут функционировать в составе АСУ или САУ. Под **информационной ЛВС** (ИЛВС) будем понимать ЛВС, предназначенную для сбора, обработки, передачи, хранения и отображения информации, представленной базами и банками данных.

ЛВСОН будем подразделять по виду реализуемых ими служб связи. Под **службой связи** понимается предоставляемые ЛВС услуги, связанные с обработкой и передачей конкретных видов информации, представляемой в

канале связи соответствующими физическими сигналами. Примерами **служб связи реального масштаба времени**, под которыми понимаются службы, характеризуемые реальным масштабом времени поступления и передачи по сети связи сигналов, при невозможности их промежуточной буферизации без искажения (потери) информации, является телефонная связь (служба передачи речи), передача подвижных изображений, радиосигналов и т.д. Примерами **служб связи оперативной обработки**, отличающихся возможностью промежуточного хранения передаваемой информации с целью эффективного использования каналов связи ЛВС, являются службы передачи данных, текстов, неподвижных изображений и т. д. С учетом сказанного, ЛВСОН будем делить на ЛВС, реализующие службы связи реального масштаба времени (**ЛВССРВ**) и службы связи оперативной обработки (**ЛВССОО**). ЛВСКО в данных приложения называются ЛВС с интеграцией служб (**ЛВСИС**), сочетающие службы связи реального масштаба времени и оперативной обработки.

Т.к. под ЛВС понимается распределенная ВС, «узким местом» которой, ограничивающим производительность ВС в целом, является общий ресурс, в первую очередь, связной, задача синтеза ЛВС, по существу, сводится к задаче эффективной реализации доступа абонентов системы к общим ресурсам. При этом важнейшей проблемой проектирования ЛВС является синтез **дисциплины обслуживания** (в данных приложениях ВС - ЛВС, диспетчеризация заявок на обслуживание сводится к выбору заявок из очередей) - правила (алгоритма), на основе которого заявки выбираются из очередей при их множественном доступе к коллективно используемым ресурсам, прежде всего к общему связному ресурсу - каналу связи ЛВС. **Множественный доступ** - процедура получения доступа к коллективно используемому ресурсу. Под **управлением множественным доступом** (арбитражом заявок на обслуживание) понимается бесконфликтный, реализуемый либо с предотвращением возможного конфликта, либо с его разрешением, доступ абонентов к коллективно используемому ресурсу в соответствии с реализуемой в ЛВС дисциплиной обслуживания требований ресурса. **Конфликт** (столкновение) ситуация, при которой через один канал связи передают данные одновременно несколько абонентов, приводящая к искажению информации в канале связи. **Децентрализованное** управление множественным доступом характеризуется тем, что задача арбитража (решаемая в рассматриваемых приложениях аппаратурно) в равной мере (симметрично) распределена между абонентами системы. Децентрализованное управление множественным доступом реализуется за счет передачи между абонентами **полномочий** (прав) - разрешающих сигналов, позволяющих абонентам доступ к коллективно используемым ресурсам.

Основное внимание в монографии уделяется исследованию принципов диспетчеризации заявок на обслуживание в ЛВС реального

времени, т.е., следуя введенной классификации ЛВС, применяемых для различных приложений ЛВСРВ и ЛВСКО.

1.2. Общие принципы построения. Структура системы

В основе реализации современных распределенных ВС находятся следующие общие принципы построения:

- параллельность обработки, позволяющая создавать мощные коллективы вычислителей при различных принципах распараллеливания задач;

- модульность, обусловливающая высокую надежность и живучесть ВС, простоту резервирования блоков и наращивания вычислительной мощности;

- распределенность (или децентрализованность) управления общими ресурсами, обеспечивающая высокую надежность систем параллельной обработки данных;

- решение задачи управления распределением прав между абонентами на доступ к «узкому месту» ВС - общему ресурсу, аппаратурными средствами, что обеспечивает высокую эффективность его использования в системе.

Структура распределенной системы с общими ресурсами, в основе построения которой лежат перечисленные принципы, представлена на рис. 1.2 - в качестве связного ресурса рассмотрим применение последовательного канала связи магистральной топологии – «общая шина».

Под **шинной (магистральной) ЛВС** понимается ЛВС, в которой всегда имеется лишь один маршрут между любыми двумя абонентами, и данные, выдаваемые в канал любым абонентом ЛВС, доступны всем абонентам, включая и источник информации.

Подобная структура отличается следующим:

- в системе отсутствуют какие-либо узлы коммутации, что обеспечивает «многоточечное» подключение абонентов к каналу связи [7] и, как следствие, максимальную надежность и модульность построения, а за счет одновременной обработки абонентами информации, поступающей из канала, и максимальную эффективность межмодульных взаимодействий;

- по этой же причине в системе наиболее сложно решается задача управления доступом абонентов к общим ресурсам - задача управления доступом к ресурсам здесь формулируется и решается в общем виде.

С учетом сказанного, на сегодняшний день это наиболее перспективная структура распределенных ВС, используемая, как правило, в системах реального времени, а нас интересующая наиболее общей постановкой исследуемой задачи управления множественным доступом к общим ресурсам. Поэтому приведенная структура ВС и будет нами

рассматриваться далее (отметим, что позволяя решить исследуемую задачу в общем случае - при магистральном канале связи, излагаемые методы при незначительных модификациях могут применяться и для частных, с точки зрения управления доступом к ресурсам, приложений, например при кольцевой топологии ЛВС).

Замечание. На рис. 1.2 кроме связного приведены другие ресурсы ВС, которые присоединяются к каналу связи также по правилу многоточечного подключения. Поэтому взаимодействие абонента с любым ресурсом осуществляется по единому каналу связи, следовательно задача управления доступом к ресурсам системы здесь опять же сводится к задаче управления доступом к связному ресурсу.

Задача диспетчеризации здесь решается **арбитром**, под которым понимается устройство, реализующее в ЛВС децентрализованное управление множественным доступом абонентов к общим ресурсам в соответствии с требуемой дисциплиной обслуживания.

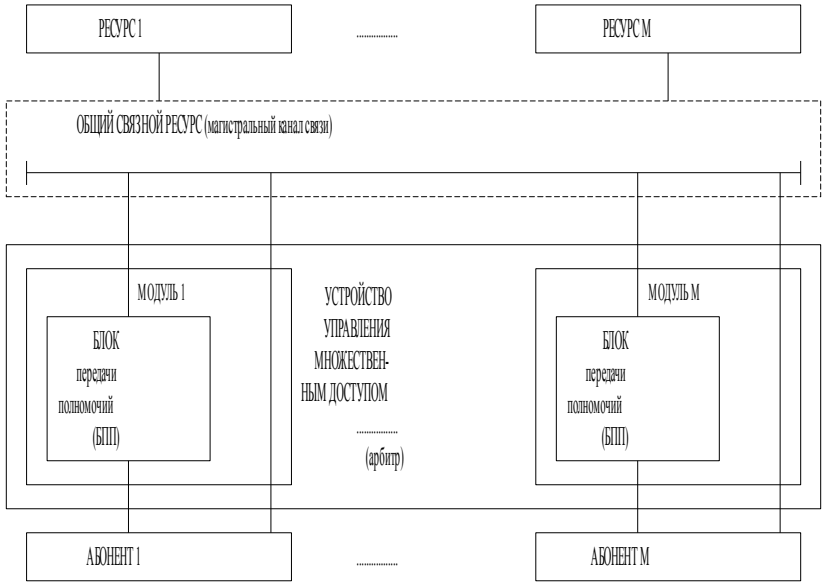


Рис. 1.2

1.3. Задачи и методы управления множественным доступом к общим ресурсам

В задачи управления множественным доступом (при аппаратурном ее решении - в задачи распределенного арбитра, см. рис. 1.2) входит:

- диспетчеризация заявок на обслуживание - в распределенных ВС при функциональном распараллеливании, при котором каждый абонент территориально приближен к источнику нагрузки и решает задачи исключительно по обработке поступающих от него заявок, сводится к задаче выбора заявок из очередей или к реализации дисциплины обслуживания (ДО);

- передача прав между абонентами системы на доступ к ресурсу в соответствии с задаваемой дисциплиной обслуживания очередностью.

Классификация наиболее широко используемых сегодня в ЛВС методов диспетчеризации [1] представлена на рис. 1.3, реализующих их методов управления множественным доступом на рис. 1.4 [8, 10]. Среди ДО можно выделить бесприоритетные, реализующие случайный выбор одной из поступивших заявок, и обслуживание в циклическом порядке (ОЦП), когда все очереди заявок опрашиваются последовательно одна за другой в циклическом порядке; приоритетные ДО с относительными приоритетами (ОП), при которых после освобождения ресурса его занимает активная заявка с наиболее высоким ОП. На практике реализуются два подхода к заданию ОП заданием фиксированным значением кода адреса в системе, либо местом подключения абонента к связному ресурсу.

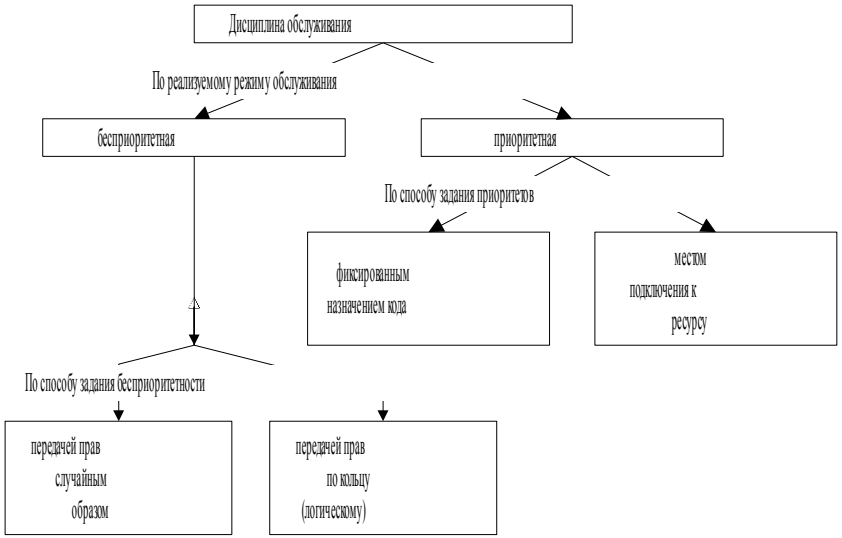


Рис. 1.3

Методы управления множественным доступом, реализующие соответствующие ДО, различаются видом сигнала передачи прав на занятие ресурса - либо асинхронным (по выделенной линии арбитража - эстафетный,

специальным кодовым словом по каналу связи - маркерный), либо временным интервалом (синхронный). Методы, реализующие ДООП, различаются способом задания относительного приоритета, либо кодом адреса в системе, либо местом подключения к каналу. Занятие канала здесь осуществляется за счет последовательного «отключения» менее приоритетных абонентов в результате сравнения разрядов кода ОП, либо при распространении отключающего сигнала по однонаправленной линии приоритета.

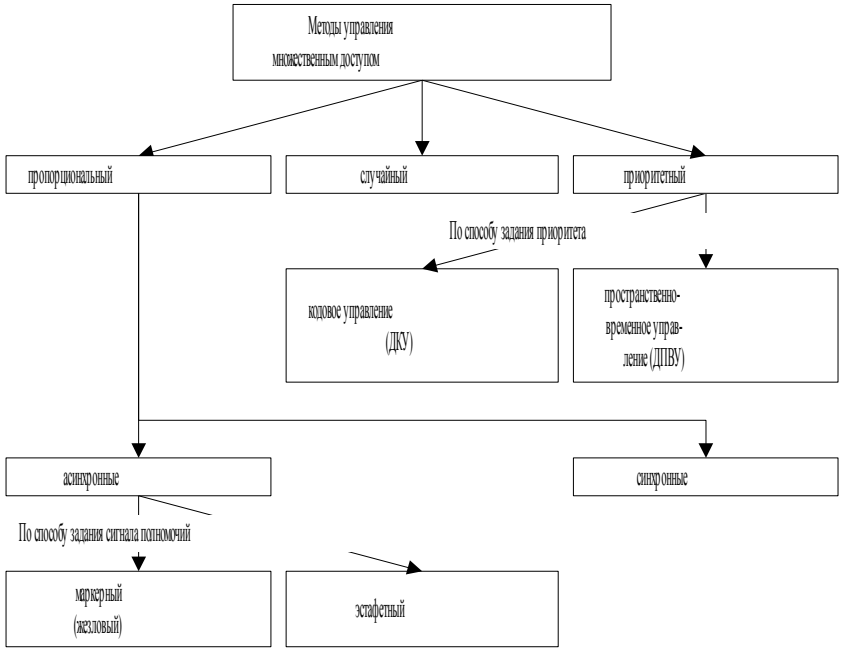


Рис. 1.4

Следуя классификации интерфейсов ЛВС реального времени [7] можно выделить: магистрально-модульные ВС, где последовательный канал связи имеет длину, как правило, до 10 метров: стандарт Bitbus, SSB (в интерфейсах Multibus), NUBUS, Fastbus, I^2C, D^2B и др.; рассредоточенные ВС - здесь в качестве передающей среды уже используется последовательный канал связи, имеющий длину от сотен метров до километров при возможности ее наращивания активными ретрансляторами. К таким интерфейсам относятся: интерфейс распределенной магистрали (ИРМ), интерфейс линейной связи с последовательной передачей информации (ИЛПС), MAP, PROWAY, (MAP, PROWAY - регламентируют правила построения ЛВС реального времени), MIL-STD-1553B и др.

Методы же управления множественным доступом, реализуемые для рассматриваемых интерфейсов с учетом представленной на рис. 1.4 классификации, следующие: маркерный (в частности сюда относится и предельный случай эстафетной смены задатчика - метод Polling - при смене задатчика или центрального узла каждым сигналом передачи прав) - наиболее широко применяемый метод, например в Bitbus, ИРМ, ИЛПС, PROWAY, MAP, MIL; и метод ДКУ, используемый в I^2C и D^2B [7]. Метод ДПВУ нашел применение лишь в сосредоточенных ВС с параллельной системной магистралью (И-41, VME-bus и др.), где приоритет абонентов убывает по мере удаления модулей «по гирлянде». Здесь применяется и метод ДКУ (в данных приложениях можно рассматривать как комбинированные ДКУ и ДПВУ), причем разряды кода, либо признаки занятия ресурса абонентами группы, подаются по специальным выделенным для этой цели управляющим линиям шины (И-41, Multibus I, II и др.) [7].

Т.о. современными стандартами, регламентирующими правила построения ЛВС реального времени, по существу, рекомендуется к использованию только маркерный метод управления множественным доступом, реализующий ОЦП посредством опроса очередей в соответствии с бесприоритетным расписанием. В монографии будет показано насколько неэффективен данный подход при реализации рассматриваемых ЛВС различных приложений.

В ЛВС, используемых в современных информационных системах, широкое использование находит случайный метод управления множественным доступом, в частности CSMA-CD (с обнаружением наложения) [8, 10, 11], реализующий бесприоритетную ДО с передачей прав случайным образом. Однако, данный подход уже не удовлетворяет современным сетевым технологиям, реализующим интеграцию служб связи, в том числе и реального времени, как при асинхронной передаче данных - ATM - в ЛВС технологии ATM используется маркерный метод управления доступом к магистральному каналу [3], так и при синхронной передаче - ISDN. Для этих приложений необходимы эффективные методы управления множественным доступом реального времени, реализующие приоритетные расписания. Такие методы могут быть получены лишь в том случае, если удастся отказаться от механизма опроса очередей при реализации расписаний, в том числе и приоритетных.

В монографии подробно рассматривается современная концепция обслуживания заявок в реальном времени, в основе которой находится оригинальный принцип обслуживания с динамическими ОП, изменяемыми по расписаниям, объединяющая в себе новые возможности обслуживания в реальном масштабе времени, со смешанными и с многоуровневыми (учитывающими приоритеты заявок) приоритетами, при высокой эффективности управления множественным доступом к ресурсам и открывающихся возможностях по унификации механизмов передачи прав на занятие ресурса между абонентами системы.

1.4. Модель системы реального времени

Так как основными требованиями к качеству построения и функционирования ВС реального времени является выполнение временных ограничений в обслуживании заявок, то распределенную ВС реального времени можно описать детеминированной моделью. К временным параметрам обслуживания заявок относятся величины: T_{p_m} - продолжительность занятия ресурса $m-M$, $m=1,\ldots,M$ абонентом для информационного с ним взаимодействия, T_{nn_m} - продолжительность передачи прав *m*-му абоненту после освобождения ресурса в системе, $K_{i(m)}$ - коэффициент частоты занятия ресурса *i*-м абонентом $i=1,\ldots,M$ относительно *m*-го $i \neq m$. Качество обслуживания заявок можно описать характеристиками: T_{a_m} - продолжительность арбитража требования *m*-го абонента (с момента появления заявки до момента предоставления абоненту права занять ресурс) или соответственно, продолжительность ожидания заявкой обслуживания, T_{o_m} - продолжительность обслуживания заявки системой. Для систем реального времени интерес представляют граничные (худшие для любой заявки) значения рассмотренных характеристик, которые соответственно обозначим: $T_{гp_m}$, $T_{гnn_m}$, $K_{гi(m)}$, $T_{га_m}$, $T_{го_m}$, $i,m=1,\ldots,M$, $i \neq m$, откуда получаем параметры обслуживания заявок в системе реального времени: $T_{га_m}$, $T_{го_m}$.

С учетом сказанного, получаем модель системы реального времени

$$\begin{cases} T_{га_m} = T_{гnn_m} + \sum_{i=1}^{M} K_{гi(m,i\neq m)}\left(T_{гnn_i} + T_{гp_i}\right) \\ T_{го_m} = T_{га_m} + T_{гp_m} \\ \forall m, m=1,\ldots,M \quad T_{гp_m} \neq \infty \end{cases} \qquad (1.1)$$

Утверждение. Обслуживание заявки в реальном масштабе времени корректно, если для любого абонента системы *m*, $m=1,\ldots,M$ выполняются условия (1.1).

Доказательство. Если условие (1.1) хотя бы для одного абонента системы не выполняется нельзя считать, что его заявки обслуживаются в

реальном масштабе времени, т.к. его параметры обслуживания T_{a_m} и T_{o_m} в этом случае не могут быть ограничены сверху и, следовательно, всегда найдутся условия функционирования системы, при которых $T_{za_m} < T_{a_m}$ и

$T_{го_m} < T_{o_m}$ или заявка будет обслужена не в реальном времени, т.е. для системы потеряна.

Т.к. особенностью обслуживания заявок в реальном времени будет то, что каждая заявка гарантированно должна быть обслужена за время $T_{го_m}$, то в данном случае приоритет заявки нельзя трактовать как преимущественное право одной заявки перед другой быть обслуженной (как, например в случаях относительных и абсолютных приоритетов)[1], здесь приоритеты заявок представляют собой (численно определяются) отношение гарантированных продолжительностей их обслуживания: $T_{го_m}$.

Будем говорить, что в системе реализована **бесприоритетная дисциплина обслуживания** требований общего ресурса **реального времени**, если для всех абонентов совпадают значения параметра $T_{го_m}$, соответственно **приоритетная дисциплина обслуживания** требований общего ресурса **реального времени**, если хотя бы для двух любых абонентов не совпадают значения параметра $T_{го_m}$.

В качестве **параметра приоритетности** обслуживания заявок в реальном времени может быть введена **количественная оценка - относительный уровень приоритетности реального времени** (или относительный приоритет реального времени) двух абонентов m и m'; $m, m' = 1,\ldots, M$, под которым понимается отношение $\delta_{m-m'} = T_{го_m} / T_{го_{m'}}$. Соответственно в системе реализована бесприоритетная дисциплина обслуживания требований ресурса, если для любых двух абонентов системы m, m' выполняется: $\delta_{m-m'} = 1 (m \neq m')$, если $\delta_{m-m'} < 1$ приоритет m абонента в $\delta_{m-m'}$ выше, в противном случае - в $\delta_{m-m'}$ ниже.

Системой (1.1) в общем случае определяются три способа задания приоритетного обслуживания заявок в РМВ и соответственно их комбинации.

1. Изменением параметров $K_{гi(m)}$.

2. Изменением параметров $T_{гпп_m}$.

3. Изменением параметров $T_{гр_m}$.

При этом очевидно, что выбор способов (и их комбинаций) задания приоритетов абонентов определяется соотношением параметров $T_{гр_m}$ и $T_{гпп_m}$. При $T_{гр_m} >> T_{гпп_m}$ целесообразно использовать способы 1 и 3, при сопоставимости $T_{гр_m}$ и $T_{гпп_m}$, соответственно способы 1 и 2.

1.5. Условия эффективности приоритетного обслуживания в реальном времени

Рассмотрим, с какой целью и при каких условиях следует назначать приоритеты абонентам распределенной ВС реального времени.

Пусть в систему реального времени поступает М типов заявок на обслуживание (каждый тип заявок в распределенной системе является принадлежностью соответствующего абонента), условием корректности функционирования системы будет выполнение М неравенств: $T_{грс_m} \le T^{з}_{грс_m}$, где $T_{грс_m}$ - гарантированная продолжительность реакции системы на воздействие, реализуемое в системе, $T^{з}_{грс_m}$ - задаваемое условиями функционирования ограничение на параметр $T_{грс_m}$. Требуемая реактивность системы по каждому *m*-му воздействию определяется следующими временными характеристиками обслуживания заявок системой: $T_{грз_m}$ - продолжительность собственно решения задачи *m*-м вычислителем (абонентом) по заранее известной программе (здесь для простоты считаем, что каждым абонентом обрабатывается только один соответствующий тип воздействий); числом информационных взаимодействий с ресурсом $L_{г_m}$, имеющих максимальную продолжительность $T_{го_m}$, необходимым для выработки адекватного воздействия в соответствии с заранее известным алгоритмом (далее для простоты будем считать, что $L_{г_m} = 1$, $m = 1,\ldots,M$). Тогда гарантированная продолжительность реакции системы на m-е воздействие

$$T_{грс_m} = T_{грз_m} + L_{г_m} T_{го_m}$$

Рассмотрим следующие возможные случаи.

1. Пусть требуется обеспечить равное время реакции системы на все воздействия - $T^{з}_{грс_m}$ для всех *m*, $m = 1,\ldots,M$ воздействий совпадают. В этом случае идеальным является бесприоритетное обслуживание - в реальном времени это передача полномочий в циклической очередности.

2. Пусть $T^{з}_{грс_m}$ заданы различными (различен функциональный смысл воздействий, что имеет место при функциональном распараллеливании). Тогда при бесприоритетной дисциплине обслуживания требований система может функционировать корректно лишь при выполнении для всех *M* абонентов самого жесткого для системы ограничения:

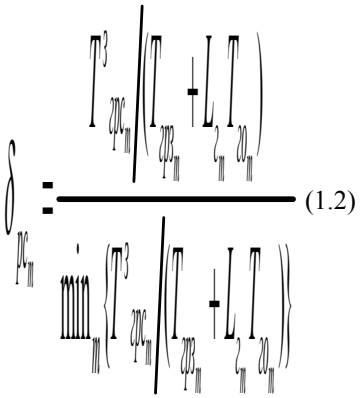
$$T_{\text{срв}_m}+L_{2_m}T_{\text{ao}_m}\le \inf\{T^3_{\text{срв}_m}, m=1,\dots,M\}$$

Например, ЛВСИС, в частности АТМ, по одним и тем же каналам связи поступает речь, радиосигнал и подвижное изображение. Для речевого сигнала, полоса пропускания частот которого F = 0 - 4000 Гц по теореме Котельникова (T = 0, 5/F) [4] получаем $T^{з}_{грс_m}$ = 125 мкс, для передачи радио сигнала F = 0 - 15000 Гц имеем $T^{з}_{грс_m}$ = 34 мкс, для качественной же передачи подвижного изображения (телевидение) соответственно F = 0 - 5 МГц уже имеем $T^{з}_{грс_m}$ = 104 нс. В УЛВС параметры $T^{з}_{грс_m}$ для различных заявок могут различаться еще существеннее.

Предположим, что рассматриваемое жесткое ограничение в системе при выбранной производительности технических средств не выполняется для некоторых абонентов. При этом в общем случае некоторые ограничения в системе могут выполняться с большим запасом по производительности ресурса. Тогда можно сформулировать задачу об эффективном перераспределении производительности вычислительных средств системы с учетом выполнения требований к корректности функционирования системы в целом. Критерием оптимальности задания очередности передачи прав (соответственно дисциплины обслуживания) будет **относительный коэффициент избыточности в эффективности обслуживания**

$$\delta_{рс_m} = T^{з}_{грс_m} \Big/ \min_m T^{з}_{грс_m}$$

При этом очевидно, что наиболее эффективно система с общими ресурсами будет реализована в том случае, если выполняются условия $\delta_{рс_m} = 1$, $m = 1,\ldots, M$, что можно считать **условием оптимальности дисциплины обслуживания реального времени**, а параметр $\delta_{рс_m}$ - **критерием оптимальности**. В общем случае для критерия оптимальности $\delta_{рс_m}$ (характеристики $T_{грз_m}$ и $L_{г_m}$ для различных воздействий не совпадают) имеем



(1.2)

Оценим, какой выигрыш для распределенных систем может дать реализация в системе приоритетного обслуживания, где приоритет вводится с целью эффективного использования производительности общего ресурса (при связном ресурсе - эффективного использования пропускной способности канала связи). С учетом $L_{г_m} = 1$ для всех M абонентов, для системы с бесприоритетной дисциплиной обслуживания (характеристика обозначена БП) требований ресурса имеем

$$T_{го_m}^{БП} = MT_{гр_m},$$

где $T_{гр_m}$ - продолжительность занятия ресурса абонентом с учетом потерь времени на передачу ему прав системой занять ресурс, считаем что $T_{гр_m}$ совпадают для всех *M* абонентов (очередей заявок).

Пусть требования к времени реакции системы на 1-е входное воздействие существенно выше, чем требования к любому иному входному воздействию $m = 2,\ldots,M$: $T^{з}_{грс_{m=1}} << T^{з}_{грс_{m=2,\ldots,M}}$. Реализуем для рассматриваемой системы приоритетную дисциплину обслуживания требований, в которой очередность предоставления прав абонентам системы на занятие ресурса выглядит следующим образом (1, 2, 1, 3, 1, 4, 1, . . , 1, *M*-1, 1, *M*, 1, 2, . .). Для данной ДО имеем следующие характеристики обслуживания заявок

$$T^{ВП}_{го_{m=1}} = 2T_{гр_m}$$

$$T^{НП}_{го_{m=\overline{2,M}}} = 2(m-1)T_{гр_m} ,$$

(где ВП, НП - соответственно характеристики абонентов с высоким и низким приоритетом), откуда выигрыш для более приоритетного абонента составит

$$\delta T^{ВП}_{го_m} = \frac{T^{БП}_{го_m}}{T^{ВП}_{го_m}} = \frac{M}{2}$$ (для всех m $T_{гр_m} = T_{гр}$)

Однако, наряду с выигрышем для более приритетного абонента имеем и проигрыш для менее приоритетных абонентов

$$\delta T^{НП}_{го_m} = \frac{T^{НП}_{го_m}}{T^{БП}_{го_m}} = \frac{2(M-1)}{M} = 2 ,$$

откуда результирующий выигрыш в производительности ресурса

$$\delta T_{грс_m} = \frac{T^{ВП}_{го_m}}{T^{НП}_{го_m}} = \frac{M}{4}$$

Оценим количественную оценку получаемого выигрыша. Пусть $T_{грз_m} = kT_{го_m}$, где $0 < k < \infty$. В предположении, что $L_{гт} = 1$, $m = 1,\ldots,M$ для бесприоритетного обслуживания имеем

$$T^{БП}_{грс_m} = (M+k)T_{гр_m} ,$$

соответственно, для приоритетного обслуживания при двух уровнях приоритетности получаем

$$T^{ВП}_{грс_m} = (2+k)T_{гр_m}$$

$$T^{НП}_{грс_m} = (2(M-1)+k)T_{гр_m}$$

соответственно получаем результирующий выигрыш в производительности системы

$$\delta T_{грс_m} = \frac{(M+k)^2}{(2+k)(2(M-1)+k)}. \qquad (1.3)$$

Зависимости $\delta T_{грс_m} = f(k)$ для различных M, при M=16;64 представлены на рис. 1.5, из которого делаем следующие выводы.

1. При существенном различии ограничений $T^{з}_{грс_m}$ для M абонентов при условии $k \le 1$ появляется возможность существенного повышения производительности системы реального времени, за счет реализации эффективного управления использованием общего ресурса абонентами, являющегося «узким местом» системы, в результате введения приоритетного обслуживания. Получаемый выигрыш в производительности системы в целом здесь может составить десятки раз, возрастая при увеличении числа абонентов в системе M.

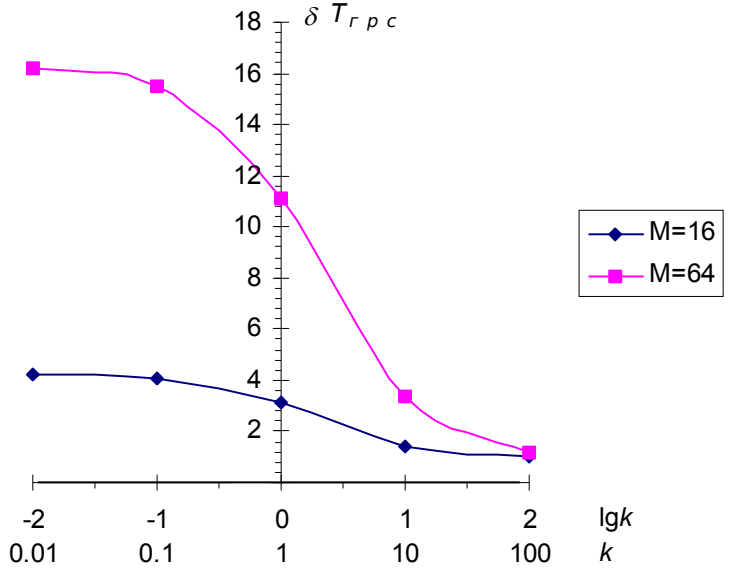


Рис. 1.5

2. Получаемый от реализации приоритетного расписания выигрыш снижается при увеличении коэффициента k, или при опережающем росте $T_{грз_m}$ над $T_{го_m}$, что определяет переход к классу сосредоточенных систем.

3. При $k > 1$ из (1.2) имеем $\delta T_{грс_m} = k^2 / k^2 = 1$ - отсутствует выигрыш как таковой, что может иметь место для некоторых приложений

сосредоточенных систем, когда $T_{грз_m} >> T_{го_m}$, где теряет актуальность задача приоритетного обслуживания заявок с целью эффективного использования ресурса абонентами системы. В данных приложениях систем реального времени применяются иные методы параллельной обработки, в частности крупноблочного распараллеливания, при бесприоритетном обслуживании заявок на доступ к ресурсу (в реальном времени - обслуживание в циклическом порядке), который здесь уже не является «узким местом» системы. Как следствие, здесь, как правило, реализуется принцип однородности структуры.

Замечание. Нетрудно показать, что предельным случаем системы, «узким местом» которой будет общий ресурс, можно считать многозадачную операционную систему реального времени (кстати говоря, сосредоточенную). Действительно, общим ресурсом здесь является квант процессорного времени, который и является «узким местом» многозадачной системы. Поэтому при использовании в системе распараллеливания по функциям (что, как правило, имеет место в специализированных операционных системах, реализуемых в задачах управления) возникает аналогичная задача приоритетного обслуживания с целью эффективного использования ресурса. Здесь также имеют место два способа задания приоритетов - изменением величины процессорного кванта для приоритетной задачи, изменением частоты (не ОЦП) опроса очередей заявок от отдельных задач. Первый подход ограничен нарушением параллельности обработки, что недопустимо в системах реального времени (кстати говоря, именно эти же соображения приводят к уменьшению длины пакета данных в сетевой технологии ATM до 53 байтов), второй имеет недостатком наличие больших временных затрат на опрос очередей по расписанию. Поэтому как и в ЛВС сегодня на практике здесь используется бесприоритеный опрос очередей (например, используемая в ОС QNX дисциплина RR), т.е. в системах практически отсутствуют возможности эффективной диспетчеризации в реальном времени при распараллеливании задач по функциям. Другими словами, рассматриваемые в монографии принципы диспетчеризации могут эффективно использоваться и в этих приложениях систем реального времени параллельной обработки - в различных приложениях как распределенных, так и сосредоточенных ВС.

С учетом сказанного, цель реализации приоритетного обслуживания в ЛВС реального времени, в которых ресурс является «узким местом», состоит в перераспределении прав между абонентами на доступ к ресурсу в соответствии с заданными ограничениями на время реакции системы на входное воздействие. Выполнение данных ограничений при минимальной производительности ресурса достигается при выполнении условия: $\delta T_{рс_m} = 1$, $m = 1,\ldots,M$, которое можно считать условием оптимальности приоритетных ДО реального времени. Соответственно, количественное

задание приоритетов абонентов может быть получено с использованием выражения (1.2).

Замечание. Приведенное исследование иллюстрирует и общий подход к синтезу алгоритма функционирования распределенной ВС ($k \le 1$), состоящий в минимизации числа межмодульных взаимодействий при обслуживании заявок, что формально записывается следующим образом: $L_{г_m} \to \min$, $m = 1,\ldots,M$.

Выше речь шла о ЛВСПО реального времени, относительно ЛВСОН реального времени отметим, что здесь реализуется аналогичная идея приоритетного обслуживания, однако, т.к. отсутствует алгоритм функционирования - невозможно задать параметр $L_{г_m}$, то в качестве ограничений следует уже рассматривать не параметр $T^{з}_{грс_m}$, задаваемый в ЛВС для всех *M* абонентов, а параметр $T^{з}_{го_m}$. При различии $T^{з}_{го_m}$ для *M* абонентов могут вводиться приоритеты с целью эффективного использования «узкого места» - связного ресурса ЛВС.

1.6. Модель системы оперативной обработки

Очевидно, что по аналогии с введенной выше моделью системы реального времени может быть введена и модель системы оперативной обработки, отличающаяся тем, что этой моделью должен учитываться параметр P_m - вероятность занятия ресурса m-ым абонентом (в общем случае $P_m < 1$) при предоставлении ему права занять ресурс. Качество обслуживания заявок в этом случае можно описать следующими характеристиками: W_{am} - средняя продолжительность арбитража требования m-го абонента, W_{Om} - средняя продолжительность обслуживания заявки системой.
При этом из (1.1) можем получить модель системы оперативной обработки, которая здесь принимает вид

$$\begin{cases} W_{am} = P_m T_{ГППm} + \sum_{i=1}^{M} K_{Гi(m,\, i \ne m)} P_i (T_{ГППi} + T_{ГРi}) \\ W_{Om} = W_{am} + P_m T_{ГРm} \end{cases}$$

Замечание. При $P_m \to 1$, $m = \overline{1, M}$ имеем детерминированную модель реального времени (1.1), где $W_{am} = T_{Гam}$, $W_{Om} = T_{ГОm}$. Т.о. приведенная здесь модель представляет собой более общий случай обслуживания, для которой справедливо следующее утверждение.
Утверждение. Обслуживание заявки в режиме оперативной обработки корректно, если обслуживание номера поступающей в систему заявки можно охарактеризовать следующим величинами: W_{Om} , $T_{ГОm}$, где $W_{Om} = T_{ГОm}$

при P_m =1 $\forall$ m, $m = \overline{1, M}$ и $W_{Om} < T_{ГОm}$ если $\exists$ m для которых P_m <1 (соответственно W_{am} и Tam).

Доказательство. При описании системы данной моделью для системы характерно, что вся зависимость от ее загрузки $P_m \leq 1$ заявки, поступающие от абонентов обслуживаются за исключением времени $T_{ГОm}$ (n* $T_{ГОm}$ для n-ой
заявки в очереди m-го абонента) т.е. любой абонент является элементом системы, т.к. при любых условиях функционирования с системой не производится отключения от ресурса. Т.о. здесь как и в системе реального времени особенностью обслуживания заявок будет то, что каждая заявка гарантированно должна быть обслужена за время $T_{ГОm}$ при среднем времени обслуживания, на которое и расчитывается система оперативной обработки,
W_{Om}. Или в данном случае приоритет заявки, как и в системах реального времени нельзя трактовать, как преимущественное право одной заявки перед другой быть обслуженной, а приоритет заявок представляют собой (численно определяются) отношением средних и гарантированных продолжительностей их обслуживания: W_{Om}, $T_{ГОm}$.

Однако в данном случае уже имеет смысл говорить не только о дисциплине обслуживания заявок, которая как и для системы реального времени характеризуется параметрами $T_{ГОm}$ = f ($K_{Гi}$ (m), $T_{ГППm}$, $T_{ГРm}$) т.е. дисциплина обслуживания как и для системы реального времени описывается детерминированной моделью, но и о использования прав абонентами, предоставленных дисциплиной обслуживания, что определяется параметром W_{Om} = f (P_m).

Будем говорить, что предоставленые абонентам дисциплиной обслуживания права используются ими в равной мере, если для них совпадают отношения W_{Om} / $T_{ГОm}$, соответственно, права предоставляемые m-му абоненту, дисциплиной обслуживания используются им в большей мере, чем m' абонентом, где m, $m' = \overline{1, M}$, m ≠ m', если $\frac{W_m}{T_{ГОm}} > \frac{W_{m'}}{T_{Om}}$

Очевидно, что права, предоставляемые абонентам дисциплиной обслуживания используются ими в равной мере при условии: P_m =P, $m = \overline{1, M}$.

В предположении, что в проблемно-ориентированной ЛВС требуется обеспечить среднее время реакции системы на входное воздействие W^3_{PCm}, что для M абонентов, $m = \overline{1, M}$ задается условием: $W_{PCm} \leq W^3_{PCm}$, где $W_{PCm} = T_{РЗm} + L_m W_{Om}$, где $T_{РЗm}$ - продолжительность решения задачи вычисления, Lm - число информационных взаимодействий с ресурсом

средней продолжительностью W_{Om} , необходимое для выработки в системе сигнала реализации на входное воздействие, по аналогии с системой реального времени может быть введен коэффициент избыточности в эффективности обслуживания m-го абонента при реализации бесприоритетной передачи прав абонента в циклическом порядке: $\delta_{PCm} = W_{PCm} / \min_m W^3_{PCm}$,

где $\min_m W^3_{PCm} = \inf\{W^3_{PCm}, m = \overline{1,M}\}$.

Тогда формализованным условием введения приоритетного обслуживания будет: $\delta_{PCm} = 1, m = \overline{1,M}$,что можно считать условием оптимальности дисциплины обслуживания, а параметр δ_{PCm} соответственно критерием оптимальности

$$\delta_{PCm} = \frac{W^3_{PCm} / (T_{PЗm} + L_m W_{Om})}{\min_m (W^3_{PCm} / (T_{PЗm} + L_m W_{Om}))}$$

(в общем случае для δ_{PCm} имеем характеристики $T_{PЗm}$ и L_m для различных воздействий не совпадают).

Т.о. как и в системе реального времени здесь можно количественно описать приоритеты и сформулировать аналогичную задачу синтеза дисциплины обслуживания, критерием оптимальности которой является параметр δ_{PCm} , а условием оптимальности: $\delta_{PCm} = 1, m = \overline{1,M}$.

Рассмотрим, чем же отличается рассмотренная модель оперативной обработки от модели реального времени. По существу только тем, что здесь учитывается два параметра W_{Om} и $T_{ГOm}$, причем основным является W_{Om} . При этом ограничения типа $T_{ГOm} \le T^3_{ГOm}$ выполняться не должны,

выдвигаются требования к выполнению условий $W_{PCm} \le W^3_{PCm}$, другими словами, здесь реализуется обслуживание в реальном времени. Но производительность технических средств выбирается не исходя из выполнения условий: $T_{ГOm} \le T^3_{ГOm}$, а исходя из выполнения условий: $W_{PCm} \le W^3_{PCm}$.

Поэтому далее говоря о построении систем реального времени, понимаем, что аналогично могут строиться и системы оперативной обработки на основе рассмотренной модели с обслуживанием (в том числе и приоритетном) по расписаниям с тем лишь отличием, что синтез расписаний осуществляется с учетом параметров W_{Om}.

Отметим, что обслуживанию по расписанию в режиме оперативной обработки присущи как существенные достоинства, так и существенные недостатки. К основному достоинству можно отнести возможность эффективного использования ресурса при $k \le 1$ за счет синтеза расписания, максимально учитывающего ограничения W^3_{PCm}, $m = \overline{1, M}$, W_{Om} для ЛВС ОН. Однако для ЛВС ПО, соответственно данное достоинство можно реализовать лишь при эффективной передаче прав по расписанию, что не обеспечивается современными методами управления множественным доступам в ЛВС. Поэтому данное достоинство будет обеспечено лишь в том случае, когда будут предложены методы эффективной передачи прав по расписанию, эффективность которых не будет зависеть от величины загрузки системы (о таких методах речь пойдет ниже).

К недостаткам данного подхода можно отнести то, что он не обеспечивает возможность защиты от перегрузки высокоприоритетных абонентов (либо заявок), за счет «отключения» от ресурса низкоприоритетных, что реализуется обслуживанием с относительными приоритетами. Поэтому на практике целесообразно говорить об использовании обоих рассмотренных методов обслуживания в ЛВС ОО, а также ставить задачу их эффективной совместной реализации в единой технической системе.

РАЗДЕЛ 2. МЕТОДЫ ДИСПЕТЧЕРИЗАЦИИ РЕАЛЬНОГО ВРЕМЕНИ

2.1. Требования к ДО заявок в распределенных ВС реального времени

Альтернативные способы обслуживания заявок в распределенных ВС составляют ДО с относительными приоритетами (ДООП) и обслуживание заявок по расписаниям (ДОР), частным случаем последнего является опрос очередей заявок в циклическом порядке - бесприоритетное расписание.

К достоинствам первого подхода можно отнести защиту от перегрузок высокоприоритетных заявок [1], за счет «отключения» от ресурса низкоприоритетных заявок при высокой загрузке системы, к недостаткам - невозможность использования в реальном масштабе времени, т. к. для всех заявок, кроме наиболее приоритетной, в общем случае не выполняются условия (1.1). Другими словами, ДООП не может корректно быть использована в системах реального времени.

Обслуживание по расписанию представляет собою единственно возможный подход к реализации ДО реального времени, т.к., чтобы передать абоненту право на занятие ресурса, его следует внести в расписание передачи прав, а в этом случае ему всегда гарантируется некоторое ограничение $T_{го_m}$, задаваемое местом абонента в очереди.

Обслуживанию по расписанию присущи следующие недостатки:

- невозможность защиты от перегрузок высокоприоритетных заявок реального времени. Эта проблема возникает при необходимости обслуживания в одной системе как важных заявок реального времени, так и некоторых заявок оперативной обработки, обслуживание которых несет в системе некоторый второстепенный характер. Внесение этих заявок в расписание также предполагает их обслуживание в реальном времени, соответственно при снижении тем самым эффективности обслуживания высокоприоритетных заявок реального времени;

- большие временные затраты на опрос очередей, с целью выявления активности заявок в очереди. Это обусловливается тем, что право абоненту занять ресурс предоставляется вне зависимости от его готовности к этому, а в соответствии с заданной очередностью;

- невозможность учета приоритета заявки и абонента системы, т.к. по расписанию права передаются между абонентами системы в предположении, что их приоритет однозначно соответствует приоритету заявки. Если в абонент поступает несколько типов заявок, образующих несколько очередей, актуальной становится задача учета приоритета уже очереди заявок абонента.

Проиллюстрируем, в какой мере снижает эффективность обслуживания заявок опрос очередей по расписанию. В качестве исследуемого метода децентрализованного управления множественным доступом рассмотрим маркерный метод (метод эстафетной смены задатчика) и регламентируемое для него стандартами miniMAP и IEEE 802.4 [8,11] расписание обслуживания в циклическом порядке. Оценку эффективности метода проведем по параметру I_s - информативность смеси в канале межмодульного обмена [9] (характеристика коэффициента пропускной способности канала связи)

$$I_s = \frac{\sum_{i=1}^{S} \delta_i n_i}{\sum_{i=1}^{S} \delta_i N_i}$$

где δ_i - частота посылок *i*-го типа в общем потоке сообщений, *S* - общее число типов посылок $i = 1,\ldots,S$, n_i - число информационных бит в посылке *i*-го типа, N_i - общее число бит в посылке *i*-го типа. Влияние на эффективность межмашинного взаимодействия передачи полномочий между абонентами системы на занятие ими общего ресурса по расписанию проиллюстрируем на изменении параметра I_s, позволяющего оценить информативность смеси при одном - информационном кадре в системе $I_{s=1}$ и соответственно при двух кадрах - информационный кадр и маркер $I_{s=2}$. При этом будем говорить, что в системе одновременно может быть *m* из *M*

абонентов $m = 1,\ldots, M$ активизированных на информационное взаимодействие. В этом случае маркер в среднем будет выдан в канал $\delta_{i=2}$

раз для предоставления абоненту права занять ресурс, в то время, как информационный кадр в этом случае будет передан единожды, где

$$\delta_{i=2} = \frac{m}{M} \sum_{a=1}^{M/m} a \, .$$

Корректность использования приведенных моделей в рассматриваемых приложениях - ЛВС, обусловливается тем, что можно пренебречь величиной временных потерь, связанных с продолжительностью распространения сигнала по каналу связи между наиболее удаленными абонентами. В этом случае коэффициент пропускной способности канала в полной мере определяется информативностью смеси в канале связи.

Зависимости $I_{s=1} = f(n)$ для случая $n = 1,\ldots,10$ представлены на рис. 2.1 и для случая $n = 10,\ldots,1000$ - на рис. 2.2. В обоих случаях представлены характеристики для информационных кадров для ЛВС реального времени, регламентируемых стандартами IEEE 802.4 и miniMAP. Длина информационного кадра (в байтах) для IEEE 802.4 определяется следующим образом [7, 10] $N_1 = n/(R+31)$, где $R = 43$, если $n \le 43$, $R = n$, если $n > 43$ (используются заполняющие байты в поле данных). Для miniMAP, ориентированного на применение в управляющих системах, структура кадра более экономична, в частности, отсутствуют заполняющие байты в поле данных, не требуется передачи 8 байт преамбулы для синхронизации приемников (задана только одна скорость передачи по каналу связи), сокращено адресное пространство кадра. На рис. 2.2 кроме того, приведена оценка информативности кадра, применяемого в рамках сетевой технологии асинхронной передачи данных с интеграцией служб ATM [3], где длины информационных кадров фиксированы и составляют $n_1 = 48$ (байтов) при $N_1 = 53$ (байта), т.е. кадр содержит всего 5 байтов управляющей информации ($I_{s=1} = 48/53 = 0{,}91$).

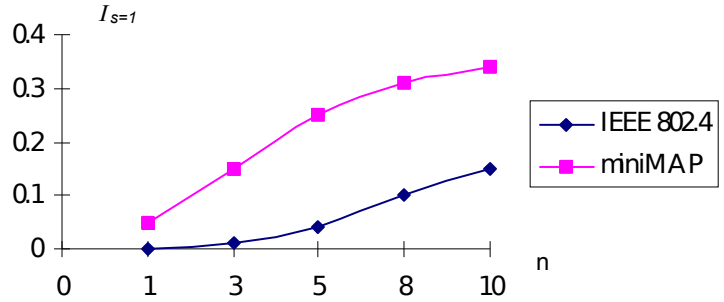


Рис. 2.1

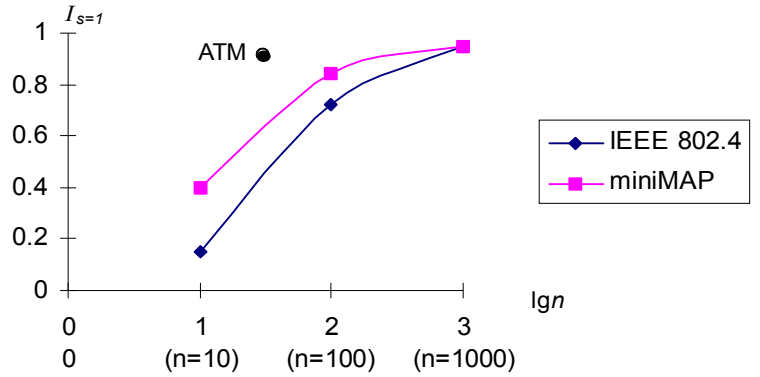


Рис. 2.2

Из рис. 2.1 и 2.2 можно сделать следующие выводы.

1. Современными ЛВС реального времени, в основу построения которых положена реализация рекомендации HDLC на уровне звена передачи данных, крайне неэффективно используется связной ресурс при передаче коротких сообщений, что делает неэффективным их применение в УЛВС (кстати говоря, и стандарт miniMAP не отличается высокой эффективностью использования связного ресурса). Здесь отметим, что альтернативным к HDLC можно считать подход, как раз и разработанный для задач управления, регламентируемый стандартами MIL-STD-1553A/B, где передача данных ориентирована не на передачу байтов, а на передачу слов [7,9], где каждое слово содержит признак команда/данные, код команды, что позволяет идентифицировать принятое слово без учета порядка его поступления в пакете.

2. Для ИЛВС реального времени, где требуется в реальном масштабе времени обмениваться большими массивами информации, обеспечивая открытость системы, ЛВС IEEE 802.4 и miniMAP, как следует из рис. 2.2, могли бы быть эффективны. Однако реальный масштаб времени можно обеспечить в системе, если оградить связной ресурс от продолжительного занятия парой абонентов (в противном случае не обеспечить обслуживание в реальном времени заявок от других абонентов). Это приводит к необходимости уменьшения длины информационного пакета для таких приложений ЛВС. В частности, современная концепция ATM задает фиксированную длину пакета (ячейки) 53 байта, что уже требует реализации и альтернативных подходов к построению информационных кадров, в частности кадр ATM содержит лишь 5 управляющих байтов, включая адресное пространство. Рис. 2.2 иллюстрирует преимущество ATM по данному параметру над другими асинхронными сетевыми технологиями, в

основе звена передачи данных которых находится HDLC, что обеспечивает эффективность данного стандарта для ЛВССРВ и ЛВСИС.

Зависимости $I_{s=2} = f(m/M)$ - информативности смеси, учитывающей потери пропускной способности канала, связанные с передачей маркера для опроса очередей по расписанию (в рассматриваемом случае - бесприоритетному) для ЛВС IEEE 802.4 и для сетевой технологии ATM представлены соответственно на рис. 2.3 и 2.4. Для ЛВС IEEE 802.4 учитываем стандартную длину маркера 17 байтов [8], для ATM исследуем некоторую гипотетическую модель - принимаем минимально возможную длину, определяемую длиной адреса абонента [$\log_2 M$], т.е. в предположении, что $M = 256$, составляющую 1 байт. Из приведенных рисунков видим, что даже при гипотетически идеальных параметрах системы передача маркера, необходимая для реализации передачи прав по расписанию, существенно сказывается на эффективности межмодульных взаимодействий, что подтверждает низкую эффективность ДОР для ЛВСРВ.

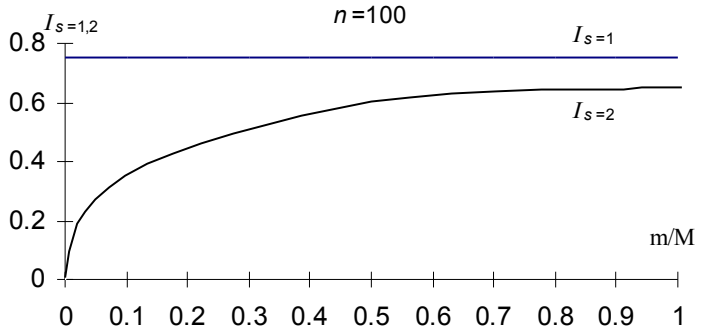


Рис. 2.3

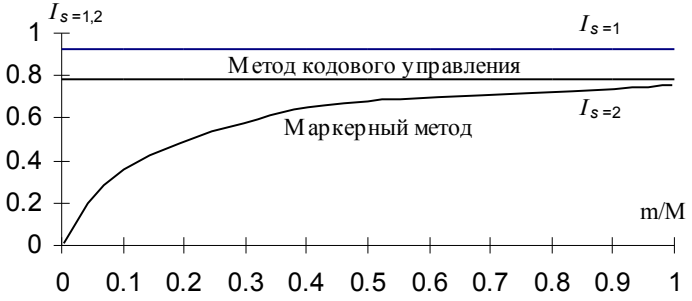


Рис. 2.4

С учетом сказанного очевидно, что в общем случае (если приоритет абоненту назначается не в соответствии с интенсивностью потока требований ресурса заявками абонента) при реализации приоритетных расписаний потери на передачу прав станут еще существенней. Это объясняет использование сегодня лишь одного способа учета приоритета в рамках маркерного метода, реализующих расписания - по параметру $T_{гр_m}$ при бесприоритетном обслуживании по остальным двум параметрам [11]. Однако, для масштаба реального времени использование данного подхода крайне ограничено, т.к. в этом случае «теряется» преимущество параллельной обработки - несколько приоритетных абонентов могут монопольно использовать ресурс в течение довольно продолжительного времени. Поэтому на практике методы, реализующие опрос абонентов по расписанию (маркерный), реализуют ДО в циклическом порядке.

Таким образом, к приоритетной ДО реального времени, реализуемой в распределенной ВС, в общем случае выдвигаются следующие требования:

- реализация ДО со смешанными режимами обслуживания для различных классов заявок - ОР для заявок реального времени, ОП для остальных заявок;
- эффективная реализация расписаний реального времени, которая позволит получать эффективные приоритетные расписания;
- возможность учета приоритета абонента и приоритета заявки при управлении множественным доступом к общим ресурсам, причем один абонент может иметь несколько очередей заявок различных классов (приоритетов).

Основу реализации обслуживания по приоритетным расписаниям в ЛВС сегодня составляет построение SPT- расписаний (короткая работа здесь обслуживания заявки общим ресурсом вперед).

Под длиной работы $T_{ДРm}$ здесь понимаются суммарные затраты $T_{ППm}$ и T_{Pm} $T_{ДРm} = T_{ППm} + T_{Pm}$.

При сопоставимости $T_{Pm} = T_P$ для различных абонентов длина работы определяется параметром $T_{ППm}$, причем $T_{ДРm}$ тем меньше, чем меньше $T_{ППm}$. Параметр $T_{ППm}$ можно уменьшить за счет передачи прав занять ресурс чаще тому абоненту, который чаще требует ресурс, т.е. за счет предоставления преимущественных прав (приоритета) абоненту, характеризуемому постоянной интенсивностью занятия ресурса λ_m. Например, $\lambda_{m=1} >> \lambda_{m=2} = \ldots = \lambda_M$. Реализуем следующее расписание - прием расписания представим в скобках(1,2,1,3,1,.....1,M-1,1,M). Если затраты времени на передачу прав (например, маркера) между двумя абонентами, то имеем

$T_{ДР1} = 2T_{ПП} + T_P$

$T_{ДР2} = \ldots = T_{ДРМ} = 2(M-1)T_{ПП} + T_P$

или при сопоставимости $T_{ПП}$ и T_P

$\delta T_{ДР} = T_{ДР1} / T_{ДР2} \approx 1 / M$

При условии $\lambda_1 \to \infty, \lambda_2, \ldots, \lambda_M \to 0$

$$I^{ПР}_{S=2} \approx T_P/(T_P+2T_{ПП})$$

Заметим, что, если для данной задачи не будут введены приоритеты, получим

$$I^{БП}_{S=2} \approx T_P/(T_P+MT_{ПП})$$

соответственно при сопоставимости T_P и $T_{ПП}$ имеем

$$\delta I_{S=2} = I^{ПР}_{S=2} / I^{БП}_{S=2} \approx \frac{1}{3} M$$

т.е. появляется возможность существенного повышения пропускной способности канала связи за счет реализации приоритетного обслуживания.

Однако, на практике данная возможность практически не используется, что вызвано следующим. При построении информационных систем (в том числе и на средствах передачи информации ЛВС) приоритетность обслуживания определяется не интенсивностью поступления заявок на обслуживание, а важностью обрабатываемой информации. Если же обратиться к стахостической мере количества информации I [. . .], для которой характерно уменьшение обратно пропорционально вероятности ее появления P (на практике используется логарифмическая мера) $I = -\log P$ можем сделать вывод, что более высокий приоритет следует назначать абонентам, обрабатывающим более важную информацию, реже поступающую систему, что противоречит изложенному выше подходу к назначению приоритетов абонентам ЛВС. Это объясняет то, что данный подход (реализация SPT-расписаний в ЛВС), несмотря на внешнюю привлекательность, на практике не используется.

Очевидно, что условием эффективного приоритетного обслуживания заявок по расписанию в ЛВС будет независимость затрат $T_{ППm}$ от интенсивности поступления заявок в систему, что формализовано можно представить следующим образом: $T_{ППm} = T_{ПП}, \; m = \overline{1, M}$.При этом (в случае $T_{Pm} = T_P, \; m = \overline{1, M}$) все работы имеют равную длину

$T_{ДРm} = T_{ДР}, \; m = \overline{1, M}$

что обуславливает эффективность обслуживания в первую очередь (с приоритетом) наиболее важной заявки.

2.2. Концепция обслуживания в реальном времени с динамическими приоритетами

2.2.1. Основа построения приоритетных расписаний

Идея излагаемой концепции состоит в реализации дисциплин обслуживания реального времени с передачей прав по расписанию (ДОР), за счет смены относительных приоритетов (ОП) в рамках реализуемой в системе дисциплины обслуживания с относительными приоритетами (ДООП) по расписанию при каждом занятии ресурса абонентами системы.

Утверждение. В любой момент времени t_s ОП не должен совпасть у заявок из нескольких очередей.

Доказательство. Если данное условие не будет выполнено, то в системе неминуем конфликт при занятии ресурса, т.к. несколько абонентов одновременно получат право занять ресурс после его освобождения.

Для описания ДООП используем матрицу приоритетов (МП), представляющую собою квадратную матрицу $Q = [q_{ij}]$ размерности $M \times M$ по числу M абонентов [1]. Элемент матрицы q_{ij} задает ОП абонента i по отношению к j: 0 - нет приоритета, 1 - есть. Для описания ДОР (в общем случае ДОСП) используем граф изменения матрицы ОП в моменты времени t_s занятия ресурса в соответствии с расписанием. Пример графа бесприоритетной ДОР, реализуемой методом динамической смены ОП, для случая $M = 4$, цикл расписания которой имеет вид (1, 2, 3, 4), представлен на рис. 2.5. Бесприоритетность расписания обеспечивается тем, что каждый абонент входит в расписание равное число раз, в общем случае может быть более одного, например (1, 1, 2, 2, 3, 3, 4, 4).

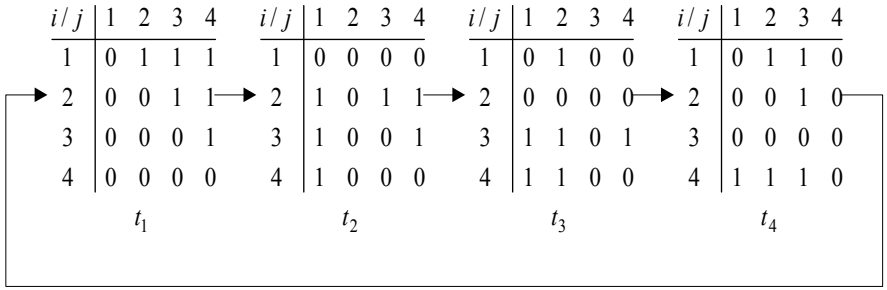


Рис. 2.5

Требования к МП. Элементы МП должны удовлетворять следующим требованиям:

- $q_{i=j} = 0$, т.к. между заявками одного класса не могут быть установлены приоритеты;

- если $q_{ij} = 1$, то $q_{ji} = 0$, т.е., если заявки класса i имеют приоритет по отношению к заявкам класса j, то последние не могут иметь приоритет по отношению к заявкам i;

- в МП не должны совпасть не любые две строки i, i', не любые два столбца j, j'; $i = 1,\ldots, M$, $i \neq m$, $j, j' = 1,\ldots, M$, $j \neq j'$.

Требования к графу смены МП ДОР РМВ. В графе смены МП ДОР РМВ (в цикле расписания) по крайней мере по одному разу должны присутствовать МП, задающие высший ОП каждого из М абонентов системы.

Утверждение. Для реализации приоритетной ДОР в цикле расписания по крайней мере двум абонентам системы высший ОП должен присваиваться различное число раз, например (1, 2, 1, 3, 1, 4).

Доказательство. В противном случае получим совпадение значений T_{o_m}, т.е. при совпадении значений параметров $T_{рз_m}$ получаем равный приоритет заявок - совпадают $T_{рс_m}$.

Изменение ОП заявок по расписанию в процессе функционирования системы должно быть реализовано по следующему правилу.

Правило изменения ОП. ОП в рамках ОР однозначно задаются расписанием, где в каждый момент времени t_s приоритет заявок соответствует порядку передачи полномочий, исключая повторные передачи прав одной очереди в цикле ОР, например для расписания (1, 2, 1, 3) в момент t_1 - ОП [1, 2, 3], в t_2 - [2, 1, 3], в t_3 - [1, 3, 2], в t_4 - [3, 1, 2].

Для ДО с динамическими ОП, изменяемыми по расписанию, функция приоритетности заявки *m*, $m = 1,\ldots, M$ имеет вид

$$\varphi_m(t) = \alpha_{ms}(t_s \le t^n < t_{s+1}) + \Delta\alpha_{md}(t_d, d = 1,\ldots,G),$$

где α_{ms} - исходный ОП заявки поступающей в момент t^n, соответствующий *s*-му состоянию цикла расписания, длиной *G*: $t_s\ s = 1,\ldots,G$; $\Delta\alpha_{md}$ - приращение (может иметь отрицательные значения) приоритета заявки, получаемое при смене состояний цикла расписания $t_d\ d = 1,\ldots,G$. Для заявок, обслуживаемых с ОП, для $\forall t_d$, $\Delta\alpha_{md} = 0$ и для $\forall t^n = t_s, s = 1,\ldots,G$, $\alpha_{ms} = const$.

2.2.2. Принципы эффективной реализации приоритетного обслуживания в распределенной системе

В основе излагаемого подхода находятся следующие принципы кодового управления реального времени.

1. В любой момент времени функционирования системы матрице приоритетов ставится во взаимно однозначное соответствие матрица кодов ОП (МКП). Строки МКП соответствуют строкам МП, а столбцы задают код ОП очереди заявки. Пример МКП, в предположении, что «1» в разряде кодового слова приоритетнее «0» (по «1» права передаются, по «0» - нет), и что приоритет разряда кода убывает по мере возрастания его порядкового номера, приведен на рис. 2. 6.

МП

i/j	1	2	3	4
1	0	1	1	1
2	0	0	1	1
3	0	0	0	1
4	0	0	0	0

$\leftrightarrow$

МКП

n/m	1	2	3
1	1	1	0
2	1	0	1

Рис. 2.6

2. В процессе функционирования системы при каждом занятии ресурса МКП изменяется в соответствии со сменой исходной МП (заменой МП на МКП во все моменты времени t_s), в результате чего получается граф смены МКП, полностью соответствующий графу смены МП (чем реализуется передача прав по расписанию). Пример взаимно однозначного отображения графов МП в МКП для случая, представленного на рис. 2. 5 (бесприоритетное обслуживание), приведен на рис. 2. 7.

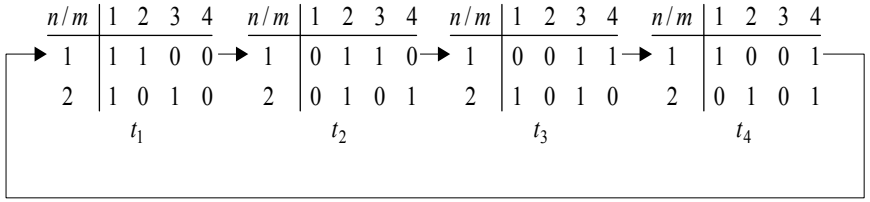


Рис. 2.7

3. При построении кодов ОП всех очередей заявок в процессе функционирования системы применяется приведенное ранее правило задания ОП заявок в графе смены МП.

4. При управлении доступом к ресурсу должно осуществляться поразрядное (начиная с более приоритетных разрядов) сравнение кодов ОП заявок , затребовавших к данному моменту времени ресурс, с «отключением» по каждому разряду менее приоритетных заявок - при однозначном кодировании заявок получаем бесконфликтное занятие ресурса.

Утверждение. Метод кодового управления на всем интервале изменения загрузки системы имеет эффективность не ниже, чем методы опроса очередей.

Доказательство. Утверждение доказывается тем, что при ином подходе (опрос очередей) права передаются поочередно от одной очереди заявок к другой, опрашивая их, и при стационарности системы для занятия ресурса может потребоваться несколько передач кодового слова - маркера, имеющего длину не менее [$\log_2 M$] разрядов. При кодовом управлении при любой

загрузке системы потребуется лишь единожды выдать в канал разряды кода приоритета той же длины, что позволяет говорить о том, что эффективность кодового управления совпадает с идеальной характеристикой опроса очередей, в случае активности на занятие ресурса заявок из всех очередей.

Зависимость изменения информативности смеси $I_{s=2} = f(m/M)$ для ЛВС, реализуемой в рамках технологии АТМ, при использовании кодового управления в реальном времени множественным доступом к каналу связи при $M = 256$ приведена на рис. 2.4, который иллюстрирует высокую эффективность кодового управления по сравнению с методом передачи маркера, реализуемого при опросе очередей.

2.2.3. Дополнительные возможности обслуживания по расписаниям в рамках концепции кодового управления

Прежде всего, рассмотрим возможности построения ДО со смешанными приоритетами (ДОСП) или комбинированных ДО - ОР и с ОП. Реализация подобных возможностей позволит совместить в единой системе альтернативные подходы к обслуживанию, достигающие совершенно противоположные цели, соответственно - обслуживание по расписанию с целью реализации обслуживания заявок в реальном времени, обслуживание с ОП, позволяющее обеспечивать защиту от перегрузок высокоприоритетных заявок, что необходимо для эффективной реализации альтернативных приложений ЛВСКО. Данные возможности в одной системе позволяет получать изложенный принцип реализации ДОР, отличающийся тем, что в любой момент функционирования системы реализуется ДООП, с условием, что ОП заявок изменяются при каждом занятии ресурса системы. Однако, в процессе функционирования системы могут изменяться ОП не всех заявок, причем как низкоприоритетных, так и высокоприоритетных, либо группы очередей заявок могут организовывать свои очередности (расписания) смены ОП. Данные возможности, открываемые реализацией ДОР, посредством смены ОП в процессе функционирования системы, положены в основу идеи реализации ДОСП, соответственно получаемого в его рамках ряда ДО.

Примеры графов ДОСП, иллюстрирующих альтернативные способы задания ОП, соответственно для защиты от перегрузок заявок реального времени (ОП неизменяем для низкоприоритетных заявок) и с целью выделения внеочередных заявок (ОП неизменяем для низкоприоритетных заявок), для случая $M = 4$ представлены на рис. 2.8, на рис. 2.8.а - 1 и 2 очереди заявок имеют ОП над 3 и 4, а 3 над 4, 1 и 2 образуют бесприоритетное ОР (1, 2), на рис. 2.8.б - 1 очередь заявок имеет ОП над остальными, 2 и 3 над 4, а 2 и 3 бесприоритетное ОР между собой (2, 3). Обозначим, соответственно ДОСП [(1, 2), 3, 4] и [1, (2, 3), 4], где в круглых

скобках отмечен цикл ОР, в квадратных (будем называть это циклом ДОСП) - ОП заявок, упорядоченный в порядке записи. В общем случае цикл ДОСП

может содержать несколько циклов ДОР, например для $M = 8$ [(1, 2, 1, 3), 4, 5, (6, 7, 8)]. При отсутствии в ДО цикла ОР, например [1, 2, 3, 4], имеем ДООП.

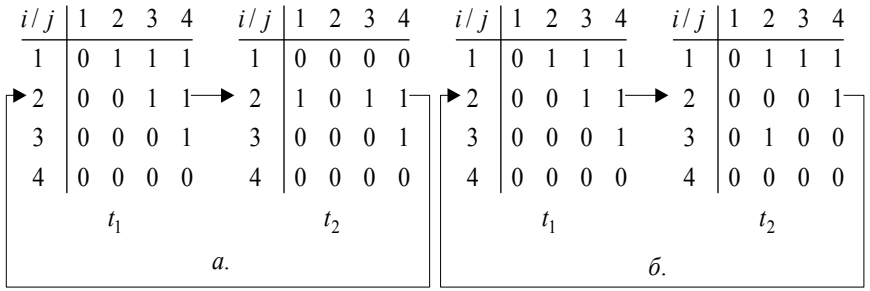


Рис. 2.8

Замечание. Предоставление высокого ОП внеочередным заявкам можно рассматривать как подход, альтернативный обслуживанию системных заявок с абсолютным приоритетом, применительно для ЛВС реального времени. Это обусловливается тем, что при небольших размерах информационного кадра и децентрализации управления множественным доступом к ресурсам, пропадает смысл прерываний взаимодействий с ресурсом низкоприоритетных абонентов при поступлении системной заявки (это приводит к большим временным потерям, чем при обслуживании с ОП, при существенном усложнении процедуры управления множественным доступом к ресурсам).

С учетом сказанного, построение ДОСП (ОР и ОП) в общем случае реализуется следующим образом:

- заявки подразделяются на три большие группы: внеочередные, РМВ, бесприоритетные;

- внеочередным заявкам присваиваются максимальные ОП - при необходимости обслуживать их с равным приоритетом, либо в какой-нибудь очередности, для них реализуется расписание (ОР);

- для заявок реального времени строится ДОР (приоритетная, либо бесприоритетная), их ОП в любой момент времени ниже ОП внеочередных заявок;

- низкоприоритетным заявкам оперативной обработки (нереального времени) присваиваются ОП более низкие, чем заявкам реального времени (чем последние при ООП защищаются от перегрузок), которые, в свою очередь, могут обслуживаться как в режиме ОП, так и между собою по расписанию.

Замечание. В общем случае в системе могут присутствовать не все три возможных группы заявок, но для исследуемого класса систем всегда присутствуют заявки реального времени.

2.2.4. Дополнительные возможности обслуживания с многоуровневыми приоритетами

Другой широкий класс возможностей в обслуживании заявок в реальном времени, открывающийся в рамках рассматриваемой концепции приоритетного обслуживания в распределенных системах, состоит в возможности учета не только приоритета абонента, но и собственно приоритета поступающей заявки на обслуживание, в предположении, что требования к временным параметрам обслуживания заявки определяются не только приоритетом абонента. Реализация данной возможности обслуживания является непременным условием построения ЛВСИС, в частности в рамках технологии АТМ, где по одним и тем же каналам связи поступают сигналы реального времени различного функционального назначения, например при передаче речи и подвижных изображений, что естественно определяет и различные требования к обслуживанию данных типов заявок. В этом случае реализуется **многоуровневый приоритет**, под которым понимаем реализацию нескольких функциональных уровней ОП, где соответствующий приоритет учитывается только в рамках соответствующего уровня. В рассматриваемых приложениях, прежде всего, выделяются разряды приоритета заявки и разряды приоритета абонента, причем старшинство разрядов определяется следующими соображениям:

- если в основе приоритетного обслуживания лежит учет, в первую очередь, приоритета заявки, младшие (более приоритетные) $[\log_2 Q]$ разрядов кодового слова, где Q - число уровней приоритета заявок, отводятся для кодирования приоритета заявок, старшие $[\log_2 M]$ разрядов под кодирование приоритета абонента;

- если в основе приоритетного обслуживания лежит учет, в первую очередь, приоритета абонента, младшие $[\log_2 M]$ разрядов кодового слова, отводятся для кодирования приоритета абонентов, старшие $[\log_2 Q]$ разрядов - под кодирование приоритета заявки.

При этом, с учетом реализации в системе ДОСП, возникают возможности как при обслуживании абонентов, так и при обслуживании заявок, получать ДОР и ДООП. Пример ДО с учетом приоритетов заявок и вычислителей приведен на рис. 2.9. Здесь в первую очередь учитывается приоритет заявки (младший разряд), причем два вида заявок обслуживаются с ОП. Требования абонентов системы для рассматриваемого случая обслуживаются в реальном времени по бесприоритетному расписанию. Получаем следующую ДО - высокоприоритетные заявки обслуживаются между собою для различных вычислителей бесприоритетно. Другой класс заявок имеет более низкий ОП, т.е. поступают на обслуживание только в

отсутствие высокоприоритетных заявок, между собою низкоприоритетные заявки также бесприоритетны.

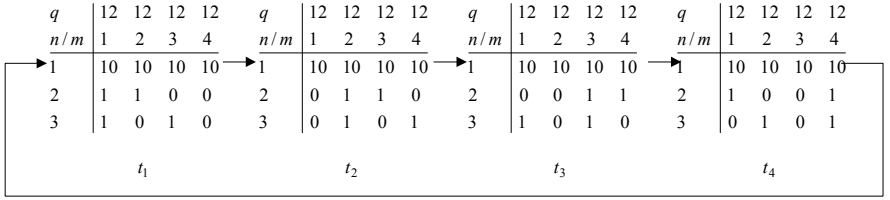


Рис. 2.9

Замечание. С целью уменьшения длины кода приоритета можно кодировать приоритеты заявок и абонентов без выделения отдельных разрядов в кодовом слове, однако это несколько усложнит алгоритм управления множественным доступом к ресурсу, за счет усложнения алгоритмов кодирования и декодирования приоритетов, однако идея многоуровневости приоритета сохранится и в этом случае.

В общем случае в системе может присутствовать несколько уровней приоритетов L, $i = 1,\ldots,M$. Если уровень с меньшим номером характеризуется более высоким приоритетом, в системе J, $i = 1,\ldots,M$, относительный приоритет каждого l-го уровня кодируется n_l разрядами кода, получаем матрицу приоритетов и соответствующую ей матрицу кодов приоритетов, приведенные на рис. 2.10.

	МП			*МКП*
l / j	1 ... J		n_l / m	1 ... J
1	*ОП уровня* 1		n_1	*КОП уровня* 1
	–			–
.	...		.	...
.	...	$\leftrightarrow$	.	...
.	...		.	...
	–			–
L	*ОП уровня L*		n_L	*КОП уровня L*

Рис. 2.10

Ранее отмечалось, что на практике в ЛВС не реализуются ДО с абсолютными приоритетами, что вызвано существенными временными потерями и увеличением сложности аппаратурной реализации прерывания взаимодействий по каналу связи в распределенной ВС. В сосредоточенных же системах, где эффективна реализация приоритетного обслуживания (например, в операционных многозадачных системах реального времени), как правило, возникает необходимость реализации в системе, наряду с

рассмотренными возможностями, обслуживание с абсолютными приоритетами. Совместить такие альтернативные способы обслуживания опять же возможно с применением концепции многоуровневых приоритетов. Выделим старшие уровни (младшие разряды кода приоритета) для кодирования типа приоритета, например «1» - абсолютный приоритет, «0» - относительный. Если многоуровневый приоритет и так учитывает несколько видов относительного приоритета, можно ввести и несколько уровней типов приоритетов (абсолютный/относительный), например, в соответствии с МКП, приведенной на рис. 2.11, где приоритет заявки выше чем приоритет абонента ВС, и приоритет заявки и приоритет абонента могут быть как относительными, так и абсолютными, что задается в разрядах «тип приоритета» кода приоритета. Совмещение обслуживания с относительными и абсолютными приоритетами становится возможным благодаря тому, что при кодовом управлении при арбитраже по каждому разряду кода всеми абонентами фиксируется с каким кодом приоритета абонент (или заявка) занимает ресурс. В частности, при арбитраже по разрядам кода, задающим тип приоритета, абонентами фиксируется был ли в канале «0» или «1», что будет основанием (если канал занят заявкой, либо абонентом с «0» значением в соответствующем разряде кода) для прерывания взаимодействия при поступлении соответственно заявки с абсолютным приоритетом, либо заявки у абонента, имеющего абсолютный приоритет.

МКП

n_l / m	1 ... J
n_1	*Тип приоритета заявки*
	–
n_2	*Тип приоритета абонента*
	–
n_3	*Приоритет заявки*
	–
n_4	*Приоритет абонента*

Рис. 2.11

Заметим, что при изложенном принципе обслуживания, появляется новая возможность учета приоритета, либо бесприоритетного обслуживания между собою заявок/абонентов, имеющих абсолютный приоритет по сравнению с остальными заявками/абонентами системы.

С использованием рассматриваемой концепции обслуживания с многоуровневыми приоритетами при кодовом управлении доступом к общим ресурсам системы могут быть реализованы сложные многоуровневые алгоритмы обслуживания, действия которых основаны на применении понятий уровней достигнутого обслуживания [2]. Дисциплины

обслуживания, реализующие многоуровневые алгоритмы, сегодня находят широкое использование в многозадачных операционных системах, в частности реального времени. Частным случаем многоуровневых алгоритмов является обслуживание с передним и задним планом - ПЗП. Согласно этому алгоритму все поступающие заявки становятся в некоторую внешнюю очередь, из которой уже отправляются на обслуживание. Если фиксированного времени $T_{гр_m}$ (кванта), предоставляемого заявке системой, недостаточно для ее обслуживания, заявка отправляется во внутреннюю очередь, из которой уже поступает на обслуживание вновь, получая следующий квант времени занятия общего ресурса. Здесь возможны два варианта обслуживания:

- внешняя очередь имеет приоритет перед внутренней, т.е. пока есть заявки во внешней очереди, именно они, а не заявки из внутренней очереди, поступают на обслуживание;

- внутренняя очередь имеет преимущество перед внешней.

Возможны также некоторые комбинации, когда для одних заявок всегда приоритетнее внешняя очередь, для других - внутренняя.

Обобщая алгоритм ПЗП, в предположении, что система может содержать некоторое число i, $i > 1$ внутренних очередей, приходим к так называемому многоуровневому алгоритму - МА, действие которого основано на применении понятия уровней достигнутого обслуживания, где каждый уровень характеризуется некоторым промежутком времени полученного обслуживания. В этом случае чем больше квантов обслуживания получила та или иная заявка, тем на более высокий уровень она попадает, из которого выбирается на обслуживание реже/чаще.

МА в рамках рассматриваемой концепции обслуживания с многоуровневыми ОП реализуется следующим образом. Вводится ОП уровня обслуживания, который кодируется в рамках реализации кодового управления доступом к общим ресурсам в приоритетных разрядах кода ОП. Менее приоритетные разряды многоуровневого кода содержат код ОП заявки. При переходе заявки на следующий уровень изменяется (увеличивается/уменьшается) ОП (соответственно код ОП) уровня.

Классификация методов обслуживания с многоуровневыми относительными приоритетами при кодовом управлении доступом к общим ресурсам представлена на рис.2.12.

Из приведенной классификации методов обслуживания с многоуровневыми приоритетами могут быть сделаны следующие выводы.

1. В рамках рассматриваемой концепции обслуживания с динамическими приоритетами может быть реализовано практически все используемое сегодня на практике многообразие ДО. Вместе с тем, могут быть получены принципиально новые ДО, прежде всего для применения в системах реального времени.

2. Все реализуемое сегодня в альтернативных приложениях ВС многообразие ДО и их комбинаций может быть унифицировано в рамках метода кодового управления доступом к общим ресурсам.

3. Механизм кодового управления доступом к общим ресурсам можно рассматривать как единый высоко эффективный унифицированный механизм реализации ДО для альтернативных приложений ВС.

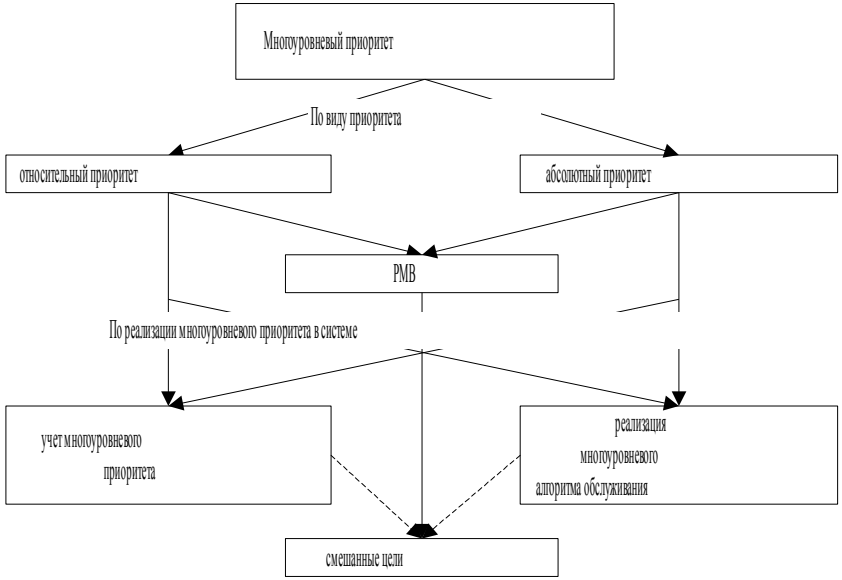


Рис. 2.12

4. Возможность унификации механизмов обслуживания в рамках принципа кодового управления множественным доступом является теоретической предпосылкой, открывающей широкие возможности в комбинировании альтернативных ДО в единой ВС.

Кстати говоря, в рамках концепции кодового управления может рассматриваться и случайный метод управления множественным доступом (подробнее речь об этом пойдет в четвертом разделе), отличия которого состоят в том, что относительный приоритет (соответственно и код приоритета) присваиваются абоненту (заявке), при его активизации на занятие ресурса, случайным образом.

Замечания.

1. В общем случае приоритеты уровней обслуживания могут изменяться по весьма сложным законам, что возможно реализовать в рамках рассмотренных выше способов кодового управления, например с целью реализации режима "фоновой задачи" (широко используется в операционных системах), где приоритет уровней задачи, решаемой на фоне других задач в

системе, всегда ниже. Кроме того, здесь также возможна реализация как режима реального времени, если приоритеты уровней изменять по расписанию, так и режима оперативной обработки.

2. Применительно к ЛВС отметим, что обслуживание с абсолютными приоритетами и с многоуровневым алгоритмом здесь предполагают обмен большими массивами данных абонентами ЛВС, что противоречит принципам функционирования ЛВСРВ. Вместе с тем, рассматриваемые возможности могут использоваться в ЛВСОО и ЛВСКО, где реализуются режимы оперативной обработки, допускающие передачу больших массивов данных по каналу связи ЛВС.

В завершении изложения методов обслуживания заявок в ВС с многоуровневыми приоритетами, в основе которых находится реализация исследуемых принципов обслуживания с динамическими приоритетами и кодовым управлением доступом к общим ресурсам, рассмотрим возможности адаптивного управления множественным доступом. В этих приложениях также может использоваться концепция обслуживания с многоуровневыми приоритетами. Идея адаптивного управления состоит в возможности изменения ДО (в частности изменение расписания передачи прав, изменение в назначении смешанных приоритетов, например вывод/ввод заявки в/из расписание) при превышении некоторого заданного функционированием системы, прежде всего в реальном времени, порогового значения времени ожидания обслуживания заявками. С целью адаптивного управления в многоуровневом ОП выделяются приоритетные разряды адаптивного управления. При функционировании системы с исходными приоритетами, в этих разрядах кодов ОП «0» значения. Если в системе появляются заявки (не зависимо у каких из абонентов ВС), для которых превышено установленное для них пороговое значение в продолжительности обслуживании, при очередном арбитраже требований ресурса в рассматриваемом(мых) разряде кода ОП появляется «1» значение, удерживаемое в течение всего времени функционирования системы с превышением заданных пороговых значений для продолжительности ожидания обслуживания заявок. При появлении «1» в разряде(ах) адаптивного уровня кода ОП, что является необходимым условием для всех абонентов использовать при арбитраже текущее значение кода ОП уже не для исходной, а для некоторой дополнительной(ых) ДО, которая реализуется (вырабатываются текущие значения кодов ОП при каждом занятии ресурса) в ВС одновременно с исходной, а используется по мере необходимости. Число разрядов в уровне адаптивного управления задается максимальным числом пороговых значений для продолжительности ожидания обслуживания заявок в системе $K^{\max}$, следующим образом [$\log_2 K^{\max}$].

Рассмотренная концепция адаптивного управления множественным доступом, в первую очередь, может эффективно использоваться в ЛВСРВ и ЛВСКО. Вместе с тем, данный подход может применяться и в ЛВСОО с

целью реализации требуемого режима функционирования ВС в условиях перегрузок.

Таким образом, можем выделить два подхода к адаптивному управлению множественным доступом к общим ресурсам, реализуемых в рамках рассмотренной концепции обслуживания с многоуровневыми динамическими приоритетами: адаптивно к продолжительности обслуживания, реализуемое многоуровневым алгоритмом обслуживания, и адаптивно к продолжительности ожидания обслуживания. Получаемая, с учетом сказанного, классификация методов адаптивного управления множественным доступом к общим ресурсам ВС, представлена на рис.2.13.

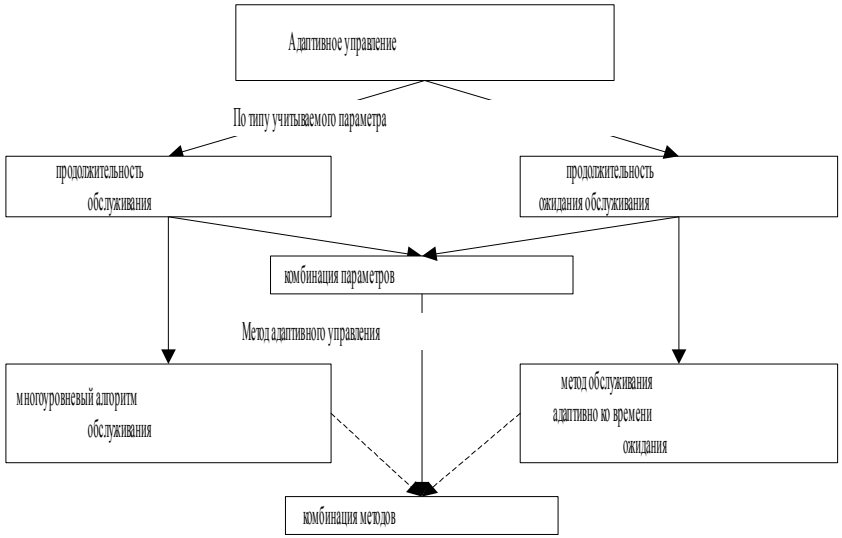


Рис. 2.13

2.2.5. Классификация ДО с динамическими приоритетами для ЛВС

Классификация возможных ДО для ЛВС, получаемых в рамках изложенной концепции кодового управления обслуживанием с динамическими приоритетами, изменяемыми по расписанию, представлена на рис. 2.14. Отметим, что здесь выделены классификационные признаки и соответственно приведена классификация ДО именно с позиции исследуемой концепции обслуживания. В рамках же каждой полученной дисциплины можно рассмотреть известные классификационные признаки - по виду

стратегии обслуживания (вентильная, ординарная, исчерпывающая) и др. [1, 5, 6].

Кроме того, для ЛВСОО и ЛВСКО может быть реализована ДО с многоуровневым приоритетом, учитывающим абсолютные приоритеты и многоуровневый алгоритм обслуживания для заявок оперативной обработки, и практически для любых приложений ЛВС – обслуживание с адаптивным управлением множественным доступом.

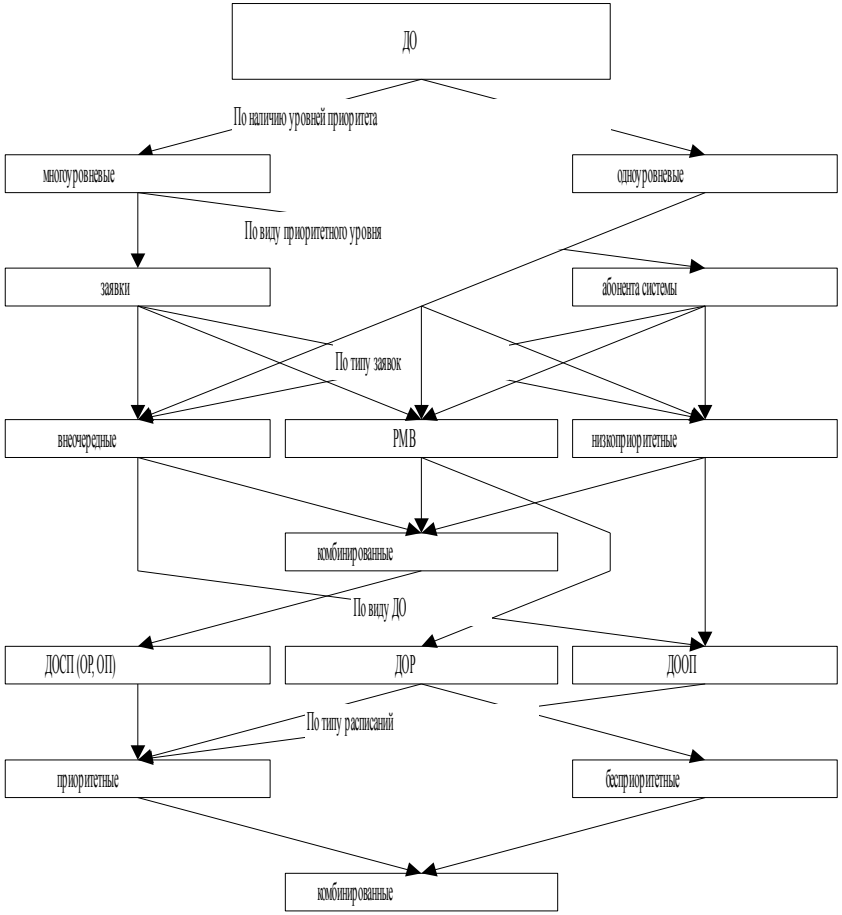


Рис. 2.14

2.3. Понятие и свойства канонического расписания реального времени

Пусть имеем R различных приоритетов абонентов в системе, обслуживаемых по расписанию, соответственно $r = 1,\ldots, R$, и M_r абонентов системы имеют равный r приоритет $m_r = 1,\ldots, M_r$: $\delta_{m_r - m'_r} = 1$, $m_r \neq m'_r$, $m_r, m'_r = 1,\ldots, M_r$. Таким образом система содержит M абонентов, которым соответствует R уровней приоритетов.

Детерминированная задача синтеза расписаний реального времени может быть сформулирована следующим образом : $M|1|T_{от_r} \to \max$, $T_{от_r} \le T_{гот_r}$, $m_r = 1,\ldots, M_r$, $r = 1,\ldots, R$, где $T_{от_r}$ - характеристика обслуживания m-го абонента r-го приоритета, $T_{гот_r}$ - ограничение, накладываемое на время обслуживания заявки абонента системой, функционирующей в реальном масштабе времени.

С учетом сформулированной задачи синтеза расписаний реального времени дадим определение канонического (в данном случае - оптимального) расписания. Под **каноническим расписанием** будем понимать расписание, обеспечивающее минимальные значения характеристик $T_{гот_r}$ в рамках заданного разбиения M абонентов на R уровней приоритета, получаемое поочередной передачей прав абонентам в рамках радиального графа.

Под **радиальным графом передачи прав** понимается граф, содержащий M_r вершин приоритета r, $r = 1,\ldots, R$, причем в каждую вершину нижестоящего приоритета входит только одна дуга от одной вершины вышестоящего приоритета, из каждой вершины всех приоритетов, кроме R-го выходит [M_{r+1} / M_r] дуг, где [a] - есть большее целое числа a, все вершины приоритета R соединяются дугой с вершиной приоритета 1 (из вершин R выходит только одна дуга) - в этом случае дуга направлена в сторону вершины 1, в то время как в остальных случаях дуга направлена в вершину, соответствующую пользователю с более низким приоритетом. Радиальный граф передачи прав представлен на рис. 2.15.

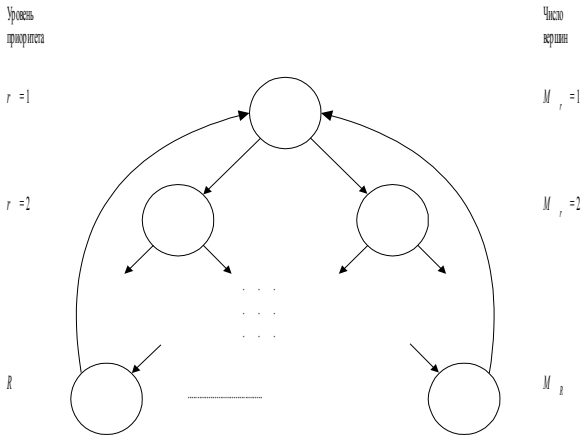


Рис. 2.15

Радиальный граф, у которого по крайней мере одна вершина *r*-го (не *R*-го) уровня приоритета соединена только с одной вершиной приоритета уровня $r+1$ назовем **вырожденным**. Соединение двух вершин одной дугой является альтернативным способом задания равного приоритета двум абонентам системы. Под **предельно вырожденным** понимаем граф, в котором из каждой вершины только одна дуга выходит и в каждую вершину только одна дуга входит - этот граф задает бесприоритетное расписание, соответствующее обслуживанию в циклическом порядке.

Под **поочередной** будем понимать такую передачу прав на занятие ресурса абонентами, при которой абоненты одного приоритета получают полномочия на доступ к ресурсу с равной частотой или бесприоритетно друг относительно друга. Для радиального графа такая очередность может быть определена в результате выполнения следующей итерационной процедуры. В общем случае реализуется *R* итераций.

1. Произвольным образом через одну вершину каждого приоритета проводится первый путь.

2. Второй путь должен пройти через вершины всех приоритетов, кроме первого(через первый путь проходит всегда), не соединенные первым путем, третий, не соединенные первыми двумя путями и т.д.

3. Если не осталось на рассмотрении вершин r -го приоритета через которые не прошел по крайней мере один путь (для невырожденного графа прежде всего это приоритет 2) из проведенных *r* путей (2 путей), $r+1$ дуга проходит как и первая, но уже через другие вершины приоритета $r+1$ и т.д., вплоть до соединения путями вершины первого приоритета со всеми вершинами приоритета *R*.

4. Все вершины *R* приоритета соединяются дугой с вершиной высшего приоритета.

5. Нумеруются пути в соответствии с очередностью их получения. Именно в соответствии с этой нумерацией (с этим порядком) в системе должны передаваться полномочия на доступ к ресурсу.

6. Для каждого полученного пути определяется очередность передачи полномочий между вершинами графа (абонентами) перечислением очередности появления соответствующих вершин на пути.

Иллюстрация получения поочередной передачи прав, с использованием приведенной процедуры представлена на рис. 2.16.

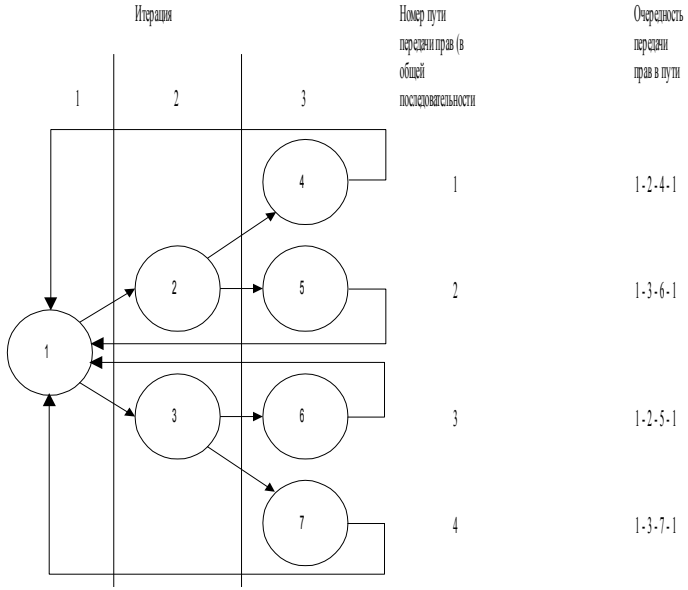
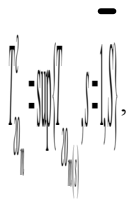


Рис. 2.16

Свойства канонических расписаний реального времени.

1. В общем случае для каждого абонента в системе в цикле расписания может быть несколько очередностей передачи прав на доступ к ресурсу $s = 1,\ldots, S$, например, (1, 2, 1, 3, 4). Тогда параметры обслуживания заявок абонента определяются гарантированной продолжительностью обслуживания $T_{го_{m(S)}}$



и средней продолжительностью гарантированного обслуживания требований с учетом различных очередностей S

$$T^{*}_{го_{m(S)}} = \sum_{s=1}^{S} pT_{го_{m(S)}} \ .$$

Таким образом в общем случае - при нескольких очередностях в системе дисциплина обслуживания реального времени задается параметрами $T^{г}_{го_{m(S)}}$ и $T^{*}_{го_{m(S)}}$. Для канонического расписания значения $T^{г}_{го_{m(S)}}$ и $T^{*}_{го_{m(S)}}$ соответственно совпадают для абонентов всех *R* приоритетов.

2. В системе реализуются минимально возможные значения $T_{го_{m(S)}}$ для всех M_r абонентов всех *R* приоритетов.

3. Относительный уровень (коэффициент) приоритетности абонентов m_r, m_r : $\delta_{r-r'}$ вырожденного графа - при минимальных $T_{го_m}$ составляет

$$\delta_{r-r'} = \frac{(RM_r - 1)}{(RM_{r'} - 1)},$$

соответственно, ограничением на общность построения радиального графа будет

$$\frac{M_{r+1}}{M_r} = 2,3,4,... ;\ r = 1,...,R$$

Из анализа свойств канонических расписаний, можем сделать следующие выводы.

1. В основе синтеза расписаний реального времени должно находиться построение канонического расписания, с последующей его модификацией, учитывающей особенности конкретной задачи синтеза.

2. При минимизации значений параметра $T_{го_{m_r}}$ для всех абонентов всех уровней приоритета системы *R* (эффективное использование ресурса) строго задается соотношение уровней коэффициента приоритетности абонентов системы.

3. Исходное соотношение коэффициентов приоритетности может быть изменено либо подбором соответствующих значений двух других параметров $T_{гр_m}$ и $T_{гпп_m}$ в рамках канонического расписания, либо изменением значений $T_{го_m}$ (либо $T^{*}_{го_m}$, в общем случае они могут не совпадать), путем их увеличения (уменьшить их уже невозможно, что следует из второго свойства канонических расписаний) для абонентов отдельных уровней приоритета.

Аналогично может быть построено расписание и для ЛВС ОО.

Следует отметить, что в этом случае расписание также может синтезироваться по параметрам $T_{ГРСm}$, $m = \overline{1, M}$. При этом значения $T_{ГРСm}$ должны выбираться таким образом, чтобы для синтезированного по этим параметрам расписания выполнялось: $W_{РСm} \le W^{3}_{РСm}$, $m = \overline{1, M}$.

Методика синтеза расписания в данном случае состоит в следующем.

1.Назначаются $T_{ГРСm},\ m=\overline{1,M}$, где $T_{ГРСm} > W^{3}_{РСm},\ m=\overline{1,M}$.

2.Для заданных параметров $T_{ГРСm},\ m=\overline{1,M}$ с использованием описанной выше методики синтезируется расписание

3.Строится модель (аналитическая или имитационная) системы массового обслуживания с полученным расписанием с целью определения значений $W_{РСm},\ m=\overline{1,M}$.

4.Проводится сравнение значений $W_{РСm}$ и $W^{3}_{РСm},\ m=\overline{1,M}$. В результате которого значения уменьшаются для тех параметров m для которых $W_{РСm} > W^{3}_{РСm}$ соответственно увеличиваются в случае, если $W_{РСm} < W^{3}_{РСm}$.

5.С новыми заданными значениями $T_{ГРСm}$ переход к п.п.2.

Синтез расписания продолжается до выполнения условия:

$W^{3}_{РСm} - W_{РСm} \le \Delta W_{РСm},\ m=\overline{1,M}$,

где значения $\Delta W_{РСm},\ m=\overline{1,M}$ задаются перед началом решения задачи, исходя из условий задачи.

Замечание. Задача синтеза расписания для ЛВС ОО может решаться и за один этап, однако это требует решения задачи определения значения $T_{ГРСm}$ для заданного $W^{3}_{РСm}$ т.е. должны быть определены и формализованы зависимости $T_{ГРСm} = f(W^{3}_{РСm}),\ m=\overline{1,M}$.

Данная задача на сегодняшний день не решена и в книге лишь обозначается в постановочной части.

2.4. Модель системы обслуживания с кодовым управлением множественным доступом

При моделировании систем с передачей прав по расписанию, реализуемой опросом очередей, в частности при маркерном методе управления множественным доступом, возникает серьезная проблема, вызванная тем, что получение аналитических зависимостей, связывающих время ожидания заявок с числом обращений к очередям по расписаниям,

каждое из которых требует учета затрат времени на передачу прав - сложная задача, которая не может быть решена классическими методами теории массового обслуживания [1]. Поэтому при проектировании таких систем используют некоторые приближения - аппроксимирующие функции, достаточно полный обзор которых приведен в [5, 6]. При кодовом же управлении множественным доступом в основе находится ДООП, что позволяет применить при моделировании систем известные формулы теории массового обслуживания.

В общем случае можем выделить две большие группы заявок на обслуживание - реального и нереального масштабов времени. Заявки первой группы обслуживаются по расписанию и параметром обслуживания здесь будет $T_{го_m}$ - гарантированная продолжительность обслуживания

$$T_{го_m} = K_{г_m}(T_{гпп} + T_{гр}),$$

где $T_{гпп}$ - гарантированные затраты времени на передачу прав абоненту для занятия системы (для кодового управления при равномерном кодировании они фиксированы и составляют [$\log_2 M$]), $T_{гр}$ - гарантированные затраты времени при занятии ресурса, которые также, как правило, фиксированы (например, связанные с передачей 53 байтов информации в рамках сетевой технологии АТМ), $K_{г_m}$ - коэффициент, задающий гарантированное предоставление ресурса (более высокий приоритет) другим абонентам системы между очередными предоставлениями прав занять ресурс *m*-му абоненту (приоритет абонента тем выше, чем меньше $K_{г_m}$).

Заявки оперативной обработки обслуживаются либо бесприоритетно, либо в режиме ОП. В последнем случае для расчета характеристик обслуживания можно воспользоваться моделью однолинейной СМО с *H* классами (приоритетами) очередей неограниченной длины с ординарной стратегией обслуживания заявок, образующих простейшие потоки с интенсивностями $\lambda_1,\ldots,\lambda_H$, и случайными длительностями обслуживания с известными первыми γ_i и вторыми начальными моментами $\gamma_i^{(2)}$, $i = 1,\ldots,H$, R_{k-1} , R_k - соответственно загрузки, создаваемые различными потоками; в которой для среднего времени ожидания обслуживания заявки класса *k* W_k имеем [1]

$$W_k = \frac{\sum_{i=1}^{H} \lambda_i \gamma_i^2 (1+\upsilon_i^2)}{2(1-R_{k-1})(1-R_k)},\ \upsilon_i = \sqrt{\frac{\gamma_i^{(2)} - \gamma_i^2}{\gamma_i^2}}$$

где

$$R_{k-1} = \sum_{i=1}^{k-1} \lambda_i \gamma_i \text{ , a } R_k = R_{k-1} + \lambda_k \gamma_k \text{ ,}$$

причем т.к. $\gamma_m = T_{гпп} + T_{гр}$ или $\gamma_m = const$, $\upsilon_m = 0$.

Исследуем эффективность ДОСП РМВ в альтернативных приложениях использования ОП. Эффективность использования ОП с целью защиты от перегрузок заявок реального времени, с учетом известного закона сохранения времени ожидания для ДО $\sum_{l=1}^{L} R_l W_l = const$ [1], где $l = 1,\ldots, L$ - уровни ОП, R_l - загрузки уровней *l*, W_l - среднее время ожидания заявок класса *l*, дает соответственно выигрыш для заявок РМВ - ОП 1 (для них имеем ОЦП) и проигрыш для заявок ОП 2 (пусть между собою для них также ОЦП) или ДО [(1,2,…,*S*), (*S* +1,...,*M*)]

$$\Delta W = W_2 - W_1 = C\left[\frac{1}{(1-R_1)(1-R_2)} - \frac{1}{1-R_1}\right],\ C = 0{,}5\sum_{l=1}^{L=2} \lambda_l \gamma_l^{(2)},$$

при обработке заявок с ОП по схеме, приведенной на рис. 2.17, где в систему поступает *M* простейших потоков с интенсивностями $\lambda_1,\ldots,\lambda_M$ и средними длительностями обслуживания $\gamma_1,\ldots,\gamma_M$ со вторыми начальными моментами $\gamma_1^{(2)},\ldots,\gamma_M^{(2)}$; $R_1 \sum_{m=1}^{S} R_m$, $R_1 \sum_{m=S+1}^{M} R_m$ - соответственно загрузки, создаваемые потоками различных ОП (обслуживание заявок каждого класса реализуется на основе дисциплины FIFO). Исследуемая характеристика для случая $L = 2$, $S = M/4$, $S = M/2$, $M = 256$ приведена на рис. 2.18, где видно, что времена ожидания заявок монотонно убывают с ростом ОП.

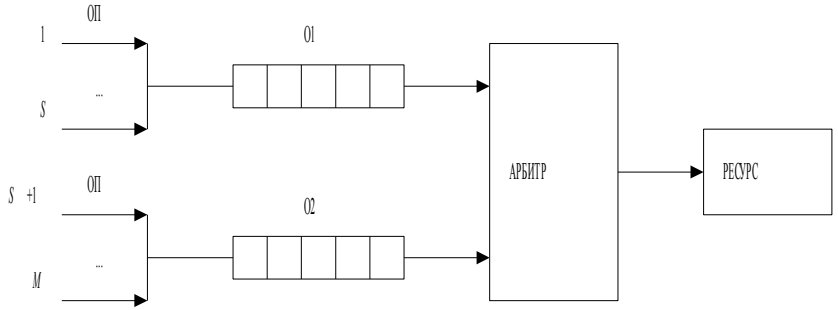


Рис. 2.17

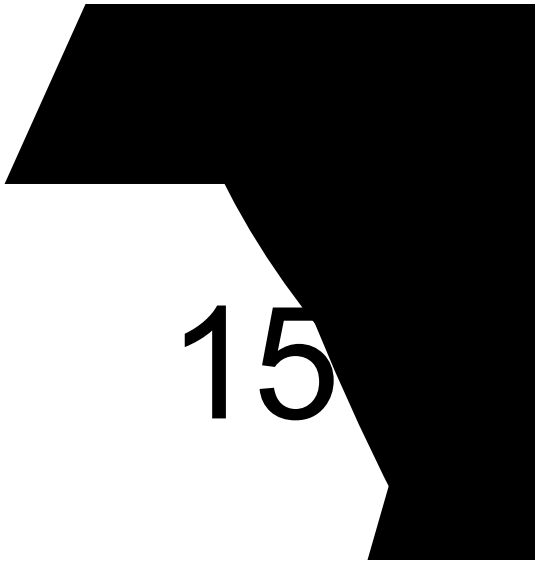


Рис. 2.18

Важной является оценка выигрыша заявок реального времени именно по параметру $T_{го_m}$. В предположении, что такие заявки имеют бесприоритетное обслуживание, причем $T_{гпп_m}$, $T_{гр_m}$ соответственно совпадают для всех M (бесприоритетное расписание) абонентов, для схемы, приведенной на рис. 2.17, имеем выигрыш, получаемый для заявок реального времени

$$\Delta T_{го_m} = (M - S)(T_{гпп} + T_{гр})$$

или, соответственно в относительных единицах $\delta T_{го_m} = M/S$. Это говорит о том, что в предположении $T_{го_m} > T_{грз_m}$, за счет введения ОП для обслуживания заявок оперативной обработки, требования к качеству обслуживания заявок реального времени можно обеспечить с экономией производительности ресурса системы (в рассматриваемом случае – «узкого места») в M/S раз, причем в отличие от метода построения приоритетного расписания здесь выигрыш может быть получен и для системы реального времени с бесприоритетным обслуживанием.

В предположении, что $T_{грс_m} = T_{грз_m} + L_{г_m} T_{гд_m}$ и $T_{грз_m} = kT_{го_m}$ имеем $\delta T_{грс_m} = (M + k)/(S + k)$. Зависимости $\delta T_{грс_m} = f(M - S)$ представлены на рис. 2.19, который иллюстрирует целесообразность

приоритетного (в смысле ДОСП) обслуживания заявок реального времени при условиях, во-первых, $k \leq 1$, во-вторых, $S << M$.

Рис. 2.19

Эффективность введения в систему внеочередных заявок ОП (как отмечалось, можно рассматривать как подход, альтернативный способу обслуживания с абсолютными приоритетами для распределенных ВС, для которых прерывание взаимодействия абонента с ресурсом связано с большими временными потерями и дополнительной сложностью децентрализованного управления доступом к ресурсу) имеет смысл оценивать мерой внеочередности их обслуживания или предоставляемым для них значением параметра $T_{го_m}$, составляющим для одной заявки ОП $T_{го_m} < 2T_{гпп} + T_{гзр}$ (при двух и более заявок ОП уже необходимо учитывать реализуемую для них ДО).

Для оценки характеристик обслуживания заявок при бесприоритетном кодовом управлении множественным доступом в рассмотренных ранее предположениях можно использовать известное выражение [1]

$$W = \frac{\sum_{m=1}^{M} \lambda_m \gamma_{m_c}^2 (1 + \upsilon_{m_c}^2)}{2(1-R)}, \quad \upsilon_{m_c} = \sqrt{\frac{\gamma_{m_c}^{(2)} - \gamma_{m_c}^2}{\gamma_{m_c}^2}}$$

Это обусловлено тем, что в отличие от методов опроса очередей здесь отсутствует передача прав в явном виде, а γ_{m_c} складывается из двух составляющих T_{p_m} и T_{nn_m} (где в рассматриваемом случае $T_{nn_m} = const$), при этом на параметр W не сказывается несимметричность загрузки системы.

Очевидно, что при реализации кодового управления множественным доступом к ресурсу, возможности повышения эффективности доступа заложены в решении задачи оптимального кодирования приоритетов

абонентов системы и в выборе эффективных способов передачи прав на занятие ресурса между абонентами системы. Данным аспектам

диспетчеризации заявок при кодом управлении множественным доступом будут посвящены следующие разделы.

РАЗДЕЛ 3. МЕТОДЫ КОДИРОВАНИЯ ПРИОРИТЕТОВ ЗАЯВОК НА ОБСЛУЖИВАНИЕ

3.1. Принципы оптимального кодирования ОП абонентов

Данный раздел посвящен исследованию проблемы повышения эффективности кодового управления множественным доступом к связному ресурсу, что актуально для альтернативных приложений ЛВСРВ и ЛВСКО. Это обусловливается небольшими размерами передаваемых пакетов данных абонентами при занятии канала связи (а в ряде приложений, например в технологии АТМ, и отказом от контроля передаваемых данных - контролируется только управляющая часть ячейки), что делает затраты времени на арбитраж даже при кодовом управлении множественным доступом к связному ресурсу сопоставимыми с продолжительностью взаимодействия абонентов при получении доступа к каналу связи.

Замечание. Здесь и далее говорим о кодировании ОП абонентов ЛВС, однако аналогично задача формулируется и решается при кодировании ОП заявок на обслуживание.

Ранее показано, что при кодовом управлении множественным доступом осуществляется арбитраж заявок по каждому разряду кода ОП n_m, $n_m = 1,\ldots, N_m$ (число которых при равномерном коде $[\log_D M]$, где D - основание кода), на каждый из которых приходятся удельные затраты $T_{уп_m}$, откуда получаем

$$T_{гпп_m} = N_m T_{уп_m} \quad (3.1)$$

Выражение (3.1) определяет альтернативные пути уменьшения затрат времени на управление множественным доступом при кодовом управлении $T_{гпп_m}$, задаваемые изменением параметров N_m и $T_{уп_m}$. Первый из них может быть сформулирован в постановке задачи оптимального кодирования: для заданного вероятностного ансамбля $\{ M, p_m, \sum_{m=1}^{M} p_m = 1 \}$ [4], где p_m - распределение вероятностей заявок на множестве M, найти множество

кодов ОП длиной N_m, $m = 1,\ldots,M$, обращающих в минимум среднюю скорость кодирования

$$R = \sum_{m=1}^{M} p_m N_m$$

Для решения рассмотренной классической задачи теории информации может быть использован метод Хаффмена [4]. Альтернативный подход - снижение потерь $T_{гпп_m}$ за счет уменьшения $T_{уп_m}$ связан с выбором механизма передачи прав на доступ к ресурсу между абонентами системы. Этот вопрос будет рассмотрен в следующем разделе.

Особенностью иной постановки задачи оптимального кодирования ОП будет необходимость учета более высокого приоритета уровня «1» над уровнем «0» (пусть исходно задали таким образом) в каждом разряде кода ОП, т.к. значением «1» в разряде кода приоритета абоненту предоставляется право занять ресурс, по значению «0» не предоставляется. Множество абонентов с заданной на нем вероятностью *p* требований общего ресурса, одинаковой для всех абонентов системы, называется **математической моделью равномерного поэтапного кодирования/декодирования** (или просто поэтапного кодирования) пользователей общих ресурсов: {*M*, *N*, $0 \le p \le 1$, $m = \overline{1, M}$, $n = \overline{1, N}$}. Величина *p* здесь является характеристикой интенсивности поступления в систему требований ресурса.

Под величиной $P_n(p)$ будем понимать вероятность однозначного декодирования *m*-го абонента по *n* из *N* разрядам кода приоритета при вероятности p обращения абонентом к общему ресурсу системы. Рассмотрим изменение величины $P_n(p)$ при $n = 1;2;3$, соответственно $P_{n=1,2,3}(p)$ при изменении *p*. Зависимости $P_n(p) = f(p)$ для случая $M = 8$ представлены на рис. 3.1 (формула для расчетов $P_n(p)$ приведена ниже), из которого следует, что при невысоких загрузках системы вероятность однозначного декодирования абонента не сильно изменяется с увеличением длины кодового слова, откуда может быть сделан вывод, что способ кодирования по всем *N* (имеется в виду равномерный код) разрядам кода приоритета в общем случае избыточен, особенно при малых загрузках системы.

$$P_n(p) = \frac{M}{2^n} p(1-p)^{\frac{M}{2^n}-1} \sum_{a=0}^{2^n-1} (1-p)^{\sum_{b=0}^{a} \frac{bM}{2^n}} \sum_{r=0}^{\sum_{c=0}^{2^n-1-a} \frac{cM}{2^n}} C^r_{\sum_{c=o}^{2^n-1-a} c} \frac{M}{2^n} p^r (1-p)^{\sum_{c=0}^{2^n-1-a} \frac{cM}{2^n} - r}$$

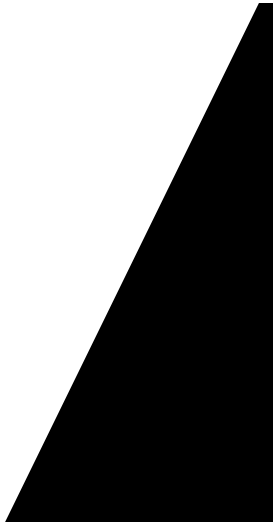


Рис. 3.1

Другими словами, выше рассмотрен **принцип поэтапного кодирования**, состоящий в управлении множественным доступом поэтапно - по частям кода приоритета, с увеличением длины кода на каждом этапе, и, как следствие, при возможности неоднозначного декодирования абонента по используемой части кода приоритета - на всех этапах, кроме последнего. Нетрудно показать, что скорость поэтапного кодирования максимальна при использовании всех *N* разрядов в коде приоритета.

Замечание. Особенностью реализации принципа поэтапного кодирования является возможность предоставления права занять ресурс одновременно нескольким абонентам, что требует обнаружения конфликта информации в канале, что, в свою очередь, будет являться условием необходимости увеличения длины кода приоритета. Принцип поэтапного кодирования предполагает конфликтную передачу прав между абонентами системы на занятие ресурса.

Влияние условия - ОП «1» над «0» при занятии ресурса, которое необходимо учитывать при кодировании приоритетов абонентов, рассмотрим на простом примере. Определим вероятность однозначного декодирования абонента по одному разряду кода приоритета (имеется в виду - только по первому разряду $n = 1$, естественно в предположении, что здесь возможен конфликт), например для случая $M = 4$, т.е. рассмотрим следующие кодовые комбинации «0011», «0001», «0111». Имеем: - для комбинации «0011» $P_{n=1}(p) = 2p(1-p)$, - для комбинации «0001» $P_{n=1}(p) = p$, - для комбинации «0111» $P_{n=1}(p) = 3p(1-p)^2$.

Зависимости $P_{n=1}(p) = f(p)$ для рассматриваемых случаев представлены на рис. 3.2, из которого могут быть сделаны следующие выводы.

1. Эффективность способа кодирования зависит от изменения *p*.

2. Могут иметь место интервалы изменения величины *p*, когда неравномерное кодирование абонентов (неравное число единиц и нулей в кодовой комбинации) эффективнее равномерного кодирования (соответственно, число нулей и единиц в кодовой комбинации совпадает).

Из рис. 3.2 в частности видим, что на интервале изменения параметра $0 \le p \le p_1$ эффективнее код «0111», на интервале изменения $p_1 < p \le p_3$ - код «0011», на интервале изменения $p_2 \le p \le 1$ - код «0001».

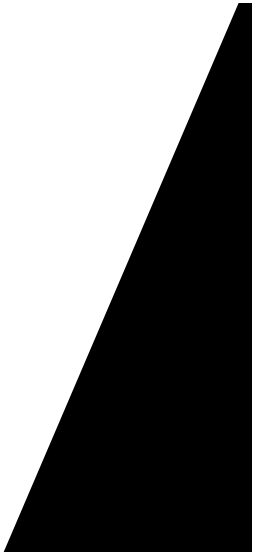


Рис. 3.2

Предельными случаями неравномерных кодов являются соответственно предельно неравномерные по «0» и по «1» коды, примеры которых для случая $M = 4$ представлены соответственно на рис. 3.3.а и 3.3.б, первый из которых эффективен при $p \to 1$, второй при $p \to 0$, причем предельно неравномерный по «0» код обеспечивает поэтапный бесконфликтный арбитраж (бесконфликтное кодирование). Особенностью этих кодов будет максимальная эффективность соответственно при низкой и высокой загрузках системы. Другими словами, **предельно неравномерные коды**, по существу представляют собою предельные случаи кодирования, ориентированные, соответственно на низкую и высокую загрузки системы, и

потому содержащие соответственно максимальное (по «1») и минимальное (по «0») число единиц в каждом кодовом слове кода, а любой применяемый

на практике неравномерный код будет занимать промежуточное положение между равномерным кодом ОП и соответствующим предельно неравномерным по «0» или по «1» кодом. Равномерный код отличается совпадением числа 1 и 0 в каждом кодовом слове.

n/m	1	2	3	4
1	1	1	1	1
2	0	1	1	1
3	0	0	1	1
4	0	0	0	1

а. Предельно неравномерный код по "1"

n/m	1	2	3	4
1	0	0	0	1
2	0	0	1	0
3	0	1	0	0
4	1	0	0	0

б. Предельно неравномерный код по "0"

Рис. 3.3

Эффективность неравномерного поэтапного кодирования можно оценить по критерию «скорость однозначного поэтапного неравномерного кодирования по N_m разрядам»

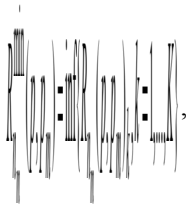
$$R_{N_m} = \sum_{n_m=1}^{N_m} n_m (P_{n_m}(p) - P_{n_m-1}(p))$$

где $P_{n_m}(p)$ - вероятность однозначного декодирования абонентов по разрядам n_m кода ОП.

Под **полной математической моделью поэтапного кодирования** приоритетов абонентов понимается математическая модель равномерного поэтапного кодирования, учитывающая параметр p_m или, соответственно, кодовая вероятностная модель, учитывающая параметр p. Или особенностью полной математической модели будет наличие уже двух параметров p и p_m различно влияющих на выбор оптимального способа кодирования: p - задает возможность эффективного поэтапного кодирования, в результате в качестве оптимального может быть получен неравномерный код; p_m - задает возможность эффективного полного кодирования, при этом в результате также в качестве оптимального может быть получен неравномерный код, причем в общем случае эти оптимальные неравномерные коды могут не совпадать.

В качестве критерия оптимальности также как и для модели поэтапного кодирования должен использоваться параметр R_{n_m}, учитывающий параметр p_m: $R_{n_m}(p, p_m)$ - математическое ожидание скорости однозначного декодирования пользователя. Условием

оптимальности выбора способа кодирования будет $R_{n_m}(p, p_m) \to \min$ или с учетом дискретности исходного набора альтернативных вариантов кодирования имеем



где $R_{n_m}(p, p_m)_k$ - параметр $R_{n_m}(p, p_m)$ для *k*-го способа кодирования из сравниваемого множества вариантов *K*, где

$$R_{n_m}(p, p_m) = \sum_{n_m=1}^{N_m} (P_{n_m}(p, p_m) - P_{n_m-1}(p, p_m))$$

Заметим, что в отличие от соответствующих формул, приведенных ранее, здесь уже учитывается не только число единиц в кодовых комбинациях разряда кодов ОП абонентов, но и то, каким абонентам (соответственно, с какими значениями параметра p_m) принадлежат значения единиц.

Утверждение. Всегда можно построить оптимальный код, который будет эффективнее случайного назначения кодов ОП.

Доказательство. Нетрудно показать, что метод случайного задания кодов ОП в среднем имеет характеристики обслуживания ниже, чем метод равномерного поэтапного кодирования, который неэффективен при низких и высоких загрузках системы. В области эффективного использования равномерного кода случайное задание кодов ОП неэффективно, т.к. не гарантирует однозначного декодирования абонента по N_m разрядам кода.

Данное утверждение приведено с той целью, чтобы показать, что в рамках концепции кодового управления, в частности при оптимальном кодировании ОП абонентов, всегда можно получить конфликтный способ управления множественным доступом в реальном времени, который будет эффективнее для конкретных приложений, чем широко используемые в ЛВСОО случайные методы [8, 11], основным достоинством которых можно считать простоту реализации. Действительно, здесь не требуется поддерживать и, как следствие, восстанавливать в случае искажения какой-либо очереди передачи прав. Это делает данный метод, регламентируемый, например стандартами IEEE 802.3, TOP [8, 11] сегодня наиболее широко применяемым в ЛВСОО и ЛВССОО где отсутствуют жесткие ограничения на продолжительности обработки заявок.

3.2. Принципы приоритетного кодирования ОП абонентов

Определяющим при реализации оптимального способа кодирования является эффективное использование общего ресурса абонентами ВС. Однако, как отмечалось ранее - это следует из системы (1.1), изменением значения T_{nn_m} можно задавать (изменять) приоритет абонента, влияя при

этом на величину гарантированного времени обслуживания заявки $T_{гo_m}$. Это актуально при небольшой величине информационного кадра. Однако, как отмечалось выше, сегодня выполнение данного условия характерно как для управляющих, так и для информационных ВС. Например, при $M = 256$, в рамках концепции АТМ получаем, что затраты T_{nn_m} составляют 15% затрат времени на передачу информационного кадра (8/53). Очевидно, что в рамках кодирования ОП абонентов может быть реализован принцип приоритетного кодирования, который состоит в кодировании ОП более важного абонента более коротким кодовым словом, с целью минимизации продолжительности доступа к ресурсу именно более приоритетного пользователя. Такой подход, справедлив из соображений учета важности абонентов, но в общем случае не дает оптимального кода в смысле эффективного использования ресурса, т.к. при кодировании приоритетов здесь уже учитываются не параметры потока требований ресурса, а исключительно важность абонентов друг относительно друга. Таким образом под **приоритетным** будем понимать такое кодирование приоритетов абонентов в системе с множественным их доступом к общим ресурсам, при котором целью кодирования является учет относительной важности абонентов.

Замечание. По существу, это также метод оптимального кодирования, но эффективность использования ресурса здесь уже определяется не характеристиками входного потока его требований, а возможностью назначения приоритетов реального времени (назначением приоритетов в соответствии с детерминированной моделью обслуживания в реальном времени).

Постановка задачи приоритетного кодирования выглядит следующим образом. Пусть задано множество абонентов системы $\{M,\ m = 1,\ldots, M\}$ характеризуемых различной важностью, которую требуется учитывать реализацией заданных соотношений длин кодов ОП N_m, $m = 1,\ldots, M$. Требуется осуществить кодирование ОП - сопоставить множеству абонентов M множество кодов приоритетов $\{X_{mN_m},\ m = 1,\ldots, M\}$, каждое из которых имеет исходно заданную длину N_m, обеспечивающее взаимно однозначное отображение множества абонентов во множество кодовых слов и наоборот.

Рассмотрим метод приоритетного кодирования или метод построения в ВС приоритетных кодов ОП абонентов. Очевидно, что прежде всего при решении рассматриваемой задачи кодирования необходимо определить допустимые соотношения длин кодов приоритетов в системе, что позволит проверить возможность построения кода приоритета с заданным соотношением длин кодов, а при невозможности, скорректировать задаваемое соотношение. Другими словами, имея исходно заданные соотношения длин

кодов, необходимо определить возможность построения приоритетного кода в принципе.

Например, возможно ли осуществить приоритетное кодирование с основанием 2 ОП семи абонентов системы кодами приоритетов с длинами (числом разрядов): 2,2,3,3,3,3,4 ? Имея положительный ответ на данный вопрос, уже можно приступать к решению собственно задачи кодирования.

Возможность построения приоритетного кода с заданными длинами кодов ОП абонентов (в общем случае неравномерного) в полной мере определяется свойствами однозначного кодирования. Поэтому здесь можно использовать неравенство Крафта для префиксных кодов [4], которое гласит, что для существования префиксного кода в алфавите объема D с длинами кодов N_m, необходимо и достаточно, чтобы выполнялось условие:

$$\sum_{m=1}^{M} D^{-N_m} \le 1 \qquad (3.2)$$

С использованием этого неравенства может быть сформулировано **условие допустимости соотношений длин кодов приоритетов абонентов** при их приоритетном кодировании кодом с основанием D - может быть выбрано соотношение длин кодовых слов N_m, $m = 1,\ldots,M$, при котором выполняется неравенство (3.2).

Вернемся к нашему примеру и проверим выполняемость неравенства Крафта. Имеем $2^{-2}+2^{-2}+2^{-3}+2^{-3}+2^{-3}+2^{-3}+2^{-4} > 1$, т.е. префиксный код не может быть построен. Изменим условие, рассмотрим соотношение длин кодовых слов: 2,2,3,3,3,4,4, имеем $2^{-2}+2^{-2}+2^{-3}+2^{-3}+2^{-3}+2^{-4}+2^{-4} = 1$, неравенство Крафта выполняется, или может быть построен префиксный код (равенство 1 означает, что данный код не может быть улучшен).

Метод приоритетного кодирования, учитывающий задаваемое соотношение длин кодовых слов, отражающее приоритетность абонентов системы основывается на следующих положениях:

- кодирование приоритетов абонентов начинается с младшего разряда с переходом к более старшему разряду кодового слова;

- анализируемому разряду кода приоритета с меньшей заданной длиной кодового слова присваивается значение «1» (более приоритетное значение для арбитража требований ресурса), с большей длиной кодового слова - значение «0»;

- при кодировании учитывается, что Q абонентов с равной заданной длиной кода приоритетов при кодировании должны различаться в $[\log_D Q]$ старших разрядах кодового слова, где $[a]$ - большее целое числа a.

Применение метода для получения кода ОП абонентов для рассмотренного выше примера проиллюстрировано на рис. 3.4.

n/m	1	2	3	4	5	6	7
1	1	1	0	0	0	0	0
2	0	1	1	1	0	0	0
3			0	1	1	0	0
4						0	1

Рис. 3.4

3.3. Задачи и методы динамического кодирования ОП абонентов

Будем говорить, что кодирование приоритетов абонентов, предполагающее неизменное соответствие кодов приоритетов абонентам в течение всего времени функционирования системы, называется **статическим кодированием** ОП абонентов ВС. Однако изложенная концепция обслуживания по расписаниям предполагает смену ОП в процессе функционирования системы, а следовательно и смену по расписанию кодов ОП абонентов при реализации децентрализованного кодового управления множественным доступом к ресурсу. Будем говорить, что кодирование ОП абонентов, предполагающее изменение ОП в соответствии с какой-либо заданной очередностью в процессе функционирования системы, называется **динамическим кодированием** ОП абонентов ВС. Другими словами, статический код с заданными очередностью и моментами времени изменения его при функционировании системы называется динамическим кодом.

Выше речь шла об оптимальном (либо в смысле эффективности использования ресурса, либо в смысле приоритетности доступа к ресурсу отдельными абонентами) статическом кодировании, где в качестве критерия оптимального выбора статического кода использовалась скорость (средняя скорость) кодирования. При изменении же кодов ОП в процессе функционирования ВС возникает проблема оптимального изменения исходных кодов ОП, реализующих достоинства исходных оптимальных кодов в процессе функционирования системы, или принципов оптимального динамического кодирования ОП абонентов ВС.

Очевидно, что динамический код можно рассматривать как конечное множество статических кодов, где каждый из статических кодов (пусть в динамическом коде их M, $m = 1,\ldots, M$) характеризуется определенной средней скоростью кодирования R_m . С учетом этого, можем ввести понятие **средней скорости динамического кода** R_∂ , которая может рассматриваться как математическое ожидание скоростей кодирования применяемых статических кодов с учетом вероятности их появления в динамическом коде,

замкнутый граф которого содержит M вершин - моменты времени t_s, $s = 1,\ldots,M$, каждая

из M вершин в замкнутом графе динамического кода встречается ровно K_m раз

$$R_д = \sum_{m=1}^{M} p_m R_m \ ,$$

где

$$p_m = \frac{K_m}{K_m^{\max}} \ , \ K_m^{\max} = \sup\{K_m, m = 1,...,M\} \ ,$$

причем с целью эффективного использования ресурса оптимальный динамический код следует выбирать из условия: $R_д \to \min$.

С учетом сказанного, может быть сформулирован следующий **принцип оптимальности динамического кода** - динамический код будет реализован в системе оптимальным образом в том случае, если в каждый момент времени t_s приоритеты абонентов системы закодированы оптимальными статическими кодами. С учетом сформулированного критерия оптимальности и предложенных ранее методов статического кодирования рассмотрим методы оптимального динамического кодирования.

В общем случае можно выделить следующие альтернативные способы изменения статического кода: сквозное и групповое динамическое изменение кодов. Под **сквозным изменением кодов** приоритетов абонентов понимается циклическое изменение (уменьшение , либо увеличение) на единицу значений кодов приоритетов при каждом изменении момента времени t_s . Пример графа динамического кода со сквозным кодированием для случая $M = 4$ представлен на рис. 3.5. Средняя скорость динамического кода $R_д$ в рассматриваемом случае совпадает со скоростью равномерного статического кода R.

Под **групповым изменением кодов** приоритетов абонентов понимается такое изменение кодов, при котором сохраняются исходные группы (по каждому отдельно взятому разряду) абонентов, имеющие одинаковые значения этого разряда кода приоритета, причем, если значение разряда изменяется на противоположное, оно изменяется у всех абонентов группы.

Рассмотрим алгоритм группового изменения статических кодов, регламентирующий очередность смены кодовых слов.

1. Инвертируется первый (младший) разряд кода приоритета.
2. Инвертируется второй разряд кода приоритета.
3. Инвертируется первый разряд кода приоритета.
4. Если код приоритета содержит два разряда, инвертируется второй разряд, переход к п.п. 1.
5. Инвертируется третий разряд кода приоритета.
6. Выполняются п.п. 1 - п.п. 3 алгоритма.

7. Если код приоритета содержит три разряда, инвертируется третий разряд, переход к п.п. 1.

8. Инвертируется четвертый разряд кода приоритета.

. . . . . .

k-1. Если *k* разрядов в коде приоритета, инвертируется *k* разряд, переход к п.п. 1.

k. Инвертируется *k*-й разряд.

k+1. Выполняются п.п. 1 - п.п k алгоритма.

k+2. Если $k+1$ разрядов в коде приоритета, инвертируется $k+1$-й разряд, переход к п.п. 1.

. . . . . .

Пример графа динамического кода при групповом кодировании для случая $M=4$ представлен на рис. 3.6.

Рассмотрим условия применения альтернативных способов динамического кодирования, с учетом того, что в зависимости от исходного задания входного потока требований ресурса в качестве оптимальных могут быть получены следующие статические коды: равномерный и неравномерный, полный и неполный (поэтапное кодирование/декодирование).

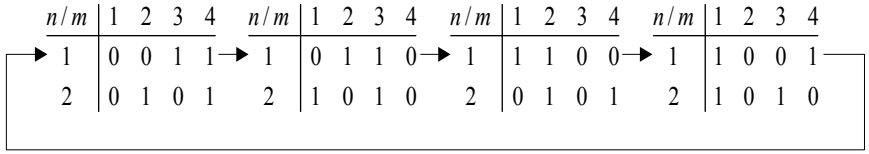


Рис. 3.5

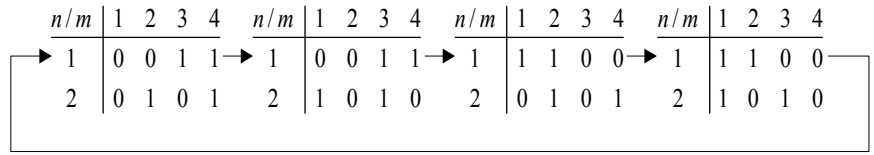


Рис. 3.6

Утверждение. Способ сквозного изменения кода при динамическом кодировании позволяет получить оптимальный динамический код при использовании в качестве статического равномерного полного или неполного кодов.

Доказательство. При циклическом изменении кода из одного полного равномерного кода всегда получается другой полный равномерный код. В случае неполного равномерного кода при циклическом изменении кода из равномерного кода получается равномерный код, причем в результате такого изменения сохраняется число единиц и нулей в каждой кодовой комбинации любого отдельно взятого разряда кодов приоритетов абонентов (сохраняется

число и изменяется их принадлежность, но для неполного кодирования важным является именно сохранение числа единиц и нулей).

Утверждение. Способ группового изменения кода при динамическом кодировании позволяет получить оптимальный динамический код при использовании в качестве статического неравномерного полного кода.

Доказательство. Особенностью группового изменения кода является то, что сохраняются исходные группы абонентов, одинаково закодированных по отдельно взятому разряду. Это означает, что сохраняется соотношение длин приоритетов между абонентами в различные моменты времени t_s. В рассматриваемом случае (входной поток требований определен параметром p_m) важно сохранение соотношения длин между группами абонентов , что здесь и обеспечивается.

Пример графа динамического кода для случая $M = 6$, при использовании в качестве оптимального статического неравномерного полного кода, представлен на рис. 3.7.

Замечание. В данном способе динамического кодирования внутри отдельных групп может использоваться как сквозное, так и групповое кодирование.

Способ группового динамического кодирования, позволяющий сохранять исходные длины кодов приоритетов групп абонентов в процессе функционирования ВС будем называть **кодированием с сохранением длины кодов групп**.

Утверждение. Способ группового изменения кода при динамическом кодировании позволяет получить оптимальный динамический код при приоритетном статическом кодировании ОП абонентов.

Доказательство. Как отмечалось, особенностью группового изменения кода является то, что здесь сохраняются исходные соотношения длин между группами абонентов. Именно это свойство кода должно быть реализовано при приоритетном кодировании, т.к. исходной длиной кода здесь реализуется приоритет абонента.

Рассмотрим случай, когда в качестве оптимального выбирается неполный (поэтапное кодирование) код. Для реализации динамического кода в этом случае не может эффективно использоваться ни способ сквозного изменения кодов, ни рассмотренный способ группового изменения кодов приоритетов, т.к. особенностью рассматриваемого статического кода будет следующее - в предположении, что различные абоненты имеют равные вероятности занятия ресурса и поток заявок определяется лишь его интенсивностью, важным становится уже не длина кодов приоритета групп абонентов , а то, сколько абонентов имеют значение «1» в каждом отдельно взятом разряде кода приоритета (или число единиц в каждом кодовом слове).

Здесь также должен быть реализован принцип группового кодирования, однако особенность получения динамического кода здесь состоит в том, что групповое кодирование осуществляется и внутри каждой

исходной группы абонентов (в групповом коде, приведенном выше, допускалось сквозное кодирование внутри групп).

Пример группового динамического кода, сохраняющего число единиц в отдельно взятом разряде для случая $M = 6$ приведен на рис. 3.8.

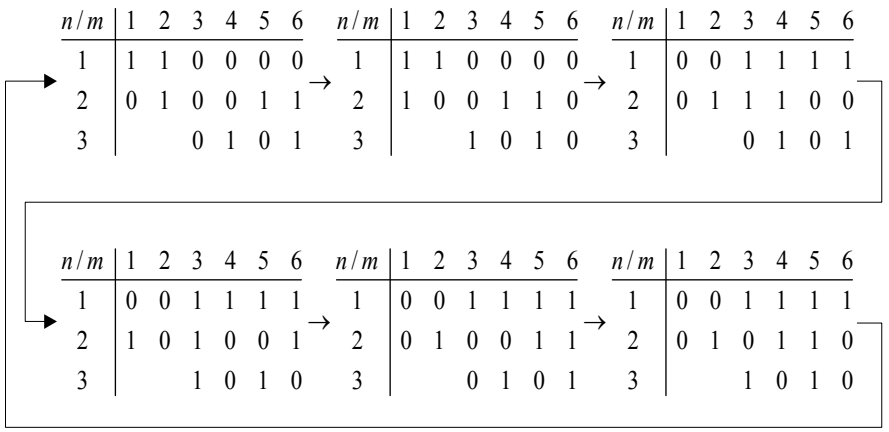


Рис. 3.7

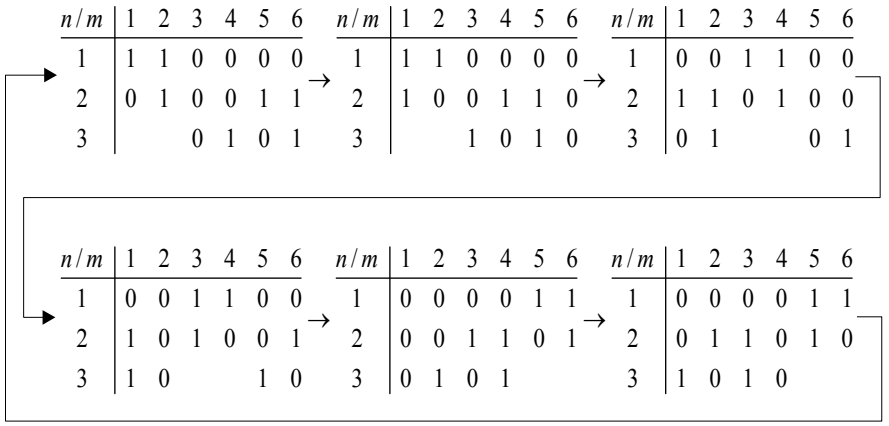


Рис. 3.8

Способ группового динамического кодирования, позволяющий сохранять исходное число единиц в отдельно взятом разряде будем называть **кодированием с сохранением числа единиц в отдельно взятом разряде**.

Утверждение. Способ группового изменения кода при динамическом кодировании с сохранением числа единиц в отдельно взятом разряде позволяет получить оптимальный динамический код при использовании в качестве статического неравномерного неполного кода при задании потока требований параметром *p*.

Доказательство. Рассмотренный способ кодирования позволяет сохранять число единиц в каждом отдельно взятом разряде кода приоритета (или в каждом кодовом слове), соответственно, исходную неравномерность кода, определяемую соотношением числа единиц и нулей в каждом отдельно взятом разряде. Следовательно, в любой момент времени статический код здесь будет оптимальным.

Кроме рассмотренных методов динамического кодирования, учитывающих либо свойства входного потока требований ресурса, либо приоритетные свойства абонентов, на практике может быть применен метод кодирования, обеспечивающий минимальную гарантированную скорость динамического кодирования. **Под гарантированной скоростью динамического кодирования** приоритетов абонентов $R_г$ будем понимать суммарную скорость статического кодирования приоритетов, определяемую в моменты времени t_s

$$R_г = \sum_{s=1}^{M} R_{г_s} .$$

Физический смысл гарантированной скорости динамического кода следующий - задает максимальное число разрядов, по которым должен быть произведен арбитраж при предоставлении прав занять ресурс *m*-му абоненту, при наличии требований ресурса у всех *M* абонентов системы. С учетом коэффициентов повторяемости моментов времени в цикле расписания t_s : K_s , $s = 1,\ldots, M$, гарантированная скорость динамического кодирования приоритета *m*-го абонента $R_{г_m}$

$$R_{г_m} = R_{г_{s=m}} + \sum_{s=1}^{M} [K_s R_{г_s}, s \neq m] .$$

Метод динамического кодирования с минимальной гарантированной скоростью основывается на следующих утверждениях.

Утверждение. Выполнение условия $R_г(R_{г_m}) \to \min$ достигается в том случае, когда $R_г = M$, соответственно

$$R_{г_m} = M(K_{s=m} + \sum_{s=1}^{M} [K_s, s \neq m]) .$$

Доказательство. Для того, чтобы один код мог быть отделен от другого минимум должен быть произведен арбитраж требований по одному разряду кода приоритета. Подставив $R_{г_s} = R_{г_m} = 1$ имеем

$$R_г = M ,$$

$$R_{г_m} = M(K_{s=m} + \sum_{s=1}^{M}[K_s, s \neq m]).$$

Утверждение. Условие $R_г(R_{г_m}) \to \min$ выполняется в том случае, когда динамический код реализуется по следующему правилу: в любой

момент времени в старшем разряде кода приоритета (именно с него начинается арбитраж требований) только наиболее приоритетного в этот момент времени абонента находится «1», у остальных абонентов системы в этом разряде «0».

Доказательство. Утверждение доказывается тем, что при рассмотренных условиях арбитраж требований ресурса может быть осуществлен только по одному (старшему) разряду кода приоритета, т.е. выполняется условие: $R_{c}(R_{c_m}) \to \min$.

Пример построенного с учетом приведенных правил динамического кода для случая $M = 4$ представлен на рис. 3.9. Получили неравномерный код со значением параметра $R_c = 4 \times 1 = 4$ (бита) или в 2 раза меньше, чем для равномерного кода.

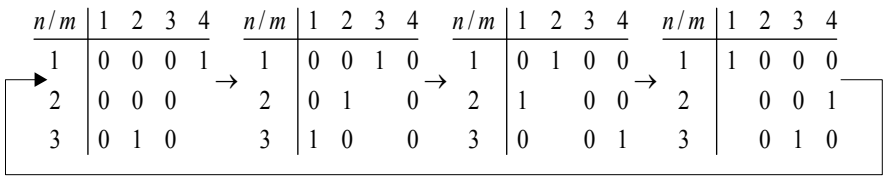


Рис. 3.9

Очевидно, что средняя скорость кодирования при реализации данного способа увеличивается не более чем на 1 (вводится еще один разряд).

Достоинством рассмотренного метода динамического кодирования будет то, что при весьма незначительном ухудшении статических скоростей кодирования в моменты времени t_s (в общем случае здесь имеем уже неоптимальные коды) не более чем на 1, может быть достигнуто весьма существенное улучшение параметра R_m, крайне важного для систем реального масштаба времени, определяющего возможность построения системы.

Оценим получаемый выигрыш от использования метода. Если длина наиболее приоритетного кода в моменты времени t_s, $s = 1,\ldots,M$ составляет N_s (в общем случае неравномерный код) имеем

$$M\Delta R_c = \sum_{s=1}^{M}(N_s - 1)$$

или

$$M\Delta R_{c_m} = \sum_{s=1}^{M}[K_s(N_s - 1) + K_m(N_m - 1), s \neq m] .$$

Замечание. При реализации приоритетной дисциплины обслуживания требований ресурса рассмотренный метод динамического кодирования

реализуется также. Отличие здесь состоит лишь в другой очередности и частоте смены моментов времени t_s , $s = 1,\ldots, M$.

В завершение исследования вопросов кодирования приоритетов абонентов ВС сформулируем и докажем весьма важное утверждение.

Утверждение. При реализации в системе ДОСП динамическая смена ОП абонентов ВС по расписанию должна происходить как при занятии ресурса заявкой реального времени, так и при занятии ресурса заявкой оперативного обслуживания (т.е. при занятии ресурса любой заявкой).

Доказательство. Утверждение доказывается тем, что в противном случае - при изменении ОП только занятием ресурса заявкой реального времени, параметр $T_{го_m}$ приоритетной заявки реального времени увеличится на ($T_{гпп} + T_{гр}$) - считаем их соответственно совпадающими для всех абонентов ВС.

Реализация передачи прав на занятие ресурса в ЛВС в соответствии с оптимальными кодами ОП рассматривается в следующем разделе.

РАЗДЕЛ 4. МЕТОДЫ КОДОВОГО УПРАВЛЕНИЯ МНОЖЕСТВЕННЫМ ДОСТУПОМ. ПРИНЦИПЫ УНИФИКАЦИИ

4.1. Конфликтные методы кодового управления

Как ранее отмечалось, в общем случае принцип кодового управления предполагает поразрядное сравнение кодов ОП при управлении доступом к ресурсу при котором в бит-синхронном, либо ином режиме канал, который по приему должен быть доступен всем абонентам системы, занимается передачей зонда абонентами. Поступление из канала значения зонда «1» абоненту, имеющему в анализируемом разряде менее приоритетное значение «0», исключает данного абонента из дальнейшей «борьбы за ресурс» проводимой процедуры кодовой передачи прав с целью бесконфликтного занятия ресурса.

Альтернативные способы кодирования ОП - полное и поэтапное, обусловливают целесообразность использования двух видов зонда, в качестве которого может использоваться либо разряд кода ОП (уровни логических «0» или «1») при полном кодировании, либо собственно информационный кадр при поэтапном кодировании - здесь при «борьбе за ресурс» абонентами, имеющими «1» в анализируемом разряде кода ОП, выдается в канал уже

информационный кадр (абоненты, имеющие «0» в разряде кода канал не занимают). В обоих случаях возможен конфликт данных в канале - в первом

случае - разрядов зонда, во втором - информационных сообщений. Так как поэтапное кодирование не гарантирует однозначного декодирования абонента по части кода при информационном (в общем случае конфликтном) методе реализуется процедура обнаружения конфликта, например по искажению несущей частоты - аналогично тому, как это делается для случайных методов управления множественным доступом CSMA-CD в ЛВСОО. По способу зондирования канала рассмотренные методы можно разделить на зондовый конфликтный метод и информационный конфликтный метод.

Под **зондовым конфликтным методом кодового управления множественным доступом** понимается метод, реализующий механизм зондирования канала специальными сигналами, соответствующими логическим значениям разрядов кода приоритета абонентов при полном (равномерном, либо неравномерном) кодировании, с возможностью обнаружения наложений (конфликтов) разрядов зонда в канале.

Замечание. При кодировании приоритетов абонентов кодами, имеющим основание 2 (частный, но наиболее просто реализуемый случай), разряд зонда при «борьбе за ресурс» в канал могут выдавать лишь абоненты, имеющие значение этого разряда «1». Имеющие в разряде «0» канал не занимают, а при получении из канала «1» прекращают «борьбу за ресурс».

Под **информационным конфликтным методом кодового управления множественным доступом** понимается метод, реализующий механизм зондирования канала непосредственно информационными сообщениями абонентами, имеющими в разряде кода приоритета, по которому осуществляется «борьба за ресурс», «1», при поэтапном (равномерном, либо неравномерном) кодировании, с возможностью обнаружения наложений (конфликтов) информационных кадров в канале по искажению несущей частоты.

В общем случае возможны два способа передачи полномочий по каналу связи - асинхронный и синхронный. Под **асинхронным** способом передачи полномочий понимается передача права занять ресурс между абонентами сигналом (сигналами, в частности коротким пакетом - маркером), поступающим по каналу связи, либо по специально отведенным под эти цели управляющим линиям (линиям арбитража), под **синхронным** способом передачи полномочий понимается передача прав временными интервалами, синхронизирующими начало передачи информации каждым абонентом системы.

В частности, после обнаружения конфликта в канале и соответствующем переходе к арбитражу по следующему разряду кода приоритета, при асинхронной передаче прав перед началом зондирования в канале должен появиться сигнал, означающий переход к зондированию по следующему разряду, при синхронной передаче прав переход к зондированию по следующему разряду осуществляется через фиксированный

интервал времени после завершения зондирования по предыдущему разряду кода приоритета абонентов.

Утверждение. Для зондовых конфликтных способов целесообразно применять только синхронную передачу прав.

Доказательство. Утверждение доказывается тем, что в данном случае абоненты зондируют канал в синхронном режиме, при котором интервалы отсчитываются от момента начала зондирования по соответствующему разряду, последовательно по всем разрядам кода (полный код). При этом не имеет никакого смысла вводить кроме синхронной (которая и так имеет место) еще и асинхронную передачу прав.

Утверждение. Конфликт зондов в канале - это единственный способ передачи полномочий в соответствии с полным (не предельно неравномерным по «0») кодом.

Замечание. Строго говоря, и при информационном конфликтном способе может быть реализована передача полномочий в соответствии с полным кодом. При этом в случае отсутствия конфликта при передаче информационного сообщения, выдачу сообщения в канал надлежит прекратить, с целью последующей передачи полномочий , что естественно противоречит здравому смыслу.

Доказательство. С учетом изложенного выше, можем говорить, что имеют место лишь два способа передачи полномочий: асинхронный и синхронный. В данном случае одним из них полномочия передаются сразу нескольким абонентам Возникает проблема проведения арбитража между ними. При реализации синхронного арбитража имеем предельно неравномерный код. При любом способе передачи полномочий асинхронно без конфликта также имеем предельно неравномерный код, т.е. конфликт здесь обязателен, причем этот конфликт должен решать задачу арбитража среди конфликтующих по данному разряду кодового слова абонентов. Если имеем конфликт информационных сообщений, то получаем поэтапное кодирование. Из сказанного имеем, что для управления множественным доступом в рассматриваемом случае (полное кодирование) должен быть реализован конфликт в канале, но не информационных сообщений или конфликт каких-либо сигналов (зондов) по которому могут быть различимы (проведен арбитраж) конфликтующие абоненты системы.

В отличие от информационного конфликтного способа, где возможны два состояния канала при «борьбе за ресурс» - отсутствие передачи и передача информационного кадра, для зондового конфликтного способа уместно говорить о трех состояниях канала - отсутствие передачи, передача «0», передача «1». Поэтому здесь эффективно может использоваться кодирования приоритетов абонентов кодом по основанию 3, что естественно снижает потери времени на арбитраж при занятии ресурса. Пример кодов приоритетов абонентов в алфавите A = {*, 0, 1} для случая $M = 18$ представлен на рис. 4.1.

n/m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
2	*	*	*	0	0	0	1	1	1	*	*	*	0	0	0	1	1
3	*	0	1	*	0	1	*	0	1	*	0	1	*	0	1	*	0

Рис. 4.1

Зондирование канала абонентами реализуется следующим образом: по знаку «*» зонд не выдается, по знаку «0» выдается сигнал «0», по знаку «1» выдается сигнал «1». Пусть приоритетность знаков следующая: самый приоритетный «1», затем «0», затем «*». Наложение сигналов реализуется следующим образом: если зонд не выдается, абоненту со знаком «*» в разряде кода достаточно обнаружить любой сигнал в канале для снятия своего требования ресурса; абонент со знаком «1» не анализирует состояние канала после выдачи сигнала; абонент со знаком «0» выдает сигнал «0» в канал и проверяет его искажение (при этом сигналы «0» не должны искажать друг друга), при искажении сигнала «0», которое здесь может быть осуществлено сигналом «1», абонент снимает свое требование ресурса. Замечание. В первом разряде отсутствует знак «*», этим разрядом должно производиться уведомление о занятии канала, или какой-то сигнал должен выдаваться в канал.

Для использования способа в реальном времени должны быть реализованы следующие правила динамического кодирования.

Правило 1. Код приоритета абонентов изменяется в результате циклического прибавления к нему знака *.

Правило 2. Сложение знаков * и * дает: * + * = 0.

Правило 3. Сложение знаков 0 и * дает: 0 + * = 1.

Правило 4. Сложение знаков 1 и * дает: 1 + * = ** (перенос знака * со сложением в следующий разряд).

Правило 5. При циклическом прибавлении знака * к коду 11...11 имеем: 11...11 + * = **...**.

Замечание. В рассмотренной классификации конфликтных методов кодового управления можно отвести место и известному случайному методу занятия канала с обнаружением наложения CSMA-CD - это информационный конфликтный метод (реализуется поэтапное кодирование) с синхронной передачей прав на занятие ресурса при случайном задании кодов приоритетов абонентов (в третьем разделе доказано, что этот метод всегда будет иметь более низкую эффективность, чем соответствующий конфликтный метод реального времени, для которого решена задача оптимального кодирования приоритетов абонентов).

Альтернативные варианты реализации информационных и зондовых конфликтных методов достаточно полно описаны в [12 - 18].

4.2. Бесконфликтные методы кодового управления

В предыдущем разделе было введено понятие предельно неравного по «0» кода приоритета абонентов, который может эффективно использоваться при высокой загрузке системы ($p \to 1$). Отличительной особенностью данного кода, пример которого приведен на рис. 3.3.б, является наличие только одной единицы в каждом кодовом слове, что подразумевает бесконфликтную передачу прав - при зондировании канала по каждому разряду кода приоритета лишь один абонент системы может занять канал. С учетом двух способов передачи прав можно выделить два бесконфликтных метода кодового управления - асинхронный (названный нами в [12-18] счетным) и синхронный, общими условиями реализации которых будет:

- использование предельно неравномерного по «0» кода приоритета абонентов системы;
- каждый сигнал полномочий должен поступать всем абонентам системы;
- сигнал полномочий, поступающий абоненту, передает права в соответствии со значениями кодов приоритетов абонентов.

Идея асинхронного метода состоит в том, что полномочия передаются как и при маркерном методе по каналу, но не коротким пакетом - маркером, содержащим адрес абонента, которому предоставляются права занять ресурс, а отдельным сигналом, поступающим всем абонентам. Каждый полученный из канала сигнал означает для абонента переход к следующему разряду кода приоритета. Абонент, имеющий в разряде «1» может бесконфликтно занять канал информационным сообщением. Если сообщения для выдачи в канал абонент не имеет, он выдает асинхронный сигнал передачи полномочий, осуществляя этим переход к следующему разряду кода.

В общем случае возможны два вида сигнала асинхронной передачи полномочий – специальный сигнал и собственно канальный кадр, передача которого по каналу изменяет текущие значения ОП, а окончание взаимодействия разрешает занять канал абоненту с максимальным текущим ОП.

Аналогично реализуется и бесконфликтный доступ с синхронной передачей прав. Основной проблемой здесь является обеспечение абсолютной синхронизации абонентов, причем с децентрализованным управлением. С учетом того, что без дополнительной синхронизации система может функционировать достаточно непродолжительное время, надежно функционирующую синхронную систему можно получить, если реализовать следующий подход.

Пусть абонент при реализации синхронной передачи полномочий будет действовать следующим образом - занимать канал передачей какого-то сигнала в случае, если у него отсутствует сообщение для передачи, в

противном случае - в течение какого-то интервала времени оставлять канал незанятым. По истечение же данного временного интервала может занять канал под информационное взаимодействие. В данном случае по умолчанию все абоненты системы будут знать о прекращении передачи полномочий в системе (такой подход нами назван методом с естественной синхронизацией) [12-18].

Другим очевидным недостатком бесконфликтной передачи прав является их передача и при отсутствии в системе требований ресурса, что не только бессмысленно, но и снижает надежность системы в целом. Так как потребность в передаче полномочий возникает только когда в системе появляется по крайней мере требование ресурса от одного абонента, целесообразно и начинать их передачу именно в этом случае. При этом для запуска процедуры передачи полномочий (если она еще не запущена в системе), абонентом, требующим ресурс, должен выдаваться в канал специальный сигнал уведомления (заметим, что арбитраж требований этим сигналом не проводится, поэтому в младшем разряде кода приоритета всех абонентов в этом случае находится значение «1», что увеличивает на 1 разряд длину кода приоритета всех абонентов системы). Получив из канала этот сигнал, все абоненты запускают децентрализованную процедуру передачи полномочий (при этом может быть использован любой из способов передачи полномочий, как синхронный, так и асинхронный). Данная группа бесконфликтных методов названа нами методами с предварительном уведомлением о занятии ресурса [12], а реализуемый при этом подход к управлению множественным доступом - **принципом управления с предварительным уведомлением о занятии ресурса**.

В общем случае, в зависимости от загрузки системы, могут быть использованы три стратегии передачи полномочий после уведомления о появлении в системе требования ресурса, представленные на рис. 4.2.

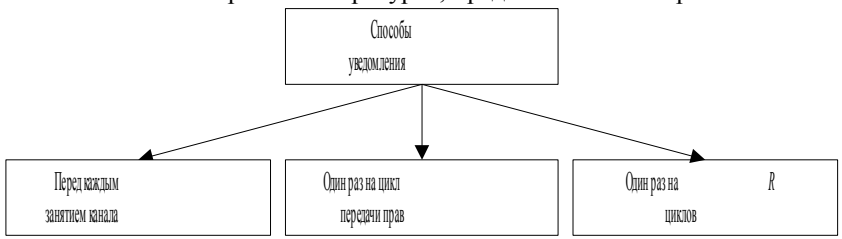


Рис. 4.2

Первая стратегия предполагает необходимость уведомления перед каждым занятием ресурса (соответственно, альтернативный ей гипотетический вариант - это работа системы без предварительного уведомления), две остальные предполагают возможность занятия ресурса без уведомления при большом числе требований - функционально это означает, что при большом числе требований передача полномочий осуществляется без

предварительного уведомления, при малом числе требований процедура передачи прав запускается только в том случае, когда в системе появляются требования ресурса (или реализуется принцип адаптивной к загрузке системы передачи полномочий на занятие ресурса).

4.3. Единая концепция комбинирования методов

На сегодняшний день с целью повышения эффективности управления множественным доступом в широком диапазоне изменения нагрузки системы реализуются комбинированные методы, при этом, как правило, в рамках единой системы пытаются объединить методы, имеющие различные области эффективного использования, соответственно конфликтные, ориентированные на функционирование при низких загрузках системы, и бесконфликтные, эффективные при высоких загрузках. На практике это приводит к попыткам совмещения методов управления множественным доступом, имеющих совершенно различные механизмы передачи прав, как правило, это случайный CSMA-CD и маркерный методы.

Принцип кодового управления позволяет комбинировать подходы в рамках единого механизма управления множественным доступом - информационного конфликтного метода с синхронной, либо асинхронной передачей прав, или зондового конфликтного метода с синхронной передачей прав. Комбинирование здесь реализуется исключительно за счет смены исходных статических кодов, адаптивно к изменению загрузки системы в процессе ее функционирования с сохранением исходного механизма передачи прав. При этом могут быть достигнуты очень высокие результаты, что обусловливается возможностью достаточно плавного (без резких скачков) учета загрузки, за счет того, что между предельными вариантами кодов - предельно неравномерного по «1» для минимальной загрузки и по «0» для максимальной загрузки (бесконфликтная передача) находится множество кодов, которые могут быть использованы с соответствующими пороговыми значениями загрузки.

Таким образом, можно говорить о существовании единой концепции получения комбинированных методов, состоящей исключительно в изменении исходных статических кодов ОП абонентов адаптивно к изменению загрузки системы. В целом рассмотренный подход намечает возможность получения единой концепции реализации управления множественным доступом, сводящей все многообразие способов управления множественным доступом к нескольким обобщающим методам, в результате чего может быть поставлен вопрос об их унификации.

4.4. Унифицированный ряд методов кодового управления

Обобщая сказанное ранее, делаем вывод о существовании следующих принципиально отличных подходов, конфликтных, различающихся способом зондирования канала и способом передачи прав к ресурсу, и бесконфликтных, отличающихся способом передачи прав. Кроме того, выше выделен класс комбинированных методов, получаемых в рамках конфликтных методов сменой статических кодов приоритетов абонентов. При этом единая концепция комбинирования методов обусловливает, что и бесконфликтные методы могут быть получены в рамках конфликтных, при использовании в них предельно неравномерных по «0» кодов приоритета абонентов системы. С учетом сказанного, делаем вывод, что все многообразие механизмов управления множественным доступом в рамках принципа кодового управления можно свести всего лишь к трем методам - информационному конфликтному с синхронной и асинхронной передачей прав на занятие ресурса, зондовому конфликтному с синхронной передачей прав. Все же многообразие методов, включая бесконфликтные и комбинированные, может быть получено в рамках трех перечисленных основных подходов, реализацией соответствующих кодов приоритетов абонентов.

Унифицированный ряд методов кодового управления множественным доступом представлен на рис. 4.3.

С целью исследования эффективности методов кодового управления множественным доступом реального времени могут быть применены вероятностные модели системы управления множественным доступом T_{nn_m}, позволяющие оценить средние затраты времени на передачу полномочий одному из $m = 1,\ldots, M$ активных абонентов системы после освобождения ресурса (или потери времени на арбитраж). Например, для ОЦП (в общем случае не представляет сложности получение вероятностных моделей и для приоритетных расписаний, однако при этом следует учитывать, что для каждого расписания необходимо получать свою модель) для зондового конфликтного метода управления множественным доступом вероятностная модель имеет вид

$$T^{з}_{nn_m}(p, p_m) = (1-(1-p)^M) \sum_{m=1}^{M} p_m N_m T_{nny},$$

для информационного конфликтного метода

$$T^{u}_{nn_m}(p, p_m) = \sum_{m=1}^{M} [(P_{n_m}(p, p_m) - P_{n_{m-1}}(p, p_m))(n_m T_{nny} + (n_m - 1)T_к)]$$

,

соответственно, для способа поочередной передачи полномочий - опрос очередей (в том числе, маркерного)

$$T^{n}_{nn_m}(p, p_m) = \sum_{l=1}^{M}[C^{l}_{M} p^{l}(1-p)^{M-l} \frac{\sum_{a=1}^{M/l} T_{nn_{m-1,m}}}{M}],$$

где T_{nny} - продолжительность передачи прав по одному разряду кода ОП, $T_{к}$ - затраты времени на обнаружение конфликта, $T_{nn_{m-1,m}}$ - затраты времени на поочередную передачу прав между очередными абонентами в расписании.

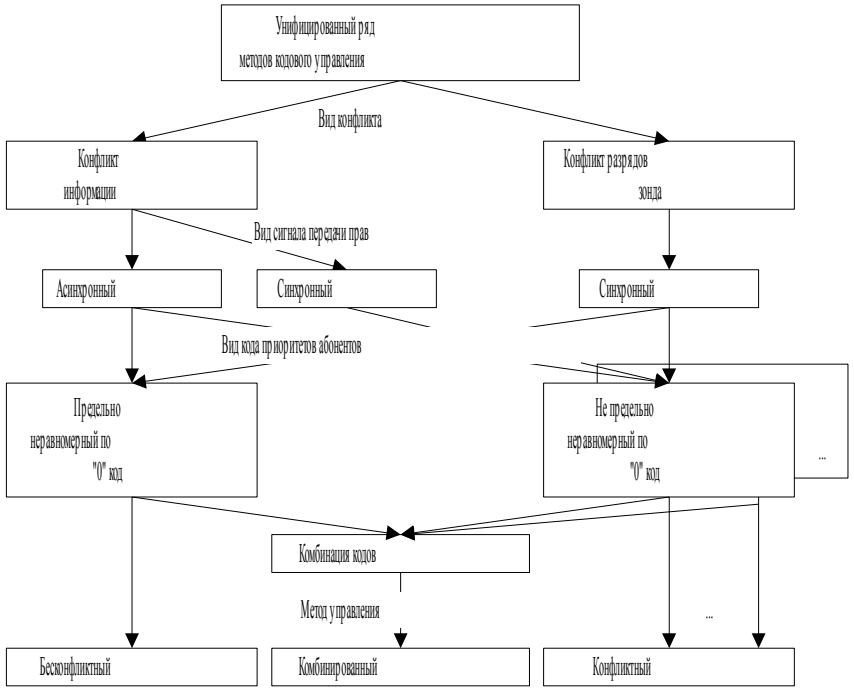


Рис. 4.3

Зависимости $T_{nn_m}(p)/\tau_u = f(p)$, в предположении, что загрузка системы симметричная (p_m совпадают для всех M абонентов системы), для альтернативных подходов представлены на рис. 4.4 (τ_u - период следования импульсов в канале связи - величина, обратная физической скорости передачи данных).

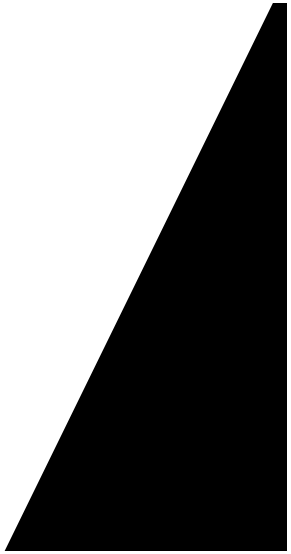


Рис. 4.4

Рис. 4.4 иллюстрирует высокую эффективность методов кодового управления при различии областей эффективного использования альтернативных подходов, что позволяет утверждать о перспективности применения изложенной концепции кодового управления множественным доступом в альтернативных приложениях ЛВСРВ и ЛВСКО. В свою очередь, предоставляемая возможность совмещения в рамках единого механизма управления доступом ДО с относительными и абсолютными приоритетами, при возможности совмещения такого обслуживания с обслуживанием по расписанию при высокой эффективности доступа к общим ресурсам при кодовом управлении позволяет утверждать и о высокой эффективности применения изложенной концепции в сосредоточенных ВС, в частности в многозадачных операционных системах реального времени, в первую очередь - специализированных.

Замечания.

1. В ЛВСРВ и ЛВСКО, как, впрочем, и в многозадачных операционных системах реального времени, при проектировании системы возникает проблема задания фиксированной продолжительности занятия ресурса абонентом $T_{гр}$, получившим право доступа, для информационного взаимодействия с ресурсом (для ЛВС определяется числом передаваемых байтов информации в пакете, для операционной системы - величиной кванта процессорного времени). При этом, с одной стороны, как отмечалось, в системах реального времени параметр $T_{гр}$ следует уменьшать (именно при этом достигается эффект параллельности обработки, т.к. тогда в течение

меньшего интервала времени общий ресурс, являющийся «узким местом», монопольно используется одним абонентом), кстати говоря, именно сказанное ограничивает реализацию приоритетного обслуживания заявок изменением значений $T_{гр_m}$, с другой стороны, уменьшение $T_{гр}$ ограничено потерями производительности ресурса (пропускной способности связного ресурса), связанными с затратами времени на передачу прав $T_{пп}$. Если параметры $T_{гр}$ и $T_{пп}$ сопоставимы (что имеет место в распределенных системах) уменьшение $T_{гр}$ приводит к обратному эффекту - в системе так велики потери производительности, связанные с управлением множественным доступом к ресурсу (в ОС - переключение задач), что в ней теряет всяческий смысл параллельная обработка. Сказанное объясняет необходимость задания $T_{гр}$ сегодня достаточно большим, что связано с низкой эффективностью управления множественным доступом, в частности при реализации расписаний опросом очередей (так в АТМ каждой ячейкой переносится 48 информационных байтов, что в какой-то мере приводит к нарушению реального масштаба времени передачи соответствующих сигналов, т.к. имеем уже некоторое накопление информации - идеальной была бы передача 1 байта данных при 256 уровнях квантования сигнала [4]; продолжительность временного кванта в ОС QNX может достигать 100 мкс, что также может приводить к нарушению условий функционирования в реальном времени). Поэтому из сказанного делаем вывод, что эффективное решение задачи управления множественным доступом, позволяющее уменьшать потери $T_{пп}$ без снижения эффективности обслуживания заявок, даст еще один эффект - позволит уменьшить $T_{гр}$, что, с одной стороны, позволит повысить эффект параллельности обработки для решаемых сегодня распределенными системами задач (т.е. качество обслуживания в реальном масштабе времени), с другой стороны, обеспечит эффективное решение новых задач в существующих системах, в частности реализовать в ЛВС новые виды служб связи.

2. К сосредоточенным ВС относятся системы, в которых абоненты находятся на небольших расстояниях друг от друга. В этом случае в шину могут дополнительно вводиться управляющие линии арбитража - здесь кодового управления - куда абоненты выдают разряды кодов ОП абонентов. При этом также могут быть реализованы все изложенные выше принципы кодового управления, однако здесь уже целесообразно использование в системе только зондового конфликтного метода управления с синхронной передачей прав, причем задача арбитража в данных приложениях может решаться одновременно с передачей данных, так используются различные линии шины для передачи кода приоритета и данных.

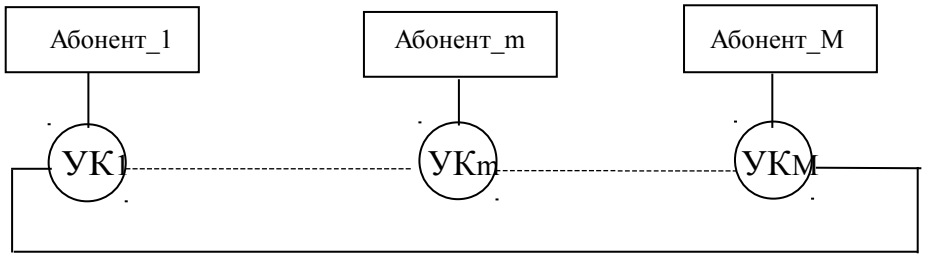
Таким образом, можем сделать вывод, что в рамках изложенной в монографии концепции диспетчеризации с децентрализованным кодовым

управлением множественным доступом, можно реализовать эффективное управление распределением ресурсов для различных приложений ВС и ЛВС реального времени и комбинированного обслуживания, что достигается как за счет предоставляемых широких возможностей учета приоритетов заявок и абонентов системы, так и за счет возможности реализации альтернативных способов передачи прав на занятие ресурса. При этом рассмотренная концепция позволяет унифицировать как альтернативные подходы к обслуживанию заявок в рамках метода обслуживания с динамическими ОП, так и собственно механизмы передачи прав в рамках метода децентрализованного кодового управления множественным доступом.

4.5. Унификация методов для альтернативных топологий сетей связи.

4.5.1.Принципы реализации и классификация методов для альтернативных топологий ЛВС.

Ранее отмечалось, что особенностью шинной топологии ЛВС является отсутствие узлов коммутации в сети связи. Это обуславливает то, что информация, вызываемая любым абонентом, практически одновременно ЛВС набирает всем остальным абонентам сети (нами данный случай выделен как наиболее общий). Другая топология (например, кольцо и звезда) предполагает последовательный анализ поступающей через узлы коммутации информации, тем не менее, и в данных случаях возможно (и целесообразно) использование предложенных нами методов диспетчеризации. Однако отличие состоит в том, что здесь (ввиду последовательности обработки информации) уже возможны два подхода и решения задачи управления множественным доступом, которые рассмотрим на примере ЛВС кольцевой топологии, структура которой представлена на рис.4.5. Отличие подходов состоит в том - устанавливается ли



приоритетность занятия связного ресурса перед передачей информационного кадра (как это делается при шинной топологии зондовыми методами), либо сразу предается соотношения. При незанятости канала без предварительного арбитража. Однако, в данном случае, ввиду последовательности арбитража

узлами коммутации ($УКm$, $m = \overline{1, M}$), появляется возможность избежать конфликтов при реализации информационных методов доступа, за счет передачи приоритета абонента, вызвавшего кадр, т.е. структура кадра в этом случае примет вид, где после открывающего флага следует код приоритета (КП) затем уже служебная информация . КП добавляется в заголовок, соответствующий МАС подуровни в соответствии с моделью ISO[. . .]. В этом случае абонент имеющий пакет для передачи канал связи с сформированным для него КП начинает выдачу пакета при отсутствии поступления другого пакета из канала на другой вход его УК. В противном случае УК сравнивает КП обоих пакетов, буферируя менее приоритетный и выдавая в канал более приоритетный пакет. При этом КП, передаваемые в пакетах, формируются по описанным выше правилам (в частности, здесь может быть реализован метод регистровых задержек).

С учетом сказанного, классификация методов кодового управления множественным доступом для топологии ЛВС, содержащих УК, принимает вид, приведенный на рис 4.6.

рис 4.6

Замечание. При зондовом способа при незанятом канала в сигнальном кадре-зонде активный источник выдает зонд, который (как и при шинной реализации) поразрядно анализируется УК других абонентов, вступивших в конфликт, которые соответствующим образом в случае необходимости изменяет значения разрядов зонда на противоположные. По содержимому КП зонда, возвратившемуся из кольца на вход УК активного абонента, он принимает решения, выдавать ли пакет в канал или пропустить зонд более приоритетному активному абоненту, чтобы тот занял канал. Структура кадра-зонда кроме КП содержит служебное поле-признак зонда, данный способ напоминает собой шарнирное кольцо, однако в отличие от последнего

позволяет реализовать рассмотренные выше преимущества принципа кодового управления.

Таким образом, предложенные в книге подходы могут быть реализованы и для других, промежуточных топологий ЛВС, причем в данном случае их реализация упрощается, за счет реализации последовательного анализа информации используемыми узлами коммутации.

4.5.2.Принципы реализации и классификация методов для глобальных сетей связи.

Отличие кольцевой топологии ЛВС (а тем более топологии «звезда») от топологии глобальной сети не так велико, как шинной. Здесь также имеют место узлы коммутации, отличие состоит лишь в том, что много (а не одна как в ЛВС звезда) и каждый в общем случае характеризуется более чем двумя входами и более чем двумя выходами, рис.4.7.

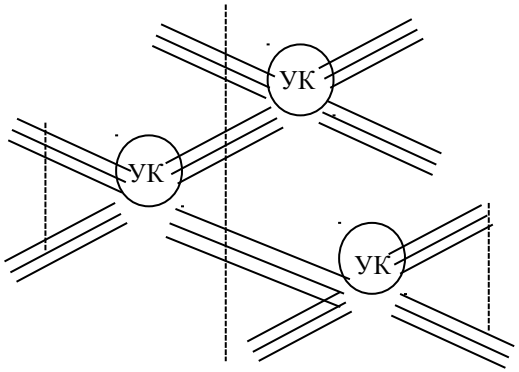


Рис. 4.7

Многочисленные каналы связи здесь уже образуют пучки линий. Однако здесь могут быть реализованы принципы кодового управления множественным доступом, аналогичные описанным выше. При этом в глобальной сети возможны два решения обслуживания: с установлением соединения и без установления соединения. В режиме установления соединения абонентом в сеть посылается специальный пакет-зонд, содержащий КП. Каждый УК анализирует КП зонда, поступающего с каждой линии(пучка), настраиваясь в соответствии с заданной для него ДО на обслуживание пакета, который затем будет передан абонентом, пославшим зонд другими словами, аналогично протоколу Х.25 здесь реализуется

согласование качества обслуживания абонентом и сетью (УК) пакета перед его передачей абонентом в сеть, за счет выдачи абонентом значения КП в служебном пакете-зонде и настройки УК сети на обслуживание поступающего с данной линии пакета в соответствии с заданной ДО. Альтернативный подход, как и для кольцевых ЛВС, состоит в выдаче абонентом в сеть КП в заголовке информационного пакета, рис. 4.5. Поступившие же в УК пакеты уже должны обслуживаться в соответствии с реализованной в нем ДО.

Замечание. КП должны передаваться в заголовках сетевого уровня протоколов доступа и подсетям. В простейшем случае КП могут передаваться и в заголовках внутрисетевых протоколов, однако в данном случае ДО, реализуемая в УК может учитывать приоритеты входящих линий и пучков линий связи.

Таким образом, как и для кольцевых ЛВС, ДО здесь реализуется УК и возможна аналогичная классификация методов, приведенная на рис.4.8.

рис 4.8

Таким образом, может быть сделан вывод о том, что рассмотренные принципы обслуживания в полном объеме могут быть реализованы как для различных топологий ЛВС, так и при построении глобальных сетей связи при альтернативных подходах и их реализации.

ЛИТЕРАТУРА

1. Алиев Т.И. Исследование методов диспетчеризации в цифровых управляющих системах. - Л.:ЛИТМО, 1984. - 82с.

2. Основы теории вычислительных систем/ Под ред. С. А. Майорова. М.:Высшая школа, 1978. - 410с.

3. Владимиров Н.А. Технология АТМ: основные положения// Сети. 1996. - № 2. - С. 62-73.

4. Колесник В.Д., Полтырев Г.Ш. Курс теории информации. - М.:Наука, 1982. - 416с.

5. Колесниченко В.Е. Основные результаты для вероятностно временных характеристик систем циклического обслуживания// Автоматика и вычислительная техника. - 1991. - № 3. - С. 59-64.

6. Колесниченко В.Е. Простые аппроксимации распределения времени ожидания для локальных вычислительных сетей с маркерным доступом// Автоматика и вычислительная техника. - 1990. - № 1. - С. 49-54.

7. Мячев А.А., Степанов В.Н., Щербо В.К. Интерфейсы систем обработки данных. - М.:Радио и связь, 1989. - 416с.

8. Овчинников В.В., Рыбкин И.И. Техническая база интерфейсов локальных вычислительных сетей. - М.:Радио и связь, 1989. - 272с.

9. Организация последовательных мультиплексных каналов систем автоматического управления/ С.Т. Хвощ, В.В. Дорошенко, В.В. Горовой; Под общ. ред. С.Т. Хвоща. - Л.:Машиностроение, 1989. - 271с.

10. Прангишвили И.В., Подлазов В.С., Стецюра Г.Г. Локальные микропроцессорные вычислительные сети. - М.:Наука, 1984. - 176с.

11. Стандарты по локальным вычислительным сетям: Справочник/ В.К. Щербо, В.М. Киреичев, С.И. Самойленко. - М.:Радио и связь, 1990. 304с.

12. Щеглов А.Ю. Принципы арбитража с предварительным уведомлением о занятии канала данных// Электронное моделирование. - 1995. - 17, № 3. С. 35 - 41.

13. Щеглов А.Ю. Способы реализации работоспособного синхронного множественного доступа// Электронное моделирование. - 1994. - 16, № 3. С. 32 - 38.

14. Щеглов А.Ю. Ускоренное зондирование канала данный - новый принцип случайного доступа в локальных вычислительных сетях// Изв. РАН. Техническая кибернетика. - 1994. - № 2. - С. 129-136.

15. Щеглов А.Ю. Высокопроизводительный метод децентрализованного кодового управления для неоднородной микропроцессорной системы// Кибернетика и системный анализ. - 1994. - № 2. - С. 176-180.

16. Щеглов А.Ю. Счетно-интервальный способ множественного доступа для ЛВС реального времени// Автоматика и вычислительная техника. 1991. - № 5. - С. 74-77.

17. Щеглов А.Ю. Алгоритм динамического изменения приоритетов для высокопроизводительного метода ДКУ реального времени// Кибернетика и системный анализ. - 1994. - № 5. - С. 179-181.

18. Щеглов А.Ю. Счетные методы множественного доступа - альтернативный способ асинхронной передачи полномочий на доступ к шинному каналу данных в локальных вычислительных сетях// Электронное моделиро вание. - 1995. -17, № 4. - С. 41-46.

19. Щеглов А.Ю. Высокопроизводительные счетные методы доступа для локальной вычислительной сети реального времени// Электронное моделирование. - 1992. - 14, № 2.- С.104-107.

20. Щеглов А.Ю. Элементы теории управления множественным доступом пользователей к общим ресурсам вычислительной системы// Вторая Международная конференция "Развитие и применение открытых систем". Тезисы докладов. - М.: Совет РАН по автоматизации научных исследований, 1995. - С.74-76.

21. Щеглов А.Ю. Ускоренный метод децентрализованного кодового управления доступом к моноканалу в реальном времени// Электронное моделирование. - 1993. - 15, № 3. - С.88-92.

22. Куконин А.Ю., Щеглов А.Ю. Методы децентрализованного управления доступом к каналу малых локальных вычислительных сетей// Управляющие системы и машины. - 1992. - № 7/8. - С.124-127.

23. Щеглов А.Ю. Приоритетные счетные методы доступа для локальных вычислительных сетей// Системы и средства телекоммуникаций. - 1992. - № 5/6. - С.40-42.

24. Щеглов А.Ю., Куконин А.Ю. Метод децентрализованного кодового управления доступом к каналу с передачей полномочий// Электронное моделирование. - 1991. - 13, № 6. - С.32-34.

СОДЕРЖАНИЕ