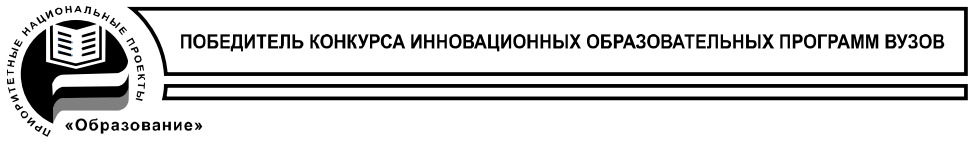
МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ



Т.И. Алиев

ОСНОВЫ МОДЕЛИРОВАНИЯ ДИСКРЕТНЫХ СИСТЕМ

Учебное пособие

Санкт-Петербург
2009

Алиев Т.И. Основы моделирования дискретных систем. – СПб: СПбГУ ИТМО, 2009. – 363 с.

В пособии излагаются математические модели и результаты анализа дискретных систем различных классов с использованием аналитических, численных и имитационных методов исследования. В качестве моделей таких систем рассматриваются модели, построенные на основе систем и сетей массового обслуживания. Аналитические методы исследования базируются на аппарате теории массового обслуживания, численные – на аппарате теории марковских случайных процессов, статистические – на методах имитационного моделирования, которое реализуется в среде GPSS World. Материал пособия сопровождается многочисленными примерами, направленными на развитие навыков и умения применять простейшие модели и методы для исследования реальных систем. Особое внимание уделяется анализу и изучению свойств систем, представляемых моделями массового обслуживания.

Пособие предназначено для студентов, обучающихся в области информационных технологий, а также для выпускников (бакалавров, магистрантов и специалистов) по направлению 230100 – «Информатика и вычислительная техника», подготавливающих выпускные квалификационные работы, в которых требуется выполнить моделирование и исследование системы с дискретным характером функционирования. Пособие может быть полезным для аспирантов и специалистов, выполняющих исследования реальных систем с использованием аналитических и имитационных методов моделирования.

Рекомендовано к печати Советом факультета компьютерных технологий и управления 10 марта 2009 г., протокол № 8

ISBN 978-5-7577-0336-7

СПбГУ ИТМО стал победителем конкурса инновационных образовательных программ вузов России на 2007-2008 годы и успешно реализовал инновационную образовательную программу «Инновационная система подготовки специалистов нового поколения в области информационных и оптических технологий», что позволило выйти на качественно новый уровень подготовки выпускников и удовлетворять возрастающий спрос на специалистов в информационной, оптической и других высокотехнологичных отраслях науки. Реализация этой программы создала основу формирования программы дальнейшего развития вуза до 2015 года, включая внедрение современной модели образования.

Введение

> «Даже если ваше объяснение настолько ясно, что исключает всякое ложное толкование, все равно найдется человек, который поймет вас неправильно» (***Законы Мэрфи***)

Математическое моделирование является мощным и эффективным инструментом исследования разнообразных объектов, систем и процессов в различных областях человеческой деятельности. Многообразие процессов, протекающих в исследуемых системах и объектах, обусловливает и многообразие математических методов и средств, используемых в теории моделирования.

Моделирование – сложнейший многоэтапный процесс исследования систем, направленный на выявление свойств и закономерностей, присущих исследуемым системам, с целью их создания или модернизации. В процессе моделирования решается множество взаимосвязанных задач, основными среди которых являются разработка модели, анализ свойств и выработка рекомендаций по модернизации существующей или проектированию новой системы.

Большинство технических систем, в том числе вычислительные системы и сети, описываются в терминах дискретных случайных процессов с использованием вероятностных методов [1, 3, 9, 11]. При этом широкое применение находят математические модели, отражающие структурно-функциональную организацию исследуемых систем, построенные на основе моделей теории массового обслуживания, анализ которых может проводиться аналитическими, численными и статистическими методами. В качестве аналитических методов используются вероятностные методы теории массового обслуживания, в качестве численных – методы теории марковских случайных процессов, в качестве статистических – методы имитационного моделирования.

При изложении аналитических методов расчёта автор, помня знаменитое высказывание, что «всякое уравнение длиной более двух дюймов, скорее всего, неверно», сознательно стремился использовать сравнительно простые математические зависимости, позволяющие рассчитывать, в первую очередь, основные характеристики функционирования систем, такие как нагрузку, загрузку и средние значения вероятностно-временны*х* характеристик исследуемой системы. По той же причине в пособии отсутствуют громоздкие выводы и сложные доказательства представленных математических зависимостей, которые позволяют, с одной стороны, достаточно просто выполнить оценочные расчёты, не прибегая к сложным вычислениям, с другой стороны, получить вполне адекватное представление о свойствах соответствующих реальных систем за счёт детального их анализа. Наиболее важные и необходимые при решении многих задач формулы заключены в рамку.

Теоретический материал пособия сопровождается многочисленными примерами и задачами, которые направлены на развитие умения и навыков применять простейшие модели и методы для расчёта нагрузки и загрузки отдельных элементов и системы целом, для проведения анализа характеристик функционирования реальных систем, представляемых моделями массового обслуживания или моделями марковских случайных процессов.

При решении задач следует иметь в виду, что при отсутствии каких-либо необходимых исходных данных могут и должны вводиться предположения и допущения, позволяющие решить (пусть и упрощённо) поставленную задачу.

В пособии для более эффективного усвоения материала фрагменты, представляющие наибольший интерес, выделяются разными шрифтами, что позволяет акцентировать внимание читателя на тех или иных аспектах, которые, по мнению автора, являются важными для понимания описанных моделей и методов.

Полужирный курсив выделяет наиболее важные и часто используемые термины и понятия, для которых дается чёткое определение или подробное описание.

Полужирным шрифтом выделяются прочие общепринятые термины и понятия, часто встречающиеся в литературе и не имеющие чёткого определения, а также вспомогательные заголовки, названия и т.д.

Курсив выделяет в тексте ключевые слова и фразы, на которые следует обратить внимание и которые раскрывают смысл излагаемого материала, а также выделяет термины и понятия, которые определены в других разделах.

Какова цель данного пособия? Для кого оно предназначено, на какой круг читателей ориентировано? Ответы на эти вопросы определяют содержание и стиль изложения материала.

Данное учебное пособие преследует две основные цели.

1. Дать неподготовленному читателю начальное представление о принципах моделирования сложных систем на примере широко используемых на практике моделей массового обслуживания и методов их расчёта с использованием трех основных подходов: аналитического, численного и имитационного. При этом излагаемый материал не должен содержать громоздкие математические выкладки и доказательства. Используемые математические выражения и формулы должны быть понятны любому техническому специалисту, имеющему базовую математическую подготовку в рамках среднего и высшего образования и самые общие представления о теории вероятностей, служащей математической основой излагаемого материала.

2. Предоставить читателю минимальный набор моделей, методов и средств для исследования несложных реальных систем в различных прикладных областях. Для достижения этой цели в пособии рассматривается большое количество примеров математических моделей, с

помощью которых иллюстрируется применение излагаемых аналитических, численных и имитационных методов расчёта.

Для достижения сформулированных целей в пособии:

- вводится четкая однозначная терминология, используемая в процессе изложения материала;
- формируется представление о моделях массового обслужива-ния, их многообразии, а также о величинах, описывающих эти модели;
- формулируются задачи моделирования как универсального инструмента исследования сложных систем, в том числе технических систем, таких как вычислительные машины, комплексы, системы и сети;
- рассматриваются методы расчёта характеристик математи-ческих моделей, представляемых в виде систем и сетей массового обслуживания;
- выполняется анализ свойств, и выявляются закономерности, присущие процессам, протекающим в моделях различных классов, а также анализируется влияние параметров модели на характеристики её функционирования.

Структура учебного пособия. Пособие содержит *6* основных *разделов*, *Заключительный раздел*, *Приложения*, *Список литературы* и *Алфавитный указатель*. Материал каждого раздела разбит на параграфы, которые имеют двойную нумерацию. Некоторые параграфы разбиты на пункты с тройной нумерацией.

Первый раздел содержит основные понятия и определения, касающиеся общих принципов моделирования, перечень параметров и характеристик дискретных систем, классификацию систем и моделей, краткую характеристику методов моделирования.

Во **втором разделе** приводятся необходимые сведения из теории вероятностей и рассматриваются законы распределений случайных величин, наиболее часто используемые в теории массового обслуживания. В этот же раздел включён материал, связанный с аппроксимацией реальных распределений случайных величин, распределениями, представляющими собой комбинацию экспоненциальных распределений.

В **третьем разделе** даются основные определения и понятия теории массового обслуживания, приводится классификация базовых моделей в виде систем массового обслуживания (СМО) и сетевых моделей в виде сетей массового обслуживания (СеМО), рассматриваются их параметры и характеристики.

В **четвёртом разделе** излагаются аналитические методы расчёта простейших одноканальных и многоканальных СМО с однородным и неоднородным потоком заявок, а также линейных разомкнутых и замкнутых однородных экспоненциальных СеМО. В этом же разделе большое внимание уделяется анализу свойств исследуемых моделей, результаты которого позволяют выявить и сформулировать ряд важных особенностей и закономерностей, присущих процессам, протекающим в

системах. Выявленные свойства могут лечь в основу рекомендаций для проектирования подобных систем.

Пятый раздел посвящен численным методам анализа моделей массового обслуживания с использованием аппарата теории марковских случайных процессов. Подробно рассматриваются примеры применения марковских случайных процессов для анализа СМО и СеМО с накопителями ограниченной ёмкости.

Методам имитационного моделирования посвящен **шестой раздел**, в котором излагаются принципы имитационного моделирования и основы моделирования в среде GPSS World. Описание системы имитационного моделирования GPSS World содержит *минимум* информации, необходимой для построения простейших моделей массового обслуживания. В частности, из 53 операторов блока в этом разделе представлены менее половины, которые используются в рассматриваемых примерах имитационных моделей СМО и СеМО. Более полное и подробное описание системы имитационного моделирования GPSS World можно найти в специальной литературе по GPSS World [4], которая в последние годы представлена в достаточном объёме, и на сайте www.gpss.ru.

Каждый раздел заканчивается тремя параграфами с одинаковыми названиями.

1. Резюме. Содержит краткое изложение представленного в разделе материала.

2. Практикум. Содержит обсуждение актуальных и наиболее часто задаваемых вопросов по изложенному материалу, а также подробные решения в качестве примеров наиболее интересных задач.

3. Самоконтроль. Содержит перечень тестовых вопросов и задач, позволяющих читателю самостоятельно выполнить проверку степени усвоения изложенного материала и закрепить полученные знания в процессе решения предлагаемых задач.

Заключительный раздел является итоговым, в котором излагаются основные принципы моделирования как систематизированная последовательность этапов, методов и средств по организации процесса моделирования.

Некоторые разделы, параграфы и пункты пособия предваряются цитатами из *законов Мэрфи*, что, по мнению автора, возможно, позволит акцентировать внимание читателя и подтолкнуть его к критическому осмыслению излагаемого материала.

В *Приложениях* представлены аббревиатуры, основные обозначения, используемые в пособии, а также приведён перечень вопросов, обсуждаемых в основных разделах пособия.

В конце пособия имеется *Предметный указатель* со ссылками на страницы, содержащие определения основных терминов и понятий, используемых в учебном пособии.

Представленный *Список литературы* не претендует на полноту и содержит ограниченный перечень литературных источников, которые в

той или иной мере использовались при написании пособия. Этот перечень включает только учебные пособия и монографии, которые условно можно разбить на три группы, содержащие материал:

- по теоретическим вопросам моделирования и математическим методам исследования систем и сетей массового обслуживания [2, 5, 6, 8, 12, 13, 17];
- по применению моделей и методов теории массового обслуживания для исследования вычислительных систем и сетей [1, 3, 9-11, 14-16];
- по имитационному моделированию систем и сетей массового обслуживания и описанию системы имитационного моделирования GPSS World [4, 7, 18].

Пособие предназначено, прежде всего, для студентов, обучающихся по направлению «Информатика и вычислительная техника», изучающих дисциплину «Моделирование» и связанные с ней дисциплины, а также для выпускников (бакалавров, магистрантов и специалистов), подготавливающих выпускные квалификационные работы, в которых требуется выполнить моделирование и исследование некоторой системы, например, компьютерной сети или её фрагмента – вычислительной системы (сервера), узла или канала передачи данных. Пособие может быть полезным в качестве введения в проблематику моделирования дискретных систем со стохастическим характером функционирования для магистрантов, аспирантов и специалистов, выполняющих исследования реальных систем с использованием аналитических и имитационных методов математического моделирования.

Автор, понимает, что, как и «в любой хорошо отлаженной программе всегда есть хотя бы одна ошибка», так и в любой многократно вычитанной книге она также наверняка присутствует. Посему автор будет благодарен за все присланные по адресу aliev@d1.ifmo. ru обнаруженные ошибки и опечатки, а также критические замечания.

Раздел 1. ОБЩИЕ ВОПРОСЫ МОДЕЛИРОВАНИЯ

«Все, что хорошо начинается, кончается плохо. Все что начинается плохо, кончается еще хуже» (***Закон Паддера***)

Моделирование – замещение одного *исходного объекта* другим объектом, называемым **моделью** (рис.1.1), и проведение экспериментов с моделью с целью получения информации о системе путем исследования свойств модели.

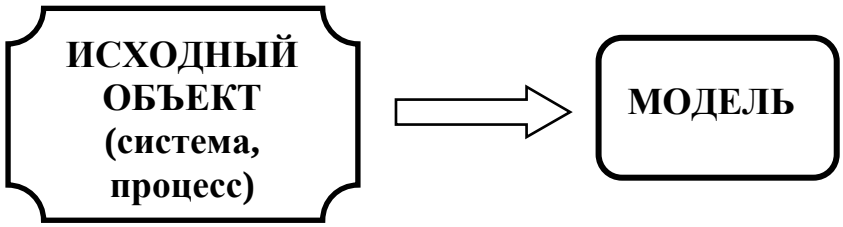


Рис.1.1. К понятию «моделирование»

Объектами моделирования в технике являются *системы* и протекающие в них *процессы*. В частности, в вычислительной технике объектами моделирования являются вычислительные машины, комплексы, системы и сети. При этом, наибольший интерес представляют **конструктивные модели**, допускающие не только фиксацию свойств (как в произведениях искусств), но и исследование свойств систем (процессов), а также решение задач проектирования систем с заданными свойствами.

Моделирование предоставляет возможность исследования таких объектов, прямой эксперимент с которыми:

- трудно выполним;
- экономически невыгоден;
- вообще невозможен.

Моделирование – важнейшая сфера применения вычислительных систем и сетей в различных областях науки и техники: в математике и физике, в авиа- и автомобилестроении, в приборо- и машиностроении, в оптике, в электронике и т.д. Все более широкое распространение моделирование находит в таких областях как экономика, социология, искусство, биология, медицина и т.п. В то же время, вычислительные системы и сети сами являются объектами моделирования на этапах проектирования новых и модернизации существующих систем, анализа эффективности использования систем в различных условиях (например, в экстремальных ситуациях, в условиях повышенных требований к надежности и живучести). Применение моделирования на этапе проектирования позволяет выполнить анализ различных вариантов предлагаемых проектных решений, определить работоспособность и оценить надежность системы, выявить узкие места и мало загруженные ресурсы, а также сформулировать рекоменда-

ции по рациональному изменению состава и структуры или способа функциональной организации системы.

1.1. Система

1.1.1. Понятия системы и комплекса

«Усложнять - просто, упрощать – сложно»
(Закон Мейера)

Система (от греч. systema – целое, составленное из частей; соединение) – совокупность взаимосвязанных элементов, объединенных в одно целое для достижения некоторой цели, определяемой назначением системы.

Элемент – минимальный неделимый объект, рассматриваемый как единое целое.

Сложная (большая) система характеризуется большим числом входящих в его состав элементов и связей между ними.

Комплекс – совокупность взаимосвязанных систем.

Элемент, система и комплекс – понятия относительные. Любой элемент может рассматриваться как система, если его расчленить на более мелкие составляющие – элементы. И наоборот, любой комплекс может рассматриваться как система, если входящие в его состав системы трактовать как элементы. В связи с этим, понятия «система» и «комплекс» часто трактуют как эквивалентные понятия. Например, вычислительную машину можно рассматривать как систему, элементами которой являются центральный процессор, оперативная память, накопители на магнитных дисках, устройства ввода-вывода. В то же время, центральный процессор можно рассматривать как систему, состоящую из таких элементов, как арифметико-логическое устройство, устройство управления, счетчик команд, регистровая память и т.д.

Для описания системы необходимо определить ее *структуру* и *функцию* и, соответственно, *структурную и функциональную организацию*.

1.1.2. Структура и функция

«Сложные системы имеют тенденцию противопоставлять себя своим же функциям»
(***Принцип Шательера***)

Структура системы задается перечнем элементов, входящих в состав системы, и связей между ними.

Способы описания структуры системы:

- ***графический*** – в форме:
  - *графа*, в котором вершины соответствуют элементам системы, а дуги – связям между ними;
  - *схем*, широко используемых в инженерных приложениях, в которых элементы обозначаются в виде специальных символов;

• ***аналитический*** – путем задания количества типов элементов, числа элементов каждого типа и матрицы связей (инцидентности), определяющей взаимосвязь элементов.

Функция системы – правило достижения поставленной цели, описывающее поведение системы и направленное на получение результатов, предписанных назначением системы.

Способы описания функции системы:

• ***алгоритмический*** – словесное описание в виде последовательностей шагов, которые должна выполнять система для достижения поставленной цели;

• ***аналитический*** – в виде математических зависимостей в терминах некоторого математического аппарата: теории множеств, теории случайных процессов, теории дифференциального или интегрального исчисления и т.п.;

• ***графический*** – в виде временных диаграмм или графических зависимостей;

• ***табличный*** – в виде различных таблиц, отражающих основные функциональные зависимости, например, в виде таблиц булевых функций, автоматных таблиц функций переходов и выходов и т.п.

1.1.3. Организация

> «Сложная система, спроектированная наспех, никогда не работает, и исправить её, чтобы заставить работать, невозможно» (***16-й закон систематики***)

Организация системы – способ достижения поставленной цели за счет выбора определенной структуры и функции системы. В соответствии с этим различают структурную и функциональную организацию системы.

Функциональная организация определяется способом порождения функций системы, достаточных для достижения поставленной цели.

Структурная организация определяется набором элементов и способом их соединения в структуру, обеспечивающую возможность реализации возлагаемых на систему функций.

Функциональная организация реализуется безотносительно к необходимым для этого средствам (элементам), в то время как структурная организация определяется функцией, возлагаемой на систему.

1.1.4. Свойства систем

> «Большая система, образованная увеличением размеров меньшей, ведет себя совсем не так, как ее предшественница» ***(Теорема о неаддитивности поведения систем)***

Любым сложным системам присущи фундаментальные свойства, требующие применения системного подхода при их исследовании методами математического моделирования. Такими свойствами являются:

• ***целостность***, означающая, что система рассматривается как единое целое, состоящее из *взаимодействующих* элементов, возможно неоднородных, но одновременно *совместимых*;

• ***связность*** – наличие существенных устойчивых связей между элементами и/или их свойствами, причем с системных позиций значение имеют не любые, а лишь *существенные* связи, которые определяют *интегративные* свойства системы;

• ***организованность*** – наличие определенной структурной и функциональной организации, обеспечивающей снижение энтропии (степени неопределенности) системы по сравнению с энтропией системообразующих факторов, определяющих возможность создания системы, к которым относятся: число элементов системы, число существенных связей, которыми может обладать каждый элемент, и т.п.;

• ***интегративность*** – наличие качеств, присущих системе в целом, но не свойственных ни одному из ее элементов в отдельности; другими словами, интегративность означает, что свойства системы хотя и зависят от свойств элементов, но не определяются ими полностью.

Таким образом, можно сделать следующие важные *выводы*:

• система не есть простая совокупности элементов;

• расчленяя систему на отдельные части и изучая каждую из них в отдельности, нельзя познать все свойства системы в целом.

1.1.5. Эффективность

«Оптимист верит, что мы живем в лучшем из миров. Пессимист боится, что так оно и есть» (***Главный парадокс***)

В общем случае моделирование направлено на решение задач:

• *анализа*, связанных с оценкой эффективности систем, задаваемой в виде совокупности *показателей эффективности*;

• *синтеза*, направленных на построение *оптимальных систем* в соответствии с выбранным *критерием эффективности*.

Эффективность – степень соответствия системы своему назначению.

Эффективность систем обычно оценивается **набором показателей эффективности**.

Показатель эффективности (качества) – мера одного свойства системы. Показатель эффективности всегда имеет *количественный* смысл.

Количество показателей эффективности технических систем во многих случаях, может оказаться достаточно большим. Обычно показатели эффективности являются противоречивыми. Это означает, что изменение структурной или функциональной организации системы приводит к улучшению одних показателей и, в то же время, к ухудшению других показателей эффективности, что существенно осложняет выбор наилучшего варианта (способа) структурно-функциональной организации

проектируемой системы. Очевидно, что желательно иметь один показатель эффективности. Таким показателем является критерий эффективности.

Критерий эффективности – мера эффективности системы, обобщающая все свойства системы в одной оценке – значении критерия эффективности. Если при увеличении эффективности значение критерия возрастает, то критерий называется ***прямым***, если же значение критерия уменьшается, то критерий называется ***инверсным***.

Критерий эффективности служит для выбора из всех возможных вариантов структурно-функциональной организации системы наилучшего (оптимального) варианта.

Оптимальная система – система, которой соответствует максимальное (минимальное) значение прямого (инверсного) критерия эффективности из всех возможных вариантов построения системы, удовлетворяющих заданным требованиям.

Анализ (от греч. análysis — разложение, расчленение) – процесс определения свойств, присущих системе. В процессе анализа на основе сведений о функциях и параметрах элементов, входящих в состав системы, и сведений о структуре системы определяются характеристики, описывающие свойства, присущие системе в целом.

Синтез (от греч. synthesis - соединение, сочетание, составление) – процесс порождения функций и структур, удовлетворяющих требованиям, предъявляемым к эффективности системы.

Таким образом, с понятием «эффективность» связаны следующие понятия:

- показатель эффективности;
- критерий эффективности;
- оптимальная система;
- анализ;
- синтез.

1.1.6. Параметры и характеристики

> «В любом наборе исходных данных самая надежная величина, не требующая никакой проверки, является ошибочной» ***(Третий закон Финэйгла)***

Количественно любая система описывается совокупностью величин, которые могут быть разбиты на два класса:

- ***параметры***, описывающие *первичные* свойства системы и являющиеся исходными данными при решении задач анализа;
- ***характеристики***, описывающие *вторичные* свойства системы и определяемые в процессе решения задач анализа как функция параметров, то есть эти величины являются вторичными по отношению к параметрам.

Множество параметров *технических систем* можно разделить на:

- **внутренние,** описывающие структурно-функциональную организацию системы, к которым относятся:

▪ ***структурные параметры***, описывающие состав и структуру системы;

▪ ***функциональные параметры***, описывающие функциональную организацию (режим функционирования) системы.

• **внешние**, описывающие взаимодействие системы с внешней по отношению к ней средой, к которым относятся:

▪ ***нагрузочные параметры***, описывающие входное воздействие на систему, например частоту и объем используемых ресурсов системы;

▪ ***параметры внешней (окружающей) среды***, описывающие обычно неуправляемое воздействие внешней среды на систему, например помехи и т.п.

Параметры могут быть:

• *детерминированными* или *случайными*;

• *управляемыми* или *неуправляемыми*.

Характеристики системы делятся на:

• ***глобальные***, описывающие эффективность системы в целом;

• ***локальные***, описывающие качество функционирования отдельных элементов или частей (подсистем) системы.

К *глобальным характеристикам технических систем* относятся:

• ***мощностные (характеристики производительности)***, описывающие скоростные качества системы, измеряемые, например, количеством задач, выполняемых вычислительной системой за единицу времени;

• ***временные (характеристики оперативности)***, описывающие временны*е* аспекты функционирования системы, например время решения задач в вычислительной системе;

• ***надежностные (характеристики надежности)***, описывающие надежность функционирования системы;

• ***экономические (стоимостные)*** в виде стоимостных показателей, например, стоимость технических и программных средств вычислительной системы, затраты на эксплуатацию системы и т.п.;

• ***прочие***: масса-габаритные, энергопотребления, тепловые и т.п.

Таким образом, параметры системы можно интерпретировать как некоторые входные величины, а характеристики – выходные величины, зависящие от параметров и определяемые в процессе анализа системы (рис.1.2).

Тогда закон функционирования системы можно представить в следующем виде:

$$\vec{H}(t) = f_c(\vec{S}, \vec{F}, \vec{Y}, \vec{X}, t),$$

где f_c – функция, функционал, логические условия, алгоритм, таблица или словесное описание, определяющее правило (закон) преобразования входных величин (параметров) в выходные величины (характеристики);

H(t) – вектор характеристик, зависящий от текущего момента времени t (t≥0): $\vec{H} = \{\vec{V}, \vec{T}, \vec{N}, \vec{C}, \vec{Z}\}$.

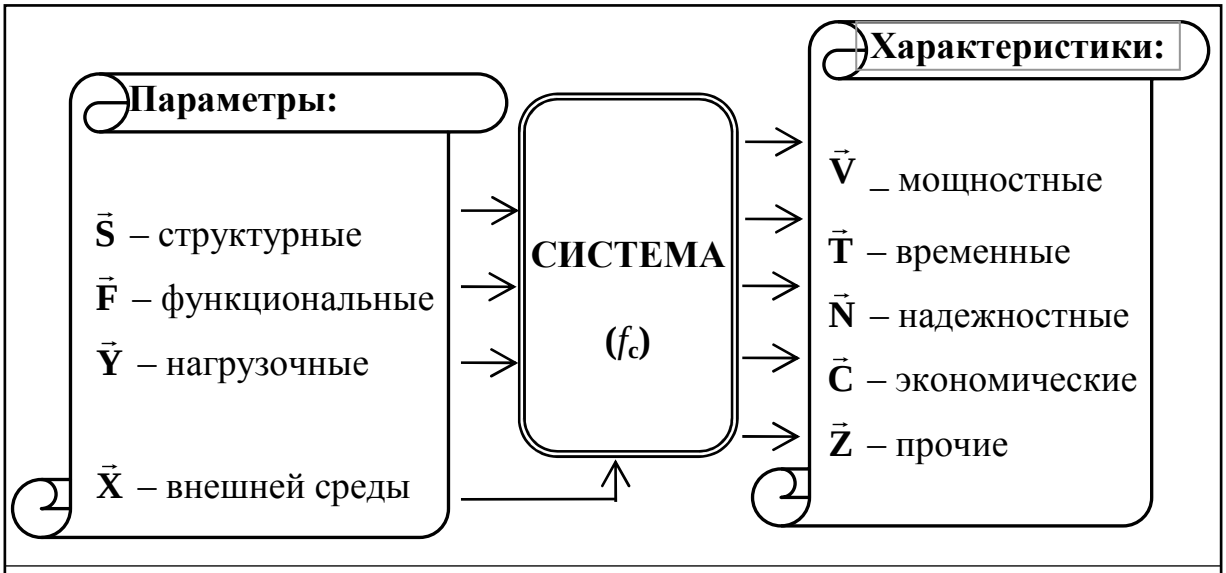


Рис.1.2. Взаимосвязь параметров и характеристик

1.1.7. Процесс

Изучение сложных систем удобно проводить в терминах процессов.

Процесс (от лат. processus – продвижение) – последовательная смена состояний системы во времени.

Состояние системы задается совокупностью значений переменных, описывающих это состояние. Система находится в некотором состоянии, если она полностью описывается значениями переменных, которые задают это состояние.

Система совершает ***переход*** из одного состояния в другое, если описывающие ее переменные изменяются от значений, задающих одно состояние, на значения, которые определяют другое состояние.

Причина, вызывающая переход из состояния в состояние, называется ***событием***.

Понятия «система» и «процесс» тесно взаимосвязаны и часто рассматриваются как эквивалентные понятия, к которым одинаково применимы термины «состояние» и «переход».

1.1.8. Классификация систем и процессов

Для унификации разрабатываемых моделей и методов исследования различных систем все многообразие существующих и возможных систем и процессов целесообразно разбить на отдельные классы, обладающие близкими свойствами и отображаемые определенными моделями, т.е. выполнить их классификацию. Обычно классификация выполняется в зависимости от конкретных признаков, в качестве которых будем использовать:

- способ изменения значений величин, описывающих состояния системы или процесса;
- характер протекающих в системе процессов;
- режим функционирования системы (режим процесса).

1. В зависимости от *способа изменения значений величин, описывающих состояния*, все системы и процессы делятся на два больших класса:

- ***с непрерывными состояниями,*** называемые также ***непрерывными системами (процессами)***, для которых характерен плавный переход из состояния в состояние, обусловленный тем, что величины, описывающие состояние, могут принимать любое значение из некоторого интервала (в том числе бесконечного), т.е. являются непрерывными;
- ***с дискретными состояниями,*** называемые также ***дискретными системами (процессами)***, для которых характерен скачкообразный переход из состояния в состояние, обусловленный тем, что величины, описывающие состояние, изменяются скачкообразно и принимают значения, которые могут быть пронумерованы, то есть являются дискретными, причем число состояний может быть как конечным, так и бесконечным.

2. В зависимости от *характера протекающих в системах процессов*, системы (процессы) делятся на:

- ***детерминированные***, поведение которых может быть предсказано заранее;
- ***стохастические (случайные, вероятностные)***, в которых процессы развиваются в зависимости от ряда случайных факторов, то есть являются случайными.

3. В зависимости от *режима функционирования*, системы (процессы) делятся на:

- системы, работающие в ***установившемся (стационарном) режиме (процесс установившийся*** или ***стационарный)***, когда характеристики системы не зависят от времени, то есть инвариантны по отношению ко времени функционирования системы;
- системы, работающие в ***неустановившемся режиме (процесс неустановившийся)***, когда характеристики системы меняются со временем, то есть зависят от времени функционирования системы; неустановившийся режим функционирования системы может быть обусловлен:
  - началом работы системы (***переходной режим***);
  - нестационарностью параметров системы (***нестационарный режим***), заключающейся в изменении параметров системы со временем;
  - перегрузкой системы (***режим перегрузки***), когда система не справляется с возложенной на нее нагрузкой.

1.2. Модель

«Если кажется, что работу сделать легко, это непременно будет трудно» (***Теорема Стакмайера***)

Модель – физический или абстрактный объект, адекватно отображающий исследуемую систему.

1.2.1. Основные требования к модели

Ко всем разрабатываемым моделям предъявляются два противоречивых требования:

- простота модели;
- адекватность исследуемой системе.

Требование *простоты модели* обусловлено необходимостью построения модели, которая может быть рассчитана доступными методами. Построение сложной модели может привести к невозможности получения конечного результата имеющимися средствами в приемлемые сроки и с требуемой точностью.

Степень сложности (простоты) модели определяется уровнем ее детализации, зависящим от принятых предположений и допущений: чем их больше, тем ниже уровень детализации и, следовательно, проще модель и, в то же время, менее адекватна исследуемой системе.

Адекватность (от лат. *adaequatus* – приравненный, равный) – соответствие модели оригиналу, характеризуемое степенью близости свойств модели свойствам исследуемой системы.

Адекватность математических моделей зависит от:

- степени полноты и достоверности сведений об исследуемой системе;
- уровня детализации модели.

При этом моделирование может проводиться:

- в условиях полной определенности, означающей наличие точной информации обо всех исходных параметрах;
- в условиях неопределенности, обусловленных:
  - неточностью сведений о параметрах;
  - отсутствием сведений о значениях некоторых параметров.

1.2.2. Классификация моделей

Многообразие систем, проявляющееся в многообразии их структурно-функциональной организации, определяет использование множества разных моделей, которые могут быть классифицированы в зависимости от:

1) *характера функционирования исследуемой системы*:

- **детерминированные**, функционирование которых описывается детерминированными величинами;

- **стохастические** или вероятностные, функционирование которых описывается случайными величинами;

2) *характера протекающих в исследуемой системе процессов*:

- **непрерывные**, в которых процессы протекают непрерывно во времени;
- **дискретные**, в которых процессы меняют свое состояние скачкообразно в дискретные моменты времени;

3) *степени достоверности исходных данных* об исследуемой системе:

- с априорно известными параметрами;
- с неизвестными параметрами;

4) *режима функционирования* системы:

- **стационарные**, в которых характеристики не меняются со временем;
- **нестационарные**, в которых характеристики изменяются со временем;

5) *назначения*:

- **статические** или **структурные**, отображающие состав и структуру системы;
- **динамические** или **функциональные**, отображающие функционирование системы во времени;
- **структурно-функциональные**, отображающие структурные и функциональные особенности организации исследуемой системы;

6) *способа представления (описания) и реализации*:

- **концептуальные** или **содержательные**, представляющие собой описание (в простейшем случае словесное) наиболее существенных особенностей структурно-функциональной организации исследуемой системы;
- **физические** или **материальные** – модели, эквивалентные или подобные оригиналу (макеты) или процесс функционирования которых такой же, как у оригинала и имеет ту же или другую физическую природу;
- **математические** или **абстрактные**, представляющие собой формализованное описание системы с помощью абстрактного языка, в частности с помощью математических соотношений, отражающих процесс функционирования системы;
- **программные (алгоритмические, компьютерные)** – программы для ЭВМ, позволяющие наглядно представить исследуемый объект посредством имитации или графического отображения математических зависимостей, описывающих искомый объект.

Соответственно различают физическое, математическое и компьютерное моделирование.

Между классами систем и моделей необязательно должно существовать однозначное соответствие. Например, дискретные системы

могут быть представлены в виде непрерывных моделей, а детерминированные системы – в виде вероятностных моделей, и наоборот.

В дальнейшем основное внимание уделяется математическому моделированию, широко используемому при исследовании сложных технических систем.

1.2.3. Параметризация моделей

Теоретические исследования сложных систем базируются на использовании моделей, отображающих объект исследования в форме, необходимой и достаточной для получения результатов, составляющих цель исследований.

Количественно любая модель, как и соответствующая ей система, описывается совокупностью величин, которые могут быть разбиты на *параметры* и *характеристики*. Состав параметров и характеристик модели определяется составом параметров и характеристик исследуемой системы и может в идеальном случае совпадать с ним. В общем случае составы параметров и характеристик модели и системы различаются, т.к. в первом случае они формулируются в терминах того математического аппарата, который используется при построении модели, а параметры и характеристики системы формулируются в терминах соответствующей прикладной области, к которой принадлежит система. В связи с тем, что, в общем случае, *параметры и характеристики системы и модели* различаются, их принято называть соответственно **системными** и **модельными**.

В связи с тем, что состав и номенклатура системных и модельных параметров и характеристик, в общем случае, различается, возникает необходимость установления соответствия между значениями системных и модельных параметров и характеристик, которое выполняется на этапе **параметризации** модели.

1.3. Задачи моделирования

> «Нет невыполнимой работы для человека, который не обязан делать ее сам» (***Закон Вейлера***)

Моделирование, как процесс исследования сложных систем, в общем случае предполагает решение следующих взаимосвязанных задач:

- разработка модели;
- анализ характеристик системы;
- синтез системы;
- детальный анализ синтезированной системы.

1.3.1. Разработка модели

Разработка модели состоит в выборе конкретного математического аппарата, в терминах которого формулируется модель, и построении модели или совокупности моделей исследуемой системы, отображающих

возможные варианты структурно-функциональной организации системы. В процессе разработки модели необходимо определить состав и перечень параметров и характеристик модели в терминах выбранного математического аппарата, и установить их взаимосвязь с параметрами и характеристиками исследуемой системы, то есть выполнить параметризацию модели.

1.3.2. Анализ характеристик

Анализ характеристик системы с использованием разработанной модели заключается в выявлении свойств и закономерностей, присущих процессам, протекающим в системах с различной организацией, и выработке рекомендаций для решения основной задачи системного проектирования – задачи синтеза.

1.3.3. Синтез системы

Синтез системы заключается в определении параметров системы, удовлетворяющих заданным требованиям к характеристикам системы.

Решение задачи синтеза связано с определением зависимостей характеристик функционирования системы от параметров, которые представляются сложными математическими конструкциями. При этом возможность получения приемлемых результатов в процессе решения задач синтеза из-за их сложности и большой трудоемкости, с учетом специфических особенностей реальных систем, превосходит возможности математических методов оптимизации, и задача синтеза в общем виде оказывается математически неразрешимой. Для того чтобы снизить сложность задачи синтеза, процесс проектирования разделяют на последовательность этапов, на каждом из которых решаются частные задачи синтеза – определяются параметры, связанные с отдельными аспектами организации системы, с использованием тех или иных моделей.

В зависимости от целей можно выделить следующие частные задачи (этапы) синтеза:

- **структурный синтез**, состоящий в выборе способа структурной организации системы, в рамках которой могут быть удовлетворены требования технического задания; структурный синтез включает в себя два этапа:
  - *элементный синтез*, состоящий в определении требований к параметрам отдельных элементов системы;
  - *топологический (конфигурационный) синтез*, состоящий в определении способа взаимосвязи элементов системы, т.е. топологии (конфигурации) системы;
- **функциональный синтез**, состоящий в выборе режима (способа) функционирования системы;
- **нагрузочный синтез**, состоящий в определении требований к параметрам нагрузки, обеспечивающим функционирование системы с заданным качеством.

На каждом из перечисленных этапов синтеза определяются значения соответствующего подмножества параметров, характеризующих структурную, функциональную организацию системы или нагрузку, возлагаемую на систему. При этом значения параметров оптимизируются лишь в отношении факторов, учитываемых на каждом из этапов синтеза, но не в отношении системы в целом. Поэтому многоэтапный синтез позволяет получить лишь приближенные оптимальные решения, качество которых проверяется путем детального анализа синтезированной системы.

1.3.4. Детальный анализ синтезированной системы

Детальный анализ синтезированной системы проводится с целью оценки качества решения задачи системного проектирования и полученных в процессе синтеза параметров системы, а также выявления предельных возможностей системы, узких мест в системе и т.д.

Поскольку задача синтеза обычно решается на моделях, использующих упрощающие решение предположения и допущения, анализ синтезированной системы, выполняемый с целью определения фактической эффективности конкретных значений характеристик, обычно проводится на основе более детальных моделей, в качестве которых чаще всего используются имитационные или комбинированные (например, аналитико-имитационные) модели.

1.4. Методы моделирования

«Все не так легко, как кажется»
(***Следствие закона Мэрфи***)

В зависимости от целей моделирование может проводиться на двух уровнях:

- *на качественном*;
- *на количественном*.

Соответственно применяются модели:

- изобразительные (наглядные);
- конструктивные.

Математическое моделирование обычно проводится на количественном уровне с использованием конструктивных моделей.

При исследовании технических систем с дискретным характером функционирования наиболее широкое применение получили следующие методы математического моделирования:

- **аналитические** (аппарат теории вероятностей, теории массового обслуживания, теории случайных процессов, методы оптимизации, …);
- **численные** (применение методов численного анализа для получения конечных результатов в числовой форме, когда невозможно получить аналитические зависимости характеристик от параметров в явном виде);

- **статистические** или **имитационные** (исследования на ЭВМ, базирующиеся на методе статистических испытаний и предполагающие применение специальных программных средств и языков моделирования: GPSS [4, 18], SIMULA, ИМСС [11] и др.).
- **комбинированные**.

1.4.1. Аналитические методы

Аналитические методы состоят в построении математической модели в виде математических символов и отношений, при этом требуемые зависимости выводятся из математической модели последовательным применением математических правил.

Достоинство аналитических методов заключается в возможности получения решения в явной аналитической форме, позволяющей проводить детальный анализ процессов, протекающих в исследуемой системе, в широком диапазоне изменения параметров системы. Результаты в аналитической форме являются основой для выбора оптимальных вариантов структурно-функциональной организации системы на этапе синтеза.

Недостаток аналитических методов – использование целого ряда допущений и предположений в процессе построения математических моделей и невозможность, в некоторых случаях, получить решение в явном виде из-за неразрешимости уравнений в аналитической форме, отсутствия первообразных для подынтегральных функций и т.п. В этих случаях широко применяются численные методы.

Аналитические методы можно разделить на:

- точные;
- приближенные;
- эвристические.

1.4.2. Численные методы

Численные методы основываются на построении конечной последовательности действий над числами. Применение численных методов сводится к замене математических операций и отношений соответствующими операциями над числами, например, к замене интегралов суммами, бесконечных сумм – конечными и т.п. Результатом применения численных методов являются таблицы и графики зависимостей, раскрывающих свойства объекта. Численные методы являются продолжением аналитических методов в тех случаях, когда результат не может быть получен в явном виде. Численные методы по сравнению с аналитическими методами позволяют решать значительно более широкий круг задач.

1.4.3. Статистические методы

В тех случаях, когда анализ математической модели даже численными методами может оказаться нерезультативным из-за

чрезмерной трудоемкости или неустойчивости алгоритмов в отношении погрешностей аппроксимации и округления, строится имитационная модель, в которой процессы, протекающие в ВС, описываются как последовательности операций над числами, представляющими значения входов и выходов соответствующих элементов. *Имитационная модель* объединяет свойства отдельных элементов в единую систему. Производя вычисления, порождаемые имитационной моделью, можно на основе свойств отдельных элементов определить свойства всей системы.

При построении имитационных моделей широко используется *метод статистических испытаний* (метод Монте-Карло). Процедура построения и анализа имитационных моделей методом статистических испытаний называется **статистическим моделированием**. Статистическое моделирование представляет собой процесс получения статистических данных о свойствах моделируемой системы.

Достоинством статистического моделирования является *универсальность*, гарантирующая принципиальную возможность проведения анализа систем любой степени сложности с любой степенью детализации.

Недостаток статистического моделирования – *трудоемкость* процесса моделирования и *частный характер результатов*, не раскрывающий зависимости, а лишь определяющий ее в отдельных точках.

Статистическое моделирование широко используется для оценки погрешностей аналитических и численных методов.

1.4.4. Комбинированные методы

Комбинированные методы представляют собой комбинацию выше перечисленных методов, в частности:

- **численно-аналитические**, в которых часть результатов получается численно, а остальные – с использованием аналитических зависимостей;
- **аналитико-имитационные**, представляющие собой имитационное моделирование в сочетании с аналитическими методами, позволяющими сократить время моделирования за счет определения значений ряда характеристик на основе аналитических зависимостей по значениям одной или нескольких характеристик, найденных путем статистической обработки результатов имитационного моделирования.

1.5. Резюме

1. Объектами моделирования в технике, в общем случае, являются *системы* и *комплексы*, обладающие *структурной и функциональной организацией*. *Структура системы* может быть задана в *графической* или *аналитической* форме. *Функция системы* может быть задана в *алгоритмической*, *аналитической*, *графической* или *табличной* форме.

Системе присущи такие *свойства* как *целостность*, *связность*, *организованность* и *интегративность*. Наличие этих свойств означает, что *систему нельзя рассматривать как простую совокупность элементов*,

поскольку, изучая каждый элемент системы в отдельности, нельзя познать все свойства системы в целом.

2. Моделирование направлено на решение задач *анализа*, связанных с оценкой *эффективности* систем, и *синтеза*, направленных на построение *оптимальных систем* в соответствии с выбранным *критерием эффективности*. Эффективность системы задается в виде совокупности *показателей эффективности*, каждый из которых служит мерой одного свойства системы. Мера эффективности, обобщающая все или некоторые, наиболее существенные, свойства системы в одной оценке называется *критерием эффективности*.

3. Для количественного описания системы используются *параметры*, описывающие *первичные* свойства системы, и *характеристики*, определяемые в процессе решения задач анализа как функция параметров.

Множество параметров технических систем можно разделить на внутренние *(структурные и функциональные)* и внешние *(нагрузочные и параметры внешней среды)*. Параметры могут быть *детерминированными* или *случайными* и *управляемыми* или *неуправляемыми*.

Основными характеристиками технических систем являются характеристики *производительности, оперативности, надежности* и *стоимости*.

4. Изучение сложных систем удобно проводить в терминах *процессов*, с которыми связаны такие понятия как *состояние, переход* из одного состояния в другое и *событие*.

Процессы и соответственно системы, в которых они протекают, могут быть квалифицированы в зависимости:

- от способа изменения значений величин, описывающих состояния (непрерывные и дискретные);
- от характера протекающих в системах процессов (детерминированные и стохастические или, что то же самое, случайные, вероятностные);
- от режима функционирования системы (с установившимся или стационарным режимом и с неустановившимся режимом).

Неустановившийся режим может быть обусловлен началом работы системы (*переходной режим*), нестационарностью параметров системы (*нестационарный режим*), перегрузкой системы (*режим перегрузки*).

5. К разрабатываемым моделям обычно предъявляются два *противоречивых* требования: *простота* и *адекватность* исследуемой системе.

Модели могут быть классифицированы в зависимости от характера функционирования исследуемой системы (*детерминированные* и *стохастические* или *вероятностные*), от характера протекающих в исследуемой системе процессов (*непрерывные* и *дискретные*), от режима функционирования системы (*стационарные* и *нестационарные*), от способа представления и реализации (*концептуальные* или

содержательные; *физические* или *материальные*; *математические* или *абстрактные*; *программные* или *компьютерные*).

Одним из важнейших этапов при разработке модели является *этап параметризации*, заключающийся в установлении соответствия между значениями системных и модельных параметров и характеристик.

6. Моделирование, как многоэтапный процесс исследования сложных систем, в общем случае предполагает решение следующих взаимосвязанных задач: разработка модели, анализ характеристик системы, синтез системы, детальный анализ синтезированной системы.

При исследовании технических систем с дискретным характером функционирования наиболее широкое применение получили *аналитические, численные, статистические* (*имитационные*) и *комбинированные* методы математического моделирования:

1.6. Практикум: обсуждение

При изложении любой научной и технической дисциплины одним из основных вопросов является формирование терминологии, служащей в дальнейшем фундаментом для изучения теоретических и практических аспектов данной дисциплины. Целью формирования терминологической основы является установление однозначного соответствия между используемым термином и вкладываемым в него смысловым содержанием.

В разделе 1 определены *базовые* понятия и термины теории моделирования, которые используются в последующих разделах.

Для того чтобы закрепить представление о введенных терминах и понятиях, попытаемся ответить на некоторые конкретные вопросы практического характера.

Вопрос 1. Можно ли персональный компьютер рассматривать как систему, элементами которого являются системный блок и связанные с ним внешние устройства – монитор, принтер и сканер?

Обсуждение. Понятие «система» широко используется в нашей повседневной жизни. Мы говорим «система знаний», «система оценок», «система взглядов» и т.д. Однако можно заметить, что приведенные термины не вполне соответствуют данному выше определению. И хотя можно попытаться найти в этих «системах» элементы и связи между ними, это всё-таки будет выглядеть несколько искусственно притянутым. Такая ситуация обусловлена тем, что в этих примерах понятие «система» используется в широком, можно сказать, общепринятом смысле. В каждой предметной области может быть введена своя трактовка понятия «система». При моделировании технических систем будем руководствоваться данным выше определением понятия «система».

Итак, если воспользоваться определением системы, как совокупности взаимосвязанных элементов, то вроде бы компьютер с внешними устройствами можно считать системой. Однако следует обратить внимание на вторую часть определения понятия «система», где сказано, что элемен-

ты, объединенные в одно целое должны обеспечивать достижение цели, определяемой назначением системы. Это означает, что система, кроме структурной организации в виде совокупности взаимосвязанных элементов, должна обладать и функциональной организацией, то есть в ней должны протекать некоторые процессы во времени, изменяющие состояние системы. С этих позиций неработающий компьютер не может трактоваться как система. В то же время, если в этом компьютере выполняется некоторая задача, его можно рассматривать как систему, обладающую структурной и функциональной организацией. Однако и здесь имеются некоторые нюансы, которые следует учитывать при выявлении соответствия рассматриваемого объекта введенному понятию «система».

Вспомним, что система должна обладать такими свойствами, как целостность, связность, организованность и интегративность. Наличие этих свойств позволяет рассматривать систему как единое целое и применять для её исследования системный подход. Особенно важным является последнее свойство – интегративность, свидетельствующее о том, что невозможно полностью познать систему, анализируя только свойства её элементов. Другими словами, система может обладать свойствами, которые не присущи ни одному из входящих в её состав элементов.

Вопрос 2. Насколько велико различие между «параметрами» и «характеристиками» системы? Могут ли характеристики быть параметрами и наоборот?

Обсуждение. В некоторых литературных источниках не акцентируется внимание на различии между параметрами и характеристиками. Более того, часто одни и те же величины называются то параметрами, то характеристиками.

Введенные выше определения четко разделяют описывающие систему величины на два класса: «параметры» и «характеристики». Характеристики системы являются функциями параметров, то есть изменение какого-либо параметра приводит к изменению характеристик системы.

В то же время следует понимать, что «параметры» и «характеристики» – понятия относительные. Это можно показать на следующем примере. Если выполняющий некоторые задачи компьютер рассматривается как система, одним из элементов которой является процессор, то производительность (быстродействие) процессора является параметром, изменение которого приведет к изменению такой величины, как время выполнения задачи, которая в данном случае представляет собой характеристику системы. Если же процессор рассматривается как система, состоящая из арифметико-логического устройства, устройства управления, регистровой памяти и т.д., то быстродействие процессора будет являться характеристикой, которая зависит от параметров входящих в ее состав элементов. Можно было бы сказать, что параметры системы в основном описывают элементы системы и их взаимосвязь (как структурную, так и функциональ-

ную), а характеристики описывают систему в целом. Однако это будет не совсем корректно, поскольку характеристики могут описывать как систему в целом (глобальные характеристики), так и её отдельные элементы и подсистемы (локальные характеристики).

Вопрос 3. Являются ли синонимами термины «показатель эффективности» и «характеристика»?

Обсуждение. Действительно, термины «показатель эффективности» и «характеристика» довольно близкие понятия по определению. Можно даже считать, что это одно и то же. И все же, между ними существует определенное различие.

Во-первых, как сказано выше, показатель эффективности всегда имеет количественный смысл, т.е. представляется в виде количественной оценки, в то время как характеристика может иметь качественный характер. Так, например, при описании вычислительных сетей широко используются такие характеристики, как открытость, масштабируемость, гибкость, информационная безопасность и т.п., количественное задание которых либо достаточно условно, либо вообще невозможно.

Во-вторых, множество показателей эффективности при исследовании некоторой системы зависит от её назначения, в то время как характеристики описывают всю совокупность свойств системы. При этом, возможно, что некоторые характеристики являются несущественными. Например, если компьютер предназначен для использования в космосе или на борту самолета, то важными показателями эффективности являются его вес и энергопотребление. Если же компьютер предназначен для решения сложных задач моделирования, оптимизации или игровых задач (например, шахматных), требующих большой вычислительной мощности, то более актуальными становятся такие показатели эффективности как производительность, время реакции, а вес и энергопотребление могут вообще не иметь особого значения.

Вопрос 4. Сколько критериев эффективности используется при синтезе оптимальной системы?

Обсуждение. Для описания системы обычно используется множество зачастую противоречивых показателей эффективности. Эти противоречия заключаются в том, что попытка улучшить какой-то один или несколько показателей эффективности за счет изменения параметров структурно-функциональной организации системы обычно приводит к ухудшению остальных показателей эффективности. Например, если мы хотим построить высокопроизводительную и сверхнадежную вычислительную систему, то, очевидно, что ее стоимость окажется чрезвычайно большой. С другой стороны, если задаться целью построить как можно более дешевую вычислительную систему, то вряд ли ее производительность и надежность окажутся на должном уровне. Таким образом, для решения задачи оптимального синтеза системы целесообразно иметь *один*

критерий эффективности, то есть одну целевую функцию, позволяющую выбрать из множества вариантов построения системы наилучший, а точнее оптимальный вариант, то есть такой, при котором критерий эффективности принимает максимальное (прямой критерий эффективности) или минимальное (инверсный критерий эффективности) значение. Существует несколько способов построения критерия эффективности при наличии множества показателей эффективности, которые рассматриваются в разделе «Технология моделирования». Это, прежде всего так называемые составные критерии эффективности (аддитивные и мультипликативные), представляющие объединение многих показателей эффективности. Однако на практике более широкое распространение получили критерии эффективности с ограничениями, которые строятся по следующему принципу: из множества показателей эффективности один выбирается в качестве Критерия эффективности, а на остальные показатели налагаются ограничения.

В то же время, следует заметить, что вид критерия эффективности зависит от назначения системы. Если система предназначена для обеспечения высокой надежности, то в качестве критерия эффективности может использоваться один из показателей надежности. Если же система должна иметь высокую производительность, то в качестве критерия эффективности следует использовать производительность системы. Возможна ситуация, когда к проектируемой системе предъявляются требования и высокой производительности и надежности. Тогда в качестве критерия эффективности можно использовать составной критерий эффективности, объединяющий два показателя эффективности – производительность и надежность.

В заключение хотелось бы заметить, что фразы типа «более оптимальная система» или «менее оптимальная система» являются некорректными, поскольку оптимальная система существует в единственном экземпляре. Это та система, которая обеспечивает экстремум (максимум или минимум) функции, задающей критерий эффективности.

Вопрос 5. В литературе часто встречается такое понятие как «многокритериальная задача». Означает ли это, что задача оптимального синтеза может решаться с использованием сразу нескольких критериев эффективности?

Обсуждение. Действительно, понятие «многокритериальная задача» достаточно широко используется в такой математической дисциплине как «Исследование операций». Задачи, в которых имеется одна целевая функция (один критерий эффективности), принимающая численные значения, относятся к задачам математического (или оптимального) программирования. Им противостоят задачи с несколькими целевыми функциями или с одной целевой функцией, но принимающей векторные значения или значения ещё более сложной природы. Эти задачи называются многокритериальными и решаются путём сведения (часто

условного) к задачам с единственной целевой функцией. Многокритериальными задачами являются задачи теории игр, изучающей формальные модели принятия оптимальных решений в условиях конфликта. При этом под конфликтом понимается явление, в котором участвуют различные стороны, наделённые различными интересами, выраженными в виде целевых функций (критериев эффективности), и возможностями выбирать доступные для них действия в соответствии с этими интересами. В условиях конфликта стремление противника скрыть свои предстоящие действия порождает неопределённость. Поэтому теория игр рассматривается также как теория принятия оптимальных решений в условиях неопределённости.

Вопрос 6. Можно ли систему, работающую в неустановившемся режиме, исследовать методами, разработанными для установившегося режима?

Обсуждение. Как сказано выше, неустановившийся режим работы системы может быть обусловлен тремя факторами: началом работы системы, нестационарностью нагрузки и перегрузками.

Большинство исследований технических систем обычно проводится в предположении, что переходной режим завершился, и в системе отсутствуют перегрузки. В противном случае следует использовать специальные методы исследования, которые в том или ином виде разработаны для переходного режима и режима перегрузок.

В то же время многие реальные системы, в том числе технические, работают в неустановившемся режиме, обусловленном нестационарностью нагрузки. Для исследования таких систем методами, разработанными для установившегося режима, могут использоваться различные подходы, основными среди которых являются следующие. Во-первых, можно попытаться выделить достаточно продолжительные интервалы времени, в течение которых нагрузка не изменяется, то есть является стационарной, или же изменение нагрузки незначительно и им можно пренебречь. Во-вторых, исследование методами, разработанными для установившегося режима, можно проводить в расчете на максимальную или некоторую среднюю нагрузку.

Вопрос 7. Каким способом достигается разумный компромисс между простотой и адекватностью модели?

Обсуждение. Достижение разумного компромисса между простотой модели и ее адекватностью исследуемой системе является одной из сложнейших проблем теории моделирования. Действительно, с одной стороны, желательно иметь модель с максимальной степенью детализации, отражающую все особенности структурно-функциональной организации. С другой стороны, такая модель может оказаться настолько сложной, что ее исследование будет невозможным или же потребует неоправданно больших материальных и временны*х* ресурсов. Следует также учитывать, что для исследования сложных моделей обычно невозможно разработать

точные математические методы, а применение громоздких приближенных методов может привести к значительным погрешностям результатов. Есть ещё один важный момент, который следует иметь в виду при разработке моделей. Это точность представления исходных данных, особенно связанных с нагрузочными параметрами. Если погрешность представления нагрузочных или структурно-функциональных параметров велика, то, очевидно, нет смысла строить сверхточную модель.

Более подробно и предметно проблема выбора уровня детализации разрабатываемой модели обсуждается в последнем разделе «Технология моделирования».

Вопрос 8. Каково значение параметризации модели в процессе исследования реальной системы?

Обсуждение. Этап параметризации модели в процессе исследования реальной системы имеет большое значение для получения корректных результатов. На этом этапе фактически закладывается фундамент адекватности модели исследуемой системе, поскольку именно в процессе параметризации определяются значения исходных параметров, которые будут использованы в модели и обеспечат достоверность получаемых результатов. Ошибки, заложенные при неудачной параметризации, не смогут быть компенсированы даже применением сверхточной (адекватной) модели и точных методов расчета. Более того, ошибки параметризации могут многократно увеличиться и привести к получению абсолютно неправильных значений исследуемых характеристик.

Следует также отметить, что на этапе параметризации устанавливается соответствие не только между значениями системных и модельных параметров и характеристик, но и терминологическое соответствие между заданными в терминах конкретной прикладной области понятиями и элементами исследуемой системы и используемыми в соответствующей математической дисциплине понятиями и элементами математической модели.

Например, в вычислительной технике при описании компьютера применяются такие понятия и элементы, как задача, программа, данные, процессор, память и т.д. Положим, что в качестве математической модели компьютера используется случайный процесс, для описания которого в теории случайных процессов используются такие термины и элементы, как состояние, переход, событие, граф переходов, матрица вероятностей переходов и т.д. Выявить и грамотно установить соответствие между указанными понятиями и элементами и является одной из задач этапа параметризации.

Фактически, параметризация – это промежуточный этап установления взаимнооднозначного соответствия между концептуальной и математической моделями.

Вопрос 9. Насколько необходим детальный анализ спроектированной системы?

Обсуждение. Настолько, насколько важно получение качественного проекта синтезируемой системы. Задача синтеза обычно решается с использованием сравнительно простых моделей, позволяющих получить решение в явной аналитической форме. При этом погрешность модели, а также методов расчета характеристик системы в случае применения приближенных аналитических зависимостей может привести к значительным различиям между расчетными и реальными значениями оптимизируемых параметров. В связи с этим возникает необходимость проверки и уточнения найденных значений параметров структурно-функциональной организации системы, для чего, естественно, необходимо использовать наиболее адекватные модели, позволяющие получить результаты, в максимальной степени соответствующие реальным. В качестве таких моделей обычно применяются имитационные модели, которые могут быть построены с максимальным приближением к реальной системе за счёт большей детализации по сравнению с аналитической моделью.

Кроме того, в процессе детального анализа синтезированной системы должны быть выявлены предельные возможности системы, узкие места в системе, а также определено, насколько хорошо (с каким запасом) выполняются заданные требования к качеству функционирования проектируемой системы.

Вопрос 10. Если, как сказано выше, статистические (имитационные) методы исследования сложных систем являются универсальными, то насколько актуально применение аналитических методов?

Обсуждение. Действительно, имитационное моделирование может рассматриваться как универсальное средство исследования сложных систем со стохастическим характером функционирования, позволяющее проводить анализ эффективности функционирования систем любой степени сложности с любой степенью детализации. Единственным фактором, ограничивающим применение имитационного моделирования, является производительность компьютера, на котором выполняются имитационные эксперименты. Естественно, чем сложнее исследуемая система, чем больше в ней элементов и связей, тем более мощный требуется компьютер, в пределе возможно даже суперЭВМ. При этом мощность компьютера подразумевает не только скорость процессорной обработки, но и большую ёмкость оперативной памяти, а в некоторых случаях – высокие требования к производительности и ёмкости внешней памяти.

В то же время, имитационное моделирование обладает недостатками, ограничивающими его применение. Одним из них является частный характер результатов, не раскрывающий зависимостей характеристик функционирования системы от параметров её структурно-функциональной организации, а лишь определяющий ее в отдельных точках.

Кроме того, имитационное моделирование может служить эффективным инструментом в процессе проектирования только в том случае, если требуется сравнить несколько вариантов построения системы и выбрать из них наилучший. Однако, практически невозможным (либо это сопряжено с большими временны*ы*ми и материальными затратами) оказывается решение задачи оптимального синтеза сложных систем, характеризующихся большой размерностью, то есть наличием большого числа структурно-функциональных и нагрузочных параметров.

Таким образом, аналитические методы моделирования следует применять в следующих случаях:

- для выполнения оценочных расчетов на этапе предварительного анализа и проектирования, не требующих высокой точности получаемых результатов;
- для изучения в широком диапазоне изменения параметров свойств и закономерностей, присущих исследуемой системе; полученные результаты могут служить основой для формирования рекомендаций по проектированию систем;
- для решения задач оптимального синтеза при проектировании новых систем.

Вопрос 11. В некоторых литературных источниках вместо понятия «оптимальная система» используется понятие «рациональная система». Каково соотношение между этими двумя понятиями?

Обсуждение. «Оптимальная система» означает, что значения параметров структурно-функциональной организации определены в процессе решения математической оптимизационной задачи и являются оптимальными, то есть обеспечивают экстремум выбранного критерия эффективности. На практике может оказаться невозможным построить систему с такими значениями параметров, что может быть обусловлено разными причинами, в том числе, дискретным характером оптимизируемых параметров.

Например, в процессе синтеза некоторой сети передачи данных получены следующие оптимальные значения пропускных способностей трёх каналов связи: 428 кбит/с, 764 кбит/с и 931 кбит/с. Положим, что реальные каналы связи могут иметь пропускные способности в 256 кбит/с, 512 кбит/с и 1024 кбит/с. Очевидно, что в качестве окончательного решения задачи проектирования будут приняты значения 512 кбит/с, 512 кбит/с (или 1024 кбит/с) и 1024 кбит/с. Поскольку эти значения отличаются от оптимальных, спроектированная система не может называться оптимальной. Такую систему обычно называют «рациональной», имея в виду, что ее параметры близки, но не равны оптимальным значениям.

Другой случай, когда в результате оптимизации получено значение пропускной способности канала 2000 кбит/с, которое существенно превышает максимально допустимое значение в 1024 кбит/с. Очевидно,

что в этом случае одно из возможных решений состоит в установке двух каналов с пропускной способностью 1024 кбит/с, что также не будет соответствовать оптимальному варианту.

Иногда под «рациональной системой» подразумевают некоторый вариант её построения, выбранный из нескольких возможных вариантов на основе анализа характеристик функционирования. Ясно, что в этом случае вообще нет речи об оптимизации.

Вопрос 12. В чем различие между понятиями «синтез» и «проектирование»?

Обсуждение. Эти понятия достаточно близкие по смыслу и часто используются как синонимы. В то же время между ними существует некоторое различие, вытекающее, прежде всего, из их иностранного происхождения.

Термин «синтез» (от греческого слова synthesis – соединение, сочетание, составление) означает соединение различных элементов в единое целое – систему и неразрывно связан с термином «анализ».

Термин «проектирование» (от латинского слова projectus, буквально означающего – брошенный вперёд) означает процесс создания проекта – прототипа новой системы.

В процессе проектирования технических систем основная задача заключается в создании проекта, на основе которого строится реальная система, а в процессе синтеза – только определяются параметры и состав проектируемой системы, которые в окончательном проекте могут значительно отличаться от «синтезированных». Таким образом, синтез можно рассматривать как один (может быть даже основной) из этапов проектирования реальных систем.

Можно также считать, что «синтез» – понятие математическое, которое часто используется в таком сочетании как «оптимальный синтез», а «проектирование» – понятие скорее техническое и не всегда предполагает применение каких-то математических методов для построения системы. Другими словами, синтез технических систем реализуется с использованием математических методов моделирования, в то время как проектирование предполагает, прежде всего, применение различных инженерно-технических решений, обоснование которых может осуществляться математическими расчетами.

1.7. Самоконтроль: перечень вопросов

1. Дать определение понятий: моделирование, элемент, система, сложная система, комплекс, структура, функция, структурная и функциональная организация, анализ, синтез, эффективность, показатель эффективности, критерий эффективности, оптимальная система.
2. В каких случаях моделирование оправдано и необходимо?

3. Перечислить и дать краткую характеристику способов описания структуры системы. Проиллюстрировать эти способы на примере персонального компьютера.

4. Перечислить и дать краткую характеристику способов описания функции системы. Проиллюстрировать эти способы на примере решения задачи в компьютере.

5. Способ достижения поставленной цели за счет выбора определенной структуры и функции системы называется ...?

6. Чем отличается реализация функциональной организации системы от структурной?

7. Что определяется в процессе анализа системы?

8. Что определяется в процессе синтеза системы?

9. Чем оценивается эффективность системы?

10. Чем инверсный критерий эффективности отличается от прямого?

11. Что понимается под оптимальной системой?

12. Свойства, присущие сложной системе, и их краткая характеристика.

13. В чем состоит различие между параметрами и характеристиками?

14. Перечислить состав параметров технической системы. Привести примеры структурных, функциональных, нагрузочных параметров.

15. Перечислить состав характеристик технической системы. Привести примеры мощностных, надежностных, стоимостных характеристик.

16. В чем состоит проблема выбора уровня детализации моделей?

17. Перечислить основные этапы моделирования систем.

18. Методы моделирования систем, их достоинства и недостатки.

19. Какой метод исследования систем является наиболее точным?

20. Какой метод исследования систем является наиболее универсальным?

21. Какой метод позволяет выполнять исследование систем на моделях любой степени детализации?

Раздел 2. ЭЛЕМЕНТЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ

«Число разумных гипотез, объясняющих любое данное явление, бесконечно»
(***Постулат Персига***)

Математическое моделирование дискретных систем со стохастическим характером функционирования предполагает использование моделей массового обслуживания, описываемых в терминах аппарата теории вероятностей. В данном разделе, не претендуя на полноту, рассматриваются некоторые элементы теории вероятностей, знание которых необходимо для понимания и усвоения материала следующих разделов, связанного с грамотным описаниием и расчётом вероятностных моделей, а также осмысленным анализом полученных результатов.

2.1. Основные понятия и определения

Базовыми понятиями в теории вероятностей являются «*событие*», «*вероятность*», *случайная величина*».

2.1.1. Событие, вероятность

«Если какая-нибудь неприятность может произойти, она случается...»
(***Закон Мэрфи***)

Событие – всякий факт, который в результате опыта может произойти или не произойти.

Вероятность события есть численная мера степени объективной возможности этого события.

Предположим, что рассматривается некоторый опыт или явление, в котором в зависимости от случая происходит или не происходит некоторое событие A.

Если условия опыта могут быть воспроизведены многократно, так что в принципе осуществима целая серия одинаковых и *независимых* друг от друга испытаний, то **вероятность** события A может быть вычислена по следующей формуле:

$$\boxed{\mathrm{P}(A) = m/n},$$

где n – общее число взаимно исключающих друг друга исходов; m – число исходов, которые приводят к наступлению события A.

Вероятность может принимать значения от 0 до 1.

Событие, вероятность которого равна 0, называется ***невозможным***, а событие, вероятность которого равна 1, называется ***достоверным***.

Несколько событий образуют ***полную группу событий***, если в результате опыта должно непременно появиться хотя бы одно из них.

Несколько событий называются ***несовместными*** в данном опыте, если никакие два из них не могут появиться вместе.

Несколько событий называются ***равновозможными*** в данном опыте, если ни одно из этих событий не является объективно более возможным,

чем другое.

События называются ***независимыми***, если появление одного из них не зависит от того, произошли ли другие события.

2.1.2. Случайная величина

Случайной величиной называется величина, которая может принимать то или иное значение, *неизвестное заранее*.

Случайные величины могут быть двух типов:

- **дискретные (прерывные)**, принимающие только отделённые друг от друга значения, которые можно пронумеровать;
- **непрерывные (аналоговые)**, которые могут принимать любое значение из некоторого промежутка.

Примерами дискретных случайных величин могут служить:

- количество задач, выполняемых вычислительной системой (ВС) за день;
- количество обращений к внешней памяти в процессе решения задачи;
- количество сообщений, переданных в компьютерной сети за единицу времени, и т.д.

Примерами непрерывных случайных величин являются:

- интервалы времени между моментами поступления в ВС запросов на решение задач или между моментами формирования сообщений, передаваемых в телекоммуникационную сеть;
- время выполнения задач в ВС и т.д.

Иногда случайные величины, имеющие дискретную природу, рассматриваются как непрерывные. Такая замена оправдана в тех ситуациях, когда случайная величина принимает большое множество значений, которые незначительно отличаются друг от друга, так что замена дискретной случайной величины непрерывной практически не влияет на результаты расчетов. Например, время передачи пакета по каналу связи в вычислительной сети, определяемое как отношение длины передаваемого пакета (в битах) к пропускной способности канала связи (бит/с), которое является дискретной случайной величиной, обычно рассматривается как непрерывная случайная величина, изменяющаяся, в общем случае, в интервале от нуля до некоторого предельного значения, определяемого максимально возможной длиной пакета.

Случайные величины часто обозначают большими буквами, а их возможные значения – соответствующими малыми буквами. Например, случайная величина X – число обращений к накопителю на магнитном диске в процессе решения задачи в вычислительной системе – может принимать значения $x_1 = 0,\ x_2 = 1,\ x_3 = 2,\ x_4 = 3, \ldots$.

2.2. Законы распределений случайных величин

«Всякая работа требует больше времени, чем вы думаете» (***Следствие закона Мэрфи***)

Математическое описание случайных величин предполагает задание закона распределения, устанавливающего соответствие между значениями случайной величины и вероятностью их появления.

Рассмотрим дискретную случайную величину X, принимающую значения $x_1, x_2, ..., x_n$. Величина X может принять каждое из этих значений с некоторой вероятностью. Обозначим через p_i $(i = \overline{1, n})$ вероятность того, что случайная величина X примет значение x_i: $p_i = \mathrm{P}(X = x_i)$. Если в результате опыта величина X принимает только одно из этих значений, то имеем *полную группу несовместных событий* и сумма вероятностей всех возможных значений случайной величины равна единице:

$$\sum_{i=1}^{n} p_i = 1.$$

Эта суммарная вероятность каким-то образом распределена между отдельными значениями. Случайная величина будет *полностью описана* с вероятностной точки зрения, если мы зададим это распределение, т.е. установим так называемый *закон распределения*.

Законом распределения случайной величины называется всякое соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями. Про случайную величину говорят, что она подчинена данному закону распределения.

2.2.1. Закон распределения дискретной случайной величины

Закон распределения дискретной случайной величины X (дискретный закон распределения), принимающей значения $x_1, x_2, ..., x_n$, может быть задан одним из следующих способов:

• ***аналитически*** в виде *математического выражения*, отражающего зависимость вероятности от значения случайной величины:

$$p_i = f(x_i) \quad (i = \overline{1, n});$$

• ***таблично*** в виде *ряда распределения* случайной величины, в котором перечислены возможные значения случайной величины и соответствующие им вероятности:

Значения случайной величины X	x_1	x_2	…	x_n
Вероятности P	p_1	p_2	…	p_n

• ***графически*** в виде *многоугольника распределения*, где по оси абсцисс откладываются возможные значения случайной величины, а по оси ординат – вероятности этих значений (рис.2.1).

Графическое представление закона распределения дискретной

случайной величины обладает наглядностью и позволяет судить о близости к тому или иному типовому закону.

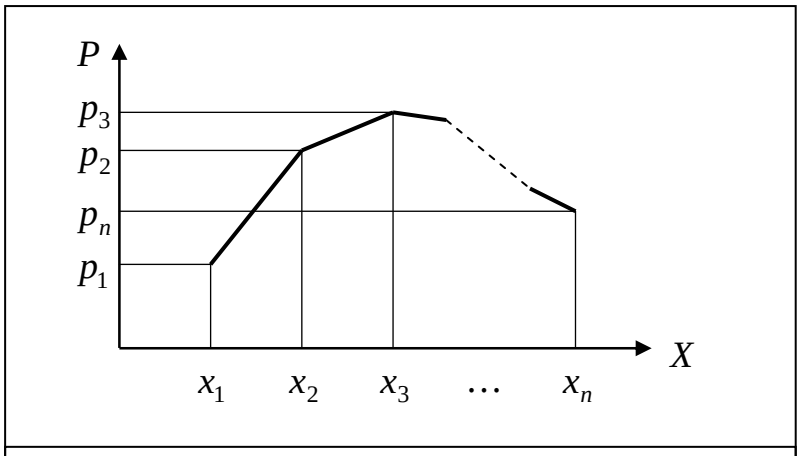


Рис.2.1. Многоугольник распределения

В качестве примеров *дискретных законов* распределения ниже рассматриваются широко используемые в теории массового обслуживания законы распределения Пуассона и геометрический.

2.2.2. Закон распределения непрерывной случайной величины

Для непрерывной случайной величины невозможно задать закон распределения в том виде, в каком он задается для дискретной величины, поскольку непрерывная случайная величина имеет *бесконечное множество возможных значений*, сплошь заполняющих некоторый промежуток, и вероятность появления любого конкретного значения *равна нулю*.

В связи с этим, для описания непрерывных случайных величин используется другой способ установления соответствия между значениями случайной величины и вероятностями их появления в виде функции распределения вероятностей.

Функция распределения вероятностей (или просто функция распределения) $F(x)$ случайной величины X представляет собой *вероятность* того, что случайная величина X примет значение меньшее, чем некоторое заданное значение x:

$$\boxed{F(x) = \mathrm{P}(X < x)}. \qquad (2.1)$$

Функция распределения непрерывной случайной величины X, принимающей любые значения из некоторого интервала, может быть представлена:

- ***аналитически*** в виде *математического выражения* (2.1), отражающего зависимость вероятности от значения случайной величины;
- ***графически*** в виде непрерывной функции (рис.2.2,а), отображающей зависимость (2.1), или в виде *гистограммы функции распределения* (рис.2.2,б), полученной экспериментально, например в процессе имитационного моделирования, и представляющей собой дискретный график, в котором по оси абсцисс откладываются частотные интервалы, охватываю-

щие все возможные значения случайной величины, а по оси ординат – накопленная частота попадания случайной величины в эти частотные интервалы.

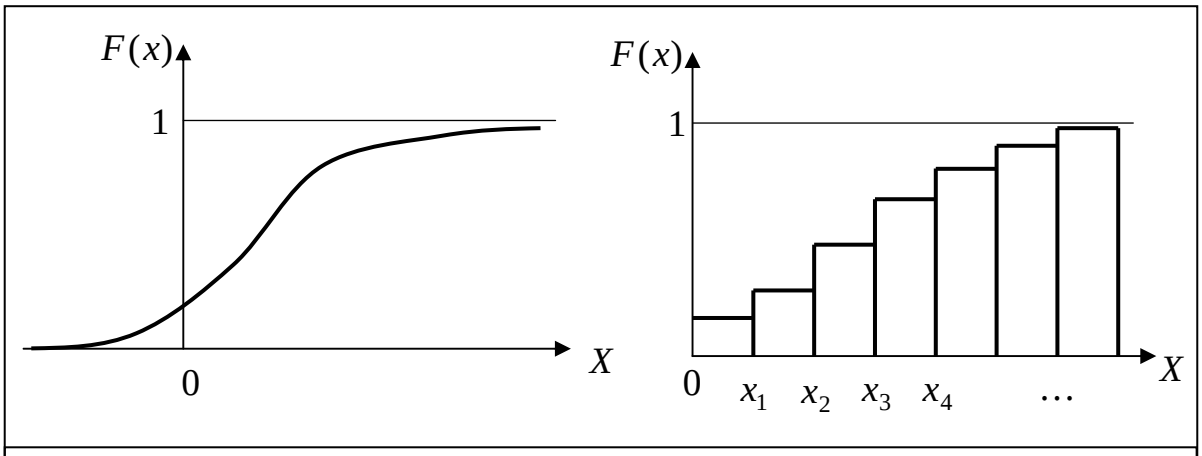


Рис.2.2. График (а) и гистограмма (б) функции распределения

Накопленная частота попадания в *i*-й частотный интервал определяется отношением количества случайных величин, значения которых находятся в интервале $(-\infty;\ x_i)$, к общему количеству случайных величин, полученных в процессе экспериментов.

Свойства функции распределения:

- функция распределения $F(x)$ есть *неубывающая* функция своего аргумента, то есть если $x_j > x_i$, то $F(x_j) \geq F(x_i)$;
- $F(-\infty) = 0$;
- $F(+\infty) = 1$.

Если случайная величина определена только в области положительных значений, ее функция распределения равна нулю на всем промежутке от минус бесконечности до нуля.

Вероятность того, что случайная величина примет значение из некоторого интервала (a, b), определяется через функцию распределения как

$$\mathrm{P}(a < x < b) = F(b) - F(a).$$

Функция распределения $F(x)$ является *универсальной* характеристикой случайной величины и существует как для *непрерывных*, так и для *дискретных* величин. Функция распределения дискретной случайной величины X, принимающей значения $x_1, x_2, \ldots, x_i, \ldots$, определяется как

$$F(x_m) = \mathrm{P}(X < x_m) = \sum_{i=1}^{m-1} p_i,$$

где p_i - вероятность того, что случайная величина X примет значение x_i.

На практике вместо функции распределения чаще используют другой способ представления закона распределения непрерывной случайной величины в виде *плотности распределения вероятностей*,

которая в отличие от функции распределения обладает большей наглядностью и позволяет получить представление о близости того или иного распределения к одному из известных теоретических распределений, имеющих аналитическое выражение.

Плотность распределения вероятностей $f(x)$ определяется как *производная* от функции распределения $F(x)$ по x:

$$f(x) = F'(x) = \frac{dF(x)}{dx}.$$

Размерность плотности распределения $f(x)$ *обратна* размерности случайной величины, в то время как функция распределения $F(x)$, как всякая вероятность, есть величина *безразмерная*.

Плотность распределения непрерывной случайной величины X, как и функция распределения, может быть представлена:

- ***аналитически*** в виде *математического выражения* $y = f(x)$;
- ***графически*** в виде непрерывной функции (графика), отображающей зависимость $y = f(x)$ (рис.2.3,а), или в виде *гистограммы плотности распределения*, в которой в отличие от гистограммы функции распределения по оси ординат откладывается частота (или число) попаданий случайной величины в каждый из частотных интервалов (рис.2.3,б).

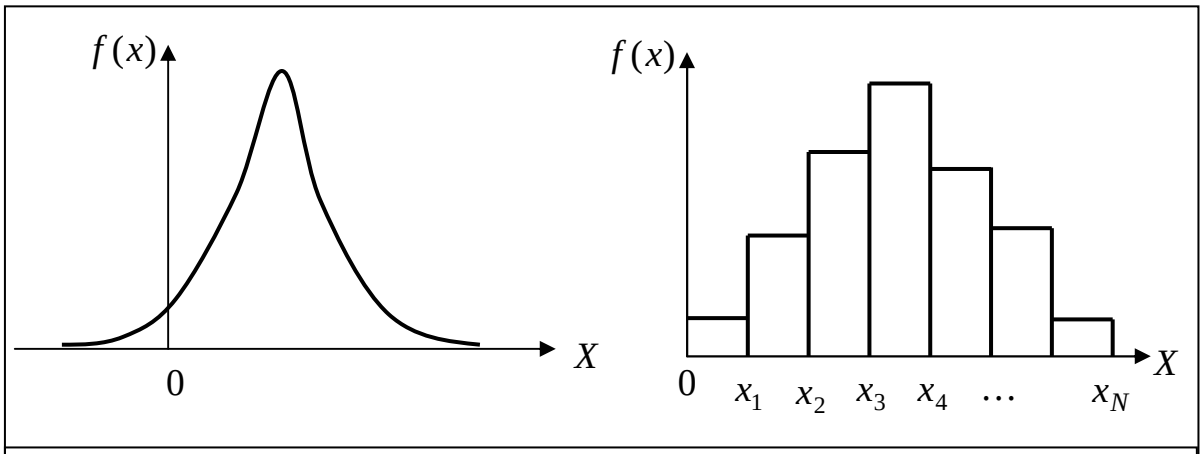


Рис.2.3. График (а) и гистограмма (б) плотности распределения

Свойства плотности распределения:

- плотность распределения есть функция *неотрицательная*: $f(x) \geq 0$;
- интеграл в бесконечных пределах от плотности распределения равен *единице*:

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1.$$

Функция и плотность распределения случайной величины однозначно связаны между собой. В частности, функция распределения определяется через плотность распределения следующим образом:

$$F(x) = \int_{-\infty}^{x} f(x)\,dx. \qquad (2.2)$$

Тогда вероятность того, что случайная величина примет значение из некоторого интервала (a, b), может быть определена через плотность распределения как

$$\mathrm{P}(a < x < b) = \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx\,.$$

Таким образом, закон распределения непрерывной случайной величины (непрерывный закон распределения) может быть задан в виде:

- ***функции распределения*** $F(x)$ случайной величины X, называемой также ***интегральным законом распределения***;
- ***плотности распределения*** $f(x)$ случайной величины X, называемой также ***дифференциальным законом распределения.***

2.3. Числовые характеристики случайных величин

«Даже маленькая практика стоит большой теории» (***Закон Букера***)

Числовые характеристики позволяют выразить в сжатой форме наиболее существенные особенности распределения случайной величины, например:

- среднее значение, около которого группируются возможные значения случайной величины;
- степень разбросанности этих значений относительно среднего;
- асимметрию (или «скошенность») плотности распределения;
- «крутость», то есть островершинность или плосковершинность плотности распределения и так далее.

В теории вероятностей используются различные числовые характеристики, имеющие разное назначение и разные области применения. Из них на практике наиболее часто применяются *начальные* и *центральные моменты* различных порядков, каждый из которых описывает то или иное свойство распределения. *Начальные моменты рассматриваются относительно начала координат*, а *центральные моменты – относительно среднего значения* (математического ожидания), то есть центра распределения.

В общем случае для описания случайной величины используется бесконечное множество начальных и центральных моментов. Между числовыми моментами и законом распределения случайной величины существует взаимное соответствие, которое означает, что, зная закон распределения, можно вычислить любые моменты, число которых бесконечно. В то же время, зная конечное число начальных или центральных моментов, можно путем аппроксимации подобрать закон распределения случайной величины в виде функции или плотности распределения, причем, чем больше известно моментов, тем точнее

аппроксимация закона распределения.

На практике обычно ограничиваются применением нескольких первых начальных или центральных моментов, что оказывается вполне достаточным для получения корректных конечных результатов.

2.3.1. Начальные моменты

Положим, что случайная величина X описывается вероятностями p_1, p_2, ..., p_n появления значений x_1, x_2, ..., x_n, если X – дискретная величина, и плотностью распределения $f(x)$ $x \in (-\infty, +\infty)$, если X – непрерывная величина.

Начальный момент s-го порядка $\alpha_s[X]$ случайной величины X определяется следующим образом $(s = 1, 2, ...)$:

$$\alpha_s[X] = \begin{cases} \sum_{i=1}^{n} x_i^s \, p_i & - \text{ для дискретной случайной величины;} \\ \int_{-\infty}^{+\infty} x^s \, f(x)\, dx & - \text{для непрерывной случайной величины.} \end{cases}$$

Первый начальный момент $\alpha_1[X]$ случайной величины X:

$$\alpha_1[X] = \begin{cases} \sum_{i=1}^{n} x_i \, p_i & - \text{ для дискретной случайной величины;} \\ \int_{-\infty}^{+\infty} x \, f(x)\, dx & - \text{для непрерывной случайной величины.} \end{cases}$$

называется ***математическим ожиданием*** или *средним значением случайной величины* и обозначается $M[X]$: $M[X] = \alpha_1[X]$. Математическое ожидание характеризует положение случайной величины на числовой оси, то есть показывает некоторое *среднее вероятностное* (не путать со средним арифметическим) значение, около которого группируются все возможные значения случайной величины.

Второй начальный момент $\alpha_2[X]$ случайной величины X характеризует *рассеивание*, то есть *разброс* (*удаленность*) значений случайной величины *относительно начала координат*, и имеет размерность квадрата случайной величины.

2.3.2. Центральные моменты

Центральный момент s-го порядка $\beta_s[X]$ случайной величины X определяется следующим образом $(s = 1, 2, ...)$:

$$\beta_s[X] = \begin{cases} \sum_{i=1}^{n} (x_i - \mathrm{M}[X])^s \, p_i & - \text{ для дискретной случайной величины;} \\ \int_{-\infty}^{+\infty} (x - \mathrm{M}[X])^s \, f(x)\, dx & - \text{для непрерывной случайной величины.} \end{cases}$$

Разность между значениями случайной величины и ее математическим ожиданием (X – M[X]) представляет собой отклонение случайной величины X от ее математического ожидания и называется ***центрированной случайной величиной***. Тогда центральный момент s-го порядка случайной величины X можно определить как математическое ожидание s-ой степени соответствующей центрированной случайной величины:

$$\beta_s[X] = \mathrm{M}[(X - \mathrm{M}[X])^s].$$

Для любой случайной величины *центральный момент первого порядка равен нулю*, так как математическое ожидание центрированной случайной величины всегда равно нулю.

Второй центральный момент называется ***дисперсией*** случайной величины и обозначается D[X]: $\mathrm{D}[X] = \beta_2[X]$.

Дисперсия вычисляется по формулам:

$$\mathrm{D}[X] = \begin{cases} \sum_{i=1}^{n} (x_i - \mathrm{M}[X])^2 p_i & - \text{ для дискретной случайной величины;} \\ \int_{-\infty}^{+\infty} (x - \mathrm{M}[X])^2 f(x)\,dx & - \text{ для непрерывной случайной величины.} \end{cases}$$

Можно показать, что дисперсия и второй начальный момент связаны следующей зависимостью:

$$\boxed{\mathrm{D}[X] = \alpha_2[X] - (\mathrm{M}[X])^2} \quad . \qquad (2.3)$$

Дисперсия случайной величины, как и второй начальный момент, характеризует разброс значений случайной величины, но, в отличие от второго начального момента, *относительно математического ожидания*, и имеет размерность квадрата случайной величины.

При решении различных задач удобно пользоваться характеристикой разброса, *размерность которой совпадает с размерностью случайной величины*. Такой характеристикой является ***среднеквадратическое отклонение*** $\sigma[X]$, которое определяется как корень квадратный из дисперсии:

$$\boxed{\sigma[X] = \sqrt{\mathrm{D}[X]}}.$$

В качестве *безразмерной* характеристики разброса случайных величин, определенных в области положительных значений, часто используют ***коэффициент вариации*** $\nu[X]$, который определяется как отношение среднеквадратического отклонения к математическому ожиданию:

$$\boxed{\nu[X] = \frac{\sigma[X]}{\mathrm{M}[X]}}$$

при условии, что $\mathrm{M}[X] > 0$.

Применение числовых характеристик существенно облегчает решение многих вероятностных задач, в частности, при решении сложных

задач, когда использование законов распределений приводит к громоздким выкладкам и не позволяет получить результаты в явном виде. Очень часто удается решить задачу до конца, оставляя в стороне законы распределения и оперируя одними числовыми характеристиками. Если в задаче фигурирует большое количество случайных величин, то для исчерпывающего суждения о результирующем законе распределения не требуется знать законы распределения отдельных случайных величин, фигурирующих в задаче, а достаточно знать лишь некоторые числовые характеристики этих величин.

Кроме того, на практике (и в повседневной жизни) редко оперируют законом распределения для описания конкретных физических величин, предпочитая использовать такие понятия как среднее значение и, в некоторых случаях, разброс значений случайной величины или минимальное и максимальное значение. Действительно, вряд ли для пассажиров, ожидающих на остановке автобус, представляет интерес закон распределения интервалов между автобусами. Более важным и понятным является указание среднего или максимального интервала. В то же время при моделировании транспортных потоков для получения корректных и достоверных результатов может потребоваться знание закона распределения или, по крайней мере, нескольких моментов распределения искомых интервалов.

Альтернативой случайной величине является неслучайная величина, называемая **детерминированной**. В некоторых задачах детерминированную величину $X = x$ рассматривают как случайную, которая с вероятностью $p = 1$ принимает одно и то же значение x.

2.4. Производящая функция и преобразование Лапласа

Аналитическое исследование сложных систем со случайным характером функционирования во многих случаях можно существенно упростить, если действия над функциями распределений заменить действиями над соответствующими производящими функциями и преобразованиями Лапласа.

Производящие функции используются для дискретных, а преобразования Лапласа – для непрерывных случайных величин.

2.4.1. Производящая функция

Производящей функцией распределения $p_k = \mathrm{P}(X = k)$ дискретной случайной величины X, принимающей неотрицательные целочисленные значения $k = 0, 1, 2, \ldots$, называется ряд

$$X^*(z) = \sum_{k=0}^{\infty} z^k p_k \qquad |z| \le 1. \tag{2.4}$$

Распределение вероятностей однозначно определяется своей производящей функцией:

$$p_k = \frac{1}{k!} X^{*(k)}(0), \qquad X^{*(k)}(0) = \frac{d^k}{dz^k} X^*(z)\Big|_{z=0} \qquad (k = 0, 1, 2, ...).$$

На основе производящей функции (2.4) могут быть вычислены начальные и центральные моменты случайной величины, в частности *математическое ожидание* и *дисперсия определяются* как

$$\mathrm{M}[X] = X^{*(1)}(1); \quad \mathrm{D}[X] = X^{*(2)}(1) + X^{*(1)}(1) - [X^{*(1)}(1)]^2. \qquad (2.5)$$

Производящая функция $X^*(z)$ суммы $X = X_1 + X_2 + \dots X_n$ независимых случайных величин равна произведению производящих функций слагаемых:

$$X^*(z) = X_1^*(z) X_2^*(z) \dots X_n^*(z).$$

2.4.2. Преобразование Лапласа

Преобразованием Лапласа плотности распределения $f(x)$ неотрицательной непрерывной случайной величины *X* называется функция

$$F^*(s) = \int_0^\infty e^{-sx} f(x) dx \qquad (s \ge 0). \qquad (2.6)$$

Плотность распределения однозначно определяется своим преобразованием Лапласа.

Дифференцируя преобразование Лапласа по *s* в точке *s*=0, можно определить *начальные моменты* случайной величины:

$$\alpha_k[X] = \frac{(-1)^k}{k!} \frac{d^k F^*(s)}{ds^k}\Big|_{s=0} \quad (k = 1, 2, ...). \qquad (2.7)$$

Преобразование Лапласа $F^*(s)$ суммы $X = X_1 + X_2 + \dots X_n$ независимых случайных величин равно произведению преобразований Лапласа слагаемых:

$$F^*(s) = F_1^*(s) F_2^*(s) \dots F_n^*(s).$$

2.5. Типовые распределения случайных величин

«Все законы – имитация реальности»
(***Метазакон Лилли.***)

Моделирование технических систем с дискретным характером функционирования предполагает применение разных законов распределений, как дискретных, так и непрерывных случайных величин. Ниже рассматриваются типовые законы распределений случайных величин, широко используемые в моделях массового обслуживания.

В качестве законов распределений **дискретных** случайных величин наиболее широко используются:

- распределение Пуассона;
- геометрическое распределение.

Поскольку в математических моделях массового обслуживания непрерывной случайной величиной обычно является *время*, наибольший

интерес представляют законы распределений *непрерывных* случайных величин, определенных в области положительных значений:

- равномерный;
- экспоненциальный;
- Эрланга;
- Эрланга нормированный;
- гиперэкспоненциальный;
- гиперэрланговский.

2.5.1. Распределение Пуассона

Дискретная случайная величина X распределена по закону Пуассона, если вероятность $P(X=k)$ того, что она примет определенное значение $x_k = k$ выражается формулой:

$$p_k = P(X = k) = \frac{a^k}{k!} e^{-a} \qquad (k = 0, 1, 2, \ldots), \tag{2.8}$$

где a – некоторая положительная величина, называемая *параметром* распределения Пуассона.

На рис.2.4 показаны многоугольники распределения Пуассона для трех значений параметра распределения: a=0,5; a=1; a=2.

Производящая функция распределения Пуассона:

$$X^*(z) = e^{-a(1-z)} \ (0 \le z \le 1).$$

2.5.2. Геометрическое распределение

Распределение дискретной случайной величины $X=k$ вида

$$p_k = P(X = k) = \rho^k (1 - \rho) \qquad (k = 0, 1, 2, \ldots), \tag{2.9}$$

где ρ - *параметр* распределения (0 < ρ < 1), называется ***геометрическим***.

Распределение (2.9) может быть записано в несколько ином виде, если параметр ρ заменить параметром γ =1-ρ:

$$p_k = \gamma(1 - \gamma)^k \qquad (0 < \gamma < 1; \quad k = 0, 1, 2, \ldots).$$

На рис.2.5 показаны многоугольники геометрического распределения для трех значений параметра распределения: $\gamma = 0{,}2$; $\gamma = 0{,}5$; $\gamma = 0{,}8$.

Производящая функция геометрического распределения:

$$X^*(z) = \frac{1 - \rho}{1 - \rho z} \ \text{ или } \ X^*(z) = \frac{\gamma}{1 - (1 - \gamma)z} \qquad (0 \le z \le 1).$$

Задание на самостоятельную работу:

1. *Определить математическое ожидание, второй начальный момент, дисперсию, коэффициент вариации для пуассоновского и геометрического распределений.*

2. *Построить многоугольники распределений для пуассоновского и геометрического законов при других значениях параметров распределений.*

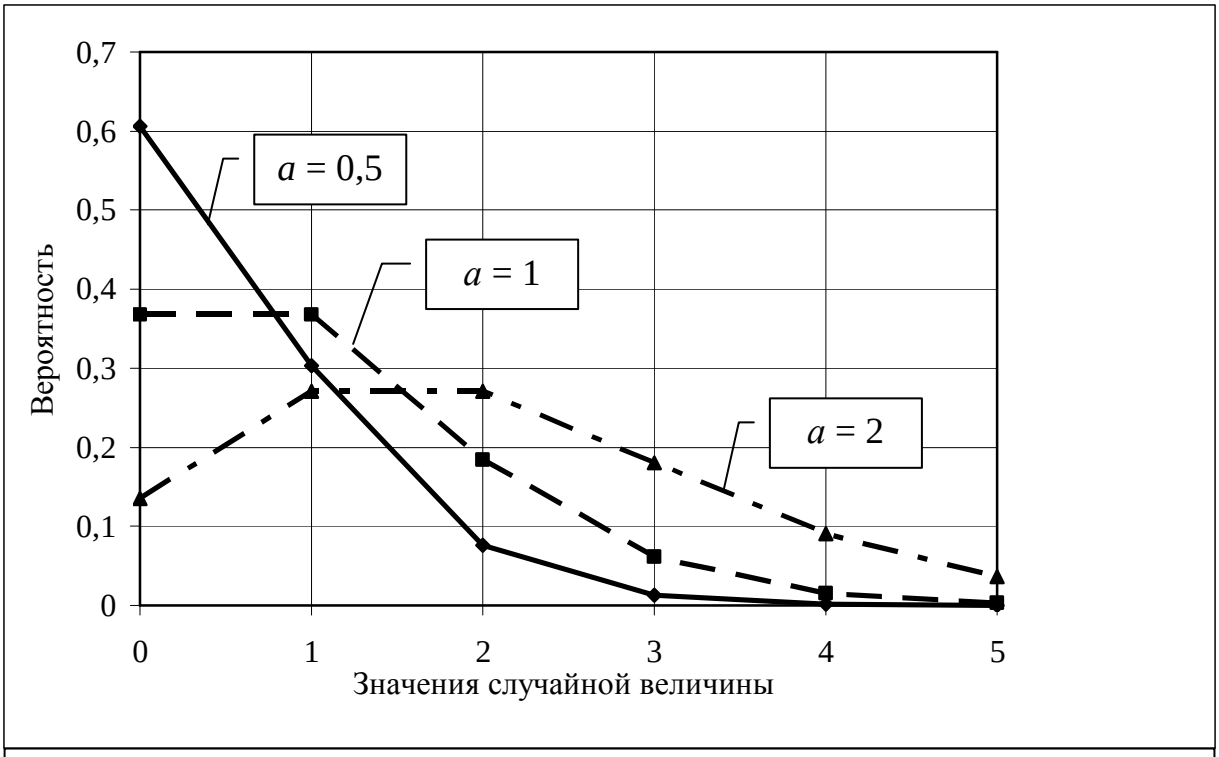


Рис.2.4. Многоугольники распределения Пуассона

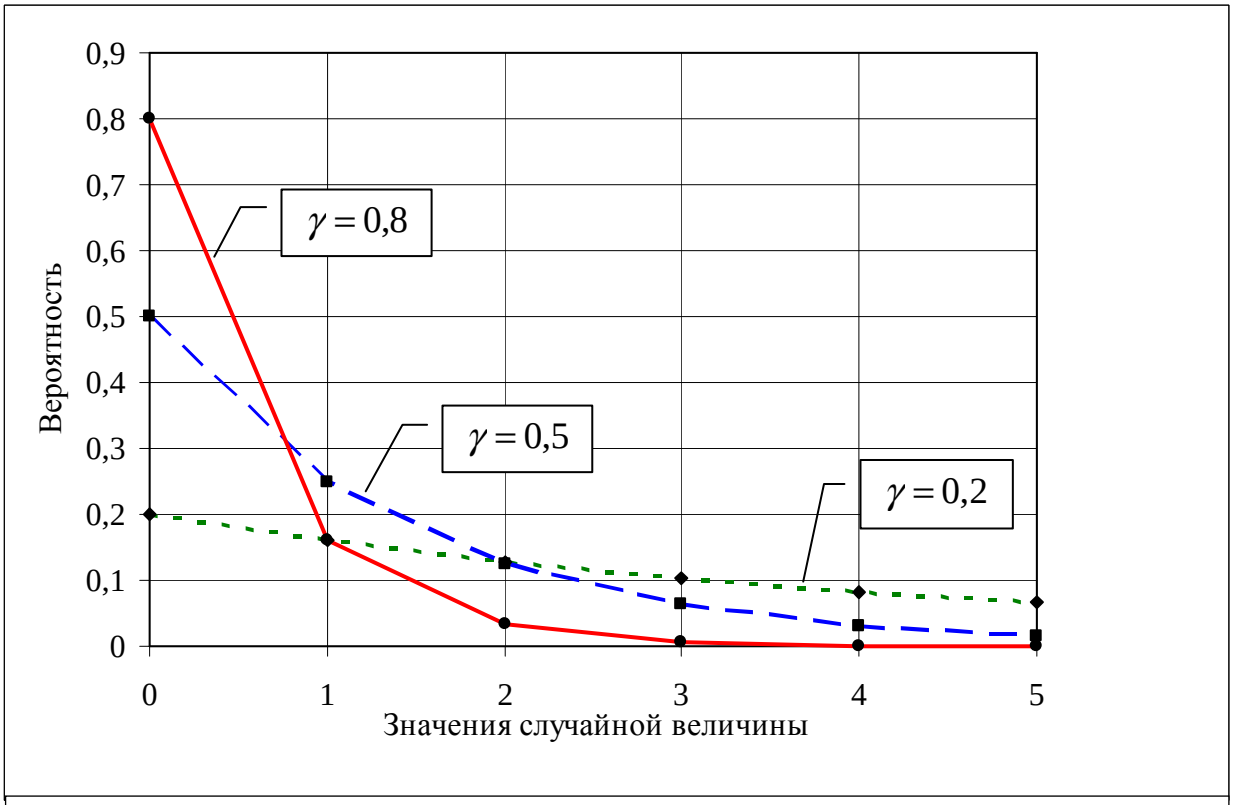


Рис.2.5. Многоугольники геометрического распределения

2.5.3. Равномерный закон распределения

Непрерывная случайная величина X распределена ***равномерно*** в интервале $(a; b)$, где $a<b$, если функция $F(x)$ и плотность $f(x)$ распределения соответственно имеют вид:

$$F(x)=\begin{cases}0 & \text{при} \quad x<a;\\ \dfrac{x-a}{b-a} & \text{при} \quad a<x<b;\\ 1 & \text{при} \quad x>b;\end{cases} \tag{2.10}$$

$$f(x)=\begin{cases}\dfrac{1}{b-a} & \text{при} \quad a<x<b;\\ 0 & \text{при} \quad x>b.\end{cases} \tag{2.11}$$

На рис.2.6 показаны функция и плотность равномерного распределения.

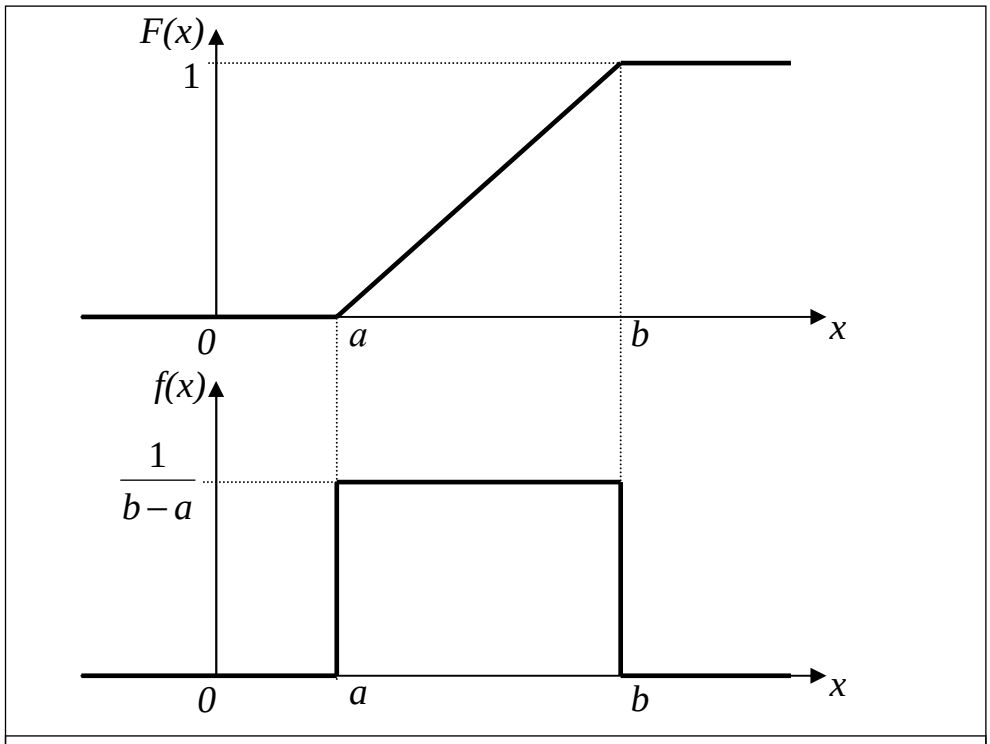


Рис.2.6. Функция и плотность равномерного распределения

Задание на самостоятельную работу:

1. *Определить математическое ожидание, второй начальный момент, дисперсию, коэффициент вариации и построить график функции и плотности равномерного распределения.*

2. *Записать выражения для функции и плотности равномерного распределения для следующих частных случаев, когда случайная величина принимает значения:*

1) *в интервале (0; b) при условии, что b>0;*

2) *в интервале (a; 0) при условии, что a<0;*

3) *в интервалах (a; b) и (c; d) при условии, что a<b<c<d. Определить математическое ожидание, второй начальный момент, дисперсию, коэффициент вариации.*

3. *Построить графики функции и плотности распределений для указанных случаев.*

2.5.4. Экспоненциальный закон распределения

Непрерывная случайная величина X, принимающая положительные значения в бесконечном интервале $(0;+\infty)$, распределена по ***экспоненциальному*** (***показательному***) ***закону***, если функция $F(x)$ и плотность $f(x)$ распределения соответственно имеют вид:

$$F(x)=1-e^{-\alpha x}, \qquad f(x)=\alpha e^{-\alpha x}, \tag{2.12}$$

где $\alpha>0$ – параметр распределения; $x\geq 0$ – непрерывная случайная величина.

Замечательной особенностью экспоненциального распределения является то, что его **коэффициент вариации** не зависит от параметра α *и* **всегда равен единице**: $\nu_{эксп}[X]=1$.

На рис.2.7 показаны функция и плотность экспоненциального распределения для трех значений параметра: $\alpha=0{,}5$; $\alpha=1$; $\alpha=2$.

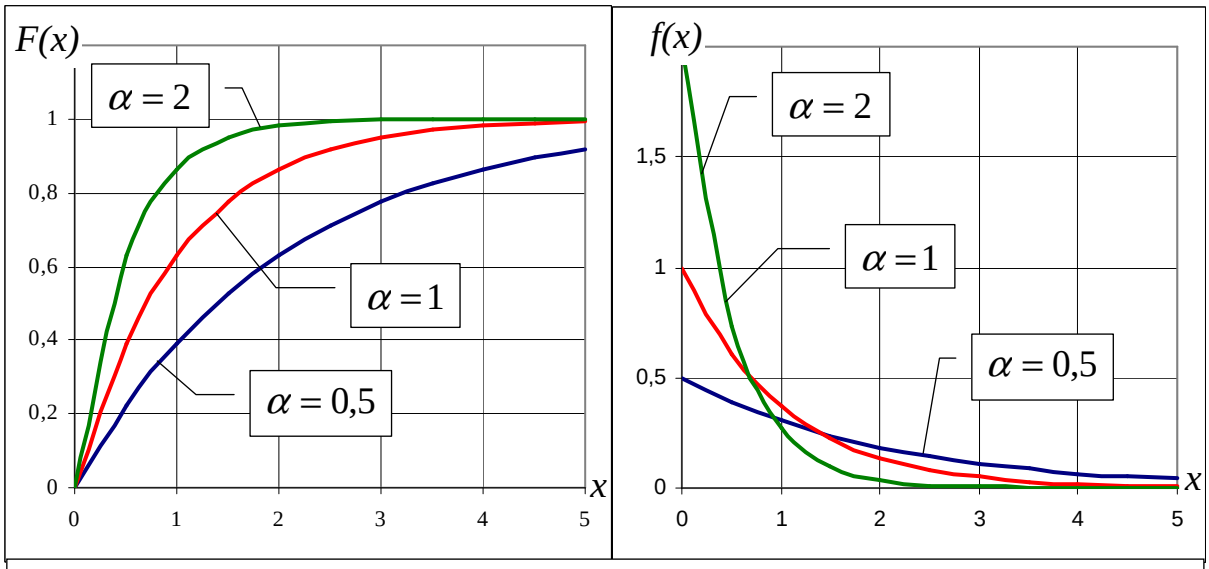


Рис.2.7. Функция и плотность экспоненциального распределения

Преобразование Лапласа экспоненциального распределения

$$F^{*}(s)=\frac{\alpha}{\alpha+s}. \tag{2.13}$$

Экспоненциальное распределение широко применяется в теории массового обслуживания при описании случайных процессов, протекающих в моделях массового обслуживания. Это объясняется тем, что экспоненциальное распределение обладает замечательным свойством, присущим только этому распределению, благодаря которому для многих моделей массового обслуживания удается получить достаточно простые аналитические результаты в явном виде. С этим же распределением тесно связан особый класс дискретных случайных процессов, называемых марковскими процессами, в которых переходы между состояниями не зависят от предыстории процесса и определяются только состоянием процесса в данное конкретное время. Это свойство иногда называют

свойством отсутствия памяти у экспоненциального распределения (точнее, у экспоненциально распределенных случайных величин), а в теории массового обслуживания используется термин «отсутствие последействия» (см. п.2.6).

Возможность получения сравнительно простых аналитических результатов при использовании предположения об экспоненциальном характере случайных процессов обусловила появление рассматриваемых ниже специфических законов распределений, представляющих собой композиции экспоненциальных распределений и позволяющих упростить решение многих задач, связанных с исследованием моделей массового обслуживания. К ним, в частности, относятся следующие распределения: Эрланга, гиперэкспоненциальное, гиперэрланговское.

Задание на самостоятельную работу:

1. *Определить математическое ожидание, второй начальный момент, дисперсию и построить график функции и плотности экспоненциального распределения.*

2. *Доказать, что коэффициент вариации экспоненциального распределения равен единице.*

3. *Вывести формулу (2.13).*

2.5.5. Распределение Эрланга

Распределением Эрланга k-го порядка называется распределение, описывающее непрерывную случайную величину *X*, принимающую положительные значения в интервале $(0;+\infty)$ и представляющую собой *сумму k независимых случайных величин, распределенных по одному и тому же экспоненциальному закону с параметром* α. Функция и плотность распределения Эрланга *k*-го порядка имеют вид:

$$F_k(x)=1-e^{-\alpha x}\sum_{i=0}^{k-1}\frac{(\alpha x)^i}{i!}; \qquad f_k(x)=\frac{\alpha(\alpha x)^{k-1}}{(k-1)!}e^{-\alpha x}, \qquad (2.14)$$

где α и k – положительные параметры распределения $(\alpha \ge 0;\ k=1,2,\ldots)$; $x \ge 0-$ непрерывная случайная величина.

На рис.2.8 показаны плотности распределения Эрланга при $\alpha=1$ для трех значений параметра: $k=1;\ k=2;\ k=4$.

При $k=1$ распределение Эрланга вырождается в экспоненциальное, а при $k \to \infty$ – приближается к *нормальному распределению*.

Преобразование Лапласа распределения Эрланга *k*-го порядка

$$F^*(s)=\left(\frac{\alpha}{\alpha+s}\right)^k. \qquad (2.15)$$

Поскольку распределение Эрланга является *двухпараметрическим*, то оно может использоваться для аппроксимации реальных распределений по двум первым моментам.

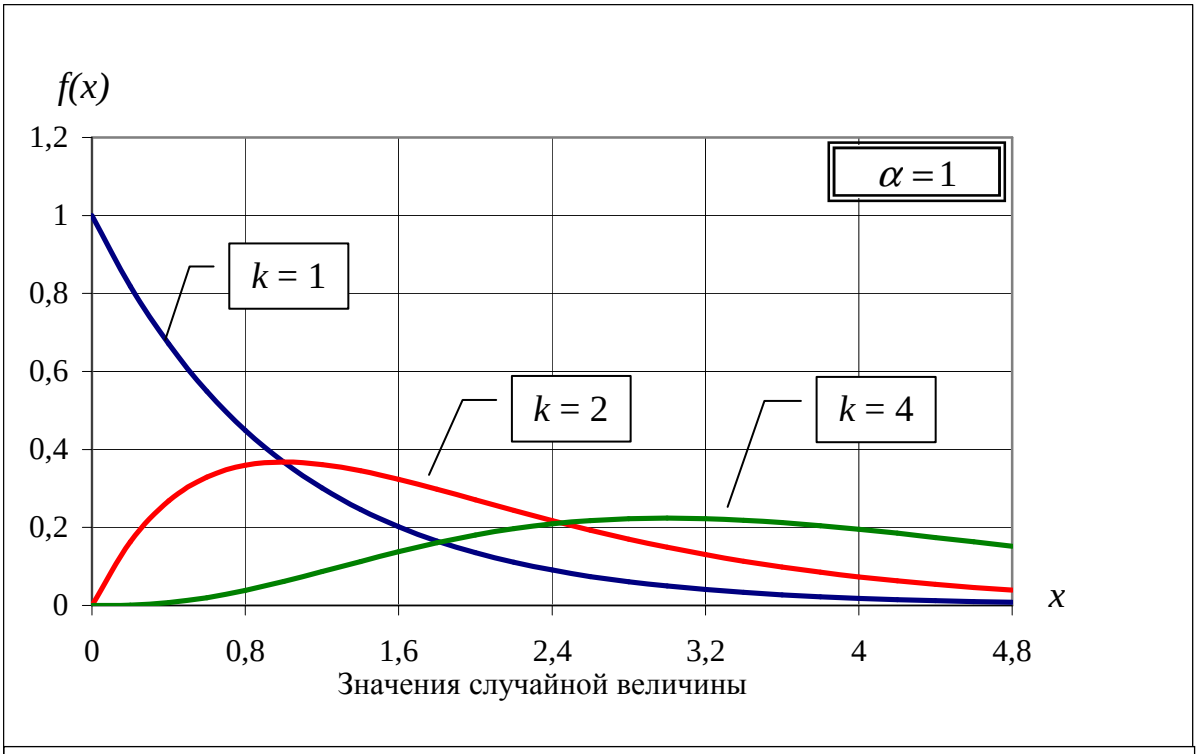


Рис.2.8. Плотность распределения Эрланга

Коэффициент вариации распределения Эрланга зависит от параметра k и принимает значения **меньшие или равное единице**:

$$\nu_{Эк}[X] = \frac{1}{\sqrt{k}} \le 1 \quad (k = 1, 2, \ldots).$$

Отметим, что математическое ожидание распределения Эрланга зависит от значения параметра k, что создаёт определенные проблемы при аппроксимации реальных распределений законом Эрланга. Эти проблемы отсутствуют при аппроксимации нормированным распределением Эрланга.

Задание на самостоятельную работу:

1. *Определить математическое ожидание, дисперсию, коэффициент вариации распределения Эрланга k-го порядка.*

2. *Доказать, что коэффициент вариации распределения Эрланга не превышает 1.*

3. *Построить графики функции и плотности распределений Эрланга 5-го и 8-го порядка.*

2.5.6. Нормированное распределение Эрланга

Нормированное распределение Эрланга представляет собой распределение суммы k независимых случайных величин, каждая из которых распределена по экспоненциальному закону с параметром $k\alpha$, зависящим от k. Другими словами, суммируются k экспоненциально распределенных случайных величин, каждая из которых имеет математическое ожидание в

k раз меньшее, чем исходное математическое ожидание реального распределения, что приводит к независимости математического ожидания нормированного распределения Эрланга от параметра k.

Математические выражения для функции и плотности нормированного распределения Эрланга можно получить из (2.14), заменив параметр α на $k\alpha$:

$$F_k(x) = 1 - e^{-k\alpha x} \sum_{i=0}^{k-1} \frac{(k\alpha x)^i}{i!}; \qquad f_k(x) = \frac{k\alpha (k\alpha x)^{k-1}}{(k-1)!} e^{-k\alpha x},$$

Коэффициент вариации нормированного распределения Эрланга так же, как и ненормированного, зависит от параметра k и принимает значения **меньшие или равное единице**: $\nu_{нЭк}[X] = \frac{1}{\sqrt{k}} \le 1 \quad (k = 1, 2, ...)$.

На рис.2.9 показаны плотности распределения Эрланга при $\alpha = 1$ для трех значений параметра: $k = 1$; $k = 2$; $k = 16$.

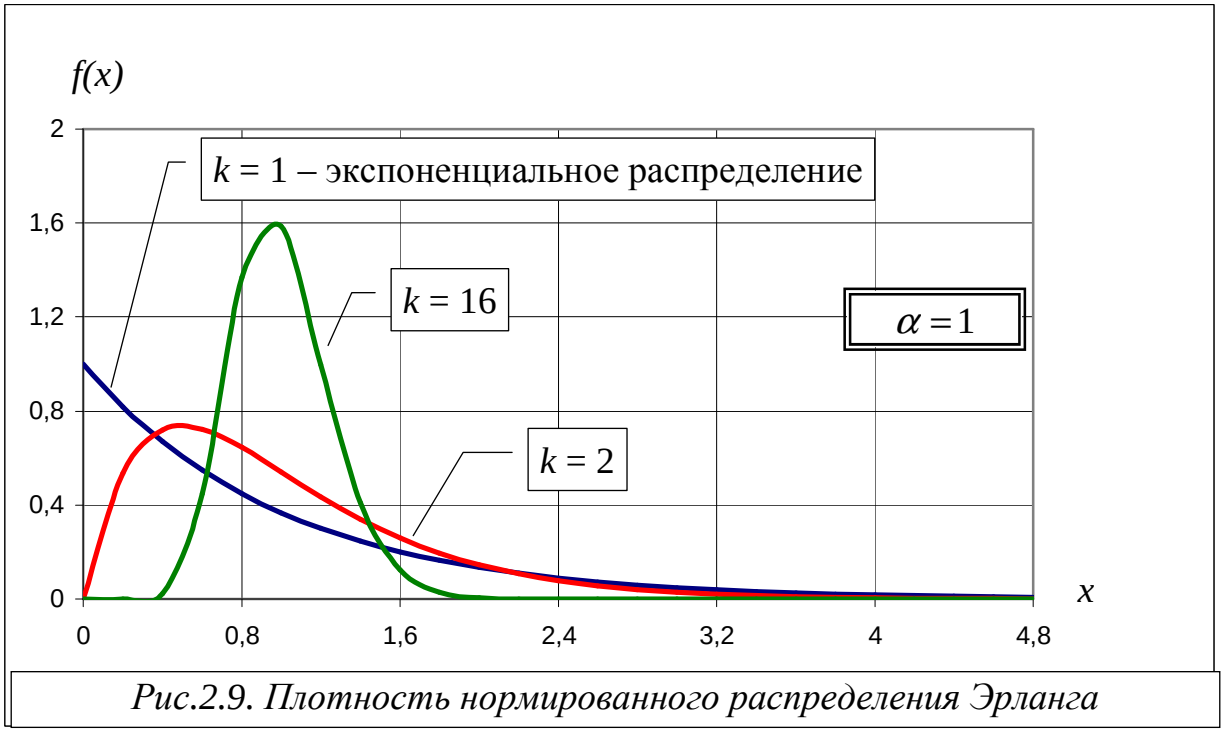


Рис.2.9. Плотность нормированного распределения Эрланга

Нормированное распределение Эрланга при $k \to \infty$, в отличие от простого распределения Эрланга, приводит к *детерминированной величине* $1/\alpha$.

Преобразование Лапласа нормированного распределения Эрланга

$$F^*(s) = \left(\frac{k\alpha}{k\alpha + s} \right)^k. \qquad (2.16)$$

Задание на самостоятельную работу:

1. *Построить графики функции и плотности нормированного распределения Эрланга 5-го порядка и сравнить с простым распределением Эрланга.*

2. *Определить математическое ожидание, второй начальный момент, дисперсию, коэффициент вариации нормированного распределения Эрланга. Доказать, что коэффициент вариации нормированного распределения Эрланга не превышает 1.*

2.5.7. Гиперэкспоненциальное распределение

В тех случаях, когда некоторое реальное распределение непрерывной случайной величины, принимающей неотрицательные значения, имеет **коэффициент вариации больше единицы**, для его аппроксимации может использоваться гиперэкспоненциальное распределение.

Как следует из названия, гиперэкспоненциальное распределение некоторым образом связано с экспоненциальным и представляет собой аддитивную смесь разных экспоненциальных распределений. Процесс формирования случайных величин с гиперэкспоненциальным распределением из экспоненциально распределенных случайных величин может быть представлен следующим образом.

Положим, что имеется n разных генераторов экспоненциально распределённых случайных величин с параметрами $\alpha_1, \ldots, \alpha_n$ соответственно (математическими ожиданиями $M_1 = 1/\alpha_1, \ldots, M_n = 1/\alpha_n$), причем $\alpha_i \neq \alpha_j$ для всех $i \neq j$ $(i, j = \overline{1, n})$. Пусть в результате одного опыта с вероятностью q_i вырабатывается только одна случайная величина i-м генератором с параметром α_i $(i = \overline{1, n})$, причем $q_1 + \ldots + q_n = 1$. Совокупность случайных величин, полученных в результате проведения множества таких опытов, будет распределена по гиперэкспоненциальному закону:

$$\left.\begin{aligned} F(x) &= \sum_{i=1}^{n} q_i (1 - e^{-\alpha_i x}) = 1 - \sum_{i=1}^{n} q_i e^{-\alpha_i x}; \\ f(x) &= \sum_{i=1}^{n} q_i \alpha_i e^{-\alpha_i x} \end{aligned}\right\} \qquad (2.17)$$

Гиперэкспоненциальное распределение (2.17) содержит $(2n-1)$ параметров: $\alpha_1, ..., \alpha_n, q_1, ..., q_{n-1}$, поскольку $q_n = 1 - \sum_{i=1}^{n-1} q_i$.

Преобразование Лапласа гиперэкспоненциального распределения

$$F^*(s) = \sum_{i=1}^{n} q_i \frac{\alpha_i}{\alpha_i + s}.$$

В простейшем варианте случайные величины с гиперэкспоненциальным распределением могут быть получены с использованием только двух экспоненциальных распределений: n=2. Тогда функция и плотность гиперэкспоненциального распределения будут иметь вид:

$$\left.\begin{aligned} F(x) &= qp(1 - e^{-\alpha_1 x}) + (1 - q)(1 - e^{-\alpha_2 x}); \\ f(x) &= q\alpha_1 e^{-\alpha_1 x} + (1 - q)\alpha_2 e^{-\alpha_2 x}. \end{aligned}\right\} \qquad (2.18)$$

Заметим, что гиперэкспоненциальное распределение (2.18) является трехпараметрическим, то есть содержит три независимых параметра: $q, \alpha_1, \alpha_2 \quad (0 < q < 1; \alpha_1 \geq 0; \alpha_2 \geq 0)$. Следовательно, аппроксимация реальных распределений гиперэкспоненциальным может осуществляться по трем моментам распределения, а не по двум, как в распределении Эрланга.

На рис.2.10 показаны плотности гиперэкспоненциального распределения случайной величины X с математическим ожиданием, равным 1, для двух значений коэффициента вариации: $\nu[X] = 2$ и $\nu[X] = 4$. Параметры распределения (2.18) имеют следующие значения:

- $\alpha_1 = 0{,}183; \quad \alpha_2 = 1{,}506$ для распределения с $\nu[X] = 2$;
- $\alpha_1 = 0{,}091; \quad \alpha_2 = 4{,}022$ для распределения с $\nu[X] = 4$,

причем параметр q одинаков для обоих распределений и равен 0,07.

Здесь же для сравнения показана плотность экспоненциального распределения с тем же математическим ожиданием M=1.

Как видно из представленных графических зависимостей, плотность гиперэкспоненциального распределения по сравнению с экспоненциальным распределением характеризуется более резким спадом в области малых значений случайной величины, причем чем больше коэффициент вариации случайной величины, тем круче эта зависимость.

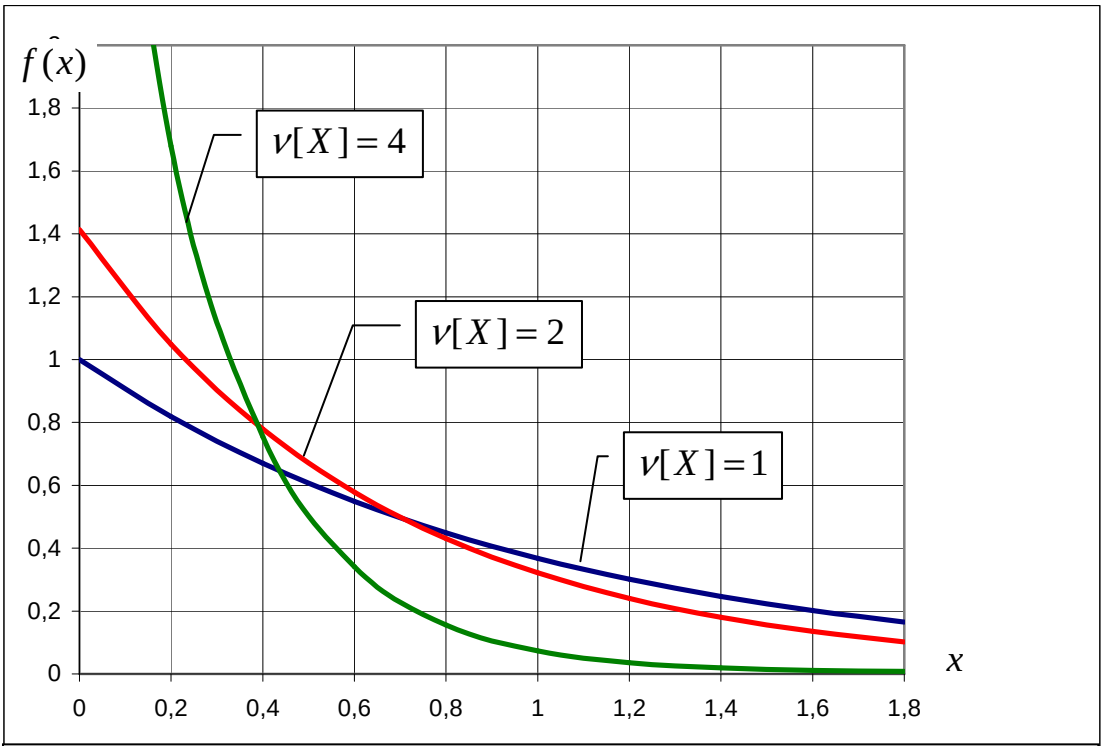


Рис.2.10. Плотность гиперэкспоненциального распределения

Можно показать, что вероятность появления маленьких значений случайной величины для гиперэкспоненциального распределения намного больше вероятности появления больших значений. Определим вероятность того, что случайная величина примет значение меньше математического ожидания M. Для этого рассчитаем значение функции распределения в

точке $x = \mathrm{M}$: $\Pr(X < \mathrm{M}) = F(x = \mathrm{M}) = F(\mathrm{M})$. Тогда для рассмотренных выше гиперэкспоненциальных распределений получим:

$$\Pr(X < \mathrm{M}) = F(\mathrm{M}) = \begin{cases} 0{,}735 & \text{для распределения с } \nu[\mathrm{X}] = 2; \\ 0.919 & \text{для распределения с } \nu[\mathrm{X}] = 4. \end{cases}$$

Таким образом, более 73% значений случайной величины, распределенной по гиперэкспоненциальному закону с коэффициентом вариации, равным 2, попадает в интервал (0; M) и только 27% значений окажутся больше математического ожидания. Для случайной величины, с коэффициентом вариации, равным 4, вероятность попадания в интервал (0; M) еще выше и составляет почти 92%. Очевидно, что чем больше коэффициент вариации, тем больше вероятность появления маленьких значений случайной величины.

Для сравнения вычислим эту же вероятность для экспоненциального распределения: $\Pr(X < \mathrm{M}) = F(\mathrm{M}) = 0{,}632$. Таким образом, вероятность появления маленьких значений экспоненциально распределенной случайной величины больше вероятности появления больших значений и составляет 63%, но при этом она значительно меньше, чем при гиперэкспоненциальном распределении.

Представление о гиперэкспоненциальном распределении будет не полным, если не обратить внимания на «хвост» этого распределения. На рис.2.11 представлен график плотности гиперэкспоненциального распределения для значений случайной величины больше 4. Напомним, что математическое ожидание случайной величины равно 1.

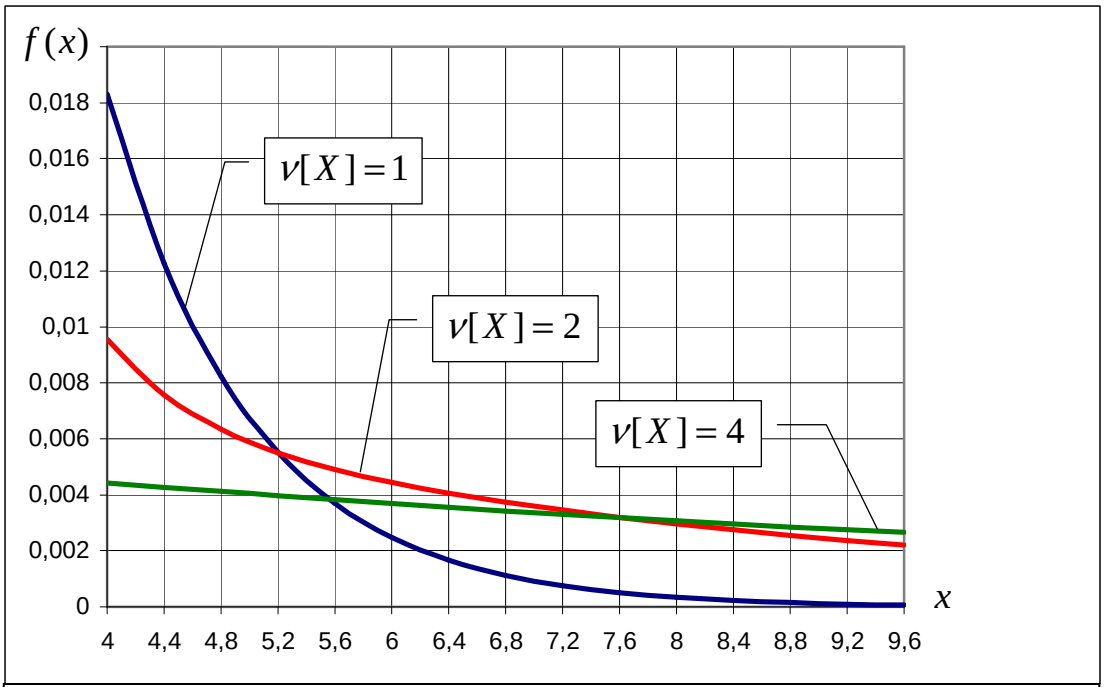


Рис.2.11. «Хвост» гиперэкспоненциального распределения

Из графика видно, что кривая плотности гиперэкспоненциального распределения с коэффициентом вариации, равным 4, имеет длинный так

называемый «тяжелый хвост», характеризующийся малым изменением. Это означает, что при гиперэкспоненциальном распределении вероятность появления больших значений случайной величины значительно выше, чем, например, для экспоненциального распределения. Таким образом, основное отличие гиперэкспоненциального распределения от экспоненциального состоит в том, что гиперэкспоненциальное распределение характеризуется б*о*льшей вероятностью появления маленьких значений случайной величины и, в то же время, б*о*льшей вероятностью появления больших значений случайной величины

Задание на самостоятельную работу:

1. *Построить графики функции гиперэкспоненциального распределения и сравнить с экспоненциальным распределением.*

2. *Определить математическое ожидание, второй начальный момент, дисперсию, коэффициент вариации гиперэкспоненциального распределения.*

3. *Доказать, что коэффициент вариации гиперэкспоненциального распределения превышает 1.*

2.5.8. Гиперэрланговское распределение

Гиперэрланговское распределение представляет собой аддитивную смесь *нормированных распределений Эрланга* и является наиболее *общим распределением* неотрицательных непрерывных случайных величин, поскольку имеет **коэффициент вариации в интервале от 0 до ∞.** Составляющими гиперэрланговского распределения, в отличие от гиперэкспоненциального, являются нормированные распределения Эрланга.

Плотность гиперэкспоненциального распределения

$$f(x)=\sum_{i=1}^{n} q_i \frac{k_i\alpha_i(k_i\alpha_i x)^{k_i-1}}{(k_i-1)!}e^{-k_i\alpha_i x} \qquad (x\ge 0). \qquad (2.19)$$

Преобразование Лапласа гиперэрланговского распределения

$$F^*(s)=\sum_{i=1}^{n} q_i\left(\frac{k_i\alpha_i}{k_i\alpha_i+s}\right)^{k_i}.$$

Задание на самостоятельную работу:

1. *Построить график плотности* гиперэрланговского *распределения и сравнить с гиперэкспоненциальным.*

2. *Определить математическое ожидание, второй начальный момент, дисперсию, коэффициент вариации гиперэрланговского распределения.*

3. *Доказать, что коэффициент вариации гиперэрланговского распределения может принимать любое значение.*

Ниже в таблице представлены основные числовые характеристики

рассмотренных распределений дискретных и непрерывных случайных величин:

- математическое ожидание $\mathrm{M}[X]$;
- второй начальный момент $\boldsymbol{\alpha_2[X]}$;
- дисперсия $\mathrm{D}[X]$;
- среднеквадратическое отклонение $\boldsymbol{\sigma[X]}$;
- коэффициент вариации $\boldsymbol{\nu[X]}$.

В графе «Примечания» указаны значения или диапазон изменения параметров соответствующих распределений.

Числовые характеристики распределений

Распределение	**M[X]**	$\boldsymbol{\alpha_2[X]}$	**D[X]**	$\boldsymbol{\sigma[X]}$	$\boldsymbol{\nu[X]}$	**Примечания**
Пуассона	a	$a(a+1)$	a	$\sqrt{a}$	$1/\sqrt{a}$	$a>0$
Геометрическое	$\frac{1-\gamma}{\gamma}$	$\frac{2(1-\gamma)^2}{\gamma^2}$	$\frac{(1-\gamma)^2}{\gamma^2}$	$\frac{1-\gamma}{\gamma}$	1	$0<\gamma<1$
Равномерное	$\frac{a+b}{2}$	$\frac{a^2+ab+b^2}{3}$	$\frac{(b-a)^2}{12}$	$\frac{b-a}{2\sqrt{3}}$	$\frac{b-a}{\sqrt{3}(a+b)}$	$b>a$
Экспоненциальное	$\frac{1}{\alpha}$	$\frac{2}{\alpha^2}$	$\frac{1}{\alpha^2}$	$\frac{1}{\alpha}$	1	$\alpha>0$
Эрланга	$\frac{k}{\alpha}$	$\frac{k(k+1)}{\alpha^2}$	$\frac{k}{\alpha^2}$	$\frac{\sqrt{k}}{\alpha}$	$\frac{1}{\sqrt{k}}$	$k=1, 2, \ldots$
Эрланга нормированное	$\frac{1}{\alpha}$	$\frac{k+1}{k\alpha^2}$	$\frac{1}{k\alpha^2}$	$\frac{1}{\alpha\sqrt{k}}$	$\frac{1}{\sqrt{k}}$	$k=1, 2, \ldots$
Гиперэкспоненциальное	$\sum_{i=1}^{n}\frac{q_i}{\alpha_i}$	$2\sum_{i=1}^{n}\frac{q_i}{\alpha_i^2}$	$\alpha_2[X]-$ $-(M[X])^2$	$\sqrt{\mathrm{D}[X]}$	$\nu[X]\ge 1$	$\sum_{i=1}^{n}q_i=1$ $\alpha_i>0$
Гиперэрлангов-ское	$\sum_{i=1}^{n}\frac{q_i}{\alpha_i}$	$\sum_{i=1}^{n}q_i\frac{k_i+1}{k_i\alpha_i^2}$	$\alpha_2[X]-$ $-(M[X])^2$	$\sqrt{\mathrm{D}[X]}$	$\nu[X]\ge 0$	$\sum_{i=1}^{n}q_i=1$ $\alpha_i>0$ $k_i=1, 2, \ldots$

Напомним, что представленные числовые характеристики связаны между собой достаточно простыми соотношениями:

$$\mathrm{D}[X]=\boldsymbol{\alpha_2[X]}-(\mathrm{M}[X])^2;$$

$$\boldsymbol{\sigma[X]}=\sqrt{\mathrm{D}[X]};$$

$$\boldsymbol{\nu[X]}=\frac{\sigma[X]}{\mathrm{M}[X]}.$$

2.6. Аппроксимация неэкспоненциальных распределений

Работая над решением задачи, всегда полезно знать ответ. (***Закон Мэрфи***)

Как было отмечено в п.2.5.4, экспоненциальное распределение обладает замечательным свойством – свойством отсутствия последействия, благодаря которому оно широко используется при описании случайных процессов, протекающих в моделях массового обслуживания. Свойство отсутствия последействия заключается в следующем (рис.2.12). Если некоторый временной интервал $\tau = t_1 - t_0$ представляет собой случайную величину, распределенную по экспоненциальному закону, то интервал $\xi = t_1 - t^*$, начинающийся от случайного момента времени t_1 до завершения данного временн*о*го интервала, распределен по тому же экспоненциальному закону с тем же параметром α (средним значением $\bar{\tau} = 1/\alpha$). Другими словами, продолжительность интервала ξ не зависит от предыстории, то есть от того, сколько времени уже прошло до момента t^*.

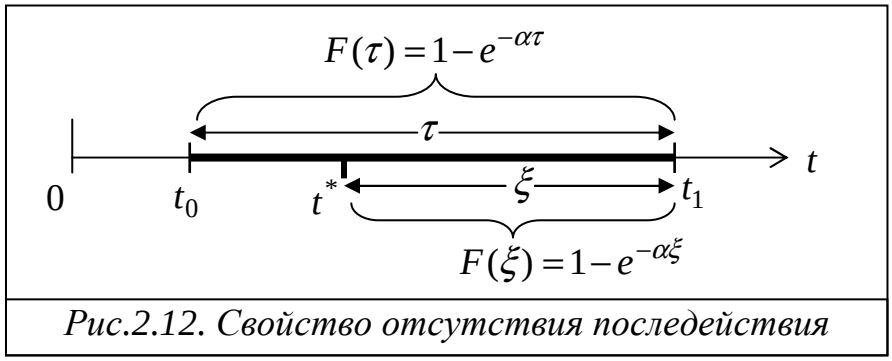


Рис.2.12. Свойство отсутствия последействия

Это замечательное свойство экспоненциального распределения используется при построении моделей марковских процессов, представляющих собой особый класс случайных процессов, развитие которых не зависит от предыстории процесса (см. п.5.1.2). Благодаря этому для многих моделей массового обслуживания удается достаточно просто получить конечные результаты, в том числе, в виде аналитических зависимостей в явном виде для расчета характеристик исследуемой системы. Поэтому часто при исследовании систем, в которых временн*ы*е процессы отличаются от экспоненциальных, стремятся свести эти процессы к экспоненциальному представлению.

Напомним, что для экспоненциального закона распределения случайных величин, определённых в области положительных значений $\tau \geq 0$, коэффициент вариации, описывающий разброс значений случайной величины, равен единице. Если реальные временн*ы*е интервалы имеют значения коэффициента вариации значительно отличающиеся от единицы, использование экспоненциального распределения может привести к большим погрешностям конечных результатов. В этих случаях в качестве аппроксимирующих функций законов распределений могут использовать-

ся вероятностные законы, представляющие собой композицию экспоненциальных распределений, а именно:

- распределение Эрланга и гипоэкспоненциальное распределение, когда коэффициент вариации временного интервала меньше единицы: $0 < \nu < 1$;
- гипреэкспоненциальное распределение, когда коэффициент вариации временного интервала больше единицы: $\nu > 1$.

При этом аппроксимация реального распределения, в простейшем случае, может выполняться по двум первым моментам распределения:

- математическому ожиданию;
- коэффициенту вариации.

2.6.1. Аппроксимация распределения с коэффициентом вариации $0 < \nu < 1$

Положим, что математическое ожидание и коэффициент вариации некоторой случайной величины τ, определенной в положительной области действительных чисел, соответственно равны t и ν, причем $0 < \nu < 1$.

Для аппроксимации закона распределения такой случайной величины в теории массового обслуживания часто используют распределение Эрланга *k*-го порядка E_k, которое может быть представлено в виде последовательности *k* экспоненциально распределенных фаз с одинаковым параметром $\alpha_i = \alpha = 1/\mathrm{M}[\tau]$ $(i = \overline{1,k})$, где $\mathrm{M}[\tau]$ – математическое ожидание экспоненциально распределенной случайной величины в одной фазе (рис.2.13).

Такое представление позволяет трактовать формирование случайных величин, распределенных по закону Эрланга, как сумму *k* случайных величин, распределенных по одному и тому же экспоненциальному закону.

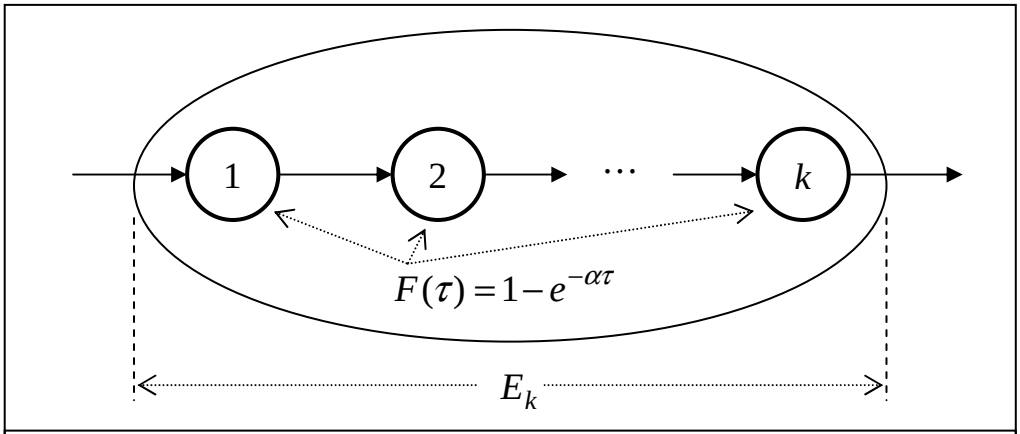


Рис.2.13. Многофазное представление распределения Эрланга

Математическое ожидание и коэффициент вариации случайной величины, распределенной по закону Эрланга *k*-го порядка:

$$\mathrm{M}_{E_k} = k\mathrm{M}[\tau]; \qquad \nu_{E_k} = \frac{1}{\sqrt{k}},$$

где $k = 1, 2, \ldots$ – параметр распределения Эрланга, принимающий только целочисленные значения.

Тогда для заданных реальных (измеренных) значений математического ожидания t и коэффициента вариации ν $(0 < \nu < 1)$ некоторой случайной величины τ, определенной в положительной области действительных чисел, параметры аппроксимирующего распределения Эрланга будут определяться следующим образом:

$$k = \left]\frac{1}{\nu^2}\right[; \qquad \mathrm{M}[\tau] = \frac{t}{k},$$

где]*x*[означает ближайшее целое, большее *x*, поскольку параметр *k* может принимать только целочисленные значения.

Нетрудно убедиться, что распределение Эрланга позволяет аппроксимировать только те реальные распределения, коэффициенты вариации которых имеют следующие значения: $\nu = 0{,}707$ при $k = 2$; $\nu = 0{,}577$ при $k = 3$; $\nu = 0{,}5$ при $k = 4$ и т.д.

Для аппроксимации распределений с любым значением коэффициента вариации, находящимся в интервале (0; 1), рассмотрим многофазное распределение с разными параметрами экспоненциальных распределений в фазах: $\alpha_i = 1/t_i \quad (i = \overline{1,k})$, где t_i – математическое ожидание экспоненциально распределенной случайной величины в *i*-й фазе. Такое распределение будем называть ***гипоэкспоненциальным распределением***.

Проанализируем свойства гипоэкспоненциального распределения на примере двухфазного распределения.

Известно, что преобразование Лапласа суммы независимых случайных величин равно произведению преобразований Лапласа слагаемых величин. Тогда преобразование Лапласа двухфазного гипоэкспоненциального распределения будет равно произведению преобразований Лапласа составляющих экспоненциальных распределений:

$$F^*(s) = F_1^*(s)\,F_2^*(s) = \frac{\alpha_1}{\alpha_1 + s} \times \frac{\alpha_2}{\alpha_2 + s} = \frac{1}{1 + st_1} \times \frac{1}{1 + st_2}.$$

Дифференцируя преобразование Лапласа по s в точке s=0, в соответствии с (2.7) найдем математическое ожидание и второй начальный момент для гипоэкспоненциального распределения:

$$\mathrm{M}_{нсЭ_2} = t_1 + t_2; \qquad \alpha^{(2)}_{нсЭ_2} = 2(t_1^2 + t_1 t_2 + t_2^2).$$

Отсюда дисперсия, среднеквадратическое отклонение и коэффициент вариации будут равны:

$$D_{нсЭ_2} = t_1^2 + t_2^2; \qquad \sigma_{нсЭ_2} = \sqrt{t_1^2 + t_2^2}; \qquad \nu_{нсЭ_2} = \frac{\sqrt{t_1^2 + t_2^2}}{t_1 + t_2}.$$

На рис.2.14 показана зависимость коэффициента вариации двухфазного гипоэкспоненциального распределения от отношения t_1 / t_2 параметров экспоненциальных составляющих.

Как видно из графика, коэффициент вариации гипоэкспоненциального распределения изменяется в пределах от 1 до 0,7, а точнее до значения коэффициента вариации распределения Эрланга 2-го порядка: $\nu = 0{,}707$, когда параметры экспоненциальных составляющих равны между собой: $t_1 = t_2$.

Очевидно, что для того, чтобы увеличить интервал изменения коэффициента вариации гипоэкспоненциального распределения, необходимо вместо двухфазного использовать многофазное представление.

Можно показать, что для гипоэкспоненциального распределения *k*-го порядка математическое ожидание, дисперсия и коэффициент вариации будут равны:

$$\mathrm{M}_{нсЭ_k} = \sum_{i=1}^{k} t_i; \qquad D_{нсЭ_k} = \sum_{i=1}^{k} t_i^2; \qquad \nu_{нсЭ_k} = \frac{\sqrt{\sum_{i=1}^{k} t_i^2}}{\sum_{i=1}^{k} t_i}. \tag{2.20}$$

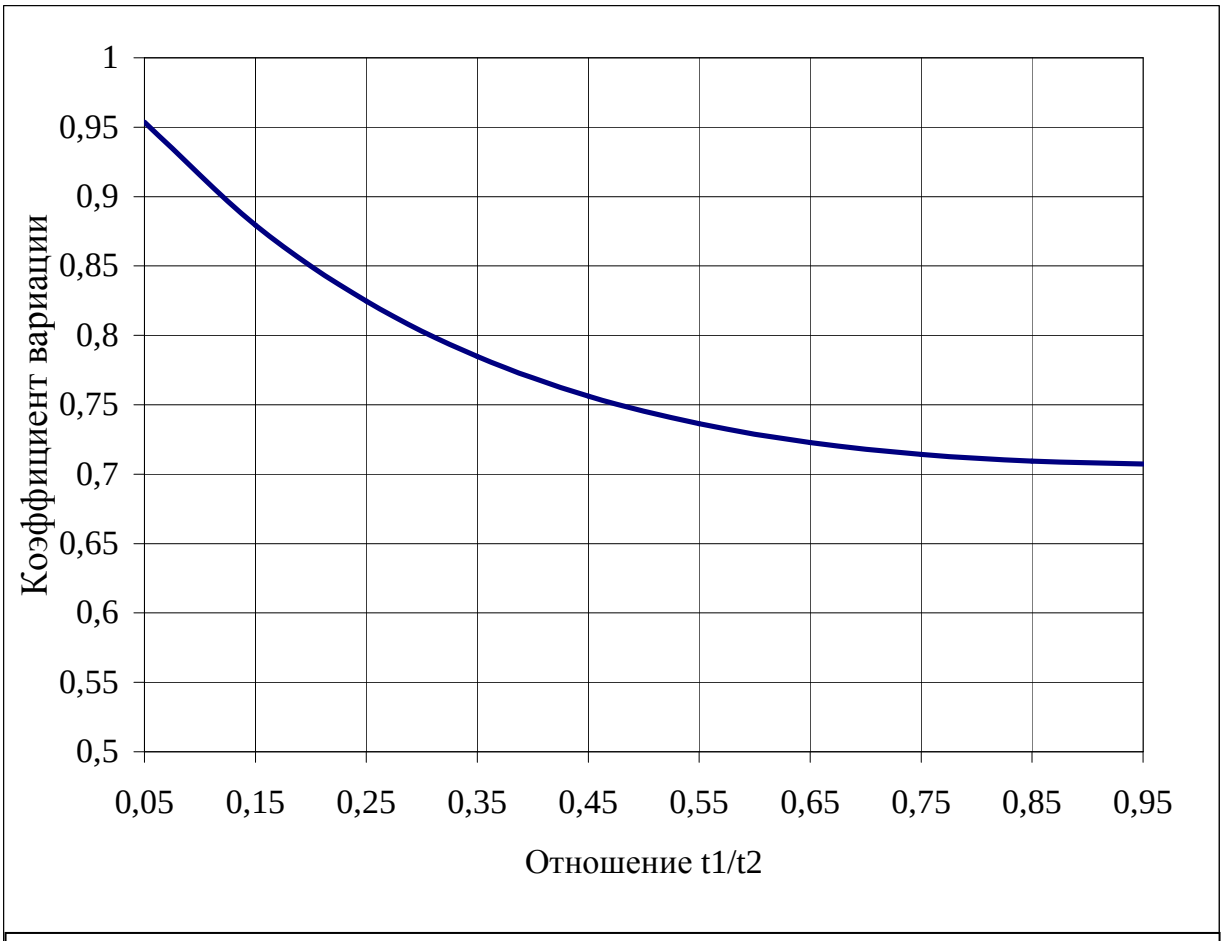


Рис.2.14. Зависимость коэффициента вариации от отношения t_1/t_2

Легко убедиться, что коэффициент вариации гипоэкспоненциального распределения лежит в интервале $(1/\sqrt{k}\,;\,1)$, причем с увеличением *k* левая

граница этого интервала приближается к нулю.

Рассмотрим задачу аппроксимации реального распределения с коэффициентом вариации $0 < \nu < 1$ гипоэкспоненциальным распределением.

Положим, что известны математическое ожидание t и коэффициент вариации ν (причем $0 < \nu < 1$) некоторой случайной величины τ, определенной в положительной области действительных чисел.

Для простоты, без потери общности, положим, что аппроксимирующее гипоэкспоненциальное распределение содержит только два типа экспоненциальных фаз: k_1 фаз с параметром $\alpha_1 = 1/t_1$ и $k_2 = k - k_1$ фаз с параметром $\alpha_2 = 1/t_2$, где t_1 и t_2 – математические ожидания экспоненциально распределенных случайных величин в фазах первого и второго типов соответственно.

Тогда из (2.20) следует, что математическое ожидание, дисперсия и коэффициент вариации будут равны:

$$\mathrm{M}_{нсЭ_k} = k_1 t_1 + k_2 t_2; \quad D_{нсЭ_k} = k_1 t_1^2 + k_2 t_2^2; \quad \nu_{нсЭ_k} = \frac{\sqrt{k_1 t_1^2 + k_2 t_2^2}}{k_1 t_1 + k_2 t_2}, \qquad (2.21)$$

причем $k_1 + k_2 = k$.

Таким образом, для аппроксимации по двум моментам необходимо, чтобы выполнялись следующие два условия:

$$\left.\begin{aligned} & k_1 t_1 + k_2 t_2 = t; \\ & \frac{\sqrt{k_1 t_1^2 + k_2 t_2^2}}{k_1 t_1 + k_2 t_2} = \nu \end{aligned}\right\},$$

где t и ν соответственно математическое ожидание и коэффициент вариации аппроксимируемого распределения.

После некоторых простых преобразований получим систему из двух линейных алгебраических уравнений:

$$\left.\begin{aligned} & k_1 t_1 + k_2 t_2 = t; \\ & k_1 t_1^2 + k_2 t_2^2 = \nu^2 t^2 \end{aligned}\right\}.$$

Полагая, что значения k_1 и k_2 заданы, решим полученную систему уравнений относительно неизвестных t_1 и t_2.

Из первого уравнения следует, что

$$t_2 = \frac{t - k_1 t_1}{k_2}. \qquad (2.22)$$

Подставляя это выражение во второе уравнение, после некоторых алгебраических преобразований получим квадратное уравнение с одним неизвестным t_1:

$$k_1 (k_1 + k_2) t_1^2 - 2 k_1 t\, t_1 + (t^2 - k_2 \nu^2 t^2) = 0.$$

Решая это квадратное уравнение, получим:

$$t_1 = \frac{t}{k}\left[1 \pm \sqrt{\frac{k_2}{k_1}\left(k\nu^2 - 1\right)}\right], \qquad (2.23)$$

где $k = k_1 + k_2$.

В качестве приемлемого решения могут быть использованы оба корня квадратного уравнения.

Для того чтобы в (2.23) под знаком квадратного корня иметь неотрицательную величину, необходимо выполнение следующего условия:

$$k \geq \frac{1}{\nu^2}. \qquad (2.24)$$

Полученное условие определяет минимальное количество фаз в аппроксимирующем гипоэкспоненциальном распределении.

Для того чтобы второе решение со знаком минус перед квадратным корнем давало $t_1 \geq 0$, дополнительно необходимо выполнение условия:

$$k_2 \leq \frac{1}{\nu^2}. \qquad (2.25)$$

Объединив условия (2.24) и (2.25), окончательно получим вполне очевидное условие:

$$k_2 \leq \frac{1}{\nu^2} \leq k. \qquad (2.26)$$

Подставим теперь (2.23) в (2.22) и найдем t_2:

$$t_2 = \frac{t}{k}\left[1 \mp \sqrt{\frac{k_1}{k_2}\left(k\nu^2 - 1\right)}\right], \qquad (2.27)$$

для которого получим условие, аналогичное (2.26):

$$k_1 \leq \frac{1}{\nu^2} \leq k. \qquad (2.28)$$

Таким образом, окончательно имеем следующие выражения для аппроксимации гипоэкспоненциальным распределением *k*-го порядка законов распределений случайных величин с коэффициентом вариации $0 < \nu < 1$:

$$k \geq \frac{1}{\nu^2}; \quad t_1 = \frac{t}{k}\left[1 + \sqrt{\frac{k_2}{k_1}\left(k\nu^2 - 1\right)}\right]; \quad t_2 = \frac{t}{k}\left[1 - \sqrt{\frac{k_1}{k_2}\left(k\nu^2 - 1\right)}\right] \qquad (2.29)$$

или

$$k \geq \frac{1}{\nu^2}; \quad t_1 = \frac{t}{k}\left[1 - \sqrt{\frac{k_2}{k_1}\left(k\nu^2 - 1\right)}\right]; \quad t_2 = \frac{t}{k}\left[1 + \sqrt{\frac{k_1}{k_2}\left(k\nu^2 - 1\right)}\right] \qquad (2.30)$$

Окончательно *алгоритм аппроксимации реального распределения с коэффициентом вариации* $0 < \nu < 1$ *гипоэкспоненциальным распределением* при заданных значениях математического ожидания t и коэффициента вариации ν (причем $0 < \nu < 1$) некоторой случайной величины τ, определенной в положительной области действительных чисел, выглядит следующим образом:

1) на основе первого выражения в (2.29) и (2.30) по заданному значению коэффициента вариации ν определяется минимально необходимое число экспоненциальных фаз k в аппроксимирующем распределении как ближайшее большее целое по отношению к $1/\nu^2$;

2) выбирается значение $k_1 \le k$ и рассчитывается $k_2 = k - k_1$;

3) на основе (2.29) или (2.30) рассчитываются значения t_1 и t_2.

Результаты аппроксимации на основе выражений (2.29) и (2.30), а также при различных значениях k_1 и $k_2 = k - k_1$, различаются значениями третьего и более высоких моментов распределения, но имеют одинаковые первые и вторые моменты.

Учет более высоких моментов при аппроксимации реальных распределений не вызывает принципиальных трудностей, но сопровождается более громоздкими математическими выкладками. К тому же, во многих случаях, влияние этих моментов на конечные результаты исследований оказывается незначительным.

Пример. Пусть математическое ожидание и коэффициент вариации аппроксимируемого выражения соответственно равны: $t = 10$ и $\nu = 0{,}4$.

В соответствии с выше изложенным алгоритмом аппроксимации:

1) минимально необходимое число экспоненциальных фаз в аппроксимирующем распределении: $k = 7$ $\left(k \ge \frac{1}{0{,}16} = 6{,}25\right)$;

2) выберем значение $k_1 = 3$, тогда $k_2 = 7 - 3 = 4$;

на основе (2.29) рассчитываются значения $t_1 = 2$ и $t_2 = 1$.

Таким образом, в качестве аппроксимирующего распределения выбираем семифазное гипоэкспоненциальное распределение, в котором *три* экспоненциальные фазы имеют математическое ожидание, равное 2, и *четыре* фазы – математическое ожидание, равное 1.

2.6.2. Аппроксимация распределения с коэффициентом вариации $\nu > 1$

Положим, что математическое ожидание и коэффициент вариации некоторой случайной величины τ, определенной в положительной области действительных чисел, соответственно равны t и ν, причем $\nu > 1$.

Для аппроксимации закона распределения такой случайной величины в теории массового обслуживания часто используют *гиперэкспоненциальное распределение*, представляющее собой композицию экспоненциальных распределений.

Гиперэкспоненциальное распределение H_r может быть представлено в виде множества параллельных фаз с экспоненциальными распределениями с параметрами $\alpha_i = 1/t_i$, где $t_i = M[\tau_i]$ $(i = \overline{1,r})$ – математическое ожидание экспоненциально распределенной случайной

величины в i-ой фазе (рис.2.15).

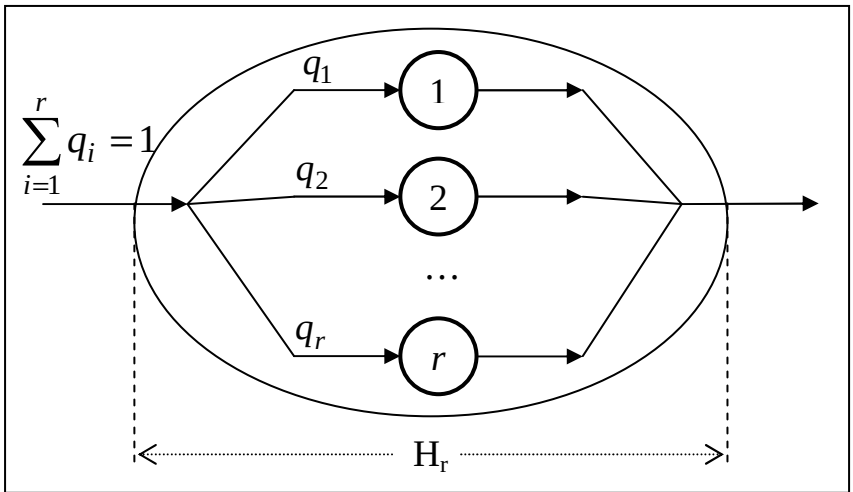


Рис.2.15. Многофазное представление гиперэкспоненциального распределения

В простейшем случае гиперэкспоненциальное распределение может быть представлено в виде двухфазного распределения (рис.2.16). Параметрами такого распределения являются: t_1 и t_2 – математические ожидания первой и второй экспоненциальных фаз соответственно; q – вероятность формирования значения случайной величины в первой фазе.

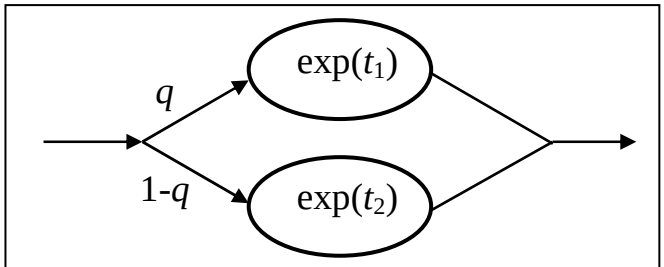


Рис.2.16. Двухфазное представление гиперэкспоненциального распределения

Полученное таким образом распределение является трехпараметрическим. Это означает, что аппроксимация таким распределением может выполняться по трем числовым моментам. Выбор значений параметров гиперэкспоненциального распределения только по двум моментам (математическому ожиданию и коэффициенту вариации) предполагает наличие некоторого произвола.

Таким образом, задача аппроксимации гиперэкспоненциальным распределением сводится к определению значений параметров t_1, t_2 и q в зависимости от известных значений математического ожидания t и коэффициента вариации ν аппроксимируемого закона распределения случайной величины τ.

Математическое ожидание и второй начальный момент гиперэкспоненциального распределения соответственно равны:

$$t = q\,t_1 + (1-q)\,t_2; \qquad (2.31)$$

$$t^{(2)} = 2[q\,t_1^2 + (1-q)\,t_2^2].$$

Тогда коэффициент вариации гиперэкспоненциального распределения:

$$\nu^2 = \frac{2[q\,t_1^2 + (1-q)\,t_2^2]}{t^2} - 1,$$

откуда:

$$2[q t_1^2 + (1-q) t_2^2] = t^2 (1+\nu^2). \qquad \text{й} \qquad (2.32)$$

Из (2.31) имеем:

$$t_2 = \frac{t - q t_1}{1-q}. \qquad \text{й ий й} \qquad (2.33)$$

Подставив последнее выражение в (2.32), после некоторых преобразований получим квадратное уравнение:

$$2q t_1^2 - 4q t t_1 + [1 + q - (1-q)\nu^2] t^2 = 0. \qquad (2.34)$$

Решая это квадратное уравнение относительно t_1, получим:

$$t_1 = t\left[1 \pm \sqrt{\frac{1-q}{2q}(\nu^2 - 1)}\right].$$

Для того чтобы гарантировать $t_1 > 0$, в качестве решения выберем корень уравнения со знаком плюс перед знаком радикала:

$$t_1 = \left[1 + \sqrt{\frac{1-q}{2q}(\nu^2 - 1)}\right] t. \qquad (2.35)$$

Подставим (2.35) в (2.33) и найдем t_2:

$$t_2 = \left[1 - \sqrt{\frac{q}{2(1-q)}(\nu^2 - 1)}\right] t. \qquad (2.36)$$

Потребуем, чтобы $t_2 \geq 0$, то есть $\frac{q}{2(1-q)}(\nu^2 - 1) \leq 1$. Отсюда:

$$q \leq \frac{2}{1+\nu^2}\ . \qquad (2.37)$$

Выражения (2.35) – (2.37) можно использовать для аппроксимации любого закона распределения с коэффициентом вариации $\nu > 1$ двухфазным гиперэкспоненциальным распределением, для чего достаточно выбрать значение вероятности q из условия (2.37) и рассчитать значения t_1 и t_2 в соответствии с (2.35) и (2.36).

Рассмотрим частный случай, когда $q = \frac{2}{1+\nu^2}$. Подставляя это выражение в (2.35) и (2.36), получим:

$$t_1 = \frac{\nu^2 + 1}{2} t; \quad t_2 = 0. \qquad (2.38)$$

Последние выражения соответствуют однофазному представлению гиперэкспоненциального распределения, показанному на рис.2.17. Заметим, что полученные для t_1 и t_2 выражения (2.35) и (2.36) – симметричны. Можно

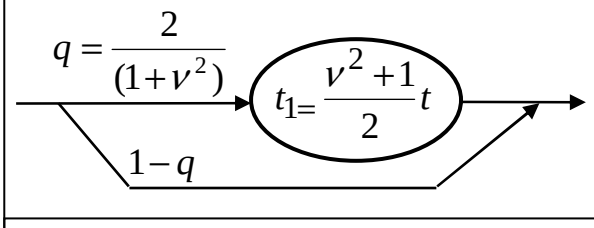


Рис.2.17. Однофазное распределение

показать, что если выбрать в качестве решения квадратного уравнения (2.34) второй корень со знаком минус перед знаком радикала и потребовать, чтобы выражение в квадратных скобках не было отрицательным, то получим:

$$t_1 = t\left[1-\sqrt{\frac{1-q}{2q}(\nu^2-1)}\right]; \qquad (2.39)$$

$$t_2 = t\left[1+\sqrt{\frac{q}{2(1-q)}(\nu^2-1)}\right], \qquad (2.40)$$

а условие (2.37) для выбора значения q примет вид:

$$q \geq \frac{\nu^2-1}{\nu^2+1}, \qquad (2.41)$$

что равносильно перестановке двух экспоненциальных фаз (см. рис.2.16) гиперэкспоненциального распределения.

Пример. Пусть $\nu = 3$, тогда в соответствии с (2.37): $q \leq \frac{2}{1+3^2} = 0{,}2$.

Рассмотрим два варианта аппроксимации.

1) Выберем $q = 0{,}1$, тогда в соответствии с (2.35) и (2.36):

$$t_1 = 7\,t; \qquad t_2 = \frac{1}{3}t.$$

Таким образом, в качестве аппроксимирующего распределения может быть выбрано двухфазное гиперэкспоненциальное распределение, в котором с вероятностью $q = 0{,}1$ случайная величина формируется в первой фазе с математическим ожиданием в 7 раз большим, чем математическое ожидание аппроксимируемой случайной величины, и с вероятностью $q = 0{,}9$ случайная величина формируется во второй фазе с математическим ожиданием в 3 раза меньшим, чем математическое ожидание аппроксимируемой случайной величины.

2) Выберем $q = 0{,}2$, тогда в соответствии с (2.38):

$$t_1 = \frac{3^2+1}{2}t = 5t; \qquad t_2 = 0.$$

В этом варианте в качестве аппроксимирующего распределения используется однофазное гиперэкспоненциальное распределение, в котором с вероятностью $q = 0{,}2$ случайная величина формируется в единственной фазе с математическим ожиданием в 5 раз большим, чем математическое ожидание аппроксимируемой случайной величины, и с вероятностью $q = 0{,}8$ случайная величина принимает значение 0. Таким образом, этот вариант аппроксимации предполагает, что 80% значений случайной величины будут нулевыми.

Рассмотренные два варианта аппроксимации обеспечивают Одинаковые значения математических ожиданий и коэффициентов вариаций, но различаются третьими и более высокими моментами распределений.

Очевидно, что второй вариант аппроксимации может оказаться более предпочтительным, например, при аппроксимации времени ожидания в системе массового обслуживания, если известно, что около 80% заявок прошли через систему с нулевым ожиданием, и коэффициент вариации времени ожидания больше единицы.

2.7. Резюме

1. Базовые понятия теории вероятностей – «*событие*», «*вероятность*», «*случайная величина*».

Вероятность – численная мера степени объективной возможности некоторого события. Вероятность может принимать только положительные значения из интервала [0; 1].

Величина, принимающая значение, *неизвестное заранее*, называется *случайной*. Различают *дискретные* (прерывные) и *непрерывные* (аналоговые) случайные величины.

2. *Закон распределения* случайной величины – соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями.

Закон распределения дискретной случайной величины может быть задан:

- *аналитически* в виде математического выражения;
- *таблично* в виде ряда распределения;
- *графически* в виде многоугольника распределения.

Закон распределения непрерывной случайной величины может быть задан в виде:

- *функции распределения* $F(x)$ случайной величины X, представляющей собой *вероятность* того, что случайная величина X примет значение меньшее, чем некоторое заданное значение x: $F(x) = \mathrm{P}(X < x)$;
- *плотности распределения* $f(x)$, определяемой как производная от функции распределения $F(x)$ по x: $f(x) = F'(x)$.

Функция распределения однозначно определяется через плотность распределения как

$$F(x) = \int_{-\infty}^{x} f(x)\,dx.$$

Свойства функции распределения:

- $F(x)$ – неубывающая функция: если $x_j > x_i$, то $F(x_j) \geq F(x_i)$;
- функция распределения принимает значения от 0 до 1, причём: $F(-\infty) = 0$ и $F(+\infty) = 1$.

Свойства плотности распределения:

- плотность распределения принимает только неотрицательные значения: $f(x) \geq 0$;

• площадь на графике, ограниченная плотностью распределения и осью абсцисс, всегда равна единице: $\int_{-\infty}^{+\infty} f(x)\,dx = 1.$

3. Числовые характеристики – *начальные* $\alpha_s[X]$ и *центральные* $\beta_s[X]$ *моменты* – позволяют выразить в сжатой форме наиболее существенные особенности распределения случайной величины

Первый начальный момент случайной величины X называется *математическим ожиданием* и характеризует *среднее значение* случайной величины: $M[X] = \alpha_1[X]$.

Второй начальный момент $\alpha_2[X]$ случайной величины X характеризует *разброс* значений случайной величины *относительно начала координат*.

Второй центральный момент называется *дисперсией* случайной величины: $\mathrm{D}[X] = \beta_2[X]$ и характеризует *разброс* значений случайной величины *относительно математического ожидания*.

Дисперсия и второй начальный момент связаны зависимостью

$\mathrm{D}[X] = \alpha_2[X] - (\mathrm{M}[X])^2$.

Среднеквадратическое отклонение $\sigma[X]$ – характеристика *разброса, размерность которой совпадает с размерностью случайной величины*:

$\sigma[X] = \sqrt{\mathrm{D}[X]}$.

Коэффициент вариации $\nu[X]$ – *безразмерная* характеристика *разброса* случайных величин, определенных в области положительных значений:

$\nu[X] = \sigma[X] / \mathrm{M}[X] \quad (\mathrm{M}[X] > 0).$

4. *Производящая функция* распределения $p_k = \mathrm{P}(X = k)$ дискретной случайной величины X:

$$X^*(z) = \sum_{k=0}^{\infty} z^k p_k \qquad |z| \le 1.$$

Математическое ожидание и дисперсия:

$\mathrm{M}[X] = p'(1); \qquad \mathrm{D}[X] = p''(1) + p'(1) - [p'(1)]^2.$

Производящая функция $X^*(z)$ суммы $X = X_1 + X_2 + \dots X_n$ независимых случайных величин:

$X^*(z) = X_1^*(z) X_2^*(z) \dots X_n^*(z)$.

Преобразование Лапласа плотности распределения $f(x)$ неотрицательной непрерывной случайной величины X :

$$F^*(s) = \int_0^{\infty} e^{-sx} f(x)\,dx \qquad (s \ge 0).$$

Начальные моменты случайной величины:

$$\alpha_k[X] = \frac{(-1)^k}{k!}\frac{d^k}{ds^k}F^*(s)\Big|_{s=0}.$$

Преобразование Лапласа суммы $Z = X + Y$ независимых случайных величин X и Y равно произведению преобразований Лапласа слагаемых:

$$Z^*(s) = X^*(s)Y^*(s).$$

5. В моделях дискретных систем наиболее широко применяются следующие законы распределений случайных величин:

• *распределение Пуассона* (дискретный закон):

$$p_k = P(X = k) = \frac{a^k}{k!}e^{-a} \qquad (k = 0, 1, 2, \ldots),$$

где a – параметр распределения ($a > 0$);

• *геометрическое распределение* (дискретный закон):

$$p_k = P(X = k) = \rho^k(1-\rho) \qquad (k = 0, 1, 2, \ldots),$$

где ρ – параметр распределения ($0 < \rho < 1$);

• *равномерное распределение* (непрерывный закон) с плотностью

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{при} \quad a < x < b; \\ 0 & \text{при} \quad x > b; \end{cases}$$

• *экспоненциальное распределение* (непрерывный закон) с функцией и плотностью

$$F(x) = 1 - e^{-\alpha x}; \qquad f(x) = \alpha e^{-\alpha x},$$

где $\alpha > 0$ – параметр распределения; $x \ge 0$; $\nu_{эксп}[X] = 1.$

• *распределение Эрланга k-го порядка* (непрерывный закон) с функцией и плотностью:

$$F_k(x) = 1 - e^{-\alpha x}\sum_{i=0}^{k-1}\frac{(\alpha x)^i}{i!}; \qquad f_k(x) = \frac{\alpha(\alpha x)^{k-1}}{(k-1)!}e^{-\alpha x},$$

где α и k –параметры распределения ($\alpha \ge 0$; $k = 1, 2, \ldots$); $x \ge 0$; $\nu_{Э_k}[X] = \dfrac{1}{\sqrt{k}} \le 1$; *математическое ожидание распределения Эрланга зависит от значения параметра k*;

• *нормированное распределение Эрланга* (непрерывный закон) с функцией и плотностью:

$$F_k(x) = 1 - e^{-k\alpha x}\sum_{i=0}^{k-1}\frac{(k\alpha x)^i}{i!}; \qquad f_k(x) = \frac{k\alpha(k\alpha x)^{k-1}}{(k-1)!}e^{-k\alpha x},$$

коэффициент вариации нормированного распределения Эрланга также меньше или равен единице: $\nu_{нЭ_k}[X] = \dfrac{1}{\sqrt{k}} \le 1$, но математическое ожидание *не зависит* от значения параметра k;

• *гиперэкспоненциальное распределение* (непрерывный закон):

$$\left.\begin{aligned} F(x) &= \sum_{i=1}^{n} q_i (1 - e^{-\alpha_i x}) = 1 - \sum_{i=1}^{n} q_i e^{-\alpha_i x}; \\ f(x) &= \sum_{i=1}^{n} q_i \alpha_i e^{-\alpha_i x} \end{aligned}\right\};$$

• *гиперэрланговское распределение* представляет собой аддитивную смесь *нормированных распределений Эрланга* и является наиболее *общим распределением* неотрицательных непрерывных случайных величин, поскольку имеет *коэффициент вариации в интервале от 0 до* ∞; плотность гиперэкспоненциального распределения:

$$f(x) = \sum_{i=1}^{n} q_i \frac{k_i \alpha_i (k_i \alpha_i x)^{k_i - 1}}{(k_i - 1)!} e^{-k_i \alpha_i x} \qquad (x \geq 0).$$

6. Если реальные временны*е* интервалы имеют значения коэффициента вариации, значительно отличающиеся от единицы, использование экспоненциального распределения может привести к большим погрешностям конечных результатов. В этих случаях в качестве аппроксимирующих функций законов распределений могут использоваться вероятностные законы, представляющие собой композицию экспоненциальных распределений, при этом аппроксимация реального распределения, в простейшем случае, выполняется по двум первым моментам: математическому ожиданию t и коэффициенту вариации ν.

В качестве таких аппроксимирующих распределений могут использоваться:

• если коэффициент вариации временного интервала меньше единицы $(0 < \nu < 1)$, *гипоэкспоненциальное распределение*, параметры которого рассчитываются по формулам:

$$k \geq \frac{1}{\nu^2}; \quad t_1 = \frac{t}{k}\left[1 + \sqrt{\frac{k_2}{k_1}\left(k\nu^2 - 1\right)}\right]; \quad t_2 = \frac{t}{k}\left[1 - \sqrt{\frac{k_1}{k_2}\left(k\nu^2 - 1\right)}\right];$$

• если коэффициент вариации временного интервала больше единицы $(\nu > 1)$, *гипреэкспоненциальное распределение*, параметры которого рассчитываются по формулам:

$$q \leq \frac{2}{1 + \nu^2}; \quad t_1 = \left[1 + \sqrt{\frac{1-q}{2q}(\nu^2 - 1)}\right] t; \quad t_2 = \left[1 - \sqrt{\frac{q}{2(1-q)}(\nu^2 - 1)}\right] t.$$

2.8. Практикум: решение задач

Задача 1. Дискретная случайная величина X принимает значения: 1; 2; 3 с вероятностями 0,2; 0,3; 0,5 соответственно.

1) Нарисовать график функции распределения дискретной случайной величины X.

2) Вычислить математическое ожидание, дисперсию, второй

начальный момент, среднеквадратическое отклонение и коэффициент вариации случайной величины *X*.

Дано: $x_1 = 1;\ \ x_2 = 2;\ \ x_3 = 1;$

$p_1 = 0{,}2;\ \ p_2 = 0{,}3;\ \ p_3 = 0{,}5.$

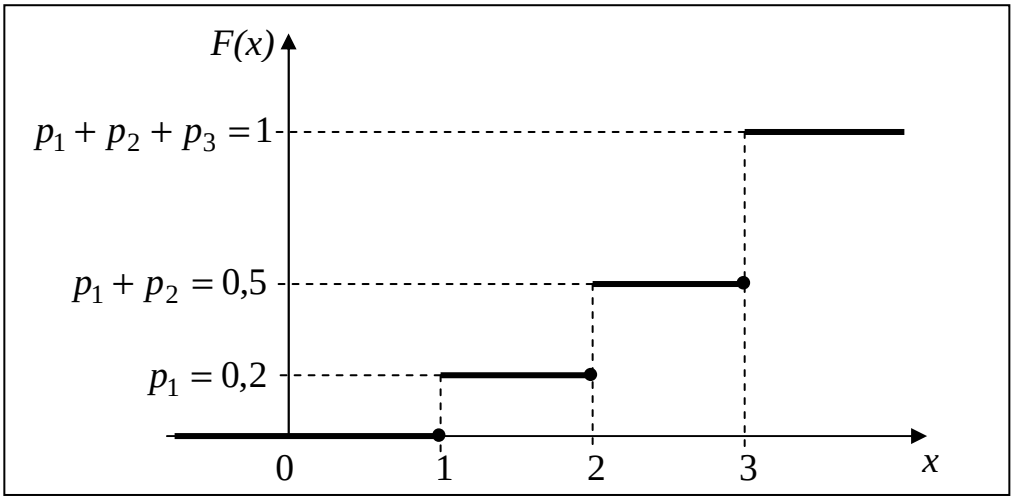
Требуется:

1) нарисовать $F(x)$;

2) вычислить $M[X]$, $D[X]$, $\alpha_2[X]$, $\sigma[X]$, $\nu[X]$.

Решение.

1) График функции распределения случайной величины *X*:



Следует отметить, что значения функции распределения $F(x)$ для каждого значения случайной величины x_i увеличиваются на величину, равную соответствующей *вероятности* p_i появления этого значения, причем самое верхнее значение *всегда равно* 1.

Отметим также, что, как показано на графике (в виде черных кружочков), значения функции распределения в точках $x = 1$, $x = 2$ и $x = 3$ соответственно равны: $F(1) = 0$, $F(2) = 0{,}2$ и $F(3) = 0{,}5$, поскольку функция распределения $F(x)$ определяется как вероятность появления случайной величины, значение которой строго меньше (а не меньше или равно) x: $F(x) = \mathrm{P}(X < x)$.

2) Математическое ожидание:

$M[X] = p_1x_1 + p_2x_2 + p_3x_3 = 0{,}2\times1 + 0{,}3\times2 + 0{,}5\times3 = 2{,}3.$

Второй начальный момент:

$\alpha_2[X] = p_1x_1^2 + p_2x_2^2 + p_3x_3^2 = 0{,}2\times1 + 0{,}3\times4 + 0{,}5\times9 = 5{,}9.$

Дисперсия: $D[X] = \alpha_2[X] - (M[X])^2 = 5{,}9 - 5{,}29 = 0{,}61.$

Среднеквадратическое отклонение: $\sigma[X] = \sqrt{D[X]} \approx 0{,}78.$

Коэффициент вариации: $\nu[X] = \dfrac{\sigma[X]}{M[X]} \approx 0{,}34.$

Задача 2. Чему равно математическое ожидание, дисперсия, второй начальный момент и коэффициент вариации детерминированной величины $x=10$? Нарисовать график функции и плотности распределения случайной величины.

Дано: детерминированная величина: $x=10$.

Требуется:

1) вычислить $M[X]$, $D[X]$, $\alpha_2[X]$, $\nu[X]$;

2) нарисовать $F(x)$ и $f(x)$.

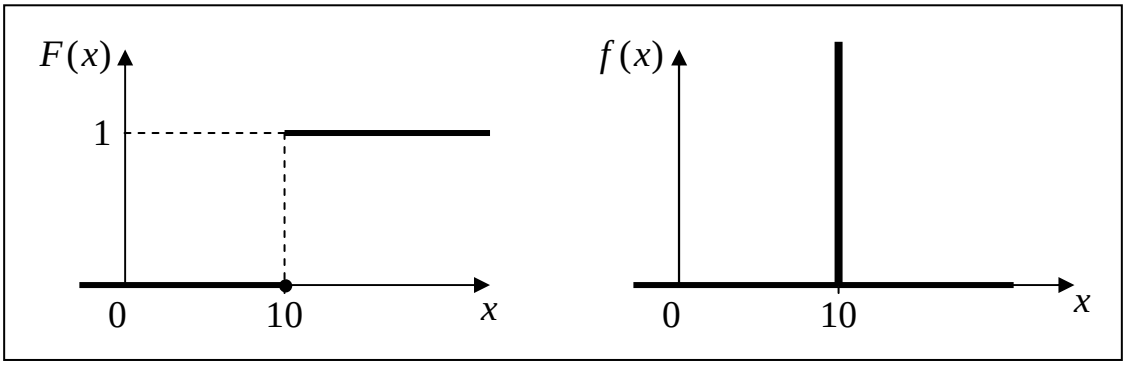
Решение.

1) Детерминированную величину можно рассматривать как случайную величину, принимающую одно и то же значение $x=10$ с вероятностью $p=1$. Тогда:

- математическое ожидание: $M[X]=px=10$;
- второй начальный момент: $\alpha_2[X]=px^2=100$;
- дисперсия: $D[X]=\alpha_2[X]-(M[X])^2=0$;
- коэффициент вариации: $\nu[X]=\dfrac{\sqrt{D[X]}}{M[X]}=0$.

Полученные достаточно тривиальные результаты становятся очевидными, если вспомнить физическое толкование представленных величин. Математическое ожидание, представляющее собой среднее значение случайной величины, естественно, совпадает с единственно возможным значением $x=10$. Дисперсия, среднеквадратическое отклонение и коэффициент вариации, определяющие разброс значений относительно математического ожидания, очевидно всегда равны нулю для детерминированной величины, поскольку разброса значений просто нет. Однако следует обратить внимание, что *второй начальный момент не равен нулю*, хотя тоже определяет разброс значений, но, в отличие от предыдущих характеристик, относительно начала координат. Действительно, единственное значение $x=10$ находится от начала координат на «расстоянии», не равном нулю и, следовательно, второй начальный момент отличен от нуля.

2) Графики функции и плотности распределения детерминированной величины:



Как следует из представленных графиков, функция распределения детерминированной величины представляет собой *функцию Хевисайда*, а плотность распределения – *дельта-функцию*:

$$F(x) = H(x - \mathrm{M}); \quad f(x) = \delta(x - \mathrm{M}),$$

где M – математическое ожидание, равное значению детерминированной величины (в нашем случае M=10).

Задача 3. Непрерывная случайная величина равномерно распределена в интервале (-30; +20). Нарисовать график плотности и функции распределения случайной величины. Определить: а) математическое ожидание случайной величины; б) вероятность того, что случайная величина принимает положительные значения.

Дано: равномерно распределённая случайная величина в интервале (-30; +20).

Требуется:

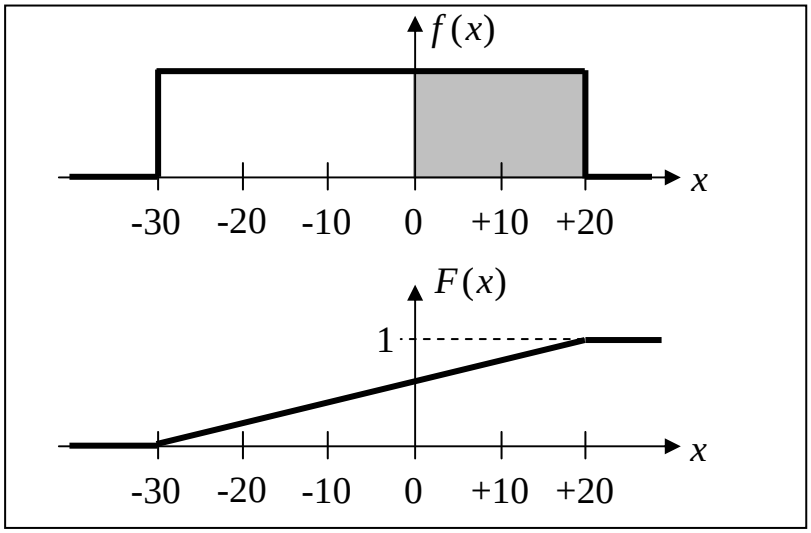
1) нарисовать $f(x)$ и $F(x)$;

2) вычислить $M[X]$;

3) определить $\Pr(X \geq 0) = 1 - F(0)$.

Решение.

1) Графики плотности и функции равномерно распределённой случайной величины:



2) Очевидно, что математическое ожидание равномерно распределённой случайной величины находится в середине заданного интервала (-30; +20) и равно: $\mathrm{M} = -5$. Этот же результат может быть получен с использованием формулы для расчета математического ожидания равномерно распределённой случайной величины: $\mathrm{M} = \frac{a+b}{2} = \frac{-30+20}{2} = -5$, где a и b - соответственно левая и правая границы интервала.

3) Вероятность того, что случайная величина принимает положительные значения, также может быть определена несколькими

способами.

Во-первых, через значение функции распределения: $\Pr(X \geq 0) = 1 - F(0) = 1 - 0{,}6 = 0{,}4$. Во-вторых, из графика плотности распределения как площадь под плотностью распределения, ограниченная слева значением $x = 0$ и справа значением $x = +20$ (на графике выделена серым цветом). Помня, что площадь под плотностью распределения на всём интервале значений случайной величины равна 1, можно сделать вывод, что площадь на интервале значений (0, +20) составляет 2/5, то есть равна 0,4.

2.9. Самоконтроль: перечень вопросов и задач

1. Что понимается под случайной величиной?
2. Приведите примеры случайных величин.
3. Является ли случайной величиной…:

… число дней в году?
… рост человека?
… количество пассажиров в автобусе?
… интервал между поездами в метро?
… температура воздуха на улице?
… напряжение в электрической сети?
… количество студентов в группе?
… оценка на экзамене?

Ответы сопроводить необходимыми пояснениями. Какие из перечисленных величин являются непрерывными?

4. Приведите примеры дискретных и непрерывных случайных величин.
5. Что характеризует вероятность?
6. Как рассчитать вероятность какого-либо события?
7. Рассчитайте вероятность того, что в тщательно перемешанной колоде из 36 карт нижней картой окажется …:

… туз пиковый?
… туз любой масти?
… любой масти туз или король?
… пиковый туз или пиковый король?
… пиковый туз или крестовый король?
… любая карта красной масти?

8. Рассчитайте вероятность того, что в тщательно перемешанной колоде из 36 карт две верхние карты окажутся …:

… тузами?
… туз и король одной масти?
… пиковый туз и пиковый король?
… пиковый туз или король любой масти?
… красной масти?

9. Из тщательно перемешанной колоды, содержащей 36 карт,

отброшена половина карт. Какова вероятность того, что в оставшейся половине колоды находится …:

… туз пиковый?

… туз любой масти?

… любой масти туз или король?

… пиковый туз или пиковый король?

… пиковый туз или крестовый король?

… любая карта красной масти?

10. Понятие и способы задания закона распределения случайной величины.

11. Понятие и свойства функции распределения случайной величины.

12. Понятие и свойства плотности распределения вероятностей.

13. Что характеризует и какую размерность имеет математическое ожидание (дисперсия; второй начальный момент; среднеквадратическое отклонение; коэффициент вариации, функция распределения, плотность распределения) случайной величины?

14. Для чего используются производящая функция и преобразование Лапласа? Для каких случайных величин используется преобразование Лапласа?

15. Назовите известные Вам дискретные и непрерывные законы распределений.

16. Чему равен коэффициент вариации: а) экспоненциального распределения; б) распределения Эрланга 9-го порядка?

17. В каком интервале находится коэффициент вариации распределения: а) Эрланга; б) гиперэкспоненциального; в) гиперэрланговского?

18. Нарисовать график плотности и функции распределения: а) экспоненциального; б) Эрланга; в) гиперэкспоненциального .

19. Показать на графике и пояснить, в чём различие между плотностями распределений экспоненциального и гиперэкспоненциального законов.

20. Показать на графике и пояснить, в чём различие между плотностями распределений Эрланга 2-го и 4-го порядка.

21. Дискретная случайная величина принимает значения 1; 2; 3 с вероятностями 0,5; 0,4; 0,1 соответственно. Вычислить математическое ожидание, дисперсию, второй начальный момент, среднеквадратическое отклонение, коэффициент вариации.

22. Чему равно математическое ожидание, дисперсия, второй начальный момент и коэффициент вариации детерминированной величины X=25?

23. Определить значение детерминированной величины X, если известно, что ее второй начальный момент равен 25?

24. Определить коэффициент вариации детерминированной величины X, если известно, что ее второй начальный момент равен 121?

25. Математическое ожидание экспоненциально распределенной

случайной величины равно 0,1. Определить среднеквадратическое отклонение, второй начальный момент и коэффициент вариации.

26. Дисперсия экспоненциально распределенной случайной величины равна 16. Определить математическое ожидание и второй начальный момент.

27. Чему равен коэффициент вариации распределения Эрланга 4-го порядка?

28. Чему равна дисперсия случайной величины, распределенной по закону Эрланга 9-го порядка с математическим ожиданием, равным 15?

29. Дискретная случайная величина принимает значения 1; 2; 3 соответственно с вероятностями 0,3; 0,1; 0,6. Нарисовать график функции распределения случайной величины. Определить математическое ожидание и второй начальный момент.

30. Непрерывная случайная величина принимает значения в интервалах (1; 2) и (3; 4), причем вероятность появления значения из интервала (3; 4) в три раза больше вероятности появления значения из интервала (1; 2). Полагая, что в пределах каждого из интервалов случайная величина имеет равновероятное распределение, построить графики функции и плотности распределения. Вычислить математическое ожидание, дисперсию, второй начальный момент, среднеквадратическое отклонение, коэффициент вариации.

31. Случайная величина может принимать только два значения 10 и 90. Какова вероятность появления этих значений, если известно, что математическое ожидание случайной величины равно 80?

32. В чём заключается свойство отсутствия последействия, присущее экспоненциальному закону распределения случайных величин?

33. Какие распределения, связанные с экспоненциальным, можно использовать для аппроксимации случайных величин с коэффициентом вариации $\nu < 1$?

34. Что представляет собой однофазное гиперэкспоненциальное распределение?

35. В каком интервале находится коэффициент вариации случайной величины, имеющей однофазное гиперэкспоненциальное распределение?

Раздел 3. МАТЕМАТИЧЕСКИЕ МОДЕЛИ ДИСКРЕТНЫХ СИСТЕМ

> «Соседняя очередь всегда движется быстрее. Как только вы перейдете в другую очередь, ваша бывшая начинает двигаться быстрее» (***Наблюдение Этторе***)

Исследование сложных систем предполагает построение абстрактных математических моделей, представленных на языке математических отношений в терминах определенной математической теории, позволяющей получить функциональные зависимости характеристик исследуемой системы от параметров. Изучение процессов, протекающих в дискретных системах со стохастическим характером функционирования, проводится в рамках *теории массового обслуживания (ТМО)* и *теории случайных процессов*. При этом многие модели реальных систем строятся на основе **моделей массового обслуживания (ММО)**, которые делятся на **базовые модели** в виде *систем массового обслуживания* и **сетевые модели** в виде *сетей массового обслуживания*, представляющие собой математические объекты, описываемые в терминах соответствующего математического аппарата.

3.1. Основные понятия

Для описания одного и того же понятия многочисленные литературные источники по моделям и методам теории массового обслуживания зачастую используют разные термины. Сама «теория массового обслуживания» часто называется «теорией очередей» (в англоязычной литературе Queue Theorie), наряду с термином «обслуживающий прибор» используются термины «устройство», «канал», «линия» и т.д. Обычно это связано с прикладной областью, в которой применяются модели массового обслуживания. Например, термины «вызов» и «линия» используются в телефонии (откуда собственно и пошла теория массового обслуживания), термин «клиент» – в моделях магазинов, банков, парикмахерских и т.д. В связи с этим, желательно иметь однозначные термины и понятия, которые будут использоваться при изложении материала в последующих разделах. Рассматривая модели массового обслуживания как абстрактные математические модели, ниже вводятся и используются термины безотносительно прикладной области применения этих моделей. Для каждого термина в круглых скобках перечислены термины-синонимы, которые могут встретиться в других источниках.

3.1.1. Система массового обслуживания

Система массового обслуживания (СМО) – математический (абстрактный) объект, содержащий один или несколько *приборов* **П** (*каналов*), обслуживающих *заявки* **З**, поступающие в систему, и

накопитель **Н**, в котором находятся заявки, образующие *очередь* **О** и ожидающие обслуживания (рис.3.1).

Заявка (требование, запрос, вызов, клиент) – объект, поступающий в СМО и требующий обслуживания в обслуживающем приборе.

Совокупность заявок, распределенных во времени, образуют ***поток заявок***.

Обслуживающий прибор или просто ***прибор (устройство, канал, линия)*** – элемент СМО, функцией которого является обслуживание заявок. В каждый момент времени в приборе на обслуживании может находиться только одна заявка.

Обслуживание – задержка заявки на некоторое время в обслуживающем приборе.

Длительность обслуживания – время задержки (обслуживания) заявки в приборе.

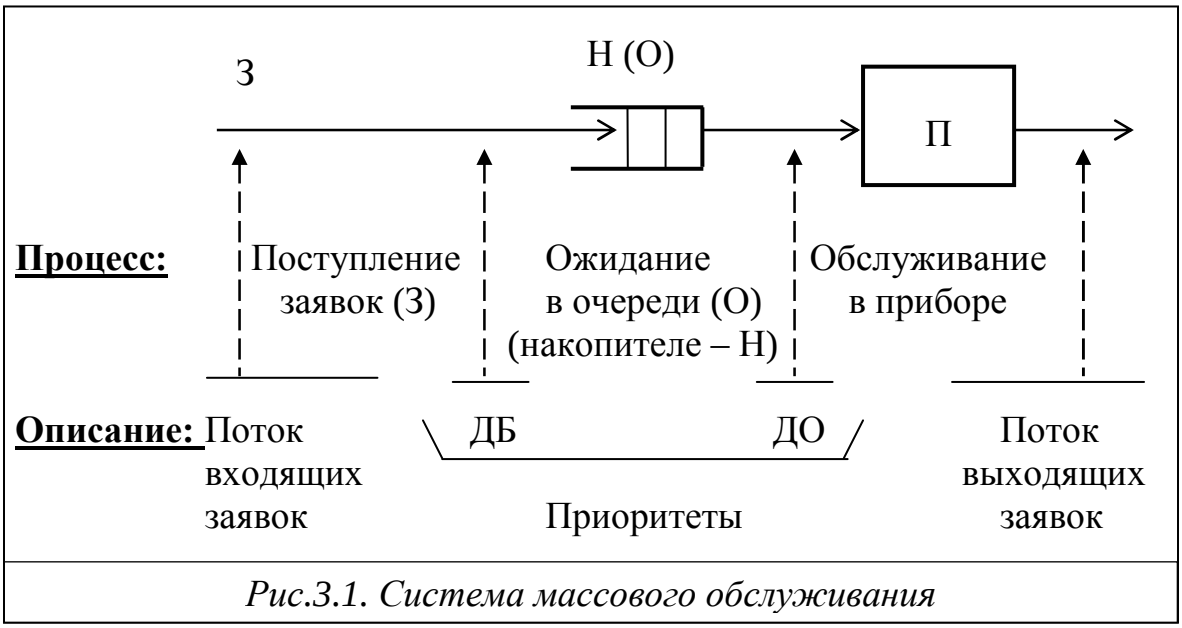


Рис.3.1. Система массового обслуживания

Накопитель (буфер) – совокупность мест для ожидания заявок перед обслуживающим прибором. Количество мест для ожидания определяет ***ёмкость накопителя***.

Заявка, поступившая на вход СМО, может находиться в двух состояниях:

- в состоянии *обслуживания* (в приборе);
- в состоянии *ожидания* (в накопителе), если все приборы заняты обслуживанием других заявок.

Заявки, находящиеся в накопителе и ожидающие обслуживания, образуют ***очередь*** заявок. Количество заявок, ожидающих обслуживания в накопителе, определяет ***длину очереди***.

Дисциплина буферизации – правило занесения поступающих заявок в накопитель (буфер).

Дисциплина обслуживания – правило выбора заявок из очереди для обслуживания в приборе.

Приоритет – преимущественное право на занесение (в накопитель) или выбор из очереди (для обслуживания в приборе) заявок одного класса по отношению к заявкам других классов.

Таким образом, СМО включает в себя:

- *заявки*, проходящие через систему и образующие *потоки заявок*;
- *очереди* заявок, образующиеся в накопителях;
- обслуживающие *приборы*.

Существует большое многообразие СМО, различающихся структурной и функциональной организацией. В то же время, разработка аналитических методов расчета характеристик функционирования СМО во многих случаях предполагает наличие ряда предположений, ограничивающих множество исследуемых СМО.

Ниже при рассмотрении СМО, если не оговорено другое, будем использовать следующие **предположения**:

- заявка, поступившая в систему, *мгновенно* попадает на обслуживание, если прибор свободен;
- в приборе на обслуживании в каждый момент времени может находиться только *одна* заявка;
- после завершения обслуживания какой-либо заявки в приборе очередная заявка выбирается на обслуживание из очереди мгновенно, то есть, другими словами, прибор *не простаивает*, если в очереди есть хотя бы одна заявка;
- поступление заявок в СМО и длительности их обслуживания не зависят от того, сколько заявок уже находится в системе, или от каких-либо других факторов;
- длительность обслуживания заявок не зависит от скорости (интенсивности) поступления заявок в систему.

3.1.2. Сеть массового обслуживания

Сеть массового обслуживания (СеМО) – совокупность взаимосвязанных СМО, в среде которых циркулируют заявки (рис.3.2,а).

Основными элементами СеМО являются узлы (У) и источники заявок (И).

Узел сети представляет собой систему массового обслуживания.

Источник – генератор заявок, поступающих в сеть и требующих определенных этапов обслуживания в узлах сети.

Для упрощенного изображения СеМО используется граф СеМО.

Граф СеМО – ориентированный граф, вершины которого соответствуют узлам СеМО, а дуги отображают переходы заявок между узлами (рис.3.2,б).

Переходы заявок между узлами СеМО, в общем случае, могут быть заданы в виде вероятностей передач.

Путь движения заявок в СеМО называется ***маршрутом***.

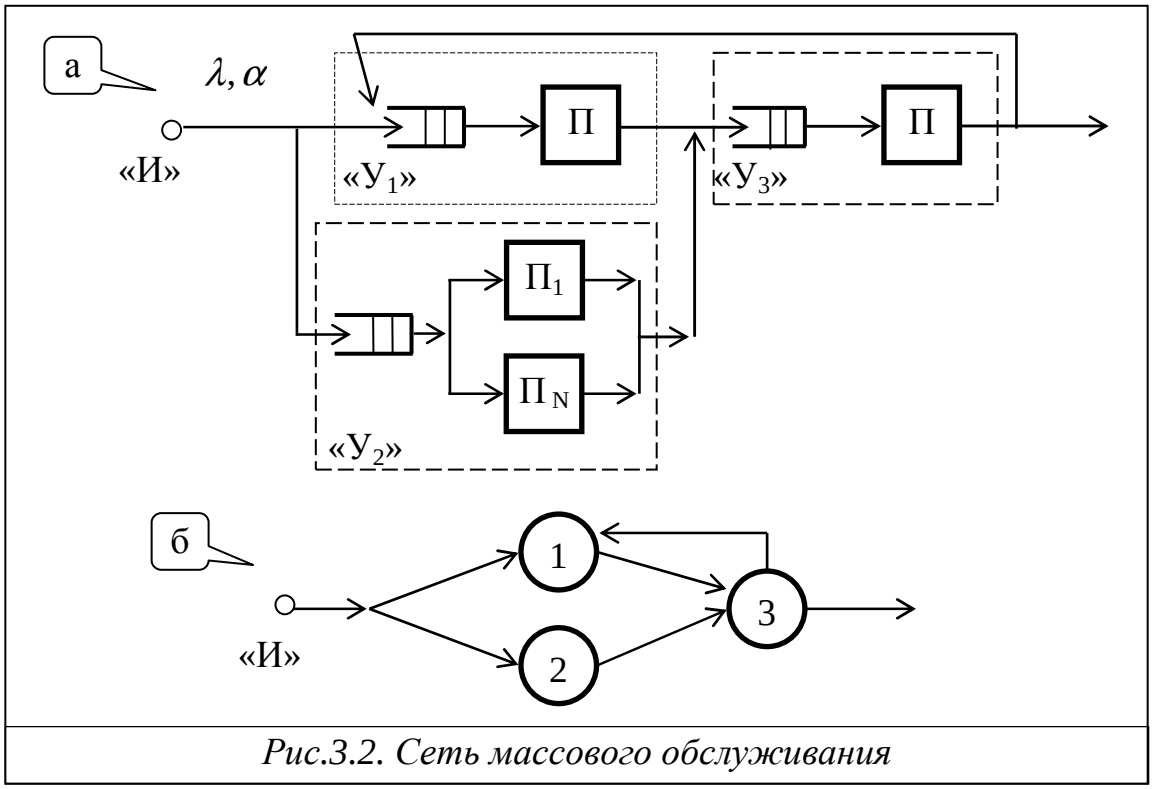


Рис.3.2. Сеть массового обслуживания

3.1.3. Поток заявок

Совокупность событий распределенных во времени называется ***потоком***. Если событие заключается в появлении заявок, имеем *поток заявок*.

Для описания потока заявок, в общем случае, необходимо задать интервалы времени $\tau_k = t_k - t_{k-1}$ между соседними моментами t_{k-1} и t_k поступления заявок с порядковыми номерами $(k-1)$ и k соответственно ($k = 1, 2, ...$; $t_0 = 0$ – начальный момент времени).

Основной характеристикой потока заявок является его ***интенсивность*** λ – среднее число заявок, проходящих через некоторую границу за единицу времени. Величина $a = 1/\lambda$ определяет *средний интервал времени между двумя последовательными заявками*.

Поток, в котором интервалы времени τ_k между соседними заявками принимают определенные заранее известные значения, называется ***детерминированным***. Если при этом интервалы одинаковы ($\tau_k = \tau$ для всех $k = 1, 2, ...$), то поток называется ***регулярным***. Для полного описания регулярного потока заявок достаточно задать интенсивность потока λ или значение интервала $\tau = 1/\lambda$.

Поток, в котором интервалы времени τ_k между соседними заявками представляют собой случайные величины, называется ***случайным***. Для полного описания случайного потока заявок, в общем случае, необходимо задать законы распределений $A_k(\tau_k)$ всех интервалов τ_k ($k = 1, 2, ...$).

Случайный поток, в котором все интервалы $\tau_1, \tau_2, \ldots$ между заявками независимы в совокупности и описываются функциями распределений $A_1(\tau_1)$, $A_2(\tau_2)$,…, называется потоком ***с ограниченным последействием***.

Случайный поток, в котором все интервалы $\tau_1, \tau_2, \ldots$ распределены *по одному и тому же закону* $A(\tau)$, называется ***рекуррентным***.

Поток заявок называется ***стационарным***, если интенсивность λ и закон распределения $A(\tau)$ интервалов между последовательными заявками не меняются со временем. В противном случае поток заявок является ***нестационарным***.

Поток заявок называется ***ординарным***, если в каждый момент времени t_k может появиться только одна заявка. Если в какой-либо момент времени может появиться более одной заявки, то имеем ***неординарный*** или ***групповой*** поток заявок.

Поток заявок называется потоком ***без последействия***, если заявки поступают *независимо* друг от друга, то есть момент поступления очередной заявки не зависит от того, когда и сколько заявок поступило до этого момента.

Стационарный ординарный поток *без последействия* называется ***простейшим***.

Интервалы времени τ между заявками в простейшем потоке распределены по *экспоненциальному закону* с функцией распределения

$$A(\tau) = 1 - e^{-\lambda\tau}, \qquad (3.1)$$

где $\lambda > 0$ – параметр распределения, представляющий собой интенсивность потока заявок.

Простейший поток часто называют ***пуассоновским***, поскольку число заявок k, поступающих за некоторый заданный промежуток времени t, распределено по *закону Пуассона*:

$$P(k,t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \qquad (3.2)$$

где $P(k,t)$ – вероятность поступления ровно k заявок за некоторый фиксированный интервал времени t; λ – интенсивность потока заявок. Здесь k – дискретная случайная величина, принимающая целочисленные значения: $k = 0, 1, 2, \ldots$, а $t > 0$ и $\lambda > 0$ – параметры закона Пуассона.

Следует отметить, что пуассоновский поток, в отличие от простейшего, может быть:

- *стационарным*, если интенсивность λ не меняется со временем;
- *нестационарным*, если интенсивность потока зависит от времени: $\lambda = \lambda(t)$.

В то же время, простейший поток, по определению, всегда является стационарным.

Аналитические исследования моделей массового обслуживания часто проводятся в предположении о простейшем потоке заявок, что обусловлено рядом присущих ему замечательных особенностей.

1. **Суммирование (объединение) потоков**. Сумма H независимых стационарных ординарных потоков с интенсивностями $\lambda_1, \ldots, \lambda_H$ образует простейший поток с интенсивностью

$$\Lambda = \sum_{k=1}^{H} \lambda_k \qquad (3.3)$$

при условии, что складываемые потоки оказывают более или менее одинаково малое влияние на суммарный поток. На практике суммарный поток близок к простейшему при $H \geq 5$. Очевидно, что *при суммировании независимых простейших потоков суммарный поток будет простейшим* при любом значении H.

2. **Вероятностное разрежение потока**. *Вероятностное* (но *не детерминированное*) разрежение *простейшего потока* заявок, при котором любая заявка случайным образом с некоторой вероятностью p исключается из потока независимо от того, исключены другие заявки или нет, приводит к образованию *простейшего потока* с интенсивностью $\lambda' = p\lambda$, где λ – интенсивность исходного потока. Поток исключенных заявок – тоже *простейший* с интенсивностью $\lambda'' = (1-p)\lambda$.

3. **Простота**. Предположение о простейшем потоке заявок позволяет для многих математических моделей сравнительно легко получить в явном виде зависимости характеристик от параметров. Наибольшее число аналитических результатов получено для простейшего потока заявок. Анализ моделей с потоками заявок, отличными от простейших, обычно усложняет математические выкладки и не всегда позволяет получить аналитическое решение в явном виде. Свое название «*простейший*» поток получил именно благодаря этой особенности.

3.1.4. Длительность обслуживания заявок

Длительность обслуживания – время нахождения заявки в приборе – в общем случае величина случайная и описывается функцией $B(\tau)$ или плотностью $b(\tau) = B'(\tau)$ распределения. В случае неоднородной нагрузки длительности обслуживания заявок разных классов могут различаться законами распределений или только средними значениями. При этом обычно предполагается независимость длительностей обслуживания заявок каждого класса.

Часто длительность обслуживания заявок предполагается распределенной по *экспоненциальному закону*, что существенно упрощает аналитические выкладки. Это обусловлено тем, что процессы, протекающие в системах с экспоненциальным распределением интервалов времени, являются *марковскими* (см. раздел 5).

Величина, обратная средней длительности обслуживания b, характеризует среднее число заявок, которое может быть обслужено за единицу времени, и называется ***интенсивностью обслуживания*****:** $\mu = 1/b$.

Во многих случаях аналитические зависимости могут быть получены для произвольного закона распределения длительности обслуживания заявок. При этом для определения средних значений характеристик обслуживания, зачастую, как будет показано ниже, достаточно задать, кроме математического ожидания b, второй момент распределения (дисперсию) или коэффициент вариации ν_b длительности обслуживания.

Время T_0, оставшееся до завершения обслуживания заявки, находящейся в приборе, от момента поступления некоторой заявки в систему, и учитывающее, что на момент поступления в системе может и не оказаться заявок, то есть учитывающее простои системы, называется ***временем дообслуживания***. Математическое ожидание этого времени [9]:

$$M[T_0] = \lambda b^2 (1+\nu_b^2)/2, \qquad (3.4)$$

где λ – интенсивность *простейшего* потока заявок, поступающих в систему.

3.1.5. Стратегии управления потоками заявок

Стратегия управления потоками заявок в моделях массового обслуживания задается в виде:

- *дисциплины буферизации (ДБ)*;
- *дисциплины обслуживания (ДО)*.

ДБ и ДО могут быть классифицированы по следующим признакам:

- наличие приоритетов между заявками разных классов;
- способ (режим) вытеснения заявок из очереди (для ДБ) и назначения заявок на обслуживание (для ДО);
- правило вытеснения или выбора заявок на обслуживание;
- возможность изменения приоритетов.

Одна из возможных **классификаций дисциплин буферизации** в соответствии с перечисленными признаками представлена на рис.3.3.

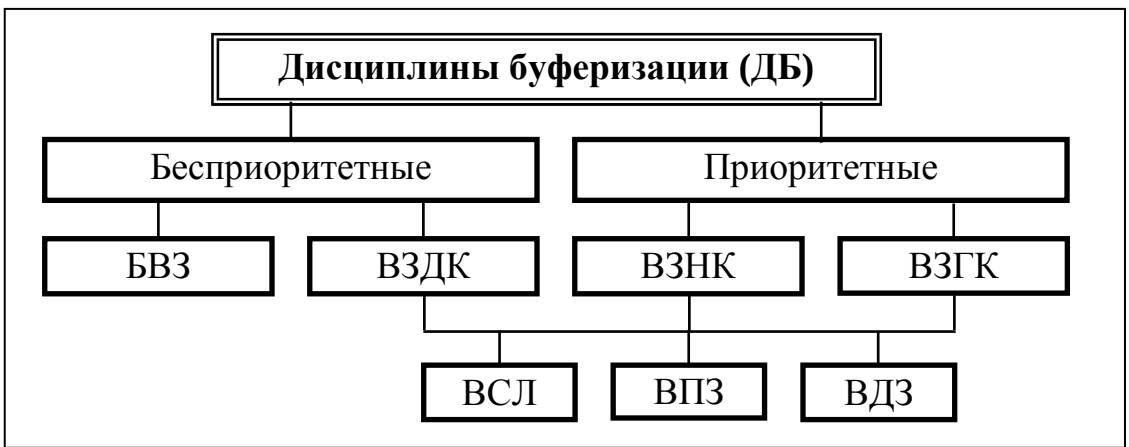


Рис.3.3. Классификация дисциплин буферизации

В зависимости от *наличия или отсутствия приоритетов* между заявками разных классов все ДБ могут быть разбиты на две группы:

- бесприоритетные;
- приоритетные.

По *способу вытеснения заявок из накопителя* можно выделить следующие классы ДБ:

- без вытеснения заявок (БВЗ) – заявки, поступившие в систему и заставшие накопитель заполненным до конца, теряются;
- с вытеснением заявки данного класса (ВЗДК), то есть такого же класса, что и поступившая;
- с вытеснением заявки самого низкоприоритетного класса (ВЗНК);
- с вытеснением заявки, принадлежащей группе низкоприоритетных классов (ВЗГК).

Два первых класса относятся к бесприоритетным ДБ, а остальные – к приоритетным.

ДБ могут использовать следующие *правила вытеснения заявок из накопителя*:

- вытеснение случайное (ВСЛ);
- вытеснение последней заявки (ВПЗ), то есть поступившей в систему позже всех;
- вытеснение «долгой» заявки (ВДЗ), то есть находящейся в накопителе дольше всех.

Часто ёмкость накопителя в моделях предполагается неограниченной, несмотря на то, что в реальной системе соответствующая ёмкость ограничена. Такое предположение оправдано в тех случаях, когда вероятность потери заявки в реальной системе из-за переполнения ограниченной ёмкости накопителя меньше 10^{-3}, поскольку в этом случае ДБ практически не влияет на характеристики обслуживания заявок.

На рис.3.4 представлена **классификация дисциплин обслуживания** заявок в соответствии с теми же признаками, что и для ДБ.

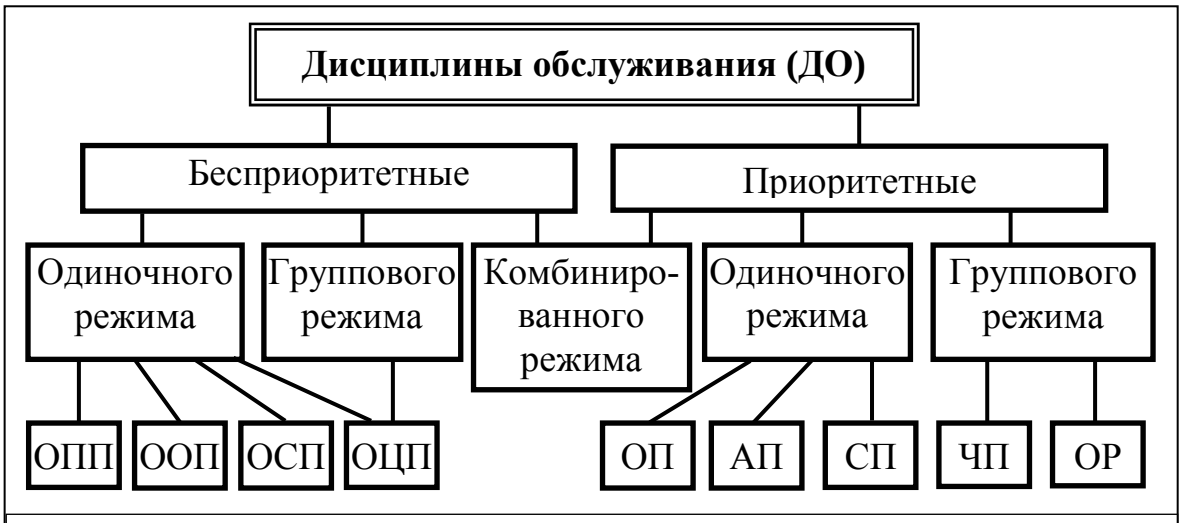


Рис.3.4. Классификация дисциплин обслуживания

В зависимости от *наличия или отсутствия приоритетов* между заявками разных классов все ДО, как и ДБ, могут быть разбиты на две группы:

- бесприоритетные;
- приоритетные.

По *способу назначения заявок на обслуживание* ДО могут быть разделены на дисциплины:

- одиночного режима;
- группового режима;
- комбинированного режима.

В ДО **одиночного режима** всякий раз на обслуживание *назначается только одна заявка* (просмотр очередей с целью назначения на обслуживание в приборе очередной заявки выполняется после обслуживания каждой заявки).

В ДО **группового режима** всякий раз на обслуживание *назначается группа заявок* одной очереди (просмотр очередей с целью очередного назначения на обслуживание выполняется только после обслуживания всех заявок ранее назначенной группы). В предельном случае назначаемая на обслуживание группа заявок может включать в себя все заявки данной очереди. Заявки назначенной на обслуживание группы *последовательно выбираются из очереди* и обслуживаются прибором, после чего на обслуживание назначается следующая группа заявок другой очереди в соответствии с заданной ДО.

Комбинированный режим – комбинация одиночного и группового режимов, когда часть очередей заявок обрабатывается в одиночном режиме, а другая часть – в групповом.

ДО могут использовать следующие *правила выбора заявок на обслуживание*:

- бесприоритетные:
  - **обслуживание в порядке поступления** (ОПП или FIFO – First In First Out), когда на обслуживание выбирается заявка, поступившая в систему раньше других;
  - **обслуживание в обратном порядке** (ООП или LIFO – Last In First Out) когда на обслуживание выбирается заявка, поступившая в систему позже других;
  - **обслуживание в случайном порядке** (ОСП), когда на обслуживание заявка выбирается случайным образом;
  - **обслуживание в циклическом порядке** (ОЦП), когда на обслуживание заявки выбираются в процессе циклического опроса накопителей в последовательности 1, 2, ..., H (H – количество накопителей), после чего указанная последовательность повторяется;
- приоритетные:

- с **относительными приоритетами** (ОП), означающими, что приоритеты учитываются только в моменты завершения обслуживания заявок при выборе новой заявки на обслуживание и не влияют на процесс обслуживания низкоприоритетной заявки в приборе; другими словами, поступление в систему заявки с более высоким приоритетом по сравнению с обслуживаемой в приборе не приводит к прерыванию обслуживаемой заявки;
- с **абсолютными приоритетами** (АП), означающими, что, в отличие от ОП, при поступлении высокоприоритетной заявки обслуживание заявки с низким приоритетом прерывается и на обслуживание принимается поступившая высокоприоритетная заявка; при этом прерванная заявка может быть возвращена в накопитель или удалена из системы; если заявка возвращена в накопитель, то её дальнейшее обслуживание может быть продолжено с прерванного места или начато заново, то есть с самого начала;
- со **смешанными приоритетами** (СП), представляющими собой любую комбинацию бесприоритетного обслуживания, ОП и АП;
- с **чередующимися приоритетами** (ЧП), являющимися аналогом ОП и проявляющимися только в моменты завершения обслуживания группы заявок одной очереди и назначения новой группы;
- **обслуживание по расписанию** (ОР), когда заявки разных классов (находящиеся в разных накопителях) выбираются на обслуживание в соответствии с некоторым расписанием (планом), задающим последовательность опроса очередей заявок, например, в случае трех классов заявок (накопителей) расписание может иметь вид: {1, 2, 1, 3, 1, 2}.

Дисциплины ОПП, ООП, ОП, АП и СП относятся к дисциплинам *одиночного режима*. Очевидно, что дисциплины *группового режима* ОЦП, ЧП и ОР, в частном случае могут быть реализованы как ДО одиночного режима, если размер назначаемой на обслуживание группы равен 1, при этом ДО ЧП вырождается в ДО ОП.

Среди представленных ДО особое место занимают дисциплины со смешанными приоритетами (СП), обладающие общностью по отношению к перечисленным ДО одиночного режима [3].

Для математического описания ДО СП используется **матрица приоритетов (МП)**, представляющая собой квадратную матрицу: $\mathbf{Q} = [q_{ij} \mid i, j = 1, \ldots, H]$, где H – число классов заявок, поступающих в систему.

Элемент q_{ij} матрицы задает приоритет заявок класса i по отношению к заявкам класса j и может принимать следующие значения:

- 0 – нет приоритета;
- 1 – приоритет относительный (ОП);
- 2 – приоритет абсолютный (АП).

Элементы МП должны удовлетворять следующим *требованиям*:

- $q_{ii} = 0$, так как между заявками одного и того же класса не могут быть установлены приоритеты;
- если $q_{ij} = 1$ или 2, то $q_{ji} = 0$, так как если заявки класса i имеют приоритет к заявкам класса j, то последние не могут иметь приоритет к заявкам класса i $(i, j = \overline{1, H})$.

В зависимости от *возможности изменения приоритетов* в процессе функционирования системы приоритетные дисциплины буферизации и обслуживания делятся на два класса:

- ***со статическим приоритетами***, которые не изменяются со временем;
- ***с динамическими приоритетами***, которые могут изменяться в процессе функционирования системы в зависимости от разных факторов, например, при достижении некоторого критического значения длины очереди заявок какого-либо класса, обладающего низким приоритетом, ему может быть предоставлен более высокий приоритет.

3.2. Классификация моделей массового обслуживания

3.2.1. Базовые модели

При моделировании реальных систем с дискретным характером функционирования широкое применение находят базовые модели в виде СМО, которые могут быть классифицированы (рис.3.5):

- по числу мест в накопителе;
- по числу обслуживающих приборов;
- по количеству классов заявок, поступающих в СМО.

1. По *числу мест в накопителе* СМО делятся на системы:

- **без накопителя**, в которых заявка, поступившая в систему и заставшая все обслуживающие приборы занятыми обслуживанием более высокоприоритетных заявок, получает отказ и теряется; такие системы называются ***СМО с отказами***;
- **с накопителем ограниченной ёмкости (*СМО с потерями*)**, в которых поступившая заявка теряется, если она застает накопитель заполненным до конца;
- **системы с накопителем неограниченной ёмкости (*СМО без потерь*)**, в которых для любой поступившей заявки всегда найдется место в накопителе для ожидания.

В дальнейшем, накопитель неограниченной ёмкости будем изображать так, как это показано на рис.3.5,а, и накопитель ограниченной ёмкости – как на рис.3.5,б.

Как уже было сказано выше, предположение о неограниченной ёмкости накопителя может использоваться для моделирования реальных систем, в которых вероятность потери заявки из-за переполнения накопителя ограниченной ёмкости меньше 10^{-3}.

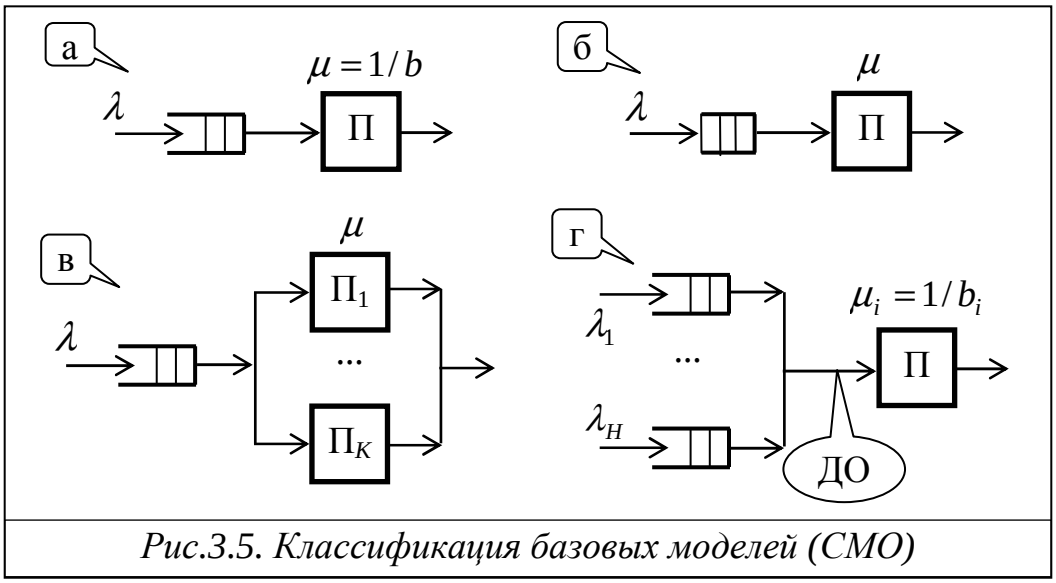


Рис.3.5. Классификация базовых моделей (СМО)

2. По *количеству обслуживающих приборов* СМО делятся на:

- **одноканальные** (рис.3.5,а, б, г), содержащие один прибор **П**;
- **многоканальные** (рис.3.5,в), содержащие K обслуживающих приборов $\mathbf{П}_1,...,\mathbf{П}_K$ ($K > 1$).

В многоканальных СМО обычно предполагается, что все приборы идентичны и равнодоступны для любой заявки, то есть при наличии нескольких свободных приборов поступившая заявка с равной вероятностью может попасть в любой из них на обслуживание.

3. По *количеству классов (типов) заявок*, поступающих в СМО, различают системы:

- **с однородным потоком** заявок (рис.3.5,а, б, в);
- **с неоднородным потоком** заявок (рис.3.5,г).

Однородный поток заявок образуют заявки одного класса, а неоднородный поток представляет собой поток заявок нескольких классов.

В СМО, представляющей собой абстрактную математическую модель, *заявки относятся к разным классам* в том случае, если они в моделируемой реальной системе различаются хотя бы одним из следующих факторов:

- *длительностью обслуживания*;
- *приоритетами*.

Если же заявки не различаются длительностью обслуживания и приоритетами, то в СМО они могут быть представлены как заявки одного класса, независимо от их физической сущности.

3.2.2. Сетевые модели

В зависимости от структуры и свойств исследуемых систем их моделями могут служить СеМО различных классов. Одна из возможных классификаций сетевых моделей приведена на рис.3.6.

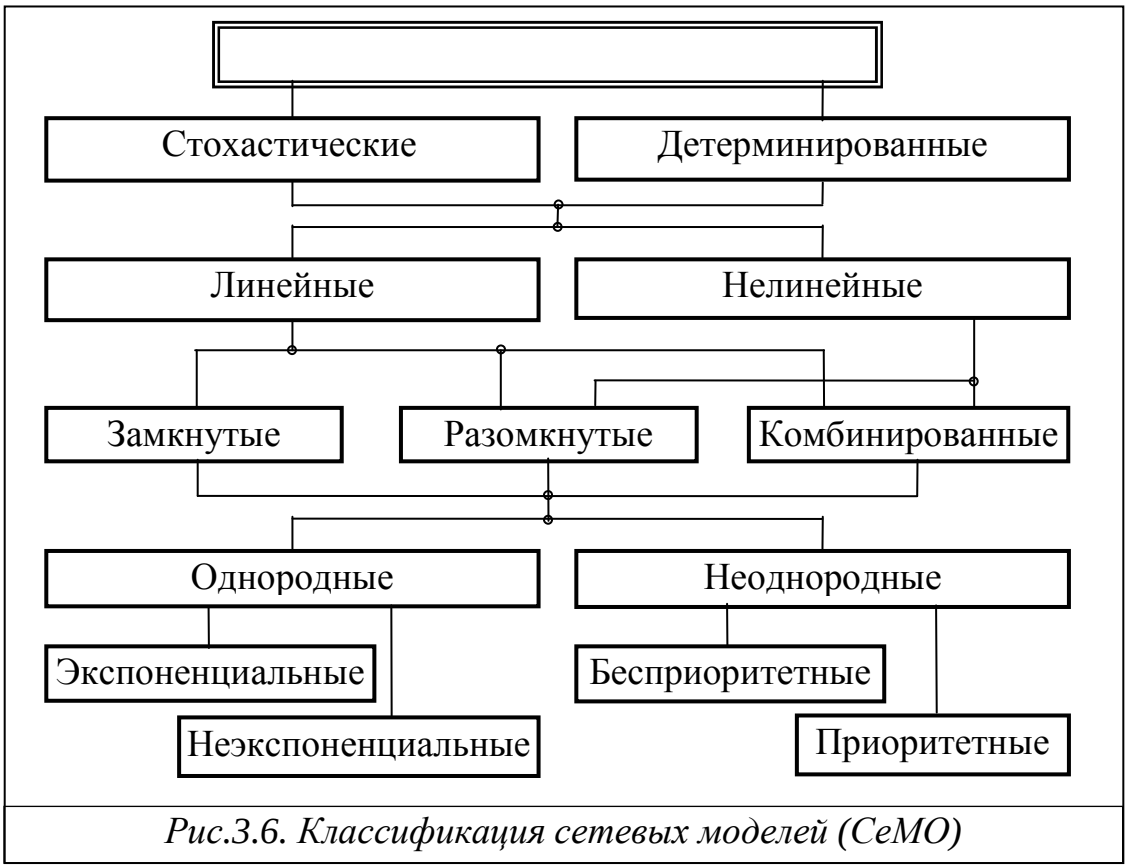


Рис.3.6. Классификация сетевых моделей (СеМО)

1. В зависимости от *характера процессов поступления и обслуживания заявок* в сети СеМО делятся на:

- **стохастические**, в которых процессы поступления и/или обслуживания заявок носят случайный характер, то есть интервалы времени между поступающими заявками и/или длительности их обслуживания в узлах представляют собой случайные величины, описываемые соответствующими законами распределений;
- **детерминированные**, в которых интервалы времени между поступающими заявками и длительности их обслуживания в узлах являются детерминированными величинами.

2. По *виду зависимостей, связывающих интенсивности потоков заявок в разных узлах*, СеМО делятся на:

- **линейные**, если эти зависимости линейные;
- **нелинейные**, если эти зависимости являются нелинейными.

В *линейных* СеМО, как это следует из определения, интенсивность потока заявок в узел j связана с интенсивностью потока заявок в узел i линейной зависимостью:

$$\lambda_j = \alpha_{ij}\lambda_i,$$

где α_{ij} – коэффициент пропорциональности, показывающий, во сколько раз отличаются интенсивности потоков заявок в узел j и в узел i $(i, j = \overline{1,n})$.

Поскольку указанная зависимость справедлива для любой пары узлов, это выражение можно записать в несколько ином виде и выразить интенсивность поступления заявок во все узлы $j = \overline{1,n}$ через одну и ту же интенсивность, например, через интенсивность λ_0 потока заявок, поступающих в СеМО из источника заявок:

$$\lambda_j = \alpha_j \lambda_0. \qquad (3.5)$$

В последнем выражении коэффициент пропорциональности $\alpha_j \geq 0$ показывает, во сколько раз интенсивность потока заявок в узел j $(i, j = \overline{1,n})$ отличается от интенсивности источника заявок, и называется ***коэффициентом передачи***. Коэффициент передачи может принимать любое положительное значение.

Коэффициент передачи играет важную роль при разработке математических зависимостей и расчете характеристик функционирования сетевых моделей. Это обусловлено тем физическим смыслом, который несет в себе коэффициент передачи.

Коэффициент передачи можно трактовать как *среднее число попаданий заявки в данный узел за время ее нахождения в сети*. Например, если коэффициент передачи узла СеМО равен 3, то это означает, что любая заявка за время нахождения в сети *в среднем* 3 раза побывает на обслуживании в данном узле. Значение коэффициента передачи, равное 0,25, будет означать, что *в среднем* только одна заявка из четырёх попадёт на обслуживание в данный узел, а три другие обойдут данный узел стороной.

В *нелинейных* СеМО интенсивности потоков заявок в узлах связаны более сложными нелинейными зависимостями, что значительно усложняет их исследование.

Нелинейность СеМО может быть обусловлена:

- *потерей заявок* в сети, например из-за ограниченной емкости накопителей в узлах;
- *размножением заявок* в сети, заключающимся, например, в формировании нескольких новых заявок после завершения обслуживания некоторой заявки в одном из узлов сети.

Таким образом, СеМО является линейной, если в ней заявки не размножаются и не теряются. Ниже рассматриваются, в основном, линейные СеМО.

3. По *числу циркулирующих в сети заявок* различают СеМО:

- разомкнутые;

- замкнутые;
- замкнуто-разомкнутые.

Разомкнутая (открытая) СеМО (РСеМО) содержит один или несколько *внешних независимых источников* заявок, которые генерируют заявки в сеть независимо от числа заявок, находящихся в сети (рис.3.7,а). В РСеМО одновременно может находиться *любое число заявок*, в том числе, и сколь угодно большое, то есть от 0 до бесконечности. С РСеМО связана внешняя среда, из которой поступают заявки в сеть и в которую они возвращаются после обслуживания в сети. Внешняя среда в РСеМО обозначается обычно как нулевой узел "0", и РСеМО, в этом случае, изображается в виде рис.3.7,б.

Замкнутая (закрытая) СеМО (ЗСеМО) не содержит *независимых внешних* источников заявок и характеризуется тем, что в ней циркулирует *постоянное число заявок M* (рис.3.7,в). На графе ЗСеМО из физических соображений, связанных с конкретным представлением процесса функционирования исследуемой реальной системы, обычно выделяется особая дуга, отображающая процесс завершения обслуживания заявок в сети и мгновенного формирования новой заявки с такими же параметрами обслуживания, что и завершившая обслуживание. Такая трактовка позволяет рассматривать завершившую обслуживание заявку как новую заявку, поступившую в сеть из *зависимого источника* заявок.

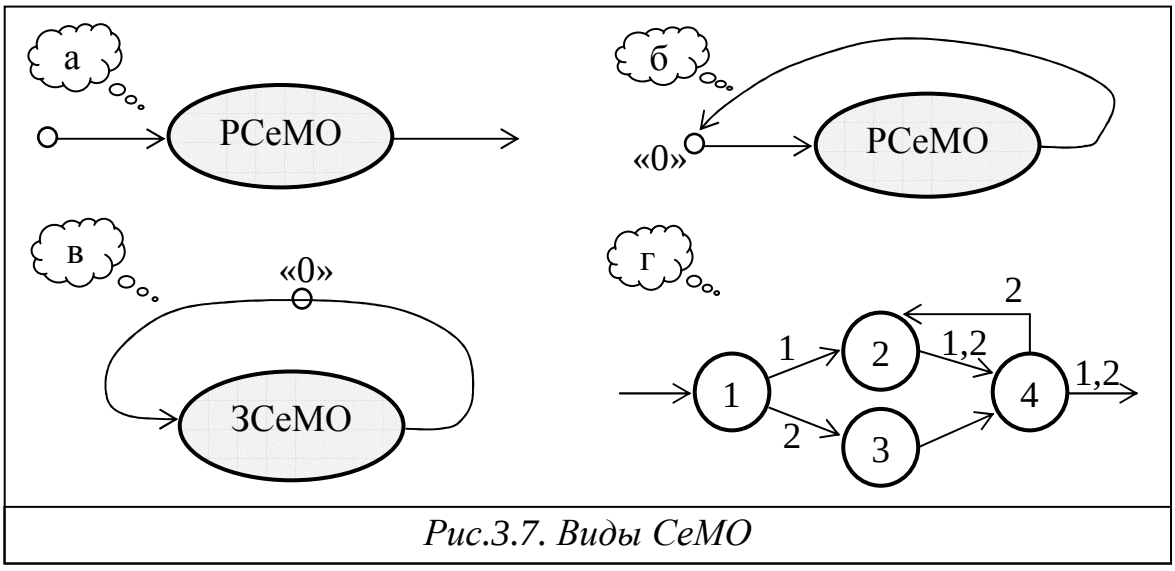


Рис.3.7. Виды СеМО

По аналогии с РСеМО на выделенной дуге ЗСеМО отмечается условная точка "0", рассматриваемая как нулевой узел и трактуемая иногда как фиктивная СМО с нулевой длительностью обслуживания или как зависимый источник заявок, генерирующий заявки только в момент поступления некоторой заявки на его вход. Выделение нулевого узла в ЗСеМО преследует двоякую цель: во-первых, достигается однозначность в представлении и математическом описании РСеМО и ЗСеМО; во-вторых, обеспечивается возможность определения временны*х* характеристик ЗСеМО относительно выделенного узла "0". В частности, ***время пребыва-***

ния заявок в ЗСеМО рассматривается как промежуток времени между двумя соседними моментами прохождения заявки через нулевой узел.

Замкнуто-разомкнутая СеМО (комбинированная) представляет собой комбинацию ЗСеМО и РСеМО, в которую, кроме постоянно циркулирующих в сети M^* заявок, из внешнего независимого источника поступают заявки такого же или другого класса, при этом суммарное число заявок в сети $M \geq M^*$.

4. По *типу циркулирующих заявок* различают СеМО:

- **однородные**, в которых циркулирует один класс заявок (однородный поток заявок);
- **неоднородные**, в которых циркулирует несколько классов заявок (неоднородный поток заявок), различающихся хотя бы одним из следующих факторов:
  - *длительностями обслуживания* в узлах;
  - *приоритетами*;
  - *маршрутами*.

Маршруты заявок разных классов задаются путем указания номеров классов заявок на соответствующих дугах сети (рис.3.7,г).

3.3. Параметры и характеристики СМО

«Чем больше ожидание, тем больше вероятность, что вы стоите не в той очереди» (***Принцип очереди***)

3.3.1. Параметры СМО

Для описания СМО используются три группы параметров:

- структурные;
- нагрузочные;
- функциональные параметры (параметры управления).

К *структурным параметрам* относятся:

- *количество обслуживающих приборов K*, равное 1 для одноканальной СМО и *K* >1 для многоканальной СМО;
- *количество k и ёмкости накопителей* E_j $(j = \overline{1,k})$;
- *способ взаимосвязи накопителей с приборами* (в случае многоканальных СМО), например в виде матрицы связей.

Нагрузочные параметры СМО включают в себя:

- количество поступающих в систему классов заявок *H*, которое равно 1 для СМО с однородным потоком заявок и *H* >1 для СМО с неоднородным потоком;
- закон распределения $A_i(\tau)$ интервалов времени между поступающими в систему заявками класса $i = \overline{1,H}$ или, по-крайней мере, первые два момента распределения, задаваемые, например, в виде интенсивности λ_i и коэффициента вариации ν_{a_i} интервалов;

- закон распределения $B_i(\tau)$ длительности обслуживания заявок класса $i = \overline{1, H}$ или, как минимум, первые два момента распределения, в качестве которых обычно используются средняя длительность b_i или интенсивность $\mu_i = 1/b_i$ обслуживания и коэффициент вариации ν_{b_i}.

Задание двух первых моментов нагрузочных параметров зачастую оказывается достаточным для оценки характеристик обслуживания заявок на уровне средних значений. Отметим, что *для описания простейшего потока достаточно задать только интенсивность поступления заявок* в систему.

Функциональные параметры задаются в виде конкретных стратегий управления потоками заявок в СМО, определяющих правило занесения заявок разных классов в накопители ограниченной ёмкости (дисциплина буферизации) и правило выбора их из очереди на обслуживание (дисциплина обслуживания).

3.3.2. Обозначения СМО (символика Кендалла)

Для компактного описания систем массового обслуживания часто используются обозначения, предложенные Д. Кендаллом [9], в виде:

A/B/N/L ,

где **A** и **B** – задают законы распределений соответственно интервалов времени между моментами поступления заявок в систему и длительности обслуживания заявок в приборе; **N** – число обслуживающих приборов в системе $(N = 1, 2, ..., \infty)$; **L** – число мест в накопителе, которое может принимать значения 0, 1, 2, … (отсутствие **L** означает, что накопитель имеет неограниченную ёмкость).

Для задания законов распределений **A** и **B** используются следующие обозначения:

G (General) – произвольное распределение общего вида;

M (Markovian) – экспоненциальное (показательное) распределение;

D (Deterministik) – детерминированное распределение;

U (Uniform) – равномерное распределение;

$\mathbf{E_k}$ (Erlangian) – распределение Эрланга *k*-го порядка (с *k* последовательными одинаковыми экспоненциальными фазами);

$\mathbf{h_k}$ (hipoexponential) – гипоэкспоненциальное распределение *k*-го порядка (с *k* последовательными разными экспоненциальными фазами);

$\mathbf{H_r}$ (Hiperexponential) – гиперэкпоненциальное распределение порядка *r* (с *r* параллельными экспоненциальными фазами);

g (gamma) – гамма-распределение;

P (Pareto) – распределение Парето и т.д.

Примеры:

M/M/1 – одноканальная СМО с накопителем неограниченной ёмкости, в которую поступает однородный поток заявок с экспоненциальным

распределением интервалов времени между последовательными заявками (простейший поток) и экспоненциальной длительностью обслуживания заявок в приборе.

M/G/3/10 – трёхканальная СМО с накопителем ограниченной ёмкости, равной 10, в которую поступает однородный поток заявок с экспоненциальным распределением интервалов времени между последовательными заявками (простейший поток) и длительностью обслуживания заявок, распределённой по закону общего вида.

D/E_2/7/0 – семиканальная СМО без накопителя (ёмкость накопителя равна 0), в которую поступает однородный поток заявок с детерминированными интервалами времени между последовательными заявками (детерминированный поток) и длительностью обслуживания заявок в приборе, распределённой по закону Эрланга 2-го порядка.

Для обозначения более сложных СМО дополнительно могут использоваться обозначения, описывающие неоднородный поток заявок и приоритеты между заявками разных классов.

3.3.3. Режимы функционирования СМО

СМО может работать в следующих режимах:

- **установившемся** или **стационарном**, когда вероятностные характеристики системы не изменяются со временем;
- **неустановившемся**, когда характеристики системы изменяются со временем, что может быть обусловлено:
  - *началом работы системы*, когда значения характеристик функционирования, меняясь со временем, стремятся в пределе к стационарным значениям (**переходной режим**);
  - *нестационарным характером* потока заявок и обслуживания в приборе (**нестационарный режим**).

Кроме этого, в некоторых системах, например в *СМО с накопителем неограниченной ёмкости*, неустановившийся режим функционирования может быть обусловлен *перегрузкой системы*, когда интенсивность поступления заявок превышает интенсивность обслуживания, и система не справляется с возлагаемой на нее нагрузкой (**режим перегрузки**). При этом характеристики функционирования СМО с течением времени растут неограниченно. В частности, длина очереди перед прибором с течением времени становится всё больше и в пределе стремится к бесконечности.

Обычно исследование СМО с накопителем неограниченной ёмкости проводится в предположении о существовании установившегося режима, непременным условием которого является требование отсутствия перегрузок, для чего необходимо, чтобы интенсивность поступления заявок была меньше, чем интенсивность обслуживания. Это требование записывается для одноканальных СМО в виде условия:

$\lambda < \mu$ или $\lambda b < 1$.

Для многоканальных СМО аналогичное условие имеет вид:

$$\lambda < K\mu \quad \text{или} \quad \frac{\lambda b}{K} < 1,$$

где K – число обслуживающих приборов, а значение $K\mu$ представляет собой суммарную интенсивность обслуживания заявок в K-канальной СМО

В СМО *с накопителем ограниченной ёмкости* превышение интенсивности поступления заявок над суммарной интенсивностью обслуживания не приводит к неограниченному росту длины очереди, что обусловлено потерей заявок. Следовательно, в СМО с накопителем ограниченной ёмкости перегрузки не приводят к работе системы в неустановившемся режиме, а приводят лишь к росту числа потерянных заявок. При этом потеря части поступающих в систему заявок при наличии накопителя ограниченной ёмкости может рассматриваться как один из механизмов борьбы с перегрузками.

3.3.4. Характеристики СМО с однородным потоком заявок

Характеристики систем со стохастическим характером функционирования являются *случайными величинами* и полностью описываются соответствующими законами распределений. На практике при моделировании часто ограничиваются определением только *средних значений* (математических ожиданий), реже – определением двух первых моментов этих характеристик.

В качестве основных характеристик СМО с однородным потоком заявок используются следующие величины:

- ***нагрузка*** системы:

$$\boxed{y = \lambda / \mu = \lambda b}; \qquad (3.6)$$

- ***коэффициент загрузки*** или просто ***загрузка*** системы, определяемая как доля времени, в течение которого система (в случае одноканальной СМО – прибор) работает, то есть выполняет обслуживание заявок; загрузка может быть рассчитана как отношение *среднего* времени T_p работы одного прибора многоканальной СМО, к общему времени наблюдения T:

$$\rho = \lim_{T \to \infty} \frac{T_p}{T}; \qquad (3.7)$$

время T_p для СМО с K обслуживающими приборами определяется путём усреднения времени работы по всем приборам:

$$T_p = \frac{1}{K} \sum_{i=1}^{K} T_i ,$$

где T_i - время работы прибора $i = \overline{1, K}$;

подставляя последнее выражение в (3.7) окончательно получим:

$$\rho = \lim_{T\to\infty} \frac{1}{KT} \sum_{i=1}^{K} T_i ;$$

очевидно, что $0 \le \rho \le 1$;

- ***коэффициент простоя*** системы:

$$\eta = 1 - \rho ; \qquad (3.8)$$

- ***вероятность потери*** заявок:

$$\pi_n = \lim_{T\to\infty} \frac{N_n(T)}{N(T)}, \qquad (3.9)$$

где T – время работы системы (наблюдения за системой); $N(T)$ – число заявок, поступивших в систему за время T; $N_n(T)$ – число потерянных заявок за время T;

- **вероятность обслуживания** заявки, то есть вероятность того, что поступившая в систему заявка будет обслужена:

$$\pi_0 = (1 - \pi_n) = \lim_{T\to\infty} \frac{N_0(T)}{N(T)}, \qquad (3.10)$$

где $N_0(T)$ – число обслуженных в системе заявок за время T, причем $N_n(T) + N_0(T) = N(T)$ и $\pi_0 + \pi_n = 1$;

- ***производительность*** системы, представляющая собой *интенсивность потока обслуженных заявок*, выходящих из системы:

$$\lambda' = \pi_0 \lambda = (1 - \pi_n)\lambda ; \qquad (3.11)$$

для *СМО с накопителем неограниченной ёмкости*, при условии отсутствия перегрузок, вероятность потери заявок $\pi_n = 0$ и, следовательно, производительность системы совпадает с интенсивностью поступления заявок в систему: $\lambda' = \lambda$;

- **интенсивность потока потерянных** (не обслуженных) **заявок** из-за ограниченной ёмкости накопителя:

$$\lambda'' = \pi_n \lambda = (1 - \pi_0)\lambda ; \qquad (3.12)$$

очевидно, что сумма интенсивностей потоков обслуженных и потерянных заявок должна быть равна интенсивности входящего в систему потока заявок: $\lambda' + \lambda'' = \lambda$;

- **среднее время ожидания** заявок в очереди: ***w***;
- **среднее время пребывания** заявок в системе, складывающееся из времени ожидания *w* и времени обслуживания *b*:

$$\boxed{u = w + b}; \qquad (3.13)$$

- **средняя длина очереди** заявок:

$$\boxed{l = \lambda' w}; \qquad (3.14)$$

- **среднее число заявок в системе** (в очереди и на обслуживании в *приборе*):

$$\boxed{m = \lambda' u}. \qquad (3.15)$$

Нагрузка и загрузка являются важнейшими характеристиками СМО, определяющими качество функционирования системы.

Нагрузка $y = \lambda b$ представляет собой интегральную оценку, объединяющую два нагрузочных параметра: частоту использования некоторого ресурса (прибора СМО), задаваемую в виде интенсивности λ поступления заявок в СМО, и время использования этого ресурса, задаваемое в виде средней длительности b обслуживания заявок в СМО. Нагрузка показывает количество работы, которую необходимо выполнить в системе. Если значение нагрузки $y < 1$, то заданная нагрузка может быть выполнена одним обслуживающим прибором, то есть одноканальная СМО будет работать без перегрузки. Если $y > 1$, то реализация заданной нагрузки в одноканальной СМО приведет к режиму перегрузки, означающему, что с течением времени всё большее число заявок будет оставаться не обслуженным, и в случае накопителя неограниченной емкости очередь заявок будет расти неограниченно. Для того чтобы система работала без перегрузок необходимо использовать многоканальную СМО, количество приборов которой должно быть больше, чем значение нагрузки: $K > y$.

В общем случае для любой СМО (с накопителем ограниченной и неограниченной ёмкости) загрузка системы может быть рассчитана через нагрузку следующим образом:

$$\boxed{\rho = \min\left(\frac{(1-\pi_n)y}{K}; 1\right)}, \qquad (3.16)$$

где K – число обслуживающих приборов в СМО; π_n – вероятность потери заявок.

Последнее выражение можно трактовать следующим образом: $\rho = \frac{(1-\pi_n)y}{K}$, если СМО работает без перегрузки, и $\rho = 1$, если СМО перегружена.

Покажем, что выражение (3.16) соответствует определению (3.7).

Рассмотрим достаточно большой промежуток времени $T \to \infty$, в течение которого работает СМО. За это время в систему поступит в среднем λT заявок, где λ – интенсивность поступления заявок в СМО, из которых будут обслужены системой $(1-\pi_n)\lambda T$ заявок ($\pi_n \lambda T$ заявок будут потеряны из-за ограниченной ёмкости накопителя). Обслуживание всех этих заявок будет длиться в течение времени $T_p = (1-\pi_n)\lambda T b$, если СМО – одноканальная, и в течение времени $T_p = \frac{(1-\pi_n)\lambda T b}{K}$, если СМО – многоканальная и содержит K обслуживающих приборов. Здесь b – средняя длительность обслуживания заявки в приборе.

Подставляя выражение для T_p в (3.7), получим:

$$\rho = \lim_{T\to\infty}\frac{T_p}{T} = \lim_{T\to\infty}\frac{(1-\pi_n)\lambda T b}{KT} = \frac{(1-\pi_n)\lambda b}{K} = \frac{\lambda' b}{K}, \qquad (3.17)$$

где $\lambda' = (1-\pi_n)\lambda$ – интенсивность обслуженных в СМО заявок.

Отметим, что загрузка системы, в отличие от нагрузки, определяется через интенсивность только *обслуженных* заявок, поскольку потерянные заявки не обслуживаются в приборах и, следовательно, не загружают систему.

Рассмотрим теперь СМО с накопителем неограниченной ёмкости и вспомним, что при возникновении перегрузок такая система не справляется с работой, что выражается в неограниченном росте очереди с течением времени.

Если $T_p < T$, то это означает, что система справляется с работой, то есть работает без перегрузок.

Если же время $T_p = \frac{\lambda T b}{K}$, которое требуется для обслуживания всех заявок, окажется больше, чем время наблюдения за системой $T_p > T$, то это означает, что система не справляется с нагрузкой, то есть работает в режиме перегрузки. В этом случае загрузка системы $\rho = 1$ (составляет 100%), а коэффициент простоя соответственно равен нулю.

Выражение (3.16) записано с учётом указанного обстоятельства.

Получим ещё одну полезную формулу для расчёта вероятности потери заявок по известному значению загрузки СМО.

Из (3.11) следует, что вероятность потери заявок в СМО с накопителем ограниченной ёмкости может быть рассчитана как

$$\pi_n = \frac{\lambda - \lambda'}{\lambda} = 1 - \frac{\lambda'}{\lambda}.$$

В то же время из (3.17) вытекает, что интенсивность обслуженных заявок

$$\lambda' = \frac{\rho K}{b}.$$

Подставляя последнее выражение в предыдущее, получим:

$$\boxed{\pi_n = 1 - \frac{\rho K}{\lambda b} = 1 - \frac{\rho}{y}K}, \qquad (3.18)$$

где $y = \lambda b$ – нагрузка системы.

Вероятность обслуживания поступившей в систему заявки:

$$\pi_0 = 1 - \pi_n = \frac{\rho}{y}K.$$

Выражение (3.18) оказывается полезным при расчёте характеристик обслуживания заявок в марковских моделях систем и сетей массового обслуживания (см. примеры в разделе 5).

Зависимости (3.14) и (3.15), связывающие средние значения временны*х* (w, u) и безразмерных (l, m) характеристик, известны как **формулы Литтла** и вместе с формулой (3.13) представляют собой *фундаментальные зависимости*, справедливые для широкого класса моделей массового обслуживания.

Из (3.15) можно получить зависимость, связывающую среднее число заявок в системе со средней длиной очереди заявок:

$$m = \lambda u = \lambda(w + b) = \lambda w + \lambda b = l + y,$$

откуда следует, что *нагрузка* $y = \lambda b$ *характеризует среднее число заявок, находящихся на обслуживании*.

При условии отсутствия перегрузок в одноканальной СМО загрузка совпадает с нагрузкой: $\rho = y = \lambda b$ и тогда $m = l + \rho$, то есть загрузку одноканальной СМО можно трактовать как среднее число заявок, находящихся на обслуживании в приборе. Отметим, что на обслуживании находится не одна заявка, как может показаться, а меньше единицы: $\rho < 1$. Это действительно так, если вспомнить, что речь идёт о *среднем* числе находящихся на обслуживании заявок, которое может быть рассчитано следующим образом. В приборе в каждый момент времени может находиться случайное число заявок, принимающее два значения: 1, если прибор работает, то есть обслуживает заявку, и 0, если прибор простаивает. Поскольку значение загрузки лежит в интервале от 0 до 1 ($0 \le \rho \le 1$) и показывает долю времени, в течение которого прибор работает, то *загрузку можно трактовать как вероятность того, что прибор работает*, а величину $\eta = (1 - \rho)$ – как вероятность простоя прибора. Тогда математическое ожидание случайной величины, принимающей значения 1 с вероятностью ρ и 0 с вероятностью $(1 - \rho)$, будет равно: $\rho \times 1 + (1 - \rho) \times 0 = \rho$, что и требовалось показать.

Обычно исследование систем проводится в предположении о стационарности входящего потока заявок и длительности обслуживания. В этом случае *условие существования установившегося режима* для СМО с накопителем неограниченной ёмкости совпадает с *условием отсутствия перегрузок* в СМО и записывается в виде: $\rho < 1$.

3.3.5. Характеристики СМО с неоднородным потоком заявок

Для СМО с неоднородным потоком заявок, в которую поступают H классов заявок с интенсивностями $\lambda_1, \ldots, \lambda_H$ и средними длительностями обслуживания $b_1, \ldots, b_H$, определяются две группы характеристик обслуживания заявок:

- характеристики по каждому классу (потоку) заявок;
- характеристики объединённого (суммарного) потока заявок.

Характеристики по каждому классу заявок $i = \overline{1, H}$ идентичны характеристикам СМО с однородным потоком:

- *нагрузка*, создаваемая заявками класса i: $y_i = \lambda_i / \mu_i = \lambda_i b_i$;
- *вероятность потери* заявок: π_{n_i};
- *вероятность обслуживания* заявки: $\pi_{0_i} = (1 - \pi_{n_i})$;
- *интенсивность* потока *обслуженных* заявок (производительность по i-му классу заявок): $\lambda_{0_i} = \pi_{0_i} \lambda_i = (1 - \pi_{n_i}) \lambda_i$;
- *интенсивность* потока *потерянных* заявок: $\lambda_{n_i} = \pi_{n_i} \lambda_i$.
- *загрузка* системы, создаваемая заявками класса i:

$\rho_i = \min\left(\dfrac{(1 - \pi_{n_i}) y_i}{K}; 1 \right)$, где π_{n_i} – вероятность потери заявок класса i из-за ограниченной ёмкости накопителя ($\pi_{n_i} = 0$, если ёмкость накопителя – неограниченная); K – число обслуживающих приборов в СМО;

- *время ожидания* заявок в очереди: w_i;
- *время пребывания* заявок в системе: $u_i = w_i + b_i$;
- *длина очереди* заявок: $l_i = \lambda_i w_i$;
- *число заявок в системе* (в очереди и на обслуживании): $m_i = \lambda_i u_i$.

Характеристики объединённого (суммарного) потока заявок позволяют определить усредненные по всем классам заявок показатели эффективности функционирования СМО:

- *суммарная интенсивность* поступления заявок в систему (интенсивность суммарного потока):

$$\Lambda = \sum_{i=1}^{H} \lambda_i ; \qquad (3.19)$$

- *суммарная нагрузка* Y и *суммарная загрузка* R системы:

$$Y = \sum_{i=1}^{H} y_i ; \qquad R = \min\left(\sum_{i=1}^{H} \rho_i ; 1 \right), \qquad (3.20)$$

причем условие отсутствия перегрузок в СМО с неоднородным потоком заявок и накопителем неограниченной ёмкости имеет вид:

$$R < 1; \qquad (3.21)$$

- *коэффициент простоя* системы: $\eta = 1 - R$;
- *среднее время ожидания* W и *среднее время пребывания* U заявок объединённого потока в системе:

$$W = \sum_{i=1}^{H} \xi_i w_i ; \qquad U = \sum_{i=1}^{H} \xi_i u_i , \qquad (3.22)$$

где $\xi_i = \lambda_i / \Lambda$ – коэффициент, учитывающий долю заявок класса i в суммарном потоке, который может трактоваться как *вероятность того, что поступившая в систему заявка принадлежит классу* i;

- *суммарная длина очереди* и *суммарное число заявок в системе*:

$$L = \sum_{i=1}^{H} l_i ; \qquad M = \sum_{i=1}^{H} m_i . \tag{3.23}$$

Можно доказать, что для характеристик объединённого (суммарного) потока справедливы те же фундаментальные соотношения (3.13) – (3.15) , что и для однородного потока:

$U = W + B$; $L = \Lambda W$; $M = \Lambda U$,

где B – среднее время обслуживания любой заявки суммарного потока:

$$B = \sum_{i=1}^{H} \xi_i b_i .$$

3.4. Параметры и характеристики СеМО

3.4.1. Параметры СеМО

Для описания *линейных разомкнутых и замкнутых однородных экспоненциальных* СеМО используется следующая совокупность параметров:

- *число узлов* в сети: n;
- *число обслуживающих приборов* в узлах сети: $K_1,...,K_n$;
- *матрица вероятностей передач*: $\mathbf{P} = [p_{ij} \mid i, j = 0,1,\ldots,n]$, где p_{ij} – вероятность передачи заявки из узла i в узел j;
- *интенсивность* λ_0 источника заявок, поступающих в **разомкнутую** СеМО (РСеМО), или *число заявок* M, циркулирующих в **замкнутой** СеМО (ЗСеМО);
- средние *длительности обслуживания* заявок в узлах сети: $b_1,\ldots,b_n$.

Заметим, что состав параметров разомкнутых и замкнутых СеМО различается только одним параметром, а именно: для ЗСеМО, в отличие от РСеМО, вместо интенсивности λ_0 поступления заявок в сеть необходимо задать число постоянно циркулирующих в сети заявок M.

Для *линейных* СеМО элементы матрицы вероятностей передач должны удовлетворять условию:

$$\sum_{j=0}^{n} p_{ij} = 1 \quad (i = \overline{0,n}). \tag{3.24}$$

Это условие отражает тот факт, что любая заявка, покинувшая некоторый узел, обязательно (с вероятностью 1) перейдёт в какой-то узел, включая тот же самый или нулевой. Переход заявки в нулевой узел означает, что заявка покинула сеть.

В случае *неэкспоненциальных* разомкнутых СеМО дополнительно необходимо задать законы распределения или, по крайней мере, вторые моменты интервалов времени между поступающими в разомкнутую сеть заявками и длительностей обслуживания заявок в узлах сети.

В случае *неоднородных* СеМО необходимо дополнительно задать количество классов заявок H в сети и для каждого класса – матрицы вероятностей передач **P(*h*)**, интенсивности $\lambda_0(h)$ или число заявок $M(h)$, а также средние длительности обслуживания $b_i(h)$ заявок класса $h=\overline{1,H}$ в узле $i=\overline{1,n}$. При необходимости могут быть заданы законы распределений интервалов между поступающими в РСеМО заявками и законы распределений длительностей обслуживания заявок разных классов в узлах сети.

3.4.2. Режимы функционирования СеМО

СеМО, как и СМО, может работать в установившемся и неустановившемся режимах. Последний может быть связан с началом работы системы (переходной режим), нестационарным характером потока заявок и обслуживания в приборе (нестационарный режим) и перегрузкой системы (режим перегрузки).

Очевидно, что для СеМО, как и для СМО, при использовании предположения о стационарности входящего потока заявок и длительностей обслуживания заявок в узлах *условие существования установившегося режима* совпадает с *условием отсутствия перегрузок*.

Рассмотрим это условие для разомкнутой и замкнутой СеМО.

Очевидно, что перегрузки в **разомкнутой СеМО** отсутствуют, если каждый узел сети работает без перегрузок. Если же хотя бы один из узлов сети не справляется с нагрузкой, то длина очереди в этом узле начнет увеличиваться до бесконечности и, следовательно, суммарное число заявок в РСеМО будет расти неограниченно.

Таким образом, для того чтобы в *разомкнутой* СеМО не было перегрузок, необходимо отсутствие перегрузок во всех узлах РСеМО, то есть загрузка ρ_j любого узла j ($j=\overline{1,n}$) должна быть строго меньше единицы:

$$\rho_j=\frac{\lambda_j b_j}{K_j}=\frac{\alpha_j\lambda_0 b_j}{K_j}<1 \qquad \text{для всех } j=\overline{1,n}.$$

Из последнего неравенства имеем:

$$\lambda_0<\frac{K_j}{\alpha_j b_j} \qquad \text{для всех } j=\overline{1,n}.$$

Это условие может быть записано также в следующем виде:

$$\boxed{\lambda_0<\min\left(\frac{K_1}{\alpha_1 b_1},\frac{K_2}{\alpha_2 b_2},...,\frac{K_n}{\alpha_n b_n}\right)}. \qquad (3.25)$$

Полученное условие налагает ограничение сверху на интенсивность поступления заявок в РСеМО из внешнего источника. Узлы, в которых указанное условие не выполняется, являются перегруженными. С течением времени это приводит к неограниченному росту числа заявок в сети, которые скапливаются в перегруженных узлах, имеющих накопители неограниченной ёмкости.

В дальнейшем при исследовании разомкнутых СеМО, если не оговорено другое, будем полагать, что в сети существует установившийся режим.

Несколько иначе дело обстоит для **замкнутых СеМО**. Поскольку в ЗСеМО циркулирует постоянное число заявок, то в узлах сети не могут образовываться очереди бесконечной длины, следовательно, в ЗСеМО всегда существует установившийся режим. Даже если в сети имеется очень «медленный» узел, в котором по сравнению с другими узлами слишком долго обрабатываются заявки, то это может привести только к тому, что все заявки будут постоянно скапливаться в очереди перед данным узлом, однако их количество будет всегда конечно и в пределе равно числу циркулирующих в сети заявок. Загрузка такого «медленного» узла будет близка к единице, поскольку постоянное наличие очереди перед этим узлом обусловливает непрерывную работу приборов узла. Такой узел обычно представляет собой так называемое «узкое место» сети.

3.4.3. Характеристики СеМО

Характеристики СеМО делятся на два класса:

- **узловые**, описывающие эффективность функционирования отдельных узлов СеМО;
- **сетевые**, описывающие функционирование СеМО в целом.

Состав *узловых характеристик* СеМО, работающей *в стационарном режиме*, такой же, как и для СМО, и для узла $j=\overline{1,n}$ включает в себя следующие характеристики:

- *нагрузка* узла: $y_j = \lambda_j b_j = \alpha_j \lambda_0 b_j$;
- *загрузка* узла: $\rho_j = \dfrac{y_j}{K_j} = \dfrac{\alpha_j \lambda_0 b_j}{K_j}$, причем $\rho_j < 1$;
- *коэффициент простоя* узла: $\eta_j = 1 - \rho_j$;
- *время ожидания* заявок в узле: w_j;
- *время пребывания* заявок в узле: $u_j = w_j + b_j$;
- *длина очереди* заявок узле: $l_j = \lambda_j w_j = \alpha_j \lambda_0 w_j$;
- *число заявок в* узле (в очереди и на обслуживании):

$m_j = \lambda_j u_j = \alpha_j \lambda_0 (w_j + b_j) = l_j + y_j$.

В приведенных выше формулах использован тот факт, что в линейных СеМО интенсивность поступления заявок в любой узел связана с интенсивностью источника соотношением (3.5).

На основе узловых характеристик рассчитываются *сетевые характеристики* СеМО:

- **суммарная нагрузка** во всех узлах, характеризующая *среднее число заявок, одновременно находящихся на обслуживании во всех узлах* сети:

$$Y = \sum_{j=1}^{n} y_j ,$$

где y_j – нагрузка узла j, причем $0 < Y \le \sum_{j=1}^{n} K_j$;

- **суммарная загрузка** всех узлов СеМО, характеризующая *среднее число параллельно работающих узлов* сети:

$$R = \sum_{j=1}^{n} \rho_j ,$$

где ρ_j – загрузка узла j, причем $0 < R \le n$;

- среднее число заявок, находящихся в очередях всех узлов сети и ожидающих обслуживания:

$$\boxed{L = \sum_{j=1}^{n} l_j}, \qquad (3.26)$$

где l_j – средняя длина очереди заявок в узле j;

- среднее число заявок, находящихся в сети:

$$\boxed{M = \sum_{j=1}^{n} m_j}, \qquad (3.27)$$

где m_j – среднее число заявок в узле j, причём для замкнутых сетей это выражение может быть использовано для проверки правильности проведенных расчетов, так как для них число заявок M в сети задано;

- среднее время ожидания заявок в сети:

$$\boxed{W = \sum_{j=1}^{n} \alpha_j w_j}, \qquad (3.28)$$

где w_j – среднее время ожидания заявок в узле j; α_j – коэффициент передачи для узла j, показывающий среднее число попаданий заявки в узел j за время её нахождения в сети; $W_j = \alpha_j w_j$ – представляет собой суммарное (полное) время ожидание заявки в узле j за время её нахождения в сети;

- среднее время пребывания заявок в сети:

$$\boxed{U = \sum_{j=1}^{n} \alpha_j u_j}, \qquad (3.29)$$

где u_j – среднее время пребывания заявок в узле j; $U_j = \alpha_j u_j$ – суммарное (полное) время пребывания заявки в узле j за время её нахождения в сети;

- **производительность замкнутой СеМО** λ_0, определяемая как интенсивность потока заявок, проходящих через выделенный нулевой узел замкнутой сети, и представляющая собой среднее число заявок, обслуженных в ЗСеМО за единицу времени; производительность ЗСеМО может быть рассчитана на основе выражения (3.5), из которого следует:

$$\boxed{\lambda_0 = \lambda_j / \alpha_j \quad (j = 1, \ldots, n)}; \tag{3.30}$$

Следует отметить, что для сетевых характеристик СеМО выполняются те же фундаментальные соотношения, что и для СМО, а именно:

$$L = \lambda_0 W; \tag{3.31}$$

$$M = \lambda_0 U; \tag{3.32}$$

$$M = L + Y; \tag{3.33}$$

$$U = W + B, \tag{3.34}$$

где $B = \sum_{j=1}^{n} \alpha_j b_j$ – суммарное время обслуживания заявки во всех узлах за время ее нахождения в сети.

Выражения (3.31) и (3.32) представляют собой формулы Литтла для расчёта сетевых характеристик СеМО.

Из (3.32) может быть получена ещё одна важная *формула для расчёта производительности ЗСеМО*:

$$\boxed{\lambda_0 = \frac{M}{U}}. \tag{3.35}$$

Для неоднородной СеМО перечисленные характеристики определяются как для каждого класса в отдельности, так и для объединенного (суммарного) потока заявок.

3.5. Резюме

1. В качестве математических моделей дискретных систем со стохастическим характером функционирования широко применяются модели массового обслуживания (ММО), которые делятся на *базовые модели* в виде одноканальных и многоканальных *систем массового обслуживания* (СМО) и *сетевые модели* в виде разомкнутых и замкнутых *сетей массового обслуживания* (СеМО).

Для описания СМО используются следующие понятия: *заявка* (требование, запрос, вызов, клиент), *поток заявок*, *обслуживающий прибор* (или просто прибор), *обслуживание*, *длительность обслуживания*, *накопитель*, *ёмкость накопителя*, *очередь*, *длина очереди*, *дисциплина буферизации*, *дисциплина обслуживания*, *приоритет*.

Для описания СеМО дополнительно используются такие понятия как *узел*, *источник*, *граф СеМО*, *маршрут*.

2. Описание потока заявок в простейшем случае предполагает задание его *интенсивности***.** Поток заявок может быть *детерминированным (регулярным)* или *случайным*, *стационарным* или нестационарным, *ординарным* или неординарным (групповым), с последействием или *без последействия*.

Стационарный ординарный поток без последействия называется простейшим (пуассоновским). Интервалы времени между заявками в простейшем потоке распределены по *экспоненциальному закону*. Аналитические исследования моделей массового обслуживания обычно проводятся в предположении о простейшем потоке заявок, что обусловлено рядом присущих ему особенностей (*суммирование потоков, вероятностное разрежение потока*), позволяющих во многих случаях получить сравнительно простые аналитические зависимости характеристик от параметров.

Длительность обслуживания заявок в приборе в простейшем случае может быть задана средним значением или величиной обратной – *интенсивностью обслуживания*, характеризующей среднее число заявок, которое может быть обслужено прибором за единицу времени.

Стратегия управления потоками заявок задается в виде *дисциплины буферизации (ДБ)* и *дисциплины обслуживания (ДО)*, которые могут быть классифицированы по следующим признакам: наличие приоритетов между заявками разных классов; способ (режим) вытеснения заявок из очереди или назначения заявок на обслуживание; правило вытеснения или выбора заявок на обслуживание; возможность изменения приоритетов.

Среди дисциплин обслуживания заявок в технических системах наибольшее распространение получили: бесприоритетная дисциплина *обслуживания в порядке поступления* (ОПП или FIFO) и приоритетные дисциплины: с *относительными* (ОП) и *абсолютными* (АП) *приоритетами*, которые могут быть *статическими* или *динамическими*.

3. Большинство СМО, используемых в качестве базовых моделей реальных систем, могут быть классифицированы: по числу мест в накопителе (*без накопителя* – СМО с отказами; *с накопителем ограниченной ёмкости* – СМО с потерями; *с накопителем неограниченной ёмкости* – СМО без потерь); по количеству обслуживающих приборов (*одноканальные* и *многоканальные*); по количеству классов заявок (с *однородным* и *неоднородным* потоком заявок).

Заявки относятся *к разным классам*, если они в моделируемой реальной системе различаются *длительностью обслуживания и/или приоритетами*.

4. Сетевые модели (СеМО) могут быть классифицированы: в зависимости от характера процессов поступления и обслуживания заявок (*стохастические*, *детерминированные*); по виду зависимостей, связывающих интенсивности потоков заявок в разных узлах СеМО (*линейные*, *нелиней-*

ные); по числу циркулирующих в сети заявок (*разомкнутые*, *замкнутые*, замкнуто-разомкнутые); по типу циркулирующих заявок (*однородные*, *неоднородные*).

В *линейных* СеМО интенсивность потока заявок в любом узле связана линейной зависимостью с интенсивностью источника через коэффициент передачи, который показывает *среднее количество попаданий заявки в данный узел за время ее нахождения в сети*.

В *нелинейных* СеМО интенсивности потоков заявок в узлах связаны нелинейными зависимостями. *Нелинейность СеМО* может быть обусловлена *потерей заявок* или *размножением заявок* в сети.

Разомкнутая СеМО содержит один или несколько *внешних независимых источников* заявок, причем в сети одновременно может находиться *любое число заявок*.

Замкнутая СеМО, в отличие от разомкнутой, не содержит *независимых внешних* источников заявок и характеризуется тем, что в ней циркулирует *постоянное число заявок M*.

5. Для компактного описания СМО используются обозначения в виде **A/B/N/L** , где **A** и **B** – задают законы распределений соответственно интервалов времени между моментами поступления заявок и длительностей обслуживания в приборе; **N** – число обслуживающих приборов в системе; **L** – число мест в накопителе.

6. Для описания СМО, в простейшем случае, используются следующие *параметры*:

- количество обслуживающих приборов *K*;
- количество *k* и емкости накопителей E_j ($j = \overline{1,k}$);
- количество поступающих в систему классов заявок *H*;
- интенсивность λ_i потока и коэффициент вариации ν_{a_i} интервалов времени между поступающими в систему заявками класса $i = \overline{1,H}$;
- среднее значение b_i и коэффициент вариации ν_{b_i} длительности обслуживания заявок класса $i = \overline{1,H}$;
- дисциплина буферизации и дисциплина обслуживания заявок.

СМО может работать в *установившемся* (*стационарном***)** режиме или в *неустановившемся* (переходном или нестационарном режиме). Кроме того, *СМО* может работать в *режиме перегрузки*, когда система не справляется с нагрузкой. При этом характеристики функционирования СМО *с накопителем неограниченной емкости* с течением времени растут неограниченно. Для того чтобы в такой СМО не было перегрузок, необходимо, чтобы нагрузка системы была меньше, чем число обслуживающих приборов, или, что то же самое, загрузка системы была

строго меньше единицы. В СМО *с накопителем ограниченной ёмкости* перегрузки не приводят к неустановившемуся режиму.

7. Характеристики систем со стохастическим характером функционирования являются случайными величинами и полностью описываются соответствующими законами распределений. На практике при моделировании часто ограничиваются определением только средних значений (математических ожиданий), реже – определением двух первых моментов этих характеристик.

В качестве основных *характеристик СМО* с однородным потоком заявок используются:

- *нагрузка* системы: $y = \lambda / \mu = \lambda b$;
- *загрузка* системы: $\rho = \min\left(\frac{(1-\pi_n)y}{K}; 1 \right)$;
- *коэффициент простоя* системы: $\eta = 1 - \rho$;
- *вероятность потери* заявок: $\pi_n = \lim\limits_{T \to \infty} \frac{N_n(T)}{N(T)}$;
- *вероятность обслуживания* заявки: $\pi_0 = (1-\pi_n);$;
- *производительность системы*: $\lambda' = \pi_0 \lambda = (1-\pi_n)\lambda$;
- *интенсивность потока потерянных заявок*: $\lambda'' = \pi_n \lambda = (1-\pi_0)\lambda$;
- *среднее время ожидания* заявок в очереди: $w = ?$ (подлежит определению для каждой конкретной СМО);
- *среднее время пребывания* заявок в системе: $u = w + b$;
- *средняя длина очереди* заявок: $l = \lambda' w$;
- *среднее число заявок* в системе: $m = \lambda' u$.

Для СМО с неоднородным потоком заявок определяются две группы характеристик обслуживания заявок: характеристики *по каждому классу заявок* и характеристики *суммарного (объединенного) потока* заявок.

8. Для описания линейных разомкнутых и замкнутых однородных экспоненциальных СеМО необходимо задать следующие *параметры*:

- число узлов в сети n;
- число обслуживающих приборов в узлах сети $K_1,...,K_n$;
- матрицу вероятностей передач $\mathbf{P} = [\, p_{ij} \mid i, j = 0, 1, \ldots, n]$;
- *интенсивность* λ_0 источника заявок, поступающих в РСеМО, или *число заявок* M , циркулирующих в ЗСеМО;
- средние длительности обслуживания заявок в узлах сети $b_1,...,b_n$.

СеМО, как и СМО, может работать в установившемся и неустановившемся режимах. Последний может быть связан с началом работы системы (переходной режим), нестационарным характером

процессов поступления и обслуживания заявок в приборе (нестационарный режим), а в разомкнутой СеМО, кроме того, перегрузкой системы (режим перегрузки). Условие отсутствия перегрузок в разомкнутой СеМО предполагает отсутствие перегрузок в каждом из узлов сети. *В замкнутой СеМО перегрузки не возникают*.

9. Характеристики СеМО делятся на узловые и сетевые. Состав узловых характеристик СеМО, работающей в стационарном режиме, такой же, как и для СМО. На основе узловых характеристик рассчитываются средние значения *сетевых характеристик* СеМО:

- суммарная нагрузка и загрузка: $Y = \sum_{j=1}^{n} y_j \quad R = \sum_{j=1}^{n} \rho_j$;
- среднее суммарное число заявок, находящихся во всех очередях сети: $L = \sum_{j=1}^{n} l_j$;
- среднее суммарное число заявок, находящихся в разомкнутой сети (во всех узлах): $M = \sum_{j=1}^{n} m_j$;
- среднее время ожидания и пребывания заявок в сети: $W = \sum_{j=1}^{n} \alpha_j w_j$; $W = \sum_{j=1}^{n} \alpha_j w_j$;
- производительность замкнутой СеМО: $\lambda_0 = \frac{M}{U}$.

Сетевые характеристики СеМО связаны между собой теми же фундаментальными соотношениями, что и характеристики СМО.

Для неоднородной СеМО перечисленные характеристики определяются как для каждого класса в отдельности, так и для объединенного (суммарного) потока заявок.

3.6. Практикум: обсуждение и решение задач

В разделе 3 рассмотрены модели массового обслуживания: СМО и СеМО, выполнена их классификация, перечислены параметры и рассчитываемые на их основе характеристики функционирования СМО и СеМО различных классов, приведены основные зависимости для расчета указанных характеристик.

Как и ранее, в процессе обсуждения представленного материала попытаемся ответить на некоторые конкретные вопросы практического характера.

Вопрос 1. Почему математическая модель называется абстрактной?

Обсуждение. Действительно, все математические модели являются абстрактными, собственно, как и сама математика. Абстрактность обусловлена переходом от параметров и характеристик реальной системы к её описанию в терминах определённого математического аппарата, например теории массового обслуживания. Затем выполняется анализ характеристик и исследование свойств этой математической модели, а полученные результаты интерпретируются применительно к реальной системе. Абстрактность математической модели состоит в том, что полученные с её помощью результаты могут быть применены к любой другой реальной системе, которая может быть представлена такой же моделью. Другими словами, одна и та же математическая модель может отображать функционирование совершенно разных по своей природе реальных систем, описываемых с помощью различных структурно-функциональных и нагрузочных параметров, состав и перечень которых определяются соответствующей прикладной областью.

Вопрос 2. Насколько предположение о простейшем характере потока заявок соответствует реальности?

Обсуждение. Простейший поток заявок является математическим представлением некоторого «идеального» потока, обладающего рядом замечательных свойств, благодаря которым для многих математических моделей удаётся получить достаточно простые аналитические зависимости, связывающие характеристики функционирования систем массового обслуживания с исходными параметрами. Одним из таких свойств является «отсутствие последействия», которое заключается в том, что поступление в систему очередной заявки не зависит от того, когда и сколько заявок поступило ранее. В реальной жизни наличие этого свойства означало бы следующее.

Представим, что вы, подходя к автобусной остановке, не успели на только что отправившийся автобус. Если поток автобусов, прибывающих на остановку, простейший, то в сложившейся ситуации это совсем не означает, что вам долго придётся ждать следующий автобус. Вполне возможно, что следующий автобус подойдет к остановке практически сразу. Точно так же, если вы пришли на автобусную остановку и застали большое число ожидающих пассажиров (что свидетельствует о том, что давно не было автобуса), то это совсем не означает, что скоро подойдет автобус. Кто-то скажет, что часто попадал в такие ситуации, и отсюда сделает вывод, что поток автобусов к остановке – простейший. В действительности же реальный поток автобусов может быть сколь угодно близок к простейшему, но не может быть простейшим по следующей причине. Если предположить, что поток автобусов к остановке – простейший, то существует (пусть и совсем ничтожная) вероятность того, что автобус вообще никогда не придёт, что, по всей видимости, невозможно (исключая случай, когда движение автобусов отменено, а все ожидающие

пассажиры не знали об этом). Наличие такой вероятности обусловлено тем, что интервалы времени между последовательными заявками (или автобусами) в простейшем потоке распределены по экспоненциальному закону, функция распределения которого ограничена слева (нулевым значением случайной величины), но не ограничена справа, то есть случайная величина, описывающая интервалы между последовательными заявками в простейшем потоке, может принимать сколь угодно большие значения, в том числе, равное бесконечности. Очевидно, что в реальных системах функция распределения обычно ограничена и справа.

Таким образом, отвечая на поставленный вопрос, можно сказать, что в реальной жизни вряд ли существует простейший поток. В то же время, многие реальные потоки могут быть достаточно близки к простейшему.

Вопрос 3. Когда оправдано использование предположения о простейшем характере потока заявок?

Обсуждение. Предположение о простейшем потоке широко используется не только из-за простоты получения математических зависимостей, но и по той причине, что многие реальные потоки близки к простейшим. Эта близость во многих случаях обусловлена следующим.

Во-первых, как сказано выше, суммирование (объединение) независимых *стационарных ординарных* потоков образует простейший поток при условии, что складываемые потоки оказывают более или менее одинаковое влияние на суммарный поток, причем на практике суммарный поток становится близким к простейшему уже при суммировании 5 потоков. Отметим, что к суммируемым потокам не предъявляется требование отсутствия последействия.

Во-вторых, можно показать, что стационарный ординарный поток заявок стремится к простейшему, если на него оказывает влияние множество случайных факторов. Именно этим можно объяснить близость потока автобусов, прибывающих на остановку, к простейшему. Действительно, если даже все автобусы отправляются с конечной остановки через одинаковые интервалы времени, то есть образуют детерминированный поток, то в процессе движения по улицам города интервалы между ними изменяются под влиянием многих, в основном случайных, факторов, таких как задержки перед светофорами, заторы и «транспортные пробки» на улицах, случайное время нахождения на остановках (зависящее от числа входящих и выходящих из автобуса пассажиров) и т.д. Всё это приводит к тому, что моменты прибытия к остановкам образуют случайный процесс, причем, чем ближе к конечной остановке, тем больше поток автобусов похож на простейший.

Предположение о простейшем характере входного потока заявок оправдано также в тех случаях, когда известно, что коэффициент вариации интервалов между последовательными заявками реального потока меньше единицы. В этом случае использование простейшего потока в модели

позволяет получить так называемые верхние оценки характеристик обслуживания заявок, гарантирующие, что в реальной системе значения характеристик будут не хуже, чем полученные на модели.

Вопрос 4. Почему в СМО с накопителем неограниченной емкости, работающей без перегрузок, возникают очереди? В каких случаях они не возникают?

Обсуждение. В СМО с накопителем неограниченной емкости перегрузки отсутствуют, если интенсивность поступления заявок меньше интенсивности обслуживания.

Рассмотрим случай, когда интенсивность поступления заявок равна 10 заявок в секунду, а интенсивность обслуживания – 1 заявка в секунду. За первую секунду в систему поступит 10 заявок, из которых будет обслужена одна заявка, а 9 – останутся в очереди. За вторую секунду в систему поступит ещё 10 заявок и одна заявка будет обслужена, в очереди окажется 18 заявок и т.д. Очевидно, что число заявок в очереди со временем будет возрастать до бесконечности, что свидетельствует о перегрузке системы, то есть система не справляется с нагрузкой.

Рассмотрим другой случай, когда интенсивность поступления заявок – 1 заявка в секунду, а интенсивность обслуживания – 10 заявок в секунду, или, что то же самое, средний интервал между последовательными заявками в потоке – 1 секунда, а средняя длительность обслуживания – 0,1 секунды. Таким образом, если заявки поступают с интервалом 1 секунда, а обслуживаются за 0,1 секунды, то возникает вопрос: откуда появляется очередь заявок?

Здесь следует обратить внимание на то, что речь идёт о *среднем* значении интервала между заявками и *среднем* значении длительности обслуживания. *Если процессы поступления и обслуживания заявок детерминированные, то очередь перед прибором не образуется*. Такие системы, естественно, не представляют интереса и не рассматриваются в теории массового обслуживания. Очередь появится только в том случае, если процесс поступления заявок в систему или процесс обслуживания их в приборе, или оба процесса – случайные. Тогда конкретное значение какого-то интервала между заявками может оказаться намного меньше среднего значения, например менее 0,1 секунды, а длительность обслуживания некоторой заявки – много больше среднего значения, например 2 секунды. Именно такие ситуации и приводят к появлению очереди перед прибором. Попутно отметим, что длина очереди – величина случайная, изменяющаяся случайным образом между нулём и некоторым максимальным значением.

Вопрос 5. Что в реальной системе может служить основанием для того, чтобы в соответствующей математической модели заявки были разделены на разные классы?

Обсуждение. Рассмотрим две модели обслуживания клиентов:

1) модель небольшого магазина, в котором только один продавец обслуживает покупателей, которыми являются и мужчины и женщины;

2) модель парикмахерской, в которой работает один мастер, делающий причёски мужчинам и женщинам.

Следует ли мужчин и женщин отнести к разным классам или же объединить их в модели в один класс?

Обе рассматриваемые модели представляют собой одноканальные СМО, в которых заявки соответствуют клиентам, а обслуживание заключается в затратах времени продавца или парикмахера на одного клиента.

В модели магазина мужчин и женщин при отсутствии у кого-нибудь из них преимущественного права (приоритета) на внеочередное обслуживание, скорее всего, можно объединить в один класс, поскольку время, затрачиваемее продавцом на одного покупателя примерно одинаково и не зависит от пола покупателя.

В парикмахерской, как известно, время, затрачиваемое на создание женской причёски много больше, чем на создание мужской причёски. В этом случае в модели парикмахерской заявки должны быть разбиты на два класса. Очевидно, что времена пребывания заявок разных классов в общем случае будут различаться, даже если их времена ожидания окажутся одинаковыми.

Вопрос 6. Когда в качестве модели реальной системы следует использовать разомкнутую, а когда замкнутую СеМО? Каким образом в замкнутой СеМО выбирается дуга, на которой отмечается точка «0»?

Обсуждение. Положим, что СеМО используется в качестве модели обслуживания покупателей в большом магазине с несколькими разными отделами, каждый из которых представляется в модели как узел сети. Покупатели в модели отображаются в виде заявок, перемещающихся между узлами СеМО.

Если количество покупателей, одновременно находящихся в магазине, может любым и принимать значения от 0 и, теоретически, до бесконечности, то в качестве модели такого магазина следует использовать разомкнутую СеМО.

Представим теперь, что мы хотим промоделировать работу этого магазина в час пик, когда в магазин стремится попасть большое число покупателей. Положим, что количество покупателей, которые могут одновременно находиться в магазине, определяется количеством корзинок или тележек, без которых вход в магазин запрещён. При отсутствии корзинок покупатели образуют очередь на входе и ожидают освобождения корзинок. Покупатель, покидающий магазин при выходе передает освободившуюся корзинку ожидающему на входе покупателю, который затем заходит в магазин. Таким образом, в магазине находится постоянное число покупателей, равное числу корзинок в магазине. Очевидно, что в

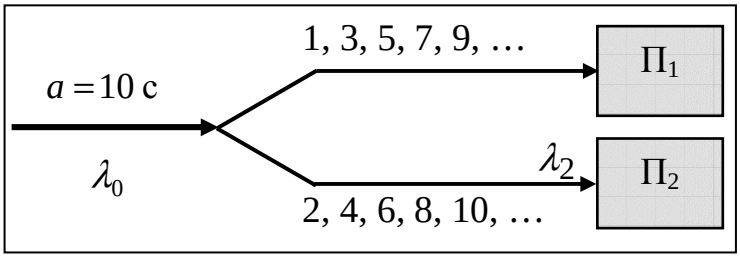
этом случае в качестве модели магазина должна использоваться замкнутая СеМО, а нулевая точка в модели должна быть выбрана на дуге, отображающей выход покупателя из магазина и вход нового покупателя.

Задача 1. В двухканальную СМО поступает простейший поток заявок со средним интервалом между соседними заявками 10 с, причем каждая вторая заявка направляется ко второму прибору. Чему равна интенсивность потока заявок ко второму прибору? Чему равен коэффициент вариации интервалов между заявками потока ко второму прибору?

Дано: СМО: $K = 2$; поток – простейший; $a = 10$ с.

Требуется:

- определить λ_2;
- определить ν_2.



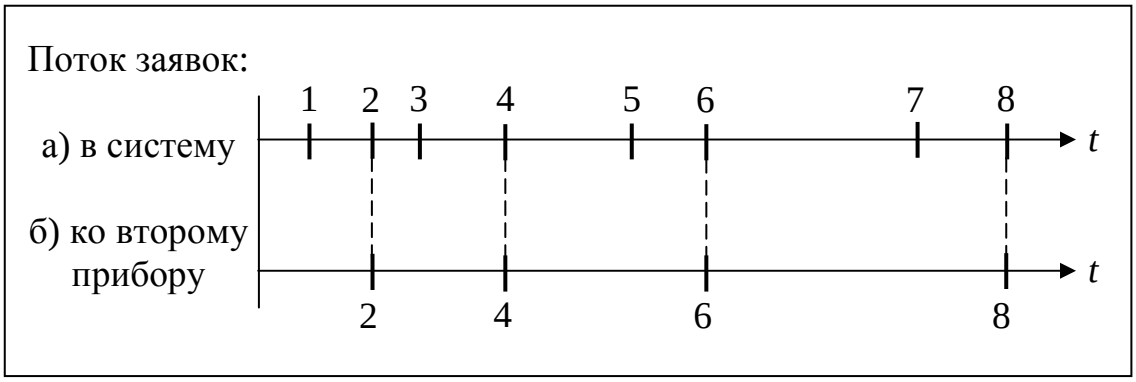
Решение.

1) Интенсивность потока заявок в СМО: $\lambda_0 = 1/a = 0{,}1\ \text{с}^{-1}$.

2) Поскольку каждая вторая заявка направляется ко второму прибору, то очевидно, что интенсивность поступления заявок ко второму прибору будет в два раза меньше, чем исходная интенсивность λ_0, то есть

$\lambda_2 = 0{,}5\lambda_0 = 0{,}05\ \text{с}^{-1}$.

2) Для определения коэффициента вариации ν_2 найдём вид закона распределения интервалов между заявками ко второму прибору, для чего построим временную диаграмму, отражающую процесс поступления заявок в систему (а) и ко второму прибору (б).



Как видно из диаграммы, интервалы между заявками ко второму прибору представляют собой сумму двух временных интервалов исходного простейшего потока заявок, поступающих в систему. Каждый такой временной интервал в случае простейшего потока представляет собой случайную величину, распределённую по экспоненциальному закону. Таким образом, интервалы между заявками ко второму прибору представляют собой случайную величину, равную сумме двух экспоненциально

распределённых величин, что соответствует распределению Эрланга 2-го порядка ($k = 2$).

Коэффициент вариации случайной величины, распределённой по закону Эрланга (см.п.2.5.5), зависит от порядка *k* и определяется по формуле:

$$\nu_2 = \nu_{Э_2} = \frac{1}{\sqrt{k}} = \frac{1}{\sqrt{2}} \approx 0{,}71.$$

Следует различать рассмотренное выше *детерминированное* разрежение потока от *вероятностного* разрежения. В случае вероятностного разрежения, когда заявки направляются ко второму прибору с вероятностью $p_2 = 0{,}5$, интенсивность поступления заявок ко второму прибору будет такой же, как и при детерминированном разрежении, то есть $\lambda_2 = p_2\lambda_0 = 0{,}5\lambda_0 = 0{,}05\ \text{с}^{-1}$. Однако коэффициент вариации в этом случае равен единице: $\nu_2 = 1$, поскольку, в соответствии с одним из сформулированных в п.3.1.3 замечательных особенностей простейшего потока, при вероятностном разрежении образуются простейшие потоки, в которых интервалы между последовательными заявками распределены по экспоненциальному закону, а не по закону Эрланга.

Задача 2. Проиллюстрировать на примере различие между дисциплинами группового и одиночного режима.

Решение. Рассмотрим следующие дисциплины обслуживания заявок:

1) одиночного режима:

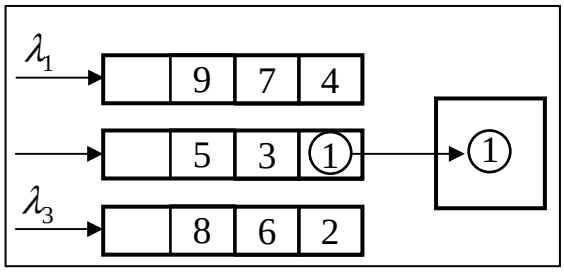
- обслуживание в порядке поступления (ОПП или FIFO);
- обслуживание в обратном порядке (ООП или LIFO);
- циклическое обслуживание в одиночном режиме (ЦО ОР), означающее, что всякий раз на обслуживание из очереди выбирается только одна заявка, после чего обслуживающий прибор переходит к следующей по порядку очереди, даже если в предыдущей очереди остались заявки;
- с относительными приоритетами (ОП), распределёнными по правилу: класс заявок с меньшим номером имеет более высокий приоритет;

2) группового режима:

- циклическое обслуживание в групповом режиме (ЦО ГР), отличающееся от одиночного режима тем, что обслуживание очереди заявок одного и того же класса осуществляется до тех пор, пока очередь не окажется пустой;
- чередующиеся приоритеты с размером группы, равным 2 (ЧП2), означающим, что из каждой очереди заявок последовательно выбирается на обслуживание не более двух заявок, после чего обслуживающий прибор переходит к непустой очереди с самым высоким приоритетом, даже если в предыдущей очереди остались заявки;

- чередующиеся приоритеты с неограниченным размером группы (ЧП), означающим, что обслуживание очереди заявок одного и того же класса осуществляется до тех пор, пока очередь не окажется пустой.

Положим, что в *некоторый фиксированный момент времени* в системе с тремя классами (очередями) заявок сложилась следующая ситуация (см. рисунок).



В системе находится 9 заявок. Номер заявки соответствует моменту поступления её в систему – чем меньше номер, тем раньше потупила заявка в систему, то есть заявка с номером 1 поступила раньше всех, а последней поступила заявка с номером 9. Все поступившие на рассматриваемый момент времени заявки распределены по классам (очередям) следующим образом: заявки самого высокоприоритетного первого класса поступили в систему в моменты 4, 7 и 9, заявки второго класса – в моменты 1, 3 и 5, заявки третьего низкоприоритетного класса – в моменты 2, 6 и 8. Положим, что в рассматриваемый момент времени на обслуживании в приборе находится заявка второго класса с номером 1. Полагая, что в систему более не поступят другие заявки, запишем последовательность обслуживания заявок при использовании перечисленных выше дисциплин обслуживания:

ОПП: 1, 2, 3, 4, 5, 6, 7, 8, 9
ООП: 1, 9, 8, 7, 6, 5, 4, 3, 2
ЦО ОР: 1, 2, 4, 7, 3, 6, 9, 5, 8
ЦО ГР: 1, 3, 5, 2, 6, 8, 4, 7, 9
ОП: 1, 4, 7, 9, 3, 5, 2, 6, 8
ЧП2: 1, 3, 4, 7, 9, 5, 2, 6, 8
ЧП: 1, 3, 5, 4, 7, 9, 2, 6, 8

Таким образом, изменение дисциплины обслуживания приводит к изменению последовательности выбора заявок на обслуживание из очередей и, следовательно, к изменению их времени ожидания. В частности, заявка с номером 9 будет иметь максимальное время ожидания при дисциплинах ОПП и ЦО ГР, а минимальное – при ООП.

Следует обратить внимание на то, что при групповом режиме заявки выбираются из очереди и обслуживаются в приборе так же по одной, как и при одиночном режиме, то есть последовательно друг за другом, а не группой. Понятие «групповой режим» лишь означает, что на обслуживание **назначается** (а не обслуживается) группа заявок (обычно одного класса), и прибор переходит к обслуживанию другой группы только после завершения обслуживания всех заявок назначенной группы.

3.7. Самоконтроль: перечень вопросов и задач

1. Выполнить классификацию СМО:
- по числу обслуживающих приборов;
- по емкости накопителя;
- по числу потоков заявок.

2. Какой поток заявок называется однородным? В каких случаях поток заявок в СМО является неоднородным?

3. В каких случаях заявки в СМО относятся к разным классам?

4. Нарисовать одноканальную СМО с неоднородным потоком заявок. Какие параметры необходимо задать для её описания? Какие характеристики функционирования СМО могут быть рассчитаны по этим параметрам?

5. Нарисовать многоканальную СМО с неоднородным потоком заявок. Какие параметры необходимо задать для её описания? Какие характеристики функционирования СМО могут быть рассчитаны по этим параметрам?

6. В чём различие между детерминированным и регулярным потоком заявок? Какой поток заявок является альтернативой детерминированного потока?

7. Как называется стационарный ординарный поток без последействия?

8. Когда поток заявок является стационарным? Привести примеры нестационарного потока заявок.

9. Какой поток заявок называется ординарным? Привести примеры неординарного потока заявок.

10. Каким является поток, в котором момент поступления очередной заявки не зависит от того, когда и сколько заявок поступило до этого момента?

11. В чём проявляется наличие последействия в потоке заявок? Привести примеры потоков заявок с последействием.

12. Понятие интенсивности потока и ее размерность. Что характеризует величина обратная интенсивности?

13. По какому закону распределены интервалы времени между заявками в простейшем потоке?

14. Какими замечательными особенностями обладает простейший поток заявок?

15. Чему равны математическое ожидание, коэффициент вариации и дисперсия интервалов времени между соседними заявками в простейшем потоке, интенсивность которого равна 2 заявки в секунду?

16. В систему поступают заявки с интервалом 80 секунд. Чему равно среднее число заявок, которые поступят в систему в течение 50-ти минут, в случае: а) детерминированного потока; б) простейшего потока; в) случайного потока?

17. В систему поступают заявки двух классов со средним интервалом между соседними заявками 0,2 с и 2 с соответственно. Определить суммарную интенсивность поступления заявок в систему. По какому закону распределены интервалы между заявками суммарного потока?

18. В систему поступают заявки трех классов со средним интервалом между соседними заявками 0,1 с; 0,2 с и 2 с соответственно. Определить суммарную интенсивность поступления заявок в систему. Чему равен коэффициент вариации интервалов между заявками суммарного потока?

19. В двухканальную СМО поступает простейший поток заявок со средним интервалом между соседними заявками 0,2 с, причем каждая третья заявка направляется ко второму прибору. Чему равна интенсивность потока заявок ко второму прибору? По какому закону распределены интервалы между заявками потока ко второму прибору?

20. В двухканальную СМО поступает простейший поток заявок с интенсивностью 15 заявок в секунду, причем с вероятностью 1/3 заявка направляется ко второму прибору. Чему равна интенсивность потока заявок к первому прибору? Чему равен коэффициент вариации интервалов между заявками потока к первому прибору?

21. Что понимается под обслуживанием заявок в СМО? Что такое интенсивность обслуживания заявок в СМО, и какова её размерность?

22. Чему равны математическое ожидание, коэффициент вариации и дисперсия длительности обслуживания заявок в СМО, распределенной по экспоненциальному закону, если известно, что интенсивность обслуживания равна 2 заявки в секунду?

23. В СМО поступают 2 класса заявок с интенсивностями 0,06 и 0,54 заявок в минуту, длительности обслуживания которых распределены по экспоненциальному закону со средними значениями 2 и 1 секунд соответственно. а) По какому закону распределена длительность обслуживания заявок суммарного (объединенного) потока? б) Чему равна средняя длительность обслуживания заявок суммарного потока?

24. Перечислить возможные дисциплины буферизации. В каких СМО не используются дисциплины буферизации?

25. Какие дисциплины обслуживания заявок относятся к бесприоритетным?

26. Краткая характеристика приоритетных дисциплин обслуживания заявок.

27. Проиллюстрировать на примере отличие дисциплин группового режима от дисциплин одиночного режима.

28. В чем отличие дисциплины с чередующимися приоритетами от дисциплины с относительными приоритетами. Проиллюстрировать на примере.

29. Что такое динамические приоритеты?

30. Что характеризуют нагрузка и загрузка? В чём отличие загрузки от нагрузки? В каких случаях нагрузка совпадает с загрузкой?

31. Перечислить факторы, обусловливающие нестационарный режим работы СМО.

32. Что такое и чем характеризуется перегрузка системы? При каких условиях возникают перегрузки системы? В каких СМО не возникают перегрузки?

33. При каком условии в одноканальной СМО отсутствуют перегрузки?

34. Раскрыть обозначение и дать краткое описание следующих СМО: а) D/M/2/3; б) M/H_2/3; в) E_3/D/2/5.

35. Привести обозначение СМО в символике Кендалла, имеющей следующее описание: двухканальная СМО с однородным простейшим потоком заявок, длительность обслуживания которых распределена по произвольному закону общего вида, с ограниченной емкостью накопителя, равной 5.

36. Перечислить характеристики одноканальной и многоканальной СМО с однородным потоком заявок и записать соотношения, устанавливающие их взаимосвязь.

37. Перечислить характеристики одноканальной СМО с неоднородным потоком заявок и записать соотношения, устанавливающие их взаимосвязь.

38. Почему в СМО, работающей в стационарном режиме, могут возникать очереди? В каких случаях они не возникают? Перечислите причины, обусловливающие возникновение очередей в СМО, работающей в стационарном режиме.

39. Какая СеМО называется линейной? Перечислить факторы, обусловливающие нелинейность СеМО.

40. Основные отличия замкнутых СеМО от разомкнутых.

41. Какая СеМО называется экспоненциальной? Перечислить факторы, обусловливающие неэкспоненциальность СеМО.

42. Какая СеМО называется неоднородной? Перечислить факторы, обусловливающие неоднородность СеМО.

43. Перечислить параметры разомкнутой и замкнутой однородной неэкспоненциальной СеМО.

44. Перечислить параметры разомкнутой и замкнутой неоднородной приоритетной СеМО.

45. Каким условиям должны удовлетворять элементы матрицы вероятно-стей передач в СеМО?

46. Узловые характеристики однородных СеМО и их взаимосвязь.

47. Сетевые характеристики разомкнутых и замкнутых однородных СеМО и их взаимосвязь.

48. Что такое "производительность замкнутой СеМО"? Какие соотношения используются для расчета производительности замкнутой СеМО?

Раздел 4. АНАЛИТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

«Всякое уравнение длиной более двух дюймов, скорее всего, неверно!» (***Автор неизвестен***)

4.1. Одноканальные СМО с однородным потоком заявок

Рассмотрим одноканальную СМО с однородным потоком заявок при следующих предположениях (рис.4.1):

1) СМО содержит *один обслуживающий прибор*, в котором в каждый момент времени может обслуживаться только одна заявка;

2) перед прибором имеется накопитель **Н** *неограниченной ёмкости*, что означает отсутствие отказов поступающим заявкам при их постановке в очередь **О**, то есть любая поступающая заявка всегда найдет в накопителе место для ожидания не зависимо от того, сколько заявок уже находится в очереди;

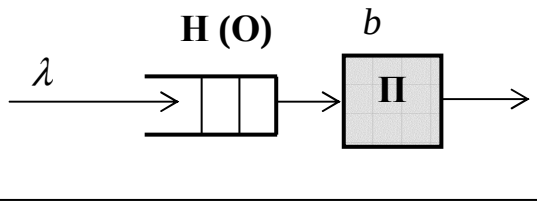


Рис.4.1. Одноканальная система массового обслуживания (СМО)

3) заявки поступают в СМО с интенсивностью λ;

4) средняя длительность обслуживания одной заявки в приборе равна b, причем длительности обслуживания разных заявок не зависят друг от друга;

5) обслуживающий прибор не простаивает, если в системе (накопителе) имеется хотя бы одна заявка, причем после завершения обслуживания очередной заявки мгновенно из накопителя выбирается следующая заявка;

6) заявки из накопителя выбираются в соответствии с бесприоритетной дисциплиной обслуживания в порядке поступления (ОПП) по правилу «первым пришел – первым обслужен» (FIFO – First In First Out).

7) в системе существует стационарный режим, предполагающий отсутствие перегрузок, то есть нагрузка и, следовательно, загрузка системы меньше 1: $y = \rho = \lambda b < 1$.

В качестве расчётной характеристики обслуживания заявок в СМО будем использовать среднее время ожидания заявок. Значения остальных характеристик функционирования СМО легко могут быть рассчитаны с использованием приведенных в разделе 3 фундаментальных соотношений (3.13) – (3.15).

Рассмотрим четыре модели СМО с однородным потоком заявок: экспоненциальную СМО M/M/1 и три неэкспоненциальные СМО типа M/G/1, G/M/1, G/G/1.

4.1.1. Характеристики экспоненциальной СМО M/M/1

Пусть заявки, поступающие в одноканальную СМО, образуют *простейший* поток с интенсивностью λ, а длительность τ_b обслуживания заявок распределена по *экспоненциальному* закону со средним значением b, причём $\rho = \lambda b < 1$, то есть система работает в установившемся режиме Такая СМО с однородным потоком заявок называется ***экспоненциальной***.

С использованием метода средних значений [2] или марковской модели (см.п.5.4.4) можно получить следующие выражения для расчета средних значений:

- времени ожидания заявок

$$\boxed{w = \frac{\rho b}{1-\rho}}; \qquad (4.1)$$

- времени пребывания заявок

$$\boxed{u = w + b = \frac{b}{1-\rho}}; \qquad (4.2)$$

- длины очереди заявок

$$l = \lambda w = \frac{\rho^2}{1-\rho};$$

- числа заявок в системе (в очереди и на обслуживании)

$$m = \lambda u = \frac{\rho}{1-\rho}.$$

Из последнего выражения вытекает, что среднее число заявок в системе $m = l + \rho$, где второе слагаемое ρ определяет среднее число заявок, находящихся на обслуживании в приборе. Кроме того, сравнивая выражения (4.1) и (4.2) получим, что $u = \rho w$.

4.1.2. Характеристики неэкспоненциальной СМО M/G/1

Пусть заявки, поступающие в одноканальную СМО, образуют простейший поток с интенсивностью λ, а длительность τ_b обслуживания заявок распределена по произвольному закону $B(\tau)$ со средним значением b и коэффициентом вариации ν_b.

С использованием метода средних значений можно показать, что среднее время ожидания заявок определяется по формуле Поллачека-Хинчина [2]:

$$\boxed{w = \frac{\lambda b^2 (1 + \nu_b{}^2)}{2(1-\rho)}}, \qquad (4.3)$$

где $\rho = \lambda b < 1$ – загрузка системы.

Среднее время пребывания заявок в системе:

$$u = w + b = \frac{\lambda b^2 (1 + \nu_b{}^2)}{2(1-\rho)} + b.$$

Следует отметить интересную особенность представленных выражений, а именно: средние значения характеристик обслуживания заявок зависят только от двух первых моментов длительности обслуживания заявок и не зависят от моментов более высокого порядка. Другими словами, для того чтобы рассчитать средние характеристики обслуживания, не обязательно знать закон распределения длительности обслуживания заявок – достаточно знать только два первых момента распределения. Можно показать, что для расчета вторых моментов характеристик обслуживания заявок достаточно задать три первых момента длительности обслуживания и т.д. Резюмируя, можно утверждать, что *для СМО* **с простейшим потоком заявок** *для расчёта первых k моментов характеристик обслуживания необходимо задать* $(k+1)$ *моментов длительности обслуживания заявок*.

4.1.3. Характеристики неэкспоненциальной СМО G/M/1

Пусть в одноканальную СМО с интенсивностью λ поступает случайный поток заявок произвольного вида, задаваемый функцией распределения интервалов между заявками $A(\tau)$, а длительность τ_b обслуживания заявок распределена по экспоненциальному закону $B(\tau)$ со средним значением b (интенсивностью $\mu = 1/b$).

СМО G/M/1 является симметричной по отношению к СМО M/G/1, рассмотренной в предыдущем пункте. Однако получение конечного результата в виде аналитического выражения для расчёта среднего времени ожидания, по аналогии с предыдущей моделью, в общем случае, оказывается невозможным. Это обусловлено тем, что среднее время ожидания, впрочем, как и другие числовые моменты, зависит не только от двух первых моментов интервалов между поступающими заявками, но и от моментов более высокого порядка, т.е. от закона распределения интервалов между заявками.

Среднее время ожидания заявок в очереди может быть рассчитано следующим образом [9]:

$$w = \varsigma b/(1-\varsigma), \qquad (4.4)$$

где ς – единственный в области $0 < \varsigma < 1$ корень уравнения

$$\varsigma = A^*(\mu - \mu\varsigma). \qquad (4.5)$$

Здесь $A^*(s)$ – преобразование Лапласа плотности распределения $a(\tau)$ интервалов между поступающими в систему заявками:

$$A^*(s) = \int_0^\infty e^{-s\tau} a(\tau) d\tau \qquad (s \geq 0).$$

Основная сложность при исследовании СМО G/M/1 заключается в том, что уравнение (4.5) для нахождения ς, в общем случае, является трансцендентным, и невозможно получить выражение для ς в явном виде. Однако в каждом конкретном случае корень уравнения (4.5) может быть найден с использованием численных методов.

Как сказано выше, средние значения характеристик обслуживания заявок зависят не только от двух первых моментов интервалов между поступающими заявками, но и от моментов более высокого порядка, причем степень влияния соответствующих моментов убывает с увеличением порядка моментов. Другими словами, влияние моментов 4-го порядка менее существенно, чем моментов 3-го порядка и т.д.

Пример 4.1. Проиллюстрируем применение описанного метода расчета к рассмотренной выше СМО M/M/1 с простейшим потоком заявок.

В простейшем потоке интервалы времени между последовательными заявками распределены по экспоненциальному закону, преобразование Лапласа которого имеет вид: $A(s) = \dfrac{\lambda}{\lambda + s}$.

Тогда уравнение (4.5) примет вид:

$$\varsigma = \frac{\lambda}{\lambda + \mu - \mu\varsigma}.$$

После некоторых преобразований получим квадратное уравнение:

$$\mu\varsigma^2 - (\lambda + \mu)\varsigma + \lambda = 0.$$

Разделив левую и правую часть этого уравнения на μ, получим:

$$\varsigma^2 - (1+\rho)\varsigma + \rho = 0.$$

Из двух корней $\varsigma_1 = 1$ и $\varsigma_2 = \rho$ последнего уравнения условию $0 < \varsigma < 1$ удовлетворяет только второй корень. Подставляя его в выражение (4.4) окончательно получим выражение для среднего времени ожидания, совпадающее с известным для СМО M/M/1 выражением (4.1).

4.1.4. Характеристики СМО общего вида G/G/1

Наиболее общим случаем одноканальных систем массового обслуживания являются СМО типа G/G/1, в которую поступает произвольный поток заявок общего вида с функцией распределения интервалов между заявками $A(\tau)$. Длительность обслуживания заявок в приборе распределена по произвольному закону $B(\tau)$.

Расчет таких систем требует задания конкретных законов распределений, что не позволяет получить аналитическое решение в общем виде. Аналитическое решение возможно только для некоторых частных распределений, связанных, например, с экспоненциальным распределением. Для большинства законов распределений интервалов между поступающими в систему заявками и длительностей их

обслуживания в приборе невозможно получить точное решение в аналитической форме.

В то же время, на практике при исследовании реальных систем редко бывают известны законы распределений указанных величин. Обычно при описании процессов поступления заявок в систему и их обслуживания в приборе ограничиваются несколькими моментами соответствующих распределений, чаще всего – двумя первыми моментами, задаваемыми в виде математического ожидания и среднеквадратического отклонения или коэффициента вариации искомой случайной величины. Однако при этом оказывается невозможным получение точного результата. Это обусловлено тем, что в случае произвольного (отличного от простейшего) потока заявок, поступающих в систему, характеристики функционирования СМО, в частности среднее время ожидания, зависят не только от двух первых моментов, но и от моментов более высокого порядка – третьего, четвёртого и т.д. Причём эта зависимость тем меньше, чем выше порядок числового момента. Таким образом, все результаты, полученные в аналитической форме при задании интервалов между поступающими в систему заявками и длительностей их обслуживания в приборе двумя первыми моментами – средними значениями $a = 1/\lambda$ и $b = 1/\mu$ и коэффициентами вариации ν_a и ν_b, представляют собой приближённые зависимости.

Как показал анализ многочисленных опубликованных результатов, одним из наиболее удачных приближений для расчета среднего времени ожидания в СМО G/G/1 является следующая формула [17]:

$$\boxed{\tilde{w} \approx \frac{\rho b(\nu_a^2 + \nu_b^2)}{2(1-\rho)} f(\nu_a)}, \qquad (4.6)$$

где $\rho = \lambda b < 1$ – загрузка системы; λ, ν_a – интенсивность потока заявок и коэффициент вариации интервалов между поступающими в систему заявками; b, ν_b – среднее значение и коэффициент вариации длительности обслуживания заявок; $f(\nu_a)$ – корректирующая функция, рассчитываемая в зависимости от значения коэффициента вариации ν_a:

$$f(\nu_a) = \begin{cases} \exp\left[-\dfrac{2(1-\rho)(1-\nu_a^2)^2}{3\rho(\nu_a^2 + \nu_b^2)}\right], & \nu_a < 1 \\ \exp\left[-(1-\rho)\dfrac{\nu_a^2 - 1}{\nu_a^2 + 4\nu_b^2}\right], & \nu_a \geq 1. \end{cases}$$

При решении многих практических задач выходящий поток заявок из одной СМО является входящим потоком в другую СМО. В этом случае для расчёта характеристик функционирования второй СМО необходимо знать характер входящего потока, наиболее полно описываемый законом распределения интервалов между последовательными заявками. В то же время, для проведения оценочных расчётов во многих случаях достаточно

знание первых двух моментов этих интервалов: математического ожидания и коэффициента вариации.

Очевидно, что в СМО *с накопителем неограниченной ёмкости*, работающей *без перегрузок*, интенсивность выходящего потока заявок равна интенсивности входящего потока, то есть математические ожидания интервалов между последовательными заявками на выходе и входе СМО совпадают.

Можно показать, что для экспоненциальной СМО M/M/1 коэффициент вариации выходящего потока равен единице.

В общем случае для СМО G/G/1 коэффициент вариации выходящего потока заявок может быть рассчитан по следующей приближённой формуле [17]:

$$\nu_c^2 \approx \nu_a^2 + 2\rho\nu_b^2 - 2\rho(1-\rho)\frac{\tilde{w}}{b}. \qquad (4.7)$$

4.1.5. Анализ свойств одноканальной СМО

«Если факты не подтверждают теорию, от них надо избавиться» (***Закон Майерса***)

Анализ свойств одноканальной СМО с однородным потоком заявок будем проводить с использованием представленных выше математических моделей в виде формул (4.1 – 4.3), определяющих зависимости характеристик обслуживания заявок от параметров поступления и обслуживания заявок для установившегося (стационарного) режима работы системы.

1. Среднее время ожидания заявок в очереди минимально при постоянной (детерминированной) длительности обслуживания заявок, когда коэффициент вариации длительности обслуживания $\nu_b = 0$, и увеличивается с ростом коэффициента вариации (дисперсии) длительности обслуживания. Заметим, что зависимость среднего времени ожидания от коэффициента вариации ν_b носит нелинейный характер. Так, например, при экспоненциально распределенной длительности обслуживания, когда $\nu_b = 1$, среднее время ожидания заявок увеличивается в 2 раза, а при $\nu_b = 2$ – в 5 раз, по сравнению с детерминированным обслуживанием.

2. Среднее время ожидания заявок существенно зависит от нагрузки y (загрузки ρ) системы (рис.4.2). При $y \geq 1$ ($\rho \to 1$) время ожидания заявок возрастает неограниченно: $w \to \infty$, т.е. заявки могут ожидать обслуживания сколь угодно долго. Отметим, что увеличение нагрузки может быть обусловлено двумя факторами: увеличением интенсивности поступления заявок в систему или увеличением длительности обслуживания заявок (например, в результате уменьшения скорости работы обслуживающего прибора).

3. Можно показать, что для бесприоритетных дисциплин обслуживания в обратном порядке (ООП) и обслуживания в случайном порядке (ОСП) *средние времена ожидания заявок будут такими же*, как и при обслуживании в порядке поступления, но *дисперсии времени ожидания будут больше*. Это обусловлено, в частности для дисциплины ООП, тем, что часть заявок, поступивших последними, будут ожидать незначительное время, в то время как заявки, попавшие в начало очереди, могут ожидать обслуживания достаточно долго, то есть увеличивается разброс значений времени ожидания относительно среднего значения.

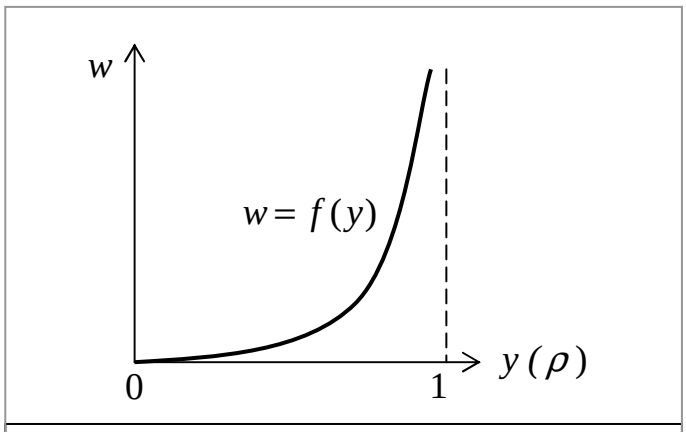


Рис.4.2. Зависимость среднего времени ожидания от нагрузки

4.2. Многоканальные СМО с однородным потоком заявок

«Работая над решением задачи, всегда полезно знать ответ» (***Правило точности***)

Рассмотрим многоканальную СМО с K идентичными обслуживающими приборами и накопителем неограниченной ёмкости, в которую поступают заявки, образующие простейший поток с интенсивностью λ (рис.4.3). Длительность τ_b обслуживания заявок распределена по экспоненциальному закону со средним значением b. Выбор заявок из очереди на обслуживание осуществляется в соответствии с бесприоритетной дисциплиной обслуживания в порядке поступления (ОПП) по правилу «первым пришёл – первым обслужен» (FIFO – First In First Out).

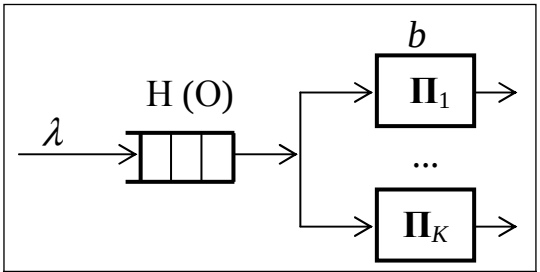


Рис.4.3. Многоканальная СМО

4.2.1. Характеристики многоканальной СМО M/M/K

В качестве основной характеристики функционирования СМО, будем использовать среднее время ожидания w заявок.

Точный метод расчета характеристик обслуживания заявок в многоканальной СМО разработан при следующих предположениях:

- поток заявок – *простейший*;
- длительность обслуживания заявок распределена по *экспоненциальному* закону со средним значением b;
- все K приборов – *идентичны*, и любая заявка может быть обслужена любым прибором;

- ёмкость накопителя – *не ограничена*;
- в системе *отсутствуют перегрузки*, то есть загрузка системы меньше 1: $\rho = \frac{\lambda b}{K} < 1$

При этих предположениях среднее время ожидания заявок определяется следующим образом:

$$w = \frac{Pb}{K(1-\rho)}\ , \qquad (4.8)$$

где P – вероятность того, что все K приборов заняты обслуживанием заявок.

Вероятность P определяется как:

$$P = \frac{(K\rho)^K}{K!(1-\rho)} P_0,$$

где P_0 – вероятность простоя многоканальной СМО, то есть вероятность того, что в системе нет заявок:

$$P_0 = \left[\frac{(K\rho)^K}{K!(1-\rho)} + \sum_{i=0}^{K-1} \frac{(K\rho)^i}{i!} \right]^{-1}.$$

4.2.2. Анализ свойств многоканальной СМО

Анализ свойств многоканальной СМО с однородным потоком заявок и накопителем неограниченной ёмкости может быть выполнен с использованием представленных выше математических моделей, определяющих зависимости характеристик обслуживания заявок от параметров поступления и обслуживания заявок для установившегося (стационарного) режима работы системы.

1. На рис.4.4 показан характер зависимости среднего времени ожидания w и среднего времени пребывания u заявок в системе от числа обслуживающих приборов K. Очевидно, что *с увеличением числа обслуживающих приборов времена ожидания и пребывания заявок уменьшаются*, при этом в пределе при $K \to \infty$ время ожидания стремится к нулю, а время пребывания достигает своего наименьшего значения, равного длительности обслуживания заявок.

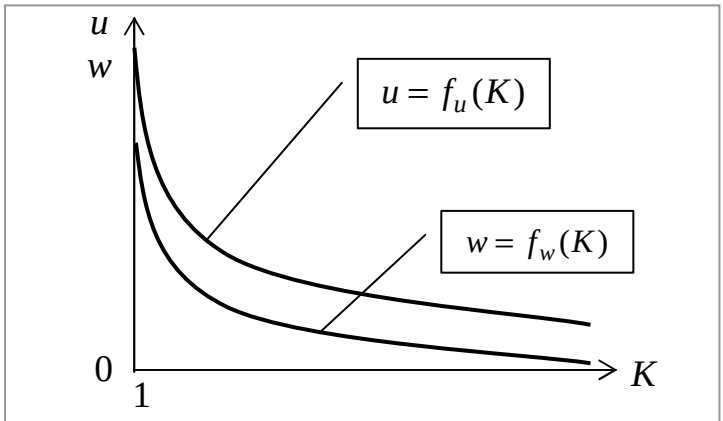


Рис.4.4. Зависимости времени ожидания и пребывания заявок от числа приборов

2. На рис.4.5 показаны аналогичные зависимости, но при условии, что при увеличении числа обслуживающих приборов K их суммарная производительность (скорость работы) остается постоянной, т.е. $V_\Sigma = KV_K = const$, где V_K – производительность одного прибора при наличии в системе K обслуживающих приборов. Из представленных графиков видно, что среднее время ожидания w заявок, как и в предыдущем случае, уменьшается с увеличением числа приборов, однако время пребывания u заявок в системе увеличивается. Последнее объясняется тем, что с увеличением числа приборов K производительность каждого из них для сохранения суммарной производительности системы уменьшается пропорционально K и, следовательно, линейно увеличивается длительность обслуживания заявки в приборе. При этом скорость увеличения длительности обслуживания больше скорости уменьшения времени ожидания, что в сумме приводит к увеличению времени пребывания заявок в системе. В пределе при $K \to \infty$ время пребывания заявок асимптотически стремится к длительности обслуживания заявок.

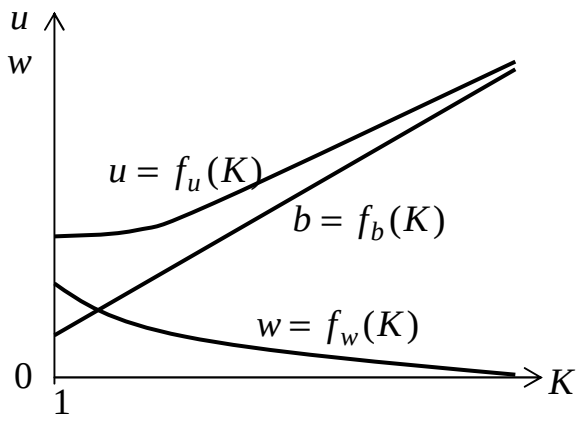


Рис.4.5. Зависимость времени пребывания заявок от числа приборов при $V_\Sigma = const$

Таким образом, при проектировании систем обслуживания следует иметь в виду, что с точки зрения задержек (времени пребывания заявок) более эффективной является одноканальная система, чем многоканальная, при равенстве суммарной производительности. Основным достоинством многоканальной системы является более высокая надёжность, проявляющаяся в том, что при выходе из строя одного или даже нескольких обслуживающих приборов система продолжает функционировать, хотя и с меньшей эффективностью, что заключается в увеличении времени пребывания заявок в системе.

3. Можно показать, что среднее время ожидания заявок, как и для одноканальных систем, существенно зависит от нагрузки y (загрузки ρ) системы. При $y \geq K$ ($\rho \to 1$) время ожидания заявок возрастает неограниченно: $w \to \infty$, то есть заявки могут ожидать обслуживания сколь угодно долго.

4.3. Одноканальные СМО с неоднородным потоком заявок

«Никогда не ставьте задачу, решение которой вам неизвестно» (***Правило Берке***)

Рассмотрим одноканальную СМО с неоднородным потоком заявок, в которую поступают H классов заявок, образующие простейшие потоки с

интенсивностями $\lambda_1,\ldots,\lambda_H$. Длительность τ_{b_k} обслуживания заявок класса k распределена по произвольному закону со средним значением b_k и коэффициентом вариации ν_{b_k}. Выбор заявок из очереди на обслуживание осуществляется в соответствии с заданной дисциплиной обслуживания, в качестве которой будем рассматривать:

- дисциплину обслуживания бесприоритетную (ДО БП), при которой заявки выбираются на обслуживание в порядке поступления;
- дисциплину обслуживания заявок с относительными приоритетами (ДО ОП);
- дисциплину обслуживания заявок с абсолютными приоритетами (ДО АП).

В качестве основной характеристики, описывающей эффективность функционирования системы, будем рассматривать средние времена ожидания заявок разных классов, на основе которых легко могут быть рассчитаны все остальные характеристики с использованием фундаментальных зависимостей, представленных в разделе 3 (п.3.3.5).

При этом следует иметь в виду, что представленные ниже формулы были получены при следующих предположениях:

1) СМО содержит *один обслуживающий прибор*, который в каждый момент времени может обслуживать только одну заявку;

2) СМО имеет *накопитель заявок неограниченной ёмкости*, что означает отсутствие отказов поступающим заявкам при их постановке в очередь, то есть любая поступающая заявка всегда найдёт в накопителе место для ожидания независимо от того, сколько заявок уже находится в очереди;

3) заявки разных классов, поступающие в СМО независимо друг от друга, образуют *простейшие потоки*;

4) длительности обслуживания заявок каждого класса в приборе распределены по *произвольному закону* и не зависят друг от друга;

5) обслуживающий прибор не простаивает, если в системе (накопителе) имеется хотя бы одна заявка любого класса, причем после завершения обслуживания очередной заявки *мгновенно* из накопителя выбирается следующая заявка в соответствии с заданной дисциплиной обслуживания;

6) при использовании ДО БП заявки разных классов выбираются на обслуживание только в зависимости от времени поступления в систему по правилу «раньше пришел – раньше обслужен», независимо от номера класса, к которому принадлежит заявка;

7) при использовании приоритетных дисциплин (ДО ОП и ДО АП) приоритеты классам заявок назначены по принципу «*класс с меньшим номером имеет более высокий приоритет*», то есть наивысшим приоритетом обладают заявки класса 1;

8) в случае ДО АП заявка, обслуживание которой прервано более высокоприоритетной заявкой, *возвращается в накопитель*, где ожидает дальнейшего обслуживания, причем ее обслуживание продолжается *с прерванного места*.

4.3.1. Характеристики и свойства ДО БП

При бесприоритетной ДО средние времена ожидания одинаковы для всех классов заявок и определяются по следующей формуле:

$$w_k^{\text{БП}} = w^{\text{БП}} = \frac{\sum_{i=1}^{H} \lambda_i b_i^2 (1+\nu_{b_i}^2)}{2(1-R)} \qquad (k=1,\ldots,H), \tag{4.9}$$

где $R = \sum_{i=1}^{H} \rho_i = \sum_{i=1}^{H} \lambda_i b_i$ – суммарная загрузка системы.

Выражение (4.5) получено в предположении, что в системе существует стационарный режим и отсутствует перегрузка: $R < 1$.

Анализ представленной аналитической зависимости (4.5) позволяет выявить **свойства ДО БП** и сформулировать следующие выводы**.**

1. *Среднее время ожидания заявок разных классов* при использовании ДО БП *одинаково* при любых интенсивностях поступления $\lambda_1,\ldots,\lambda_H$ и законах распределений $B_1(\tau),\ldots,B_H(\tau)$ длительностей обслуживания заявок: $w_k^{\text{БП}} = w^{\text{БП}}$ для всех $k=1,\ldots,H$. Отметим, что средние времена пребывания в системе заявок разных классов, в общем случае, различны, так как различны длительности обслуживания: $u_k^{\text{БП}} = w^{\text{БП}} + b_k \ \ (k=1,\ldots,H)$.

2. *Среднее время ожидания заявок* в очереди минимально при постоянной (детерминированной) длительности обслуживания заявок каждого класса, когда коэффициент вариации длительности обслуживания $\nu_{b_k} = 0$, и *увеличивается с ростом коэффициента вариации (дисперсии) длительности обслуживания*. Заметим, что зависимость среднего времени ожидания от коэффициента вариации ν_{b_k} носит нелинейный характер. Так, например, при экспоненциально распределенной длительности обслуживания, когда $\nu_{b_k} = 1$, среднее время ожидания заявок увеличивается в 2 раза, а при $\nu_{b_k} = 2$ – в 5 раз, по сравнению с детерминированным обслуживанием.

3. Среднее время ожидания заявок существенно зависит от суммарной нагрузки *Y* (загрузки *R*) системы (рис.4.6,а). При $Y \geq 1 \ \ (R \to 1)$ *время ожидания* заявок всех классов *возрастает неограниченно*: $w^{\text{БП}} \to \infty$, то есть заявки могут ожидать обслуживания сколь угодно долго. Отметим, что увеличение суммарной нагрузки может быть

обусловлено двумя факторами: увеличением интенсивностей поступления в систему заявок разных классов или увеличением длительности обслуживания заявок (например, за счет уменьшения скорости работы обслуживающего прибора).

Зависимость среднего времени пребывания в системе заявок разных классов от суммарной нагрузки аналогична зависимости времени ожидания (рис.4.6,б). Единственное отличие состоит в том, что *средние времена пребывания в системе заявок разных классов, в общем случае, различны*, то есть $u_i \neq u_j \quad (i \neq j)$, поэтому на графике, в отличие от времени ожидания, могут отображаться несколько зависимостей. Это различие обусловлено различием длительностей обслуживания заявок разных классов.

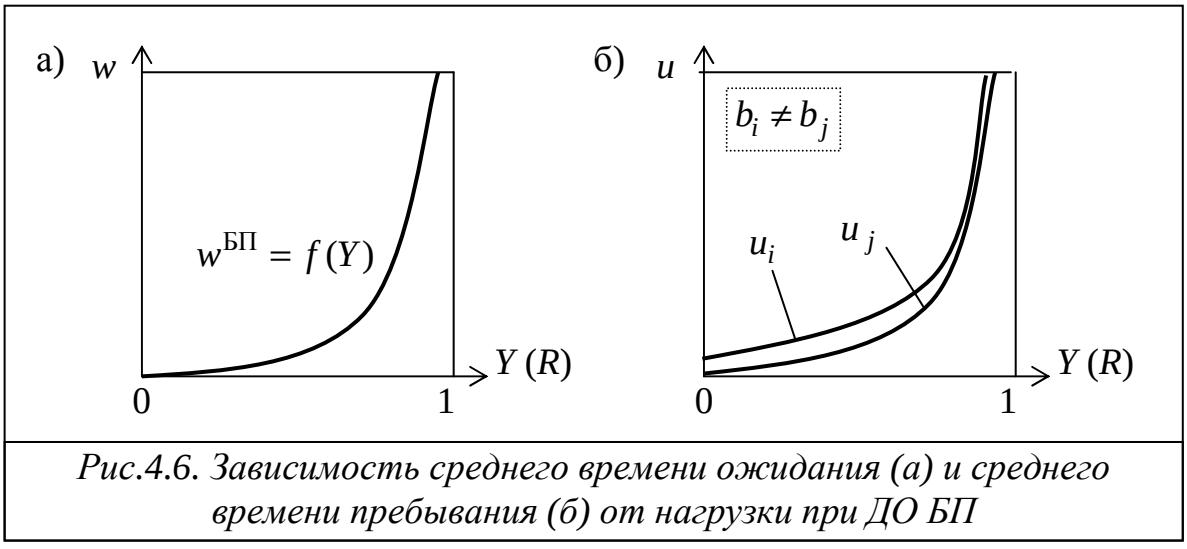


Рис.4.6. Зависимость среднего времени ожидания (а) и среднего времени пребывания (б) от нагрузки при ДО БП

Аналогично, на графиках, отображающих зависимости средних длин очередей и числа заявок в системе от суммарной нагрузки, в общем случае, будут изображаться несколько кривых, соответствующих разным классам заявок. Отметим, что *средние длины очередей заявок разных классов*, несмотря на одинаковое время ожидания, *в общем случае*, *различны* и, в соответствии с формулой Литтла ($l_i = \lambda_i w^{БП}$), совпадают только в случае равенства интенсивностей поступления заявок разных классов в систему.

4. Можно показать, что *для бесприоритетной дисциплины обслуживания в обратном порядке (ООП)*, когда заявки на обслуживание выбираются по правилу «последний пришёл – первый обслужен», *средние времена ожидания заявок будут такими же,* как и при обслуживании в порядке поступления (ОПП), *но дисперсия* времени ожидания *будет больше*. Это обусловлено тем, что заявки, поступившие последними, будут ожидать незначительное время, в то время как заявки, попавшие в начало очереди, могут ожидать обслуживания достаточно долго, что обусловливает большой разброс значений времени ожидания.

5. Аналитическое исследование дисциплины обслуживания в циклическом порядке (ДО ЦП) достаточно сложно и связано с громоздкими математическими выкладками. Поэтому, не выписывая

громоздких формул, отметим лишь наиболее характерные особенности, присущие этой ДО.

Для дисциплины обслуживания в циклическом порядке среднее время ожидания заявок разных классов в общем случае не одинаково (рис.4.7).

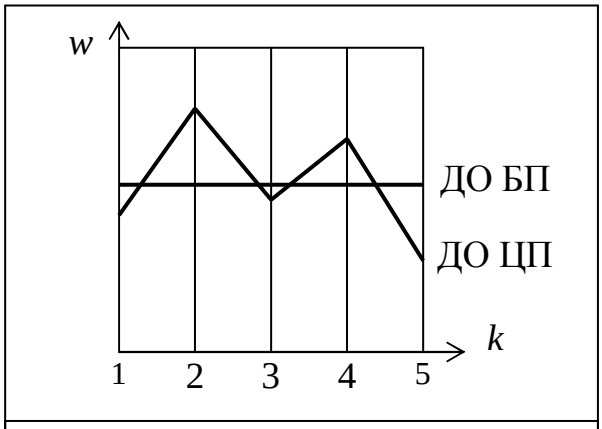


Рис.4.7. Зависимость времени ожидания от номера класса заявок

Это различие зависит от соотношения параметров потоков $(\lambda_1,\ldots,\lambda_H)$ и обслуживания $(B_1(\tau),\ldots,B_H(\tau))$ заявок разных классов. В некоторых случаях ДО ЦП позволяет обеспечить меньшую суммарную длину очереди заявок, чем ДО БП. Зависимость среднего времени ожидания заявок каждого класса от суммарной нагрузки Y имеет такой же вид, как и для ДО БП (рис.4.6).

4.3.2. Характеристики и свойства ДО ОП

Приоритеты называются ***относительными***, если они учитываются только в момент выбора заявки на обслуживание и не сказываются на работе системы в период обслуживания заявки любого класса (приоритета). Относительность приоритета связана со следующим. После завершения обслуживания какой-либо заявки из очереди на обслуживание выбирается заявка класса с наиболее высоким приоритетом, поступившая ранее других заявок этого класса (такого же приоритета). Если в процессе её обслуживания в систему поступят заявки с более высоким приоритетом, то обслуживание рассматриваемой заявки не будет прекращено, то есть эта заявка, захватив прибор, оказывается как бы более приоритетной. Таким образом, приоритет *относителен* в том смысле, что он имеет место лишь в момент выбора заявок на обслуживание и отсутствует, если прибор занят обслуживанием какой-либо заявки.

Введение относительных приоритетов (ОП) позволяет уменьшить по сравнению с ДО БП время ожидания высокоприоритетных заявок.

При описании свойств для определённости будем полагать, что относительные приоритеты назначены по правилу: «более высокий приоритет – классу заявок с меньшим номером».

Для ДО ОП среднее время ожидания заявок класса k определяется по следующей формуле:

$$w_k^{\text{ОП}} = \frac{\sum_{i=1}^{H} \lambda_i b_i^{\,2}(1+\nu_{b_i}^{\,2})}{2(1-R_{k-1})(1-R_k)} \qquad (k=1,\ldots,H), \tag{4.10}$$

где R_{k-1} и R_k – суммарные загрузки, создаваемые заявками, которые имеют приоритет не ниже $(k-1)$ и k соответственно:

$$R_{k-1} = \sum_{i=1}^{k-1} \rho_i; \qquad R_k = \sum_{i=1}^{k} \rho_i . \tag{4.11}$$

Анализ представленной зависимости (4.10) позволяет выявить **свойства ДО ОП** и сформулировать следующие выводы.

1. Введение относительных приоритетов по сравнению с ДО БП приводит к *уменьшению времени ожидания высокоприоритетных заявок* первого класса и к *увеличению времени ожидания низкоприоритетных заявок* класса H: $w_1^{\text{ОП}} < w_1^{\text{БП}}$ и $w_H^{\text{ОП}} > w_H^{\text{БП}}$.

2. При использовании ДО ОП *средние времена ожидания заявок монотонно увеличиваются с уменьшением приоритета* при любых интенсивностях поступления $\lambda_1, \ldots, \lambda_H$ и законах распределения $B_1(\tau), \ldots, B_H(\tau)$ длительностей обслуживания: $w_1^{\text{ОП}} < w_2^{\text{ОП}} < \ldots < w_H^{\text{ОП}}$.

Отметим, что для средних времён пребывания заявок разных классов последнее соотношение, в общем случае, может и не выполняться.

Свойства, сформулированные выше, иллюстрируются рис.4.8,а, показывающим характер зависимости среднего времени ожидания заявок w_k от номера класса k при использовании ДО БП и ДО ОП.

3. На рис.4.8,б показаны зависимости среднего времени ожидания заявок разных классов от суммарной нагрузки Y системы при использовании ДО ОП. Здесь же для сравнения приведена аналогичная зависимость для ДО БП (штриховая линия). Характер зависимостей свидетельствует о том, что для ДО ОП при $Y \to 1$ резко увеличивается время ожидания заявок низкоприоритетных классов, в то время как для высокоприоритетных заявок это увеличение незначительно. Более того, для высокоприоритетных заявок обеспечивается достаточно хорошее качество обслуживания, то есть небольшое время ожидания даже при возникновении перегрузок, когда суммарная нагрузка становится больше единицы: $Y \geq 1$. Это свойство, называемое ***защитой от перегрузок***, обеспечивается за счет отказа в обслуживании низкоприоритетным заявкам, время ожидания которых при этом резко возрастает. При ДО БП защита от перегрузок *отсутствует* для всех классов заявок.

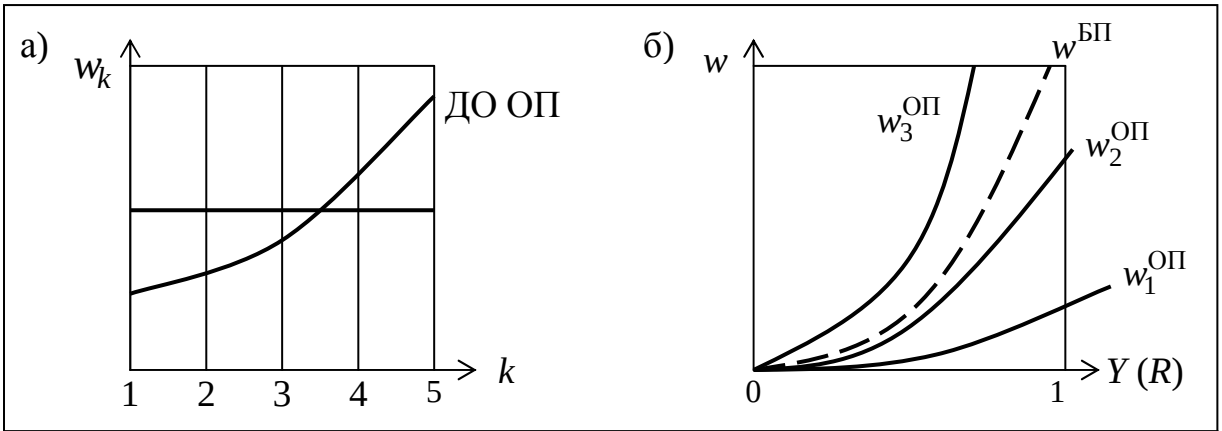


Рис.4.8. Зависимости среднего времени ожидания от номера класса (а) и от нагрузки (б) для ДО ОП и ДО БП

4. Рассмотрим более детально свойство защиты от перегрузок при ДО ОП, для чего построим зависимости среднего времени ожидания заявок трех классов при значительном росте нагрузки Y (рис.4.9).

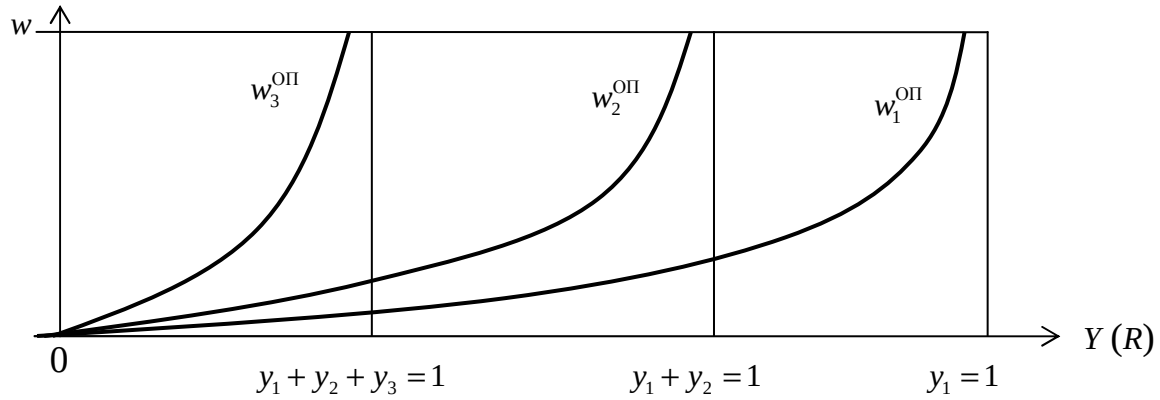


Рис.4.9. Зависимости среднего времени ожидания заявок разных классов от нагрузки

При достижении суммарной нагрузки, создаваемой заявками всех трех классов, значения 1 ($y_1 + y_2 + y_3 = 1$) время ожидания заявок 3-го класса устремляется в бесконечность, что означает отказ в обслуживании, при этом заявки классов 1 и 2 продолжают обслуживаться и имеют конечное время ожидания. Дальнейшее увеличение нагрузки приводит к отказу в обслуживании заявок второго класса при $y_1 + y_2 = 1$, то есть когда создаваемая заявками 1-го и 2-го классов нагрузка достигнет значения 1. Заявки первого класса получат отказ в обслуживании при $y_1 = 1$. Таким образом, в отличие от ДО БП при ДО ОП система полностью перестаёт обслуживать заявки, то есть функционировать, только в том случае, если нагрузка, создаваемая заявками самого высокоприоритетного (первого) класса, достигнет значения 1.

4.3.3. Характеристики и свойства ДО АП

Иногда время ожидания заявок некоторых классов необходимо уменьшить в такой степени, которая недостижима при использовании ДО ОП. Можно предположить, что время ожидания уменьшится, если при поступлении высокоприоритетной заявки обслуживание ранее поступившей заявки с низким приоритетом прерывается, и прибор переходит к обслуживанию высокоприоритетной заявки. Приоритет, прерывающий обслуживание низкоприоритетной заявки, называется ***абсолютным***, а соответствующая дисциплина – ***дисциплиной обслуживания с абсолютными приоритетами*** (ДО АП).

Прерванная заявка может быть потеряна или возвращена в накопитель, где она будет ожидать дальнейшего обслуживания. В последнем случае возможны два варианта продолжения обслуживания прерванной заявки:

- обслуживание с начала, то есть прерванная заявка будет обслуживаться заново с самого начала;
- дообслуживание, когда обслуживание прерванной заявки в приборе будет выполняться с прерванного места.

В дальнейшем, если не оговорено иное, будем предполагать *дообслуживание* прерванной заявки.

Для ДО АП среднее время ожидания заявок класса k определяется по следующей формуле:

$$w_k^{АП} = \frac{\sum_{i=1}^{k} \lambda_i b_i (1+\nu_{b_i}^2)}{2(1-R_{k-1})(1-R_k)} + \frac{R_{k-1} b_k}{1-R_{k-1}} \qquad (k=1,...,H) \qquad (4.12)$$

где R_{k-1} и R_k – суммарные загрузки, создаваемые заявками, которые имеют приоритет не ниже $(k-1)$ и k соответственно, и определяемые по формулам (4.11).

Анализ выражения (4.12) для расчета среднего времени ожидания заявок при использовании ДО АП и его сопоставление с аналогичным выражением (4.10) для ДО ОП позволяет выявить **свойства ДО АП** и сформулировать следующие выводы.

1. Выражение (4.12) содержит два слагаемых: $w_k^{\text{АП}} = s_k + z_k$, отображающих среднее ***время ожидания начала обслуживания*** s_k и среднее ***время ожидания в прерванном состоянии*** z_k соответственно:

$$s_k = \frac{\sum_{i=1}^{k} \lambda_i b_i (1+\nu_{b_i}^2)}{2(1-R_{k-1})(1-R_k)}, \qquad z_k = \frac{R_{k-1} b_k}{1-R_{k-1}} \qquad (k=1,...,H).$$

2. *Время ожидания заявок класса* k зависит только от значений параметров классов $1,\ldots,k$ заявок, имеющих более высокий или такой же приоритет, и *не зависит от параметров классов заявок* $k+1,\ldots,H$, имеющих более низкий приоритет.

3. *Для заявок класса* 1, имеющих самый высокий абсолютный приоритет, *обеспечивается минимально возможное время ожидания* по сравнению со всеми другими ДО, то есть при любой другой ДО среднее время ожидания заявок первого класса не может быть меньше, чем при ДО АП. Это объясняется тем, что в случае ДО АП заявки первого класса обслуживаются как бы в изоляции, независимо от заявок других классов.

4. Времена ожидания начала обслуживания s_k монотонно увеличиваются с уменьшением приоритета: $s_1 < s_2 < \ldots < s_H$, однако время ожидания высокоприоритетной заявки в прерванном состоянии z_k может оказаться больше времени ожидания z_{k+1} заявки с более низким приоритетом, если длительности обслуживания связаны соотношением $b_k >> b_{k+1}$, так как количество прерываний заявками более высокого

приоритета и, следовательно, время ожидания в прерванном состоянии прямо пропорционально зависит от длительности обслуживания заявок данного класса. Вследствие этого, *полное время ожидания заявок высокоприоритетного класса*, складывающееся из времени ожидания начала обслуживания и времени ожидания в прерванном состоянии, *может оказаться больше, чем у заявок класса с низким приоритетом*: $w_k^{\text{АП}} >> w_{k+1}^{\text{АП}}$. Очевидно, что $w_1^{\text{АП}} < w_2^{\text{АП}} < \ldots < w_H^{\text{АП}}$, если длительности обслуживания заявок разных классов связаны соотношением $b_1 \le b_2 \le \ldots \le b_H$.

5. Введение АП по сравнению с ОП приводит к уменьшению среднего времени ожидания самых высокоприоритетных заявок первого класса и к его увеличению для заявок класса *H*: $w_1^{\text{АП}} < w_1^{\text{ОП}}$ и $w_H^{\text{АП}} > w_H^{\text{ОП}}$.

Два последних результата иллюстрируются рис.4.10,а. Для ДО АП пунктиром показан случай, когда $w_3^{\text{АП}} >> w_4^{\text{АП}}$, из чего следует, что $b_3 >> b_4$.

Зависимость полного времени ожидания от суммарной нагрузки *Y* системы при использовании ДО АП аналогична зависимости для ДО ОП (см. рис.4.10,б) с тем лишь отличием, что при ДО АП высокоприоритетные заявки лучше защищены от перегрузок.

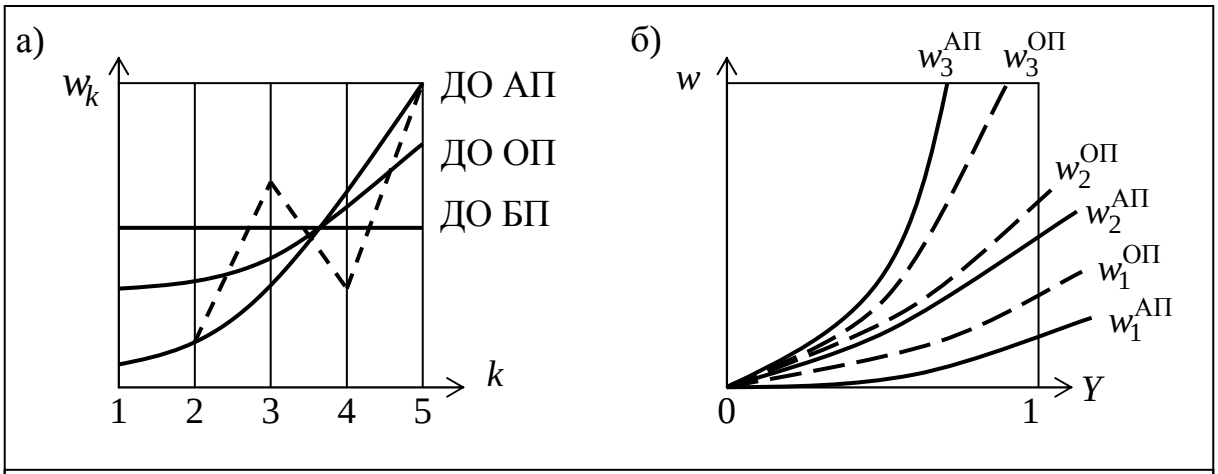


Рис.4.10. Зависимости среднего времени ожидания от номера класса (а) и от нагрузки (б) для ДО АП и ДО ОП

4.3.4. Законы сохранения

«Число законов стремится заполнить все доступное для публикации пространство» (***Закон Диджиованни***)

Изменение ДО позволяет уменьшить время ожидания высокоприоритетных заявок за счет увеличения времени ожидания низкоприоритетных заявок. Очевидно, что за счет изменения ДО нельзя добиться того, чтобы уменьшилось или увеличилось время ожидания заявок всех классов. Этот факт сформулирован в виде закона сохранения времени ожидания.

Формулировка закона сохранения времени ожидания. *Для любой дисциплины обслуживания* (ДО)

$$\sum_{i=1}^{H} \rho_i w_i = \underset{\text{ДО}}{Const}, \qquad (4.13)$$

то есть сумма произведений загрузок ρ_i на среднее время ожидания w_i $(i = \overline{1, H})$ заявок всех классов инвариантна относительно ДО.

Закон сохранения времени ожидания выполняется при следующих условиях:

- система без потерь – все заявки на обслуживание удовлетворяются;
- система простаивает лишь в том случае, когда в ней нет заявок;
- при наличии прерываний длительность обслуживания прерванных заявок распределена по экспоненциальному закону;
- все поступающие потоки заявок – простейшие, и длительности обслуживания не зависят от интенсивностей потоков заявок.

Значение константы в законе сохранения можно определить следующим образом. Поскольку закон сохранения справедлив для любых ДО, удовлетворяющих перечисленным условиям, то он справедлив и для ДО БП, для которой $w_k^{БП} = w^{БП}$ для всех (k = 1, …, H). Отсюда находим значение константы:

$$Const = w^{БП} \sum_{i=1}^{H} \rho_i = R\, w^{БП} .$$

Подставив полученное значение константы и формулу (4.9) для расчёта $w^{БП}$ в закон сохранения, окончательно получим:

$$\sum_{i=1}^{H} \rho_i w_i = \frac{R \sum_{i=1}^{H} \lambda_i b_i^{\,2} (1 + \nu_{bi}^{\,2})}{2(1 - R)} . \qquad (4.14)$$

Закон сохранения времени ожидания универсален и справедлив для всех ДО, удовлетворяющих указанным условиям. Его можно использовать для оценки достоверности приближенных результатов, полученных при исследовании сложных ДО и проведении имитационного моделирования, а также при решении задач синтеза.

Модификация закона сохранения. Закон сохранения может быть модифицирован применительно ко времени пребывания заявок в системе с учетом того, что $w_i = u_i - b_i$. Подставив это выражение в закон сохранения времени ожидания (4.14) после некоторых преобразований, *получим закон сохранения времени пребывания*:

$$\sum_{i=1}^{H} \rho_i u_i = \frac{R \sum_{i=1}^{H} \lambda_i b_i^{\,2} (1 + \nu_{bi}^{\,2})}{2(1 - R)} + \sum_{i=1}^{H} \rho_i b_i . \qquad (4.15)$$

Заметим, что изменение ДО приводит только к изменению времени ожидания и времени пребывания, а остальные величины, входящие в выражения (4.14) и (4.15), не изменяются.

Рассмотрим случай, когда средние длительности обслуживания заявок разных классов одинаковы: $b_i = b = \text{const}$ для всех $i = \overline{1, H}$. Тогда выражение (4.14) может быть преобразовано следующим образом:

$$\sum_{i=1}^{H} \lambda_i b\, w_i = b \sum_{i=1}^{H} \lambda_i\, w_i = b \sum_{i=1}^{H} l_i = b\, L = \underset{\text{ДО}}{Const},$$

откуда получим новую формулировку закона сохранения в виде **закона сохранения суммарной длины очереди заявок**:

$$\sum_{i=1}^{H} \lambda_i\, w_i = L = \underset{\text{ДО}}{Const}.$$

Таким образом, если средние длительности обслуживания заявок разных классов одинаковы, то изменение ДО не приводит к изменению суммарной длины L очередей заявок всех классов, которая остается постоянной. В то же время длины очередей l_i ($i = \overline{1, H}$) заявок каждого класса меняются с изменением ДО.

4.4. Разомкнутые экспоненциальные СеМО с однородным потоком заявок

«Чем сложнее и грандиознее план, тем больше шансов, что он провалится»
(***Производная от закона Мэрфи***)

4.4.1. Описание разомкнутых СеМО

Рассмотрим разомкнутую экспоненциальную сеть массового обслуживания (СеМО) с однородным потоком заявок при следующих предположениях:

1) разомкнутая СеМО (РСеМО) *произвольной топологии* содержит *n* узлов;

2) после завершения обслуживания в каком-либо узле передача заявки в другой узел происходит *мгновенно*;

3) в качестве узлов могут быть как *одноканальные*, так и *многоканальные* СМО;

4) все приборы многоканального узла являются *идентичными*, и любая заявка может обслуживаться любым прибором;

5) заявка, поступившая в многоканальный узел, когда все или несколько приборов свободны, направляется случайным образом в *любой свободный прибор*;

6) в каждом узле РСеМО имеется накопитель заявок *неограниченной ёмкости*, что означает отсутствие отказов поступающим заявкам при их постановке в очередь, то есть любая поступающая в узел заявка всегда

найдет в накопителе место для ожидания независимо от того, сколько заявок уже находится в очереди;

7) заявки поступают в РСеМО из *внешнего независимого* источника и образуют *простейший поток* заявок;

8) длительности обслуживания заявок во всех узлах сети представляют собой случайные величины, распределенные по *экспоненциальному закону*;

9) обслуживающий прибор любого узла *не простаивает*, если в его накопителе имеется хотя бы одна заявка, причем после завершения обслуживания очередной заявки *мгновенно* из накопителя выбирается следующая заявка;

10) в каждом узле сети заявки из накопителя выбираются в соответствии с *бесприоритетной* дисциплиной обслуживания в порядке поступления (ОПП) по правилу «первым пришел – первым обслужен» (FIFO – First In First Out).

Для описания *линейных разомкнутых однородных экспоненциальных* СеМО необходимо задать следующую совокупность **параметров**:

- *число узлов* в сети: n;
- *число обслуживающих приборов* в узлах сети: $K_1,...,K_n$;
- *матрицу вероятностей передач*: $\mathbf{P} = [\, p_{ij} \mid i, j = 0, 1, \ldots, n]$, где вероятности передач p_{ij} должны удовлетворять условию (3.23): сумма элементов каждой строки должна быть равна 1;
- *интенсивность* λ_0 источника заявок, поступающих в РСеМО;
- средние *длительности обслуживания* заявок в узлах сети: $b_1,\ldots,b_n$.

На основе перечисленных параметров могут быть рассчитаны узловые и сетевые характеристики, описывающие эффективность функционирования соответственно узлов и РСеМО в целом.

Расчет характеристик функционирования *линейных разомкнутых однородных экспоненциальных* СеМО базируется на эквивалентном преобразовании сети и проводится в четыре этапа:

- расчет коэффициентов передач α_j и интенсивностей потоков заявок λ_j в узлах $j = \overline{1,n}$ СеМО;
- проверка условия отсутствия перегрузок в СеМО;
- расчет узловых характеристик;
- расчет сетевых характеристик.

4.4.2. Расчет коэффициентов передач и интенсивностей потоков заявок в узлах РСеМО

Покажем, что интенсивности $\lambda_0,\ldots,\lambda_n$ потоков заявок, поступающих в узлы $0,...,n$ сети, однозначно определяются вероятностями передач p_{ij} $(i, j = 1,\ldots,n)$, задающими маршруты заявок в СеМО.

Будем рассматривать только установившийся режим.

Так как в линейной СеМО заявки не размножаются и не теряются, то интенсивности входящего и выходящего потоков для любого узла будут равны между собой.

Интенсивность потока заявок, входящих в любой узел j сети, равна сумме интенсивностей потоков заявок, поступающих в него из других узлов $i = \overline{0,n}$ (рис.4.11). Поскольку заявки из узла i поступают в узел j с вероятностью p_{ij}, то интенсивность потока заявок, поступающих из i в j, равна $p_{ij}\lambda_i$, где λ_i - интенсивность выходящего и, следовательно, входящего потока заявок узла i. С учетом этого, на входе узла j имеется поток с интенсивностью

$$\lambda_j = \sum_{i=0}^{n} p_{ij}\lambda_i \quad (i = 0,1,\ldots,n). \tag{4.16}$$

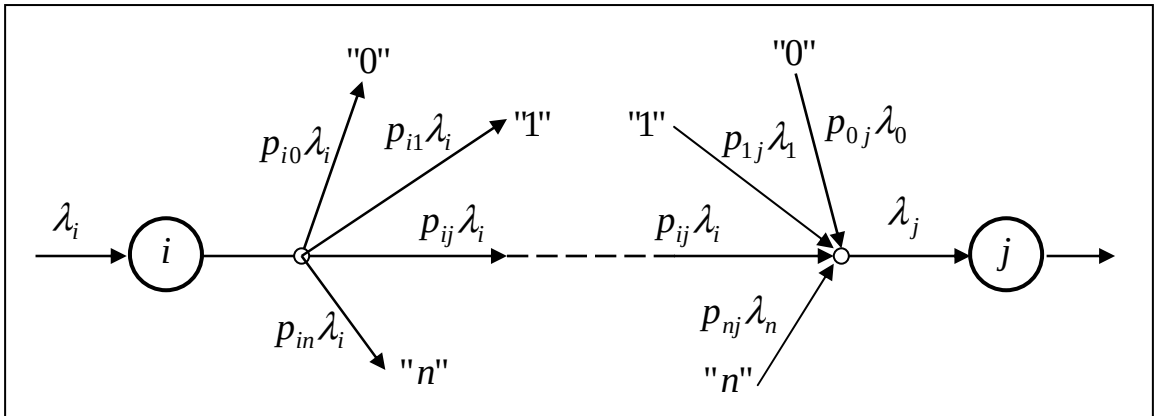


Рис.4.11. К расчёту интенсивностей потоков заявок в узлах РСеМО

Выражение (4.16) представляет собой *систему линейных алгебраических уравнений* $(n+1)$-го порядка, из которой могут быть найдены интенсивности потоков заявок в виде соотношения $\lambda_j = \alpha_j\lambda_0$ $(j = \overline{1,n})$. Коэффициент α_j называется ***коэффициентом передачи*** и определяет среднее число попаданий заявки в узел j за время ее нахождения в сети, причем $\alpha_0 = 1$.

Для разомкнутой СеМО известна интенсивность источника заявок λ_0. Можно показать, что система уравнений для расчета интенсивностей имеет единственное решение вида $\lambda_j = \alpha_j\lambda_0$, где λ_0 – заданная величина.

4.4.3. Проверка условия отсутствия перегрузок в СеМО

В п.3.4.2 показано, что в разомкнутой СеМО отсутствуют перегрузки, если выполняется условие (3.25):

$$\lambda_0 < \min\left(\frac{K_1}{\alpha_1 b_1}, \frac{K_2}{\alpha_2 b_2}, ..., \frac{K_n}{\alpha_n b_n}\right).$$

Если указанное условие не выполняется, то, как следует из него, стационарный режим в разомкнутой СеМО может быть реализован одним из следующих способов:

- уменьшением интенсивности λ_0 внешнего источника заявок до значения, при котором это условие будет выполняться;
- увеличением количества обслуживающих приборов K_j в перегруженных узлах;
- уменьшением длительностей b_j обслуживания заявок в перегруженных узлах;
- уменьшением коэффициентов передач α_j в перегруженных узлах.

4.4.4. Расчет узловых характеристик РСеМО

Один и тот же объект, рассматриваемый на разных уровнях детализации, можно представить различными моделями массового обслуживания, характеристики которых одинаковы или отличаются на величину, не превосходящую заданной погрешности. При выполнении определенных условий такие модели легко преобразуются друг в друга.

Для сетевых моделей в виде разомкнутых и замкнутых СеМО могут использоваться два вида преобразований:

- эквивалентное преобразование;
- толерантное преобразование.

Две сетевые модели *эквивалентны*, если сравниваемые характеристики этих моделей не отличаются друг от друга.

Две сетевые модели *толерантны* (подобны), если значения определенных характеристик отличаются друг от друга на величину, не превосходящую заданную.

Использование свойств эквивалентных и толерантных моделей позволяет упростить расчет характеристик моделей путем замены сложных сетевых моделей более простыми. Эквивалентными могут быть сетевые модели одного типа (например, две замкнутые сети), толерантными — модели как одного, так и разных типов [11].

Расчет характеристик функционирования линейных разомкнутых однородных экспоненциальных СеМО базируется на эквивалентном преобразовании сети, заключающемся в представлении *разомкнутой СеМО с n узлами* в виде *n независимых экспоненциальных СМО типа*

M/M/N (простейший поток заявок, длительность обслуживания распределена по экспоненциальному закону, *N* обслуживающих приборов). При этом интенсивность входящего потока заявок в СМО, отображающую узел j ($j=\overline{1,n}$) сети, определяется из системы алгебраических уравнений (4.16) через интенсивность входящего в сеть потока и коэффициент передачи узла: $\lambda_j = \alpha_j \lambda_0$, а средняя длительность обслуживания заявок в СМО равна длительности обслуживания b_j заявок в соответствующем узле СеМО.

Характеристики всех *n* СМО (время ожидания заявок в очереди и пребывания в системе, длина очереди и число заявок в системе, среднее число занятых приборов и т.д.) представляют собой узловые характеристики СеМО.

Среднее время ожидания заявок в очереди может быть рассчитано с использованием выражения (4.8) для многоканальных СМО типа M/M/N или выражения (4.1) для одноканальных СМО типа M/M/1, остальные характеристики узла j ($j=\overline{1,n}$) – с использованием фундаментальных соотношений, представленных в п.3.4.3, а именно:

- нагрузка в узле *j*, показывающая среднее число занятых приборов: $y_j = \lambda_j b_j$;
- загрузка *узла j*: $\rho_j = \min(y_j / K_j; 1)$, где K_j – число обслуживающих приборов в узле *j*;
- коэффициент простоя узла: $\pi_j = 1 - \rho_j$;
- время пребывания заявок в узле: $u_j = w_j + b_j$;
- длина очереди заявок: $l_j = \lambda_j w_j$;
- число заявок в узле (в очереди и на обслуживании в приборе): $m_j = \lambda_j u_j$.

Рассчитанные таким образом характеристики отдельных СМО в точности соответствуют узловым характеристикам исходной СеМО, то есть в отношении своих характеристик модель массового обслуживания, представляющая собой совокупность независимых СМО (каждая СМО рассматривается независимо от других), строго эквивалентна исходной разомкнутой СеМО в целом.

4.4.5. Расчет сетевых характеристик РСеМО

Сетевые характеристики, описывающие эффективность функционирования СеМО в целом, рассчитываются на основе полученных значений узловых характеристик.

В состав сетевых характеристик входят:

- среднее число заявок, ожидающих обслуживания в сети, и среднее число заявок, находящихся в сети:

$$L = \sum_{j=1}^{n} l_j ; \qquad M = \sum_{j=1}^{n} m_j ,$$

где l_j – средняя длина очереди и m_j – среднее число заявок в узле *j*;

- среднее время ожидания и среднее время пребывания заявок в сети:

$$W = \sum_{j=1}^{n} \alpha_j w_j ; \qquad U = \sum_{j=1}^{n} \alpha_j u_j ,$$

где w_j и u_j – соответственно среднее время ожидания и среднее время пребывания заявок в узле *j*; α_j – коэффициент передачи для узла *j*, показывающий среднее число попаданий заявки в узел *j* за время ее нахождения в сети.

Пример 4.2. Проиллюстрируем изложенный метод расчета характеристик функционирования линейных разомкнутых однородных экспоненциальных СеМО на примере СеМО с четырьмя узлами ($n = 4$), граф которой представлен на рис.4.12. Связи между узлами СеМО описываются следующей матрицей вероятностей передач:

$$\mathbf{P} = \begin{array}{c|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0{,}1 & 0 & 0{,}2 & 0{,}7 & 0 \\ 2 & 0 & 0 & 0 & 0 & 1 \\ 3 & 0 & 0 & 0 & 0 & 1 \\ 4 & 0 & 1 & 0 & 0 & 0 \end{array} .$$

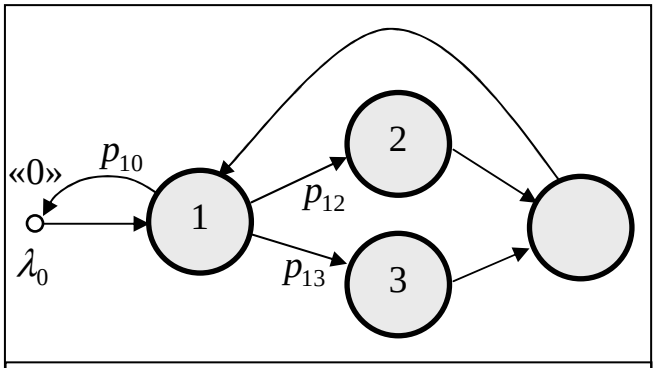


Рис.4.12. Граф разомкнутой СеМО

В РСеМО поступает простейший поток заявок с интенсивностью $\lambda_0 = 0{,}1\,\text{с}^{-1}$ Положим, что все узлы СеМО – одноканальные, а средние длительности обслуживания заявок в узлах соответственно равны: $b_1 = 0{,}8\,\text{с};\ b_2 = 2\,\text{с};\ b_3 = 0{,}4\,\text{с};\ b_4 = 0{,}3\,\text{с}.$

Система линейных алгебраических уравнений для расчёта интенсивностей потоков заявок в узлах СеМО, согласно (4.16), имеет вид:

$$\left.\begin{aligned} \lambda_0 &= p_{10}\,\lambda_1 = 0{,}1\,\lambda_1 \\ \lambda_1 &= p_{01}\lambda_0 + p_{41}\lambda_4 = \lambda_0 + \lambda_4 \\ \lambda_2 &= p_{12}\,\lambda_1 = 0{,}2\,\lambda_1 \\ \lambda_3 &= p_{13}\,\lambda_1 = 0{,}7\,\lambda_1 \\ \lambda_4 &= p_{24}\lambda_2 + p_{34}\lambda_3 = \lambda_2 + \lambda_3 \end{aligned}\right\} .$$

Решая эту систему уравнений, получим следующие значения интенсивностей: $\lambda_1 = 1\,\text{с}^{-1}$, $\lambda_2 = 0{,}2\,\text{с}^{-1}$, $\lambda_3 = 0{,}7\,\text{с}^{-1}$, $\lambda_4 = 0{,}9\,\text{с}^{-1}$. Тогда

коэффициенты передач будут равны: $\alpha_1 = \lambda_1 / \lambda_0 = 10$; $\alpha_2 = \lambda_2 / \lambda_0 = 2$; $\alpha_3 = \lambda_3 / \lambda_0 = 7$; $\alpha_4 = \lambda_4 / \lambda_0 = 9$.

Определим предельную интенсивность поступления заявок в разомкнутую СеМО, при которой в сети отсутствуют перегрузки. Для этого воспользуемся выражением (3.25), определяющим условие отсутствия перегрузок в РСеМО:

$$\lambda_0 < \min\left(\frac{K_1}{\alpha_1 b_1}, \frac{K_2}{\alpha_2 b_2}, \frac{K_3}{\alpha_3 b_3}, \frac{K_4}{\alpha_4 b_4}\right) = 0{,}125\ \text{с}^{-1}.$$

РСеМО работает без перегрузок, поскольку данное условие выполняется.

В соответствии с эквивалентным преобразованием представим рассматриваемую экспоненциальную разомкнутую СеМО в виде 4-х независимых СМО типа M/M/1, в которые поступают простейшие потоки заявок соответственно с интенсивностями: $\lambda_1 = 1\ \text{с}^{-1}$, $\lambda_2 = 0{,}2\ \text{с}^{-1}$, $\lambda_3 = 0{,}7\ \text{с}^{-1}$, $\lambda_4 = 0{,}9\ \text{с}^{-1}$, а средние длительности обслуживания заявок в СМО совпадают с длительностями обслуживания в соответствующих узлах СеМО: $b_1 = 0{,}8\ \text{с}$; $b_2 = 2\ \text{с}$; $b_3 = 0{,}4\ \text{с}$; $b_4 = 0{,}3\ \text{с}$.

Значения узловых характеристик СеМО, рассчитанные с использованием выражения (4.1) для среднего времени ожидания заявок в очереди СМО типа M/M/1 и фундаментальных соотношений, представленных в п.3.4.3, приведены в табл.4.1.

Таблица 4.1

Узловые характеристики	*Расчётные формулы*	*Узел 1*	*Узел 2*	*Узел 3*	*Узел 4*
Нагрузка	$y_j = \lambda_j b_j$	0,8	0,4	0,28	0,27
Загрузка	$\rho_j = \min(y_j / K_j; 1)$	0,8	0,4	0,28	0,27
Коэф-т простоя	$\pi_j = 1 - \rho_j$	0,2	0,6	0,72	0,73
Время ожидания	$w_j = \rho_j b_j / (1 - \rho_j)$	3,2	1,33	0,16	0,11
Время пребывания	$u_j = w_j + b_j$	4	3,33	0,56	0,41
Длина очереди	$l_j = \lambda_j w_j$	3,2	0,27	0,11	0,10
Число заявок в узле	$m_j = \lambda_j u_j$	4	0,67	0,39	0,37

В табл.4.2 представлены математические зависимости и полученные на их основе значения сетевых характеристик, рассчитанные с учётом найденных значений узловых характеристик.

Таблица 4.2

Сетевые характеристики	*Расчётные формулы*	*Значения*
Время ожидания в сети	$W = \sum_{j=1}^{n} \alpha_j w_j$	36,75
Время пребывания в сети	$U = \sum_{j=1}^{n} \alpha_j u_j$	54,25
Число заявок в состоянии ожидания	$L = \sum_{j=1}^{n} l_j$	3,68
Число заявок в сети	$M = \sum_{j=1}^{n} m_j$	5,43

4.4.6. Анализ свойств разомкнутых СеМО

Свойства разомкнутых СеМО определяются значениями узловых и сетевых характеристик, связанных между собой зависимостями, представленными в разделе 3. Наибольший интерес представляют свойства сети в целом, поскольку свойства отдельных узлов СеМО аналогичны свойствам соответствующих одноканальных и многоканальных СМО.

На рис. 4.13 показана зависимость основной сетевой характеристики РСеМО – среднего времени пребывания U заявок в сети от интенсивности λ_0 поступления заявок в сеть. Зависимость $U' = f'(\lambda_0)$ аналогична зависимости среднего времени пребывания заявок в СМО от загрузки системы, изменение которой может быть обусловлено, в частности, изменением интенсивности поступления заявок в СМО. Как и в СМО, имеется некоторое предельное значение интенсивности $\lambda'_{0\,max}$, при котором среднее время пребывания заявок в сети становится бесконечно большим, что свидетельствует о перегрузке в СеМО. Выше (см.п.3.4.2) показано, что в РСеМО отсутствуют перегрузки, если они отсутствуют во всех узлах сети, то есть перегрузка в разомкнутой СеМО наступает в том случае, когда загрузка одного из узлов сети становится равной единице. Такой узел называется «***узким местом***» и характеризуется тем, что очередь заявок перед ним со временем растёт до

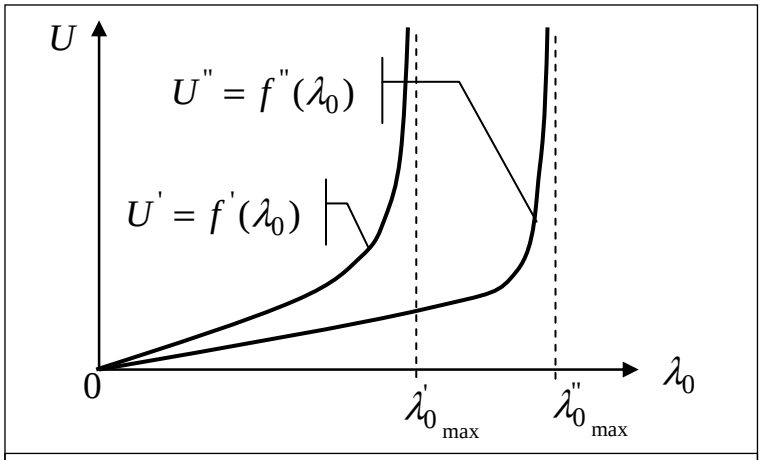


Рис.4.13. Время пребывания заявок в РСеМО

бесконечности и, как следствие, становится бесконечным число заявок в разомкнутой СеМО.

Для того чтобы избавиться в РСеМО от перегрузки, необходимо *разгрузить* «узкое место». Это может быть достигнуто следующими способами:

- увеличением скорости работы (быстродействия) обслуживающего прибора;
- увеличением числа обслуживающих приборов в узле.

Любой из этих способов позволяет увеличить производительность СеМО в целом и, как следствие, улучшить характеристики сети. Зависимость среднего времени пребывания U заявок в сети от интенсивности λ_0 поступления заявок в сеть принимает вид $U'' = f''(\lambda_0)$, то есть время пребывания заявок при одной и той же интенсивности λ_0 становится меньше (поскольку сеть имеет б*о*льшую производительность), а предельное значение интенсивности $\lambda''_{0\,\max}$, при котором наступает перегрузка СеМО, становится больше: $\lambda''_{0\,\max} > \lambda'_{0\,\max}$. При этом появляется новое узкое место в СеМО, и дальнейшее улучшение сети может быть достигнуто путём разгрузки нового узкого места. Очевидно, что если СеМО является моделью реальной технической системы, разгрузка узкого места за счёт увеличения скорости работы обслуживающего прибора или числа приборов в узле означает увеличение стоимости реальной системы.

Существует ещё один способ разгрузки узкого места СеМО, заключающийся в *уменьшении вероятности передачи* заявок к узлу, являющемуся узким местом. Этот способ часто используется в реальных системах и обычно не связан с увеличением стоимости системы. Например, в вычислительной системе изменение вероятностей передач к накопителям внешней памяти может быть достигнуто за счет перераспределения файлов между накопителями: наиболее часто используемые файлы, расположенные в наиболее загруженном накопителе, переносятся в наименее загруженный накопитель. При этом уменьшается количество обращений к загруженному накопителю (коэффициент передачи соответствующего узла СеМО).

Характер зависимостей других сетевых характеристик (времени ожидания, числа заявок в сети и в состоянии ожидания) разомкнутой СеМО от интенсивности поступления заявок аналогичен показанному на рис. 4.13.

Пример 4.3. Проиллюстрируем способы разгрузки узкого места и получаемый от этого эффект для четырёхузловой разомкнутой СеМО, рассмотренной в примере 4.2. Там же было показано, что интенсивность поступления заявок в разомкнутую СеМО, при которой в сети отсутствуют перегрузки, должна удовлетворять условию: $\lambda_0 < 0{,}125\ \text{с}^{-1}$.

1. Рассчитаем сначала характеристики РСеМО, работающей в области загрузок, близких к 1, для чего положим, что интенсивность потока поступающих в сеть заявок равна $\lambda_0 = 0{,}12\ \text{с}^{-1}$

Тогда интенсивности потоков заявок в узлы РСеМО соответственно будут равны: $\lambda_1 = \alpha_1\lambda_0 = 1{,}2\ \text{с}^{-1}$, $\lambda_2 = \alpha_2\lambda_0 = 0{,}24\ \text{с}^{-1}$, $\lambda_3 = \alpha_3\lambda_0 = 0{,}84\ \text{с}^{-1}$, $\lambda_4 = \alpha_4\lambda_0 = 1{,}08\ \text{с}^{-1}$, а средние длительности обслуживания заявок, как и ранее, будут равны: $b_1 = 0{,}8\,\text{с};\ b_2 = 2\,\text{с};\ b_3 = 0{,}4\,\text{с};\ b_4 = 0{,}3\,\text{с}$.

Рассчитанные значения узловых и сетевых характеристик СеМО приведены в табл.4.3.

Таблица 4.3

Характеристики	*Узел 1*	*Узел 2*	*Узел 3*	*Узел 4*	*СеМО*
Нагрузка	0,96	0,48	0,336	0,324	2,10
Загрузка	0,96	0,48	0,336	0,324	2,10
Время ожидания	19,2	1,85	0,202	0,144	198,4
Время пребывания	20	3,85	0,602	0,444	215,9
Длина очереди	23,04	0,44	0,170	0,155	23,8
Число заявок	24	0,92	0,506	0,479	25,9

Анализ представленных результатов показывает, что увеличение интенсивности поступления заявок в РСеМО всего лишь на 20% до значения $\lambda_0 = 0{,}12\ \text{с}^{-1}$, привело к резкому росту значений сетевых характеристик. В частности, среднее время пребывания заявок в сети выросло в 4 раза, а число заявок, находящихся в очередях – почти в 6,5 раз. Это говорит о том, что СеМО работает в области больших загрузок, где незначительное увеличение нагрузки приводит к существенному изменению характеристик обслуживания заявок. Наиболее загруженным узлом СеМО, то есть узким местом, является узел 1, загрузка которого много больше загрузок других узлов и составляет $\rho_1 = 0{,}96$. Именно в этом узле характеристики обслуживания заявок выросли наиболее существенно: среднее время пребывания заявок в 5 раз (с 4 до 20 секунд), а средняя длина очереди – более чем в 7 раз (с 3,2 до 23 заявок).

2. Для улучшения характеристик обслуживания заявок в РСеМО необходимо разгрузить узкое место сети, которым является узел 1. Для этого увеличим скорость работы обслуживающего прибора в 2 раза, что, в конечном счете, приведёт к уменьшению длительности обслуживания заявок в 2 раза, которая станет равной $b_1 = 0{,}4$ с.

Рассчитанные значения узловых и сетевых характеристик СеМО после разгрузки узкого места приведены в табл.4.4.

Анализ представленных результатов показывает, что разгрузка узкого места позволила существенно уменьшить значения сетевых характеристик: среднее время пребывания заявок в сети уменьшилось более чем в 9

раз, а число заявок, находящихся в очередях – почти в 20 раз. Отметим, что изменение длительности обслуживания заявок в узле 1 привело к изменению узловых характеристик только этого узла; узловые характеристики остальных узлов не изменились. Это является следствием независимого функционирования узлов экспоненциальной разомкнутой СеМО, что фактически и позволяет использовать метод расчёта характеристик сети, основанный на декомпозиции, то есть представлении сети в виде совокупности независимых СМО.

Таблица 4.4

Узловые характерстики	*Узел 1*	*Узел 2*	*Узел 3*	*Узел 4*	*СеМО*
Нагрузка	0,48	0,48	0,336	0,324	1,62
Загрузка	0,48	0,48	0,336	0,324	1,62
Время ожидания	0,369	1,846	0,202	0,144	10,1
Время пребывания	0,769	3,846	0,602	0,444	23,6
Длина очереди	0,443	0,443	0,170	0,155	1,21
Число заявок	0,923	0,923	0,506	0,479	2,83

3. Для сравнения выполним разгрузку узкого места другим способом, а именно: увеличим число обслуживающих приборов в узле 1 с одного до двух: $K_1 = 2$, сохранив прежнее значение длительности обслуживания одним прибором: $b_1 = 0{,}8$ с.

Рассчитанные значения узловых и сетевых характеристик СеМО после разгрузки узкого места приведены в табл.4.5.

Таблица 4.5

Узловые характерстики	*Узел 1*	*Узел 2*	*Узел 3*	*Узел 4*	*СеМО*
Нагрузка	0,96	0,48	0,336	0,324	2,10
Загрузка	0,48	0,48	0,336	0,324	1,62
Время ожидания	0,288	1,846	0,202	0,144	9,28
Время пребывания	1,088	3,846	0,602	0,444	26,78
Длина очереди	0,346	0,443	0,170	0,155	1,11
Число заявок	1,306	0,923	0,506	0,479	3,21

Сравним полученные значения сетевых характеристик со значениями, представленными в табл. 4.4 для первого способа разгрузки узкого места за счёт уменьшения длительности обслуживания заявок. При втором способе разгрузки узкого места за счёт увеличения числа обслуживающих приборов ($K_1 = 2$; $b_1 = 0{,}8$ с) среднее время ожидания заявок в сети несколько уменьшилось по сравнению с первым способом ($K_1 = 1$; $b_1 = 0{,}4$ с). В то же время, среднее время пребывания заявок в РСеМО увеличились более чем на 10%, что обусловлено большей длительностью обслуживания заявок ($b_1 = 0{,}8$ с) в каждом из приборов двухканального

узла 1 по сравнению с одноканальным узлом при первом способе ($b_1 = 0{,}4$ с). Как и в предыдущем случае, изменение числа обслуживающих приборов в узле 1 привело к изменению узловых характеристик только этого узла.

4.5. Замкнутые экспоненциальные СеМО с однородным потоком заявок

> «Во всякой формуле константы (особенно те, которые взяты из технических справочников) должны рассматриваться как переменные» (***Универсальные законы*** …)

4.5.1. Описание замкнутых СеМО

Рассмотрим замкнутую экспоненциальную сеть массового обслуживания с однородным потоком заявок при следующих предположениях:

1) замкнутая СеМО (ЗСеМО) *произвольной топологии* содержит n узлов;

2) после завершения обслуживания в каком-либо узле передача заявки в другой узел происходит *мгновенно*;

3) все узлы замкнутой СеМО *одноканальные*;

4) в СеМО циркулирует *постоянное число заявок*;

5) длительности обслуживания заявок во всех узлах сети представляют собой случайные величины, распределенные по *экспоненциальному* закону;

6) *ёмкость накопителя* в каждом узле СеМО *достаточна* для хранения всех заявок, циркулирующих в сети, что означает отсутствие отказов поступающим заявкам при их постановке в очередь любого узла (в частности, можно считать, что ёмкость накопителя в каждом узле равна числу заявок, циркулирующих в сети);

7) обслуживающий прибор любого узла *не простаивает*, если в его накопителе имеется хотя бы одна заявка, причем после завершения обслуживания очередной заявки мгновенно из накопителя выбирается следующая заявка;

8) в каждом узле сети заявки из накопителя выбираются в соответствии с *бесприоритетной дисциплиной обслуживания* в порядке поступления (ОПП) по правилу «первым пришел – первым обслужен» (FIFO – First In First Out).

Для описания линейных замкнутых однородных экспоненциальных СеМО необходимо задать такую же совокупность параметров, как и для разомкнутых СеМО, с единственным отличием, заключающимся в том, что вместо интенсивности источника заявок следует задать число заявок, циркулирующих в ЗСеМО. Таким образом, совокупность параметров для замкнутых СеМО будет иметь следующий вид:

- *число узлов* в сети: n;

- *число обслуживающих приборов* в узлах сети: $K_1,...,K_n$;
- *матрица вероятностей передач*: $\mathbf{P} = [p_{ij} \mid i, j = 0,1,\ldots,n]$, где p_{ij} – вероятность передачи заявки из узла i в узел j;
- число заявок M, циркулирующих в ЗСеМО;
- средние *длительности обслуживания* заявок в узлах сети: $b_1,\ldots,b_n$.

На основе перечисленных параметров могут быть рассчитаны узловые и сетевые характеристики, описывающие эффективность функционирования соответственно узлов и ЗСеМО в целом.

Расчёт характеристик функционирования линейных замкнутых однородных экспоненциальных СеМО с одноканальными узлами базируется на так называемой «теореме о прибытии» и проводится с использованием метода средних значений в два этапа:

- расчет коэффициентов передач в узлах замкнутой СеМО;
- расчет характеристик ЗСеМО.

4.5.2. Расчет коэффициентов передач в узлах ЗСеМО

Для замкнутой СеМО на первом этапе рассчитываются только коэффициенты передач. Интенсивности потоков заявок в узлах ЗСеМО не могут быть рассчитаны, как в РСеМО, поскольку для ЗСеМО изначально не известна интенсивность λ_0, которая является не параметром, задаваемым в составе исходных данных, а характеристикой, представляющей собой производительность ЗСеМО и определяемой в процессе анализа эффективности функционирования ЗСеМО.

Для расчёта коэффициентов передач $\alpha_1,\ldots,\alpha_n$ после некоторых преобразований можно воспользоваться той же системой линейных алгебраических уравнений (4.16). Для этого в левой и правой части выражения (4.16) представим интенсивности в виде $\lambda_j = \alpha_j \lambda_0$. Разделив левую и правую часть выражения (4.16) на λ_0, окончательно получим систему линейных алгебраических уравнений относительно $\alpha_1,\ldots,\alpha_n$:

$$\alpha_j = \sum_{i=0}^{n} p_{ij}\alpha_i \quad (i = 0,1,\ldots,n). \qquad (4.17)$$

Полагая $\alpha_0 = 1$, можно найти корни системы уравнений, численно определяющие значения. $\alpha_1,\ldots,\alpha_n$

4.5.3. Расчет характеристик ЗСеМО

Характеристики ЗСеМО могут быть рассчитаны с использованием марковских процессов, поскольку количество состояний марковского процесса, в отличие от РСеМО, не бесконечно и равно числу сочетаний C^{M}_{M+n-1}, где n – число узлов в ЗСеМО и M – число заявок, циркулирующих в ЗСеМО. При этом основная трудность заключается в определении веро-

ятностей состояний сети $P(M_1,\ldots,M_n)$ в случае большой ее размерности $(n>5; M>5)$, когда число состояний оказывается значительным. При выполнении расчетов на ЭВМ это, во многих случаях, приводит к потере значимости в процессе промежуточных вычислений и, следовательно, к невозможности получения конечных результатов.

От указанного недостатка свободен *метод средних значений*, позволяющий вычислять средние характеристики функционирования экспоненциальных СеМО на основе сравнительно простых рекуррентных соотношений.

Положим, что замкнутая однородная СеМО содержит n **одноканальных** узлов, длительности обслуживания заявок в которых распределены по экспоненциальному закону со средними значениями $b_1,\ldots,b_n$ соответственно. Пусть для каждого узла i сети известно среднее число попаданий заявки в данный узел за время ее нахождения в сети, то есть коэффициент передачи α_i, который, если конфигурация сети задана матрицей вероятностей передач $P=[p_{ij}\,|i,j=0,1,\ldots,n]$, определяется в результате решения системы линейных алгебраических уравнений (4.17).

Обозначим: u_i - среднее время пребывания заявки в узле i за время пребывания в сети; m_i – среднее число заявок в узле $i\,(i=1,\ldots,n)$; λ_0 – производительность замкнутой сети. Очевидно, что эти величины зависят от числа заявок M, циркулирующих в замкнутой сети, то есть $u_i=u_i(M)$; $m_i=m_i(M)$; $\lambda_0=\lambda_0(M)$.

Можно показать, что имеют место следующие соотношения:

$$u_i(M)=b_i[1+m_i(M-1)]; \tag{4.18}$$

$$U(M)=\sum_{i=1}^{n}\alpha_i u_i(M); \tag{4.19}$$

$$\lambda_0(M)=\frac{M}{U(M)}; \tag{4.20}$$

$$m_i(M)=\alpha_i\lambda_0(M)\,u_i(M), \tag{4.21}$$

где $U(M)$ – среднее время пребывания заявок в сети при условии нахождения в ней M заявок; $m_i(0)=0$.

Выражение (4.18) получено на основе так называемой *теоремы о прибытии* [1], утверждающей, что в замкнутой экспоненциальной сети с одноканальными узлами, в которой циркулируют M заявок, стационарная вероятность состояния любого узла в момент поступления в него новой заявки совпадает со стационарной вероятностью того же состояния рассматриваемого узла в сети, в которой циркулирует на одну заявку меньше, то есть $(M-1)$ заявок. Это означает, что в сети с M заявками среднее число заявок $m_i(M)$, находящихся в узле i в момент поступления в этот узел новой заявки, равно $m_i(M-1)$. Тогда среднее время пребывания

в узле i поступившей заявки будет складываться из среднего времени обслуживания всех $m_i(M-1)$ ранее поступивших и находящихся в узле i заявок и средней длительности обслуживания рассматриваемой заявки:

$$u_i(M) = b_i m_i(M-1) + b_i = b_i[1 + m_i(M-1)].$$

В этом выражении учтено, что среднее время дообслуживания заявки, находящейся в приборе на момент поступления рассматриваемой заявки, равно средней длительности обслуживания b_i в силу свойства отсутствия последействия, присущего экспоненциальному закону. Среднее время пребывания заявки в узле i за время ее нахождения в сети, учитывающее число попаданий α_i заявки в данный узел, равно $U_i(M) = \alpha_i u_i(M)$.

Выражения (4.19) и (4.20) представляют собой формулы Литтла для сети, а выражение (4.21) – для узла i, где $\lambda_i(M) = \alpha_i \lambda_0(M)$ – интенсивность потока заявок в узел i $(i = 1,\ldots,n)$.

На основе рекуррентных соотношений (4.18) – (4.21) последовательно для $M = 1,2,\ldots,M^*$, где M^* – заданное число заявок в замкнутой сети, могут быть рассчитаны средние значения характеристик замкнутой экспоненциальной СеМО.

Заметим, что приведенный метод расчета является *точным* для замкнутых экспоненциальных СеМО с *одноканальными* узлами.

Пример 4.4. Рассчитаем характеристики замкнутой однородной экспоненциальной СеМО, полученной путём преобразования разомкнутой СеМО (рис. 4.12), рассмотренной в Примере 4.2, в замкнутую. Положим, что «нулевая точка», отображающая завершение обслуживания заявок в сети и мгновенное формирование новой заявки, выбрана на дуге, выходящей из узла 1 и входящей снова в этот же узел (рис.4.14). Напомним, что в ЗСеМО относительно «нулевой точки» рассчитываются временны*е* сетевые характеристики: время нахождения в состоянии ожидания и время пребывания заявок в сети, а также производительность ЗСеМО.

ЗСеМО содержит $n = 4$ одноканальных узла, связи между которыми описываются той же матрицей вероятностей передач:

$$\mathbf{P} = \begin{array}{c|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0{,}1 & 0 & 0{,}2 & 0{,}7 & 0 \\ 2 & 0 & 0 & 0 & 0 & 1 \\ 3 & 0 & 0 & 0 & 0 & 1 \\ 4 & 0 & 1 & 0 & 0 & 0 \end{array}.$$

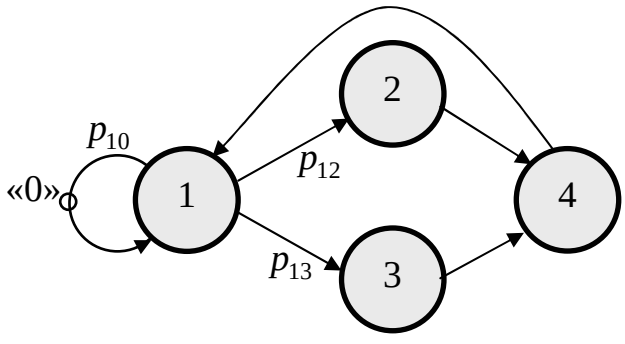


Рис.4.14. Граф замкнутой СеМО

Следовательно, коэффициенты передач для всех узлов, рассчитываемые путём решения системы линейных алгебраических уравнений (4.17), будут иметь те же самые значения: $\alpha_1 = 10$; $\alpha_2 = 2$; $\alpha_3 = 7$; $\alpha_4 = 9$.

В ЗСеМО циркулирует M заявок, средние длительности обслуживания которых в узлах равны: $b_1 = 0{,}8\,\text{с}$; $b_2 = 2\,\text{с}$; $b_3 = 0{,}4\,\text{с}$; $b_4 = 0{,}3\,\text{с}$.

Ниже в табл.4.6 представлены значения времени пребывания $u_i(M)$ и числа заявок $m_i(M)$ в узлах сети, а также среднего времени пребывания $U(M)$ заявок в сети и производительности $\lambda_0(M)$, рассчитанные на основе выражений (4.18) – (4.21), для числа циркулирующих в сети заявок $M = 1, 2, \ldots, 6$. Корректность выполненных расчетов подтверждается тем, что для всех $M = 1, 2, \ldots, 6$ выполняется проверочное условие:

$$\sum_{i=1}^{4} m_i(M) = M .$$

Таблица 4.6

M	i	$u_i(M)$	$U(M)$	$\lambda_0(M)$	$m_i(M)$
1	1	0,8			0,46
	2	2,0			0,23
	3	0,4			0,16
	4	0,3	17,5	0,057	0,15
2	1	1,17			1,02
	2	2,46			0,43
	3	0,46			0,28
	4	0,35	22,94	0,087	0,27
3	1	1,61			1,68
	2	2,86			0,59
	3	0,51			0,37
	4	0,38	28,87	0,104	0,36
4	1	2,14			2,43
	2	3,19			0,72
	3	0,55			0,44
	4	0,41	35,29	0,113	0,42
5	1	2,74			3,25
	2	3,45			0,82
	3	0,57			0,48
	4	0,42	42,14	0,119	0,45
6	1	3,40			4,14
	2	3,63			0,88
	3	0,59			0,50
	4	0,44	49,35	0,122	0,48

На рис.4.15 представлены зависимости производительности рассматриваемой замкнутой СеМО и среднего времени пребывания заявок в сети

от количества $M = \overline{1,10}$ циркулирующих заявок. Анализ полученных результатов показывает, что все характеристики, включая производительность λ_0, растут с увеличением M.

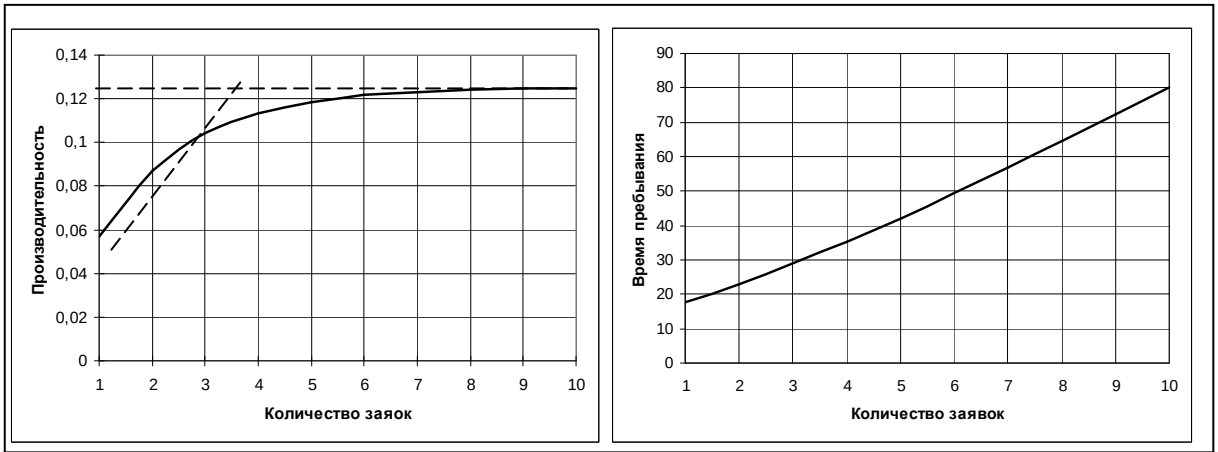


Рис.4.15. Производительность и время пребывания заявок в ЗСеМО

Производительность сети асимптотически приближается к максимально возможной производительности (пропускной способности ЗСеМО), совпадающей с предельно допустимой интенсивностью поступления заявок в аналогичной разомкнутой СеМО (см. Пример 4.1), при которой в сети отсутствуют перегрузки, и равна $\lambda_0 = 0{,}125\ \text{с}^{-1}$.

Среднее время пребывания заявок в ЗСеМО растёт неограниченно с увеличением количества заявок с сети.

Остальные характеристики замкнутой СеМО (загрузки и коэффициенты простоя узлов, время ожидания, длины очередей и число заявок в узлах сети, полное время ожидания в сети) могут быть рассчитаны с использованием фундаментальных соотношений, представленных в разделе 3 (п.3.4.3).

4.5.4. Анализ свойств замкнутых СеМО

Для замкнутых СеМО, как и для разомкнутых, наибольший интерес представляют свойства сети в целом, в частности, влияние циркулирующих в ЗСеМО числа заявок, на такие сетевые характеристики как производительность λ_0 замкнутой СеМО и среднее время пребывания U заявок в сети.

Анализ представленных на рис.4.16, зависимостей позволяет сформулировать следующие выводы.

1. Зависимость $\lambda_0 = f(M)$ производительности ЗСеМО λ_0 от числа M циркулирующих заявок вначале растёт с увеличением M до некоторого значения M_0, после которого рост производительности замедляется, а с дальнейшим увеличением M производительность сети асимптотически стремится к некоторому предельному значению $\hat{\lambda}_0$, представляющему

собой *пропускную способность* ЗСеМО. Для объяснения этой зависимости вспомним, что производительность замкнутой сети измеряется как интенсивность потока заявок, проходящих через некоторую условную точку, обозначаемую как «0» и расположенную на одной из дуг СеМО, отображающей завершение обслуживания заявок в сети и мгновенное формирование новой заявки, поступающей в сеть. Выше (см. пример 4.4) было показано, что увеличение числа заявок в замкнутой СеМО приводит к увеличению значений всех сетевых характеристик, включая производительность λ_0. В свою очередь, увеличение производительности приводит к увеличению загрузок узлов СеМО, связанных с интенсивностью λ_0 зависимостью:

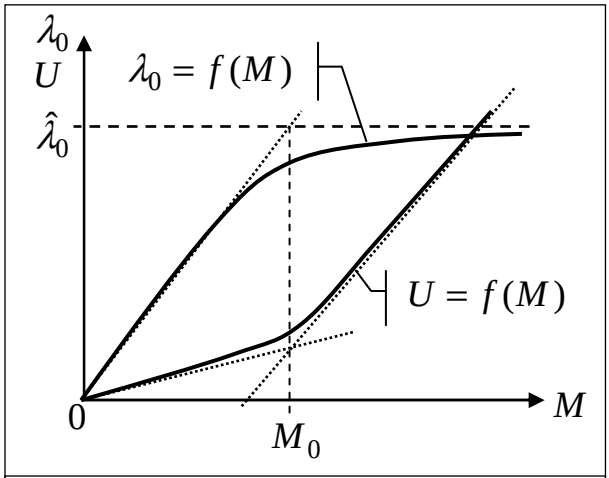


Рис.4.16. Характеристики ЗСеМО

$$\rho_j = \frac{\alpha_j \lambda_0 b_j}{K_j},$$

где α_j, b_j и K_j – соответственно коэффициент передачи, средняя длительность обслуживания и количество приборов в узле $j = \overline{1,n}$.

Когда число заявок в ЗСеМО достигает некоторого значения M_0, загрузка одного из узлов становится близкой к 1, при этом практически прекращается рост производительности, которая при $M \to \infty$ достигает своего предельного значения – пропускной способности $\hat{\lambda}_0$. Такой узел представляет собой «узкое место» сети, и значение пропускной способности $\hat{\lambda}_0$ определяется пропускной способностью узкого места из условия, что загрузка ρ_y этого узла равна 1:

$$\rho_y = \frac{\alpha_y \lambda_0 b_y}{K_y} = 1.$$

Отсюда пропускная способность замкнутой СеМО:

$$\hat{\lambda}_0 = \frac{K_y}{\alpha_y b_y},$$

где α_y, b_y и K_y – соответственно коэффициент передачи, средняя длительность обслуживания и количество обслуживающих приборов в узле, являющимся узким местом.

Правая часть последнего выражения представляет собой пропускную способность узла, являющегося узким местом сети: $\mu_y = \frac{K_y}{\alpha_y b_y}$.

Действительно, $\alpha_y b_y$ представляет собой полное время обслуживания одной заявки в данном узле с учётом того, что заявка за время нахождения в сети в среднем α_y раз побывает в данном узле. Тогда величина, обратная $\alpha_y b_y$, представляет собой интенсивность обслуживания заявок одним прибором в данном узле: $\mu_1 = 1/\alpha_y b_y$, а $\mu_y = K_y \mu_1$ – интенсивность обслуживания заявок узлом, то есть всеми приборами.

Этот же результат можно получить следующими рассуждениями. Если загрузка некоторого узла, являющегося узким местом СеМО, становится равной 1, то это означает, что все приборы данного узла постоянно обслуживают заявки, то есть не простаивают. Тогда интенсивность выходящего из этого узла потока заявок будет равна интенсивности обслуживания: $\lambda_y = \mu_y = K_y \mu_1$. Напомним, что интенсивность потока заявок в узле λ_y связана с производительностью ЗСеМО λ_0 зависимостью $\lambda_y = \alpha_y \lambda_0$. Отсюда вытекает, что производительность ЗСеМО равна $\lambda_0 = \dfrac{\lambda_y}{\alpha_y} = \dfrac{K_y \mu_1}{\alpha_y} = \dfrac{K_y}{\alpha_y b_y}$.

2. Среднее время пребывания заявок (рис.4.16) в замкнутой СеМО, как и производительность, растёт с увеличением числа M циркулирующих в сети заявок, причём вначале наблюдается незначительный рост, а затем, после значения $M = M_0$, наблюдается линейный рост времени пребывания.

Действительно, если в сети циркулирует только одна заявка, то в такой сети не может быть очередей, и время пребывания заявок в СеМО складывается только из времён обслуживания заявок в узлах с учётом коэффициентов передач:

$$U = \sum_{i=1}^{n} \alpha_i b_i .$$

С увеличением числа заявок M в узлах ЗСеМО появляются очереди, причём очевидно, что чем больше заявок в сети, тем более длинные очереди образуются в узлах и тем больше время ожидания, а, следовательно, и время пребывания заявок в ЗСеМО.

Сопоставляя зависимости производительности и среднего времени пребывания заявок от их числа в ЗСеМО, можно сделать следующий вывод: увеличение числа заявок в сети, с одной стороны, приводит к увеличению производительности, что может рассматриваться как положительный фактор, а, с другой стороны, – к увеличению времени пребывания заявок в сети, что является нежелательным фактором.

Точка $M = M_0$ характеризует некоторое граничное значение числа заявок в ЗСеМО. Дальнейшее увеличение числа заявок в сети оказывается нецелесообразным, поскольку приводит к резкому увеличению времени

пребывания заявок в ЗСеМО при незначительном увеличении производительности сети.

3. Когда загрузка узкого места становится равной единице, дальнейший рост производительности за счёт увеличения числа заявок в ЗСеМО невозможен. Для увеличения производительности ЗСеМО, как и в РСеМО, необходимо разгрузить узкое место, то есть уменьшить загрузку: $\rho_y = \dfrac{\alpha_y \lambda_0 b_y}{K_y} = 1$, что при одной и той же производительности может быть достигнуто:

- уменьшением длительности обслуживания заявок b_y, например за счет увеличения скорости работы (быстродействия) обслуживающего прибора;
- увеличением числа обслуживающих приборов K_y в узле;
- уменьшением коэффициента передачи α_y или, что то же самое, вероятности передачи заявок к узлу, являющемуся узким местом.

Если до разгрузки узкого места зависимость производительности ЗСеМО от числа заявок в сети имела вид $\lambda_0^{'} = f(M)$ (рис.4.17), а пропускная способность была равна $\hat{\lambda}_0^{'}$, то после разгрузки – зависимость производительности от числа заявок будет иметь вид $\lambda_0^{''} = f(M)$, а пропускная способность станет равной $\hat{\lambda}_0^{''} > \hat{\lambda}_0^{'}$. При этом граничное значение числа заявок в ЗСеМО увеличится: $M_0^{''} > M_0^{'}$

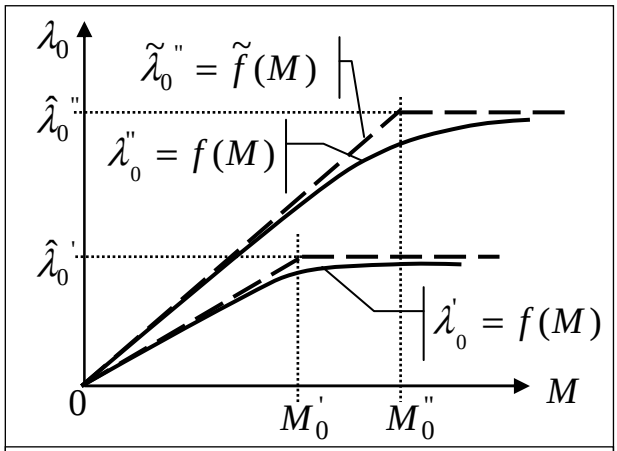


Рис.4.17. Разгрузка «узкого места»

Следует отметить, что к рассматриваемой зависимости производительности ЗСеМО λ_0 от числа M циркулирующих в сети заявок может быть применена линейная аппроксимация $\tilde{\lambda}_0^{''} = \tilde{f}(M)$, показанная на рис.4.17 в виде пунктирных линий и представляющая собой верхнюю границу производительности ЗСеМО. Последнее означает, что производительность ЗСеМО будет не больше, чем рассчитанное верхнее значение.

Нетрудно представить себе и изобразить на графике, как изменится зависимость среднего времени пребывания заявок в замкнутой СеМО от числа циркулирующих в сети заявок после разгрузки узкого места.

Отметим, что в некоторых случаях разгрузка узкого места не приводит к улучшению характеристик СеМО, в частности, к увеличению производительности. Обычно это связано с тем, что в СеМО может существовать несколько узлов, являющихся «узкими местами». Условием

этого является равенство загрузок узлов: $\rho_i = \rho_j$ или $\frac{\alpha_i \lambda_0 b_i}{K_i} = \frac{\alpha_j \lambda_0 b_j}{K_j}$, откуда окончательно получим: $\frac{\alpha_i b_i}{K_i} = \frac{\alpha_j b_j}{K_j}$ $(i \neq j)$. В этом случае для улучшения характеристик ЗСеМО необходимо одновременно разгрузить все «узкие места».

Последовательно разгружая узкие места СеМО, мы можем прийти к некоторой «идеальной» сети, в которой загрузки всех узлов одинаковы.

СеМО, в которой загрузки всех узлов равны, называется ***сбалансированной***. Сбалансированная СеМО обладает наилучшими характеристиками по сравнению с несбалансированной.

При построении реальных систем, моделями которых служат СеМО, необходимо, по-возможности, строить сбалансированные системы, хотя на практике по многим причинам достичь этого не удаётся.

4.6. Резюме

1. Одноканальная экспоненциальная СМО M/M/1 является наиболее *простой* с точки зрения аналитического расчета. Средние времена ожидания и пребывания заявок в СМО M/M/1 рассчитываются по сравнительно простым формулам:

$$\boxed{w = \frac{\rho b}{1-\rho}} \quad \text{и} \quad \boxed{u = \frac{b}{1-\rho}},$$

где $\rho = \lambda b < 1$ – загрузка системы; λ – интенсивность поступления заявок в систему; b – средняя длительность обслуживания заявок в приборе.

Для СМО M/G/1 среднее время ожидания заявок определяется по *формуле Поллачека-Хинчина*:

$$\boxed{w = \frac{\lambda b^2 (1 + \nu_b^{\,2})}{2(1-\rho)}},$$

где ν_b – коэффициент вариации длительности обслуживания.

Для общего случая одноканальных СМО типа G/G/1 с однородным потоком применяются *приближённые* аналитические методы расчёта.

Свойства одноканальной СМО с однородным потоком заявок:

- среднее время ожидания заявок в очереди *минимально при детерминированной длительности обслуживания* заявок с коэффициентом вариации $\nu_b = 0$ и увеличивается нелинейно с ростом коэффициента вариации (дисперсии) длительности обслуживания;
- среднее время ожидания заявок существенно зависит от нагрузки y (загрузки ρ) системы и при $y \geq 1$ $(\rho \to 1)$ возрастает *неограниченно*: $w \to \infty$, т.е. заявки могут ожидать обслуживания сколь угодно долго;

- для дисциплин обслуживания в обратном порядке и обслуживания в случайном порядке средние времена ожидания заявок будут такими же, как и при обслуживании в порядке поступления, но *дисперсии времени ожидания будут больше*.

2. В случае многоканальных СМО с однородным потоком заявок *точный* метод расчета среднего времени ожидания заявок разработан только для СМО типа M/M/K:

$$w = \frac{Pb}{K(1-\rho)},$$

где $\rho = \frac{\lambda b}{K}$ – загрузка системы; P – вероятность того, что все K приборов заняты обслуживанием заявок:

$$P = \frac{(K\rho)^K}{K!(1-\rho)} P_0,$$

где P_0 – вероятность простоя многоканальной СМО, то есть вероятность того, что в системе нет заявок:

$$P_0 = \left[\frac{(K\rho)^K}{K!(1-\rho)} + \sum_{i=0}^{K-1} \frac{(K\rho)^i}{i!} \right]^{-1}.$$

Свойства многоканальной СМО с однородным потоком заявок:

- *с увеличением числа обслуживающих приборов времена ожидания и пребывания заявок уменьшаются*, при этом в пределе при $K \to \infty$ время ожидания стремится к нулю, а время пребывания становится равным длительности обслуживания заявок;
- при увеличении числа обслуживающих приборов K и сохранении их суммарной производительности (скорости работы) время ожидания заявок уменьшается, однако время пребывания заявок в системе увеличивается и в пределе (при $K \to \infty$) асимптотически стремится к длительности обслуживания заявок, то есть *с точки зрения задержек (времени пребывания заявок) более эффективной является одноканальная система, чем многоканальная, при равенстве суммарной производительности; достоинством многоканальной системы является более высокая надежность*;
- среднее время ожидания заявок, как и для одноканальных систем, зависит от нагрузки y (загрузки ρ) системы и при $y \geq K$ $(\rho \to 1)$ время ожидания заявок возрастает *неограниченно*: $w \to \infty$, то есть заявки могут ожидать обслуживания сколь угодно долго.

3. Для одноканальных СМО с неоднородным потоком заявок и бесприоритетной ДО средние времена ожидания одинаковы для всех классов заявок и определяются по следующей формуле:

$$w_k^{\text{БП}} = w^{\text{БП}} = \frac{\sum_{i=1}^{H} \lambda_i b_i^2 (1+\nu_{b_i}^2)}{2(1-R)} \qquad (k=1,\ldots,H),$$

где $R = \sum_{i=1}^{H} \rho_i = \sum_{i=1}^{H} \lambda_i b_i$ – суммарная загрузка системы ($R<1$).

Свойства одноканальной СМО с бесприоритетной ДО:

- среднее время ожидания заявок разных классов при использовании ДО БП *одинаково*;
- среднее время ожидания заявок в очереди *минимально* при детерминированной длительности обслуживания заявок каждого класса и увеличивается нелинейно с ростом коэффициента вариации (дисперсии) длительности обслуживания;
- среднее время ожидания заявок зависит от суммарной нагрузки Y (загрузки R) системы и при $Y \geq 1$ $(R \to 1)$ время ожидания заявок всех классов возрастает *неограниченно*: $w^{\text{БП}} \to \infty$, однако средние времена пребывания в системе и средние длины очередей заявок разных классов, в общем случае, различны, поскольку различны длительности обслуживания и интенсивности поступления заявок разных классов;
- для бесприоритетной дисциплины обслуживания в обратном порядке средние времена ожидания заявок будут такими же, как и при обслуживании в порядке поступления, но дисперсия времени ожидания будет больше;
- для дисциплины обслуживания в циклическом порядке средние времена ожидания заявок разных классов в общем случае *не равны*.

4. Для ДО ОП средние времена ожидания заявок k–го класса определяются по следующей формуле:

$$w_k^{\text{ОП}} = \frac{\sum_{i=1}^{H} \lambda_i b_i{}^2 (1+\nu_{bi}{}^2)}{2(1-R_{k-1})(1-R_k)} \qquad (k=1,\ldots,H),$$

где R_{k-1} и R_k – суммарные загрузки системы со стороны заявок, которые имеют приоритет не ниже $(k-1)$ и k соответственно.

Свойства ДО ОП:

- введение относительных приоритетов по сравнению с ДО БП приводит к уменьшению времени ожидания высокоприоритетных заявок и к увеличению времени ожидания низкоприоритетных заявок;

- средние времена ожидания заявок при использовании ДО ОП *монотонно увеличиваются* с уменьшением приоритета: $w_1^{\text{ОП}} < w_2^{\text{ОП}} < \ldots < w_H^{\text{ОП}}$;
- ДО ОП обладает свойством *защиты от перегрузок*, заключающемся в том, что высокоприоритетные заявки даже при возникновении перегрузки ($Y \geq 1$) имеют конечное время ожидания за счёт отказа в обслуживании низкоприоритетным заявкам, время ожидания которых при этом резко возрастает и стремится к бесконечности.

5. Для ДО АП средние времена ожидания заявок *k*–го класса определяются по следующей формуле:

$$w_k^{\text{АП}} = \frac{\sum_{i=1}^{k} \lambda_i b_i (1+\nu_{b_i}^2)}{2(1-R_{k-1})(1-R_k)} + \frac{R_{k-1} b_k}{1-R_{k-1}} \qquad (k = 1, \ldots, H),$$

где R_{k-1} и R_k – суммарные загрузки системы со стороны заявок, которые имеют приоритет не ниже $(k-1)$ и k соответственно.

Свойства ДО АП:

- среднее время ожидания заявок складывается из двух составляющих: среднего *времени ожидания начала обслуживания и среднего времени ожидания в прерванном состоянии*;
- время ожидания заявок класса k зависит только от значений параметров классов заявок, имеющих более высокий или такой же приоритет, и *не зависит* от параметров классов заявок, имеющих более низкий приоритет;
- для заявок с самым высоким абсолютным приоритетом обеспечивается *минимально возможное время ожидания* по сравнению со всеми другими ДО;
- *полное время ожидания у заявок высокоприоритетного класса* k *может оказаться больше*, чем у заявок класса $(k+1)$ с более низким приоритетом, если длительности обслуживания заявок этих классов связаны соотношением $b_k >> b_{k+1}$;
- введение АП по сравнению с ОП приводит к уменьшению среднего времени ожидания самых высокоприоритетных заявок и к его увеличению для заявок класса самого низкого приоритета;
- при ДО АП высокоприоритетные заявки лучше защищены от перегрузок, чем при ДО ОП.

6. В соответствии с **законом сохранения** времени ожидания *изменение ДО позволяет уменьшить время ожидания высокоприоритетных заявок за счет увеличения времени ожидания низкоприоритетных заявок*:

$$\sum_{i=1}^{H} \rho_i \, w_i = \underset{\text{ДО}}{Const}.$$

Закон сохранения выполняется при следующих условиях:

- система без потерь;
- система простаивает лишь при отсутствии в системе заявок;
- при наличии прерываний длительность обслуживания прерванных заявок распределена по экспоненциальному закону;
- все поступающие потоки заявок – простейшие, и длительность обслуживания не зависит от параметров потоков заявок.

7. В качестве параметров линейных разомкнутых однородных экспоненциальных СеМО необходимо задать:

- *число узлов* в сети: n;
- *число обслуживающих приборов* в узлах сети: $K_1,...,K_n$;
- *матрицу вероятностей передач*: $\mathbf{P} = [\, p_{ij} \mid i, j = 0, 1, \ldots, n]$, где p_{ij} – вероятность передачи заявки из узла i в узел j;
- *интенсивность* λ_0 источника заявок, поступающих в РСеМО;
- средние *длительности обслуживания* заявок в узлах сети: $b_1,\ldots,b_n$.

Условие отсутствия перегрузок в разомкнутой СеМО:

$$\lambda_0 < \min\left(\frac{K_1}{\alpha_1 b_1}, \frac{K_2}{\alpha_2 b_2}, ..., \frac{K_n}{\alpha_n b_n} \right).$$

Расчет характеристик функционирования линейных разомкнутых однородных экспоненциальных СеМО базируется на эквивалентном преобразовании сети и проводится в три этапа:

- расчет интенсивностей потоков заявок λ_j в узлах $j = \overline{1,n}$ РСеМО путём решения *системы линейных алгебраических уравнений*:

$$\lambda_j = \sum_{i=0}^{n} p_{ij} \lambda_i \quad (j = 0,1,\ldots,n);$$

- расчет узловых характеристик:
  - нагрузка в узле j, показывающая среднее число занятых приборов: $y_j = \lambda_j b_j$;
  - загрузка узла j: $\rho_j = \min(y_j / K_j; 1)$, где K_j – число обслуживающих приборов в узле j;
  - коэффициент простоя узла: $\pi_j = 1 - \rho_j$;
  - время пребывания заявок в узле: $u_j = w_j + b_j$;
  - длина очереди заявок: $l_j = \lambda_j w_j$;

- число заявок в узле (в очереди и на обслуживании в приборе): $m_j = \lambda_j u_j$.

- расчет сетевых характеристик:
  - среднее число заявок, ожидающих обслуживания в сети, и среднее число заявок, находящихся в сети:

$$L = \sum_{j=1}^{n} l_j ; \qquad M = \sum_{j=1}^{n} m_j ,$$

где l_j - средняя длина очереди и m_j - среднее число заявок в узле *j*;

  - среднее время ожидания и среднее время пребывания заявок в сети:

$$W = \sum_{j=1}^{n} \alpha_j w_j ; \qquad U = \sum_{j=1}^{n} \alpha_j u_j ,$$

где w_j и u_j - соответственно среднее время ожидания и среднее время пребывания заявок в узле *j*; $\alpha_j = \lambda_j / \lambda_0$ $(j = \overline{1,n})$ - коэффициент передачи для узла *j*, показывающий среднее число попаданий заявки в узел *j* за время ее нахождения в сети.

Свойства разомкнутых СеМО:

- свойства отдельных узлов СеМО аналогичны свойствам соответствующих одноканальных и многоканальных СМО;
- с увеличением интенсивности λ_0 поступления заявок в сеть сетевые характеристики увеличиваются, причём имеется предельное значение интенсивности $\lambda'_{0\,\max}$, при котором, в частности, среднее время пребывания заявок в сети становится бесконечно большим, что свидетельствует о перегрузке в СеМО;
- узел сети, загрузка которого с увеличением λ_0 стремится к единице, называется «*узким местом*» и характеризуется бесконечным ростом очереди заявок перед ним и, как следствие, бесконечным ростом числа заявок в СеМО;
- способы разгрузки «узкого места»:
  - увеличение скорости работы (быстродействия) обслуживающего прибора;
  - увеличение числа обслуживающих приборов в узле;
  - уменьшение вероятности передачи заявок к узлу, являющемуся узким местом.

8. В качестве параметров линейных замкнутых однородных экспоненциальных СеМО необходимо задать:

- *число узлов* в сети: n;
- *число обслуживающих приборов* в узлах сети: $K_1,...,K_n$;

- *матрица вероятностей передач*: $\mathbf{P} = [\, p_{ij} \mid i, j = 0, 1, \ldots, n]$, где p_{ij} – вероятность передачи заявки из узла i в узел j;
- число заявок M^*, циркулирующих в ЗСеМО;
- средние *длительности обслуживания* заявок в узлах сети: $b_1, \ldots, b_n$.

В замкнутых СеМО *всегда существует установившийся режим*.

Расчет характеристик функционирования линейных замкнутых однородных экспоненциальных СеМО с одноканальными узлами проводится с использованием метода средних значений в два этапа:

- расчет коэффициентов передач в узлах замкнутой СеМО путём решения системы линейных алгебраических уравнений относительно $\alpha_1, \ldots, \alpha_n$ с учётом того, что $(\alpha_0 = 1)$:
- $\alpha_j = \sum_{i=0}^{n} p_{ij} \alpha_i \quad (i = 0, 1, \ldots, n)$;
- расчет характеристик ЗСеМО с использованием следующих рекуррентных соотношений для значений $M = 1, 2, \ldots, M^*$:

$$u_i(M) = b_i[1 + m_i(M - 1)];$$

$$U(M) = \sum_{i=1}^{n} \alpha_i u_i(M);$$

$$\lambda_0(M) = \frac{M}{U(M)};$$

$$m_i(M) = \alpha_i \lambda_0(M)\, u_i(M),$$

где M^* - заданное число заявок в замкнутой сети; $m_i(0) = 0$.

Свойства замкнутых СеМО:

- зависимость производительности ЗСеМО λ_0 от числа M циркулирующих заявок растёт с увеличением M и стремится к некоторому предельному значению $\hat{\lambda}_0$, представляющему собой пропускную способность ЗСеМО;
- среднее время пребывания заявок в замкнутой СеМО, как и производительность, растет с увеличением числа циркулирующих в сети заявок, причём рост времени пребывания вначале незначителен, а затем принимает линейный характер;
- увеличение числа заявок в сети, с одной стороны, приводит к увеличению производительности (положительный фактор), а, с другой стороны – к увеличению времени пребывания заявок в сети (нежелательный фактор).
- для каждой замкнутой СеМО существует некоторое граничное значение числа заявок в сети, после которого резко увеличивается время пребывания заявок в ЗСеМО при незначительном увеличении производительности сети;

- когда загрузка узкого места становится равной единице, дальнейший рост производительности за счёт увеличения числа заявок в ЗСеМО невозможен; для увеличения производительности ЗСеМО необходимо разгрузить узкое место одним из следующих способов:
  - уменьшением длительности обслуживания заявок (увеличением скорости работы обслуживающего прибора);
  - увеличением числа обслуживающих приборов в узле;
  - уменьшением коэффициента передачи.
- если в СеМО существует несколько узлов, одновременно являющихся «узким местом», для улучшения характеристик функционирования ЗСеМО необходимо одновременно разгрузить все узкие места;
- при построении реальных систем, моделями которых служат СеМО, следует, по-возможности, строить сбалансированные системы, в которых загрузки всех узлов одинаковы.

4.7. Практикум: решение задач

Задача 1. В одноканальную систему обслуживания поступают заявки двух классов с интенсивностями 0,3 и 1 заявок в секунду. Интенсивности их обслуживания соответственно равны 0,5 и 5 заявок в секунду.

а) Сформулировать условия, при которых время пребывания заявок 1-го класса будет равно 2 секунды?

б) Чему будет равно время пребывания заявок 1-го класса, если при тех же условиях интенсивность их поступления увеличится в два раза?

в) Чему будет равно время пребывания заявок 1-го класса, если при тех же условиях интенсивность их обслуживания увеличится в два раза?

Дано: Одноканальная СМО:

- количество классов заявок: $H = 2$;
- интенсивности потоков: $\lambda_1 = 0{,}3\, c^{-1}$; $\lambda_2 = 1\, c^{-1}$;
- интенсивности обслуживания: $\mu_1 = 0{,}5\, c^{-1}$; $\mu_2 = 5\, c^{-1}$.

Требуется: а) сформулировать условия, при которых $u_1 = 2$ с;

б) определить $u_1^{'} = ?$ при $\lambda_1^{'} = 2\lambda_1$;

в) определить $u_1^{''} = ?$ при $\mu_1^{''} = 2\mu_1$;

Решение.

а) Время пребывания заявок класса 1: $u_1 = w_1 + b_1$, где w_1 – время ожидания; $b_1 = 1/\mu_1 = 2\, c$ – длительность обслуживания. Очевидно, что $u_1 = 2$ с, если $w_1 = 0$, то есть заявки 1-го класса не должны образовывать очередь. Для этого необходимо, чтобы:

- заявки 1-го класса имели абсолютный приоритет по отношению к заявкам 2-го класса; это означает, что заявки 2-го класса не смогут влиять на характеристики обслуживания заявок 1-го класса, однако это не исключает образования очереди заявок 1-го класса;
- для того чтобы заявки 1-го класса не образовывали очередь, процессы поступления и обслуживания заявок 1-го класса должны быть детерминированными, то есть интервалы между поступающими в систему заявками 1-го класса и длительности их обслуживания должны быть детерминированными (не случайными) величинами;
- нагрузка, создаваемая заявками 1-го класса не должна превышать 1, в противном случае система будет перегружена и не сможет справиться с обслуживанием заявок 1-го класса, время ожидания которых будет расти до бесконечности.

Проверим выполнение последнего условия: $y_1 = \lambda_1 / \mu_1 = 0{,}6$ – система работает без перегрузок. Таким образом, для того чтобы $u_1 = 2\,\text{с}$, необходимо выполнение двух первых условий.

б) Определим $u_1^{'} = ?$ при $\lambda_1^{'} = 2\lambda_1$. Если интенсивность поступления заявок 1-го класса увеличится в 2 раза, то загрузка, то создаваемая заявками нагрузка тоже увеличится в 2 раза и станет равной $y_1^{'} = 2\lambda_1 / \mu_1 = 1{,}2$, что означает перегрузку системы, следовательно, время ожидания и время пребывания заявок 1-го класса вырастут до бесконечности: $u_1^{'} = \infty$.

Заметим, что если бы нагрузка не превысила значение 1, то время пребывания заявок 1-го класса осталось бы прежним: $u_1 = 2\,\text{с}$.

в) Определим $u_1^{''} = ?$ при $\mu_1^{''} = 2\mu_1$. Увеличение интенсивности обслуживания заявок 1-го класса приведёт к уменьшению нагрузки в 2 раза: $y_1^{''} = \lambda_1 /(2\mu_1) = 0{,}3$, то есть система будет работать без перегрузки. С другой стороны, длительность обслуживания заявок тоже уменьшится в 2 раза: $b_1 = 1/(2\mu_1) = 1\,c$, следовательно, время пребывания станет равным $u_1^{''} = b_1^{''} = 1\,c$.

Задача 2. Интенсивность поступления заявок в разомкнутую трехузловую СеМО равна 2 заявки в секунду. Среднее число заявок в узлах СеМО соответственно равно: 2, 4 и 6. Определить среднее время пребывания заявок в сети.

Дано: РСеМО; $n = 3$; $\lambda_0 = 2\,c^{-1}$; $m_1 = 2$; $m_2 = 4$; $m_3 = 6$.

Требуется: определить U.

Решение.

Среднее время пребывания заявок в СеМО определяется по формуле:

$$U = \sum_{j=1}^{n} \alpha_j u_j = \sum_{j=1}^{n} \alpha_j \frac{m_j}{\lambda_j} = \sum_{j=1}^{n} \alpha_j \frac{m_j}{\alpha_j \lambda_0} = \sum_{j=1}^{3} \frac{m_j}{\lambda_0} = \frac{2+4+6}{2} = 6\,c\,.$$

Здесь последовательно применены формулы (3.29), (3.15) и (3.5) для узла j.

Этот же результат можно получить, исходя из формулы Литтла (3.31), связывающей среднее время пребывания и число заявок в сети:

$$U = \frac{M}{\lambda_0} = \frac{1}{\lambda_0} \sum_{j=1}^{3} m_j = 6\,c\,.$$

4.8. Самоконтроль: перечень вопросов и задач

СМО с однородным потоком заявок

1. Как зависит среднее время ожидания заявок в СМО от коэффициента вариации длительности обслуживания? Во сколько раз изменится среднее время ожидания заявок при переходе от постоянной длительности обслуживания к экспоненциально распределенной? Во сколько раз изменится среднее время ожидания при переходе от экспоненциального распределения длительности обслуживания к гиперэкспоненциальному распределению с коэффициентом вариации, равным 2?
2. Какое распределение длительности обслуживания заявок в СМО является предпочтительным для уменьшения среднего времени ожидания заявок?
3. Изменится ли разность между средним временем пребывания и средним временем ожидания заявок в СМО при изменении: а) скорости работы (быстродействия) прибора; б) интенсивности потока заявок; в) количества приборов?
4. Изменится ли разность между средним числом заявок в системе и средней длиной очереди при изменении: а) скорости работы (быстродействия) прибора; б) интенсивности потока заявок; в) количества приборов?
5. Заявки поступают в одноканальную СМО с интервалом 2,5 минуты, длительность обслуживания заявок в приборе 45 секунд. Определить загрузку и коэффициент простоя системы.
6. Интенсивность поступления заявок в трехканальную СМО – 21 заявка в секунду, интенсивность обслуживания – 10 заявок в секунду. Определить: а) вероятность того, что обслуживающий прибор работает; б) вероятность того, что обслуживающий прибор простаивает; в) среднее число заявок, находящихся на обслуживании; г) среднее число работающих приборов?
7. Интенсивность поступления заявок в четырехканальную СМО равна интенсивности обслуживания заявок одним прибором. Определить:

а) вероятность того, что система простаивает; б) среднее число простаивающих приборов; в) на какую величину среднее число заявок в системе отличается от средней длины очереди.

8. Длительность обслуживания заявок в одном приборе четырехканальной СМО равна 4 минуты. Определить предельную интенсивность поступления заявок в систему, при которой в системе существует стационарный режим?

9. Интенсивность поступления заявок в СМО – 15 заявок в секунду, длительность обслуживания одной заявки – 5 секунд. Определить число обслуживающих приборов, при котором в системе существует стационарный режим?

10. Заявки поступают в одноканальную СМО с интервалом 0,5 секунд, интенсивность обслуживания – 2,5 заявки в секунду, среднее время пребывания заявок в системе – 2 секунды. Определить среднюю длину очереди заявок.

11. Интенсивности поступления и обслуживания заявок в СМО соответственно равны 4 и 5 заявок в секунду. Определить среднее время пребывания заявок в системе, если известно, что средняя длина очереди равна 6.

12. Интенсивность обслуживания заявок в СМО равна 5 заявок в секунду. Определить предельную интенсивность поступления заявок в систему, при которой среднее время пребывания заявок в системе не превысит 1 секунду и средняя длина очереди не превысит 2.

СМО с неоднородным потоком заявок

13. При какой дисциплине обслуживания заявок средние времена ожидания заявок разных классов одинаковы?

14. При каких условиях среднее время ожидания заявок для ДО ОП является возрастающей (убывающей) функцией от номера класса заявок?

15. При каких условиях среднее время пребывания заявок для ДО ОП является возрастающей функцией от номера класса заявок?

16. Может ли среднее время пребывания заявок для ДО ОП быть убывающей функцией от номера класса заявок, если среднее время ожидания заявок является возрастающей функцией от номера класса заявок? Ответ пояснить.

17. Может ли заявка с более высоким относительным приоритетом иметь б*о*льшее время пребывания, чем низкоприоритетная? Ответ обосновать.

18. При каких условиях среднее время ожидания заявок для дисциплины обслуживания с абсолютными приоритетами является возрастающей функцией от номера класса заявок?

19. В каких случаях среднее время ожидания заявок более высокого приоритета при ДО АП может иметь б*о*льшее значение, чем заявок низкого приоритета?

20. Может ли среднее время ожидания заявок быть убывающей функцией от номера класса (приоритета) заявок?

21. Почему среднее время ожидания заявок наивысшего приоритета при использовании ДОАП отличается от нуля?

22. При каких условиях характер зависимости среднего времени пребывания заявок от номера класса совпадает с характером зависимости среднего времени ожидания? Могут ли зависимости среднего времени пребывания и среднего времени ожидания заявок от номера класса иметь противоположный характер?

23. В каких случаях характер зависимости средней длины очереди заявок от номера класса отличается от характера зависимости среднего времени ожидания заявок? При каких условиях характер зависимости средней длины очереди заявок от номера класса совпадает с характером зависимости среднего времени ожидания заявок?

24. Может ли среднее число заявок в СМО отличаться от средней длины очереди заявок больше, чем на единицу? Изменится ли разность между ними при изменении дисциплины обслуживания?

25. Изменится ли разность между средним временем пребывания и средним временем ожидания для заявок суммарного потока при изменении дисциплины обслуживания?

26. Можно ли утверждать, что если при какой-либо дисциплине обслуживания значение одной из характеристик суммарного потока (например, время ожидания) меньше значений аналогичной характеристики при другой дисциплине обслуживания, то и значения остальных характеристик при первой дисциплине также будут меньше? Проиллюстрировать на примере.

27. Изменятся ли характеристики обслуживания всех классов заявок или только некоторых из них при ДО БП (ДО ОП, ДО АП), если изменить: а) интенсивность поступления заявок какого-либо класса; б) среднюю длительность обслуживания какого-либо класса заявок?

28. Нарисовать зависимости среднего времени ожидания и среднего времени пребывания заявок в СМО от номера класса заявок для бесприоритетной дисциплины обслуживания. Объяснить характер этих зависимостей.

29. Нарисовать зависимости среднего времени ожидания заявок в СМО от суммарной загрузки в случае трех классов заявок и дисциплины обслуживания с относительными и абсолютными приоритетами. Провести анализ этих зависимостей.

30. Нарисовать зависимости среднего времени ожидания и среднего времени пребывания заявок от номера класса заявок для дисциплин обслуживания с относительными и с абсолютными приоритетами. Объяснить характер этих зависимостей.

31. Физический смысл и математическая запись закона сохранения времени ожидания. Сформулировать условия, сопровождающие закон

сохранения. При каком условии в случае ДО АП не выполняется закон сохранения времени ожидания?

32. Как получается значение константы в законе сохранения времени ожидания?

33. При каком условии закон сохранения времени ожидания вырождается в закон сохранения: а) суммарной длины очереди заявок: L = Const; б) суммарного числа заявок в системе: M = Const; в) среднего времени ожидания; г) среднего времени пребывания?

34. Можно ли, изменив дисциплину обслуживания, изменить: а) время ожидания заявок всех классов; б) время пребывания заявок всех классов; в) длительность обслуживания заявок; г) время ожидания заявок только одного класса; д) время ожидания заявок только двух классов? Показать на примерах.

35. В каких случаях следует применять приоритетные дисциплины обслуживания заявок с относительными и абсолютными приоритетами?

36. Что такое "защита от перегрузок" и для каких дисциплин обслуживания она существует? Проиллюстрировать на графике свойство "защиты от перегрузок".

37. В одноканальную СМО поступают 2 простейших потока заявок с интенсивностями 0,1 и 0,2 заявок в секунду; длительности их обслуживания соответственно 2 и 4 секунды. Чему будет равно среднее время ожидания заявок 1-го класса при использовании бесприоритетной дисциплины?

38. В одноканальную СМО поступают 2 класса заявок с интенсивностями 0,1 и 0,2 заявок в секунду. Длительности их обслуживания соответственно 2 и 3 секунды. Среднее время ожидания заявок при использовании бесприоритетной дисциплины обслуживания – 5 секунд. После введения приоритетов среднее время ожидания заявок 1-го класса стало равно 2 секундам. Чему равно среднее время ожидания заявок 2-го класса?

39. В систему поступают заявки трех классов с интенсивностями 2, 1 и 0,5 заявок в секунду соответственно. Сформулировать условия, при которых среднее время пребывания заявок всех классов будет одинаково.

40. В одноканальную систему обслуживания поступают заявки двух классов с интенсивностями 0,5 и 2 заявки в секунду. Интенсивности их обслуживания соответственно равны 5 и 1,25 заявок в секунду. а) При каких условиях время пребывания заявок 2-го класса будет равно 0,8 секунды? б) Чему будет равно время пребывания заявок 2-го класса, если при тех же условиях интенсивность поступления заявок 1-го класса увеличится в два раза? в) Чему будет равно время пребывания заявок 2-го класса, если при тех же условиях интенсивность их поступления увеличится в два раза? г) Чему будет равно время пребывания заявок 2-го класса, если при тех же условиях интенсивность их обслуживания увеличится в два раза?

Разомкнутые и замкнутые СеМО

41. Что показывает коэффициент передачи в СеМО?

42. Дать физическое толкование значения коэффициента передачи узла СеМО, равное: а) 4; б) 0,4.

43. В чём различие между эквивалентным и толерантным преобразованиями СеМО? Привести пример эквивалентного преобразования СеМО.

44. Что такое «узкое место» и какие способы используются для разгрузки «узкого места»?

45. Показать на графике, как изменится зависимость производительности и среднего времени пребывания заявок в замкнутой СеМО от числа циркулирующих в сети заявок после разгрузки узкого места. В каких случаях разгрузка узкого места может не дать положительного эффекта?

46. В двухузловой замкнутой СеМО циркулирует 1 заявка. Определить загрузку узлов 1 и 2, если известно, что загрузка узла 1 в 3 раза больше, чем загрузка узла 2.

47. В замкнутой двухузловой СеМО циркулирует одна заявка, которая последовательно переходит из одного узла в другой. Длительность обслуживания в узлах распределена по экспоненциальному закону с одним и тем же средним значением, равным 5 минут. По какому закону распределено время пребывания заявки в сети? Определить производительность замкнутой СеМО.

48. В замкнутой двухузловой СеМО циркулирует одна заявка, которая последовательно переходит из одного узла в другой. Длительности обслуживания в узлах 1 и 2 сети соответственно равны 2 и 3 с. Определить: а) коэффициенты простоя узлов замкнутой СеМО; б) среднее число заявок, находящихся в каждом из узлов СеМО.

49. В замкнутой двухузловой СеМО циркулирует 4 заявки, которые последовательно переходят из одного узла в другой. Длительности обслуживания заявок в узлах сети одинаковы и равны 2 с. Среднее время ожидания заявок в узле 1 равно 3 с. Определить: а) производительность замкнутой СеМО; б) загрузку узлов сети; в) среднее число заявок, находящихся в состоянии ожидания.

50. В разомкнутую СеМО поступают заявки с интервалом 5 секунд. Время пребывания заявок в сети равно 15 секунд. Определить среднее число заявок в сети и интенсивность выходящего из сети потока заявок.

51. Средние времена ожидания заявок в узлах трехузловой СеМО соответственно равны: 1, 2 и 4 секунды, а коэффициенты простоя узлов равны 0,8; 0,4; 0,7. Определить среднее время ожидания заявок в сети, если известно, что длительности обслуживания заявок во всех узлах одинаковы и коэффициент передачи узла 1 равен 2.

52. Известны вероятности состояний двухузловой замкнутой СеМО: P(0,4)=0,1; P(1,3)=0,4; P(2,2)=0,2; P(3,1)=0,1; P(4,0)=0,2, где состояние (i_1, i_2) задает число заявок в одноканальном узле 1 и трехканальном узле 2

соответственно. Определить среднее число заявок в СеМО, находящихся в состоянии ожидания.

53. Известны вероятности состояний трехузловой замкнутой СеМО: P(0,0,2)=0,1; P(0,1,1)=0,2; P(0,2,0)=0,15; P(1,0,1)=0,35; P(1,1,0)=0,05; P(2,0,0)=0,15, где состояние (i_1, i_2, i_3) задает число заявок в узле 1, 2, 3 соответственно. Определить среднее число параллельно работающих узлов сети.

54. Известны вероятности состояний трехузловой замкнутой СеМО: P(0,0,2)=0,1; P(0,1,1)=0,3; P(0,2,0)=0,4; P(1,0,1)=0,05; P(1,1,0)=0,05; P(2,0,0)=0,1. Длительности обслуживания заявок во всех одноканальных узлах одинаковы. Определить значения коэффициентов передач второго и третьего узлов сети, если известно, что коэффициент передачи первого узла равен 2.

55. Известны вероятности состояний трехузловой замкнутой СеМО: P(0,0,2)=0,1; P(0,1,1)=0,3; P(0,2,0)=0,4; P(1,0,1)=0,05; P(1,1,0)=0,05; P(2,0,0)=0,1. Определить производительность СеМО, если известно, что коэффициент передачи первого узла (четырехканального) равен 2, а средняя длительность обслуживания заявок в этом узле равна 0,1 с.

Раздел 5. ЧИСЛЕННОЕ МОДЕЛИРОВАНИЕ (МОДЕЛИ СЛУЧАЙНЫХ ПРОЦЕССОВ)

«В задаче из N уравнений всегда будет N+1 неизвестная» ***(Уравнения Снэйфу)***

При изучении сложных систем со стохастическим характером функционирования полезной математической моделью является *случайный процесс*, который развивается в зависимости от ряда случайных факторов. Примерами случайных процессов могут служить процессы поступления и передачи данных в телекоммуникационной сети, процессы выполнения задач и обмена данными с внешними устройствами в вычислительной системе и т.п.

Большинство моделей дискретных систем со стохастическим характером функционирования строится на основе моделей массового обслуживания, процессы в которых являются случайным и, во многих случаях, *марковскими* или некоторым образом связанные с марковскими процессами. Поэтому для решения таких задач теории массового обслуживания может использоваться математический аппарат **теории марковских процессов**. Применение марковских процессов оказывается особенно эффективным и результативным *при исследовании систем и сетей массового обслуживания с накопителями ограниченной ёмкости*.

Математическое описание марковских процессов обычно представляется в виде систем дифференциальных (в случае нестационарного режима) или алгебраических (для стационарного режима) уравнений, решение которых, в общем случае, получить в явном виде не удается. Это обусловливает необходимость применения *численных методов* решения систем дифференциальных или алгебраических уравнений.

5.1. Понятие случайного процесса

Основными для случайных процессов являются понятия *состояния* и *перехода* из одного состояния в другое.

Случайный процесс находится в некотором ***состоянии***, если он полностью описывается значениями переменных, которые задают это состояние.

Процесс совершает ***переход*** из одного состояния в другое, если описывающие ее переменные изменяются от значений, задающих одно состояние, на значения, которые определяют другое состояние.

Случайный процесс состоит в том, что с течением времени процесс переходит из одного состояния в другое *заранее не известное состояние*.

Понятия «состояние» и «переход» используются как для описания случайного процесса, так и системы, в которой этот процесс протекает. Поэтому при моделировании реальных систем часто говорят о состоянии системы и переходе системы из одного состояния в другое.

Если множество состояний, в которых может находиться процесс

счётное, то есть все возможные состояния могут быть пронумерованы, то соответствующий процесс называется ***случайным процессом с дискретными состояниями*** или просто ***дискретным случайным процессом***. В этом случае переменные, описывающие состояния случайного процесса, принимают либо целочисленные значения, либо вполне конкретные отделённые друг от друга дискретные значения. Обычно состояния дискретного случайного процесса определяются таким образом, чтобы каждое возможное состояние могло быть обозначено порядковым номером, при этом число возможных состояний системы может быть *конечным*: $\mathbf{E}_1, \mathbf{E}_2, ..., \mathbf{E}_n$ или *бесконечным*: $\mathbf{E}_1, \mathbf{E}_2, ..., \mathbf{E}_n, ...$ (иногда состояния нумеруются, начиная с нуля: $\mathbf{E}_0, \mathbf{E}_1, ..., \mathbf{E}_n, ...$). Для случайного процесса с дискретными состояниями характерен скачкообразный переход из одного состояния в другое (рис.5.1,а). Например, случайный процесс, протекающий в простейшей СМО с однородным потоком заявок, может быть представлен количеством заявок, находящихся в системе в произвольный момент времени. Тогда состояние $\mathbf{E}_k$ случайного процесса и, следовательно, самой системы будет означать, что в СМО находится ровно $k = 0, 1, 2, ...$ заявок.

Если множество состояний не может быть пронумеровано, то имеем ***случайный процесс с непрерывными состояниями*** или просто ***непрерывный случайный процесс***, для которого характерен плавный переход из состояния в состояние и который задаётся в виде непрерывной ункции времени: $E(t)$ (рис.5.1,б). Например, процесс изменения температуры некоторого объекта может рассматриваться как случайный процесс с непрерывными состояниями.

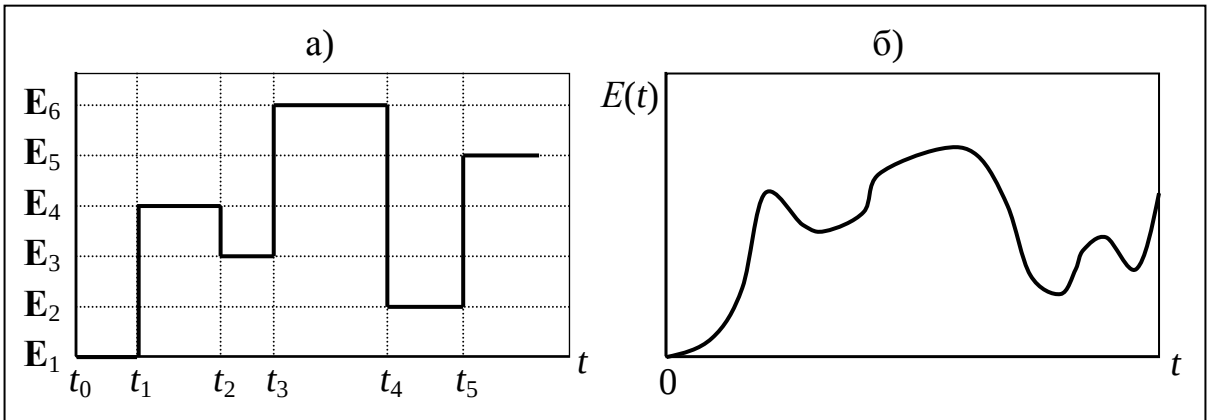


Рис.5.1. Процессы с дискретными (а) и непрерывными (б) состояниями

Поскольку модели массового обслуживания относятся к классу дискретных систем, то в дальнейшем будут рассматриваться только случайные процессы с дискретными состояниями.

При описании дискретных систем в терминах случайных процессов одним из основных этапов является этап *кодирования состояний*, заключающийся в определении состава переменных и их значений, используемых для описания состояний. Состав переменных в значительной мере

определяется назначением разрабатываемой модели, зависящим от целей исследований.

5.1.1. Случайные процессы с дискретными состояниями

Предположим, что система может находиться в одном из состояний $\mathbf{E}_1$, $\mathbf{E}_2$, ... (часто состояния обозначаются просто номерами 1, 2,...). Пусть состояние системы меняется скачкообразно в зависимости от некоторого параметра t, причем переход из состояния в состояние является случайным. Будем называть параметр t – **временем** и считать, что t пробегает либо целые, либо действительные числа. Обозначим через $Z(t)$ случайный процесс, описывающий состояние системы в момент времени t.

Случайный процесс $Z(t)$ называется ***случайным процессом с дискретным временем***, если переходы из состояния в состояние возможны только в строго *определенные заранее фиксированные моменты времени*, которые можно пронумеровать: $t_1, t_2, \ldots$.

Если промежуток времени между переходами из состояния в состояние является *случайным* и переход возможен в любой заранее не известный момент времени t, то процесс называется ***случайным процессом с непрерывным временем***.

Процесс с дискретным временем имеет место либо когда структура системы такова, что ее состояния могут изменяться только в заранее определенные моменты времени, либо когда предполагается, что для описания процесса достаточно знать состояние системы в отдельные моменты времени. Тогда эти моменты можно пронумеровать и говорить о состоянии $\mathbf{E}_i$ в момент t_k или просто в момент k ($k = 0,1,2,\ldots$).

Процессы с дискретным временем называются **стохастическими последовательностями** или **случайными цепями**.

Случайные процессы с дискретными состояниями могут изображаться в виде **графа переходов** (состояний), в котором вершины соответствуют состояниям, а ориентированные дуги – переходам из одного состояния в другое.

Граф переходов называется ***размеченным***, если на дугах графа указаны условия перехода в виде *вероятностей переходов* (для процессов с дискретным временем) или *интенсивностей переходов* (для процессов с непрерывным временем).

Состояния $\mathbf{E}_i$ могут быть:

- **невозвратными**, если процесс после какого-то числа переходов непременно покидает их;
- **поглощающими**, если случайный процесс, достигнув этих состояний прекращается.

Случайный процесс называется ***транзитивным***, если из любого состояния можно перейти за то или иное число шагов в любое другое

состояние и вернуться в исходное.

5.1.2. Понятие марковского случайного процесса

Случайный процесс называется ***марковским***, если вероятность любого состояния в будущем зависит только от его состояния в настоящем и не зависит от того, когда и каким образом процесс оказался в этом состоянии.

Описывающий поведение системы процесс $Z(t)$ называется **цепью Маркова**.

Для того чтобы случайный процесс *с непрерывным временем* был *марковским*, необходимо, чтобы интервалы времени между соседними переходами из состояния в состояние были распределены по *экспоненциальному закону*. Для доказательства последнего утверждения воспользуемся следующими рассуждениями.

Пусть время нахождения случайного процесса в некотором состоянии $\mathbf{E}_i$ до его перехода в другое состояние $\mathbf{E}_j$ распределено по экспоненциальному закону с функцией распределения $F_{ij}(\tau) = 1 - e^{-\alpha_{ij}\tau}$, где α_{ij} - параметр распределения, характеризующий частоту перехода из состояния $\mathbf{E}_i$ в состояние $\mathbf{E}_j$ и определяемый как величина, обратная среднему времени нахождения случайного процесса в состоянии $\mathbf{E}_i$ до момента его перехода в состояние $\mathbf{E}_j$. Вычислим вероятность того, что случайный процесс перейдет в состояние $\mathbf{E}_j$ в течение интервала времени $\Delta\tau$ при условии, что в состоянии $\mathbf{E}_i$ процесс уже находится в течение времени τ_0. Эта условная вероятность равна

$$P_{ij}(\Delta\tau \,|\, \tau \geq \tau_0) = \Pr(\tau_0 \leq \tau \leq \tau_0 + \Delta\tau \,|\, \tau \geq \tau_0) =$$

$$= \frac{\Pr(\tau_0 \leq \tau \leq \tau_0 + \Delta\tau)}{\Pr(\tau \geq \tau_0)} = \frac{F(\tau_0 + \Delta\tau) - F(\tau_0)}{1 - F(\tau_0)} = 1 - e^{-\alpha_{ij}\Delta\tau}.$$

Из последнего выражения следует, что вероятность перехода из одного состояния в другое зависит только от исходного состояния $\mathbf{E}_i$ и не зависит от интервала времени τ_0, то есть от того, как долго находился процесс в состоянии $\mathbf{E}_i$, а также от того, какие состояния предшествовали состоянию $\mathbf{E}_i$. Другими словами, поведение случайного процесса не зависит от предыстории и определяется только его состоянием в настоящий момент, то есть процесс является марковским.

Еще одно замечательное **свойство экспоненциального распределения** вытекает из полученного выражения, а именно: если время нахождения случайного процесса в некотором состоянии $\mathbf{E}_i$ до его перехода в другое состояние $\mathbf{E}_j$ распределено по экспоненциальному закону с параметром α_{ij}, то *интервал времени от любого случайного момента времени до момента перехода* в состояние $\mathbf{E}_j$ имеет *такое же экспоненциальное распределение с тем же параметром* α_{ij}. Эта особенность является следствием отсутствия последействия, присущего всем процессам с экспонен-

циальным распределением времени нахождения в том или ином состоянии.

Таким образом, безусловная $P_{ij}(\Delta\tau)$ и условная $P_{ij}(\Delta\tau \mid \tau \geq \tau_0)$ вероятности перехода в другое состояние за время $\Delta\tau$ для марковского процесса одинаковы и равны

$$P_{ij}(\Delta\tau) = P_{ij}(\Delta\tau \mid \tau \geq \tau_0) = 1 - e^{-\alpha_{ij}\Delta\tau}.$$

Пусть интервал времени $\Delta\tau$ достаточно мал. Тогда, разлагая $e^{-\alpha_{ij}\Delta\tau}$ в ряд по степеням $\alpha_{ij}\Delta\tau$ при $\Delta\tau \to 0$ и пренебрегая величинами высшего порядка малости, получим вероятность перехода из одного состояния в другое за бесконечно малый интервал времени:

$$P_{ij}(\Delta\tau) = 1 - (1 - \alpha_{ij}\Delta\tau) = \alpha_{ij}\Delta\tau. \qquad (5.1)$$

5.2. Параметры и характеристики марковского случайного процесса

5.2.1. Параметры марковского случайного процесса

Для описания марковского случайного процесса с дискретными состояниями используется следующая совокупность параметров:

- **перечень состояний** $\mathbf{E}_1, ..., \mathbf{E}_n$, в которых может находиться случайный процесс;
- **матрица переходов**, описывающая переходы случайного процесса между состояниями в виде:
  - *матрицы вероятностей переходов* **Q** для процессов *с дискретным временем*;
  - *матрицы интенсивностей переходов* **G** для процессов *с непрерывным временем*;
- **начальные вероятности** $p_1(0), \ldots, p_n(0)$.

Для **определения перечня состояний** случайного процесса необходимо корректно решить задачу кодирования состояний, которое зависит от смысла, вкладываемого в понятие «состояние» для каждой конкретной системы. Так, например, состояние некоторой системы массового обслуживания (а, следовательно, и случайного процесса, протекающего в ней) может быть задано числом заявок, находящихся в системе в данный момент времени, а состояние сети массового обслуживания – распределением числа заявок по всем узлам сети.

Для случайных процессов с дискретным временем изменения состояний происходят только в определенные моменты времени $t_1, t_2, \ldots, t_k, \ldots$. Переходы между состояниями описываются ***вероятностями переходов***. Если непосредственный переход из одного состояния в другое невозможен, то вероятность, соответствующая данному переходу, равна нулю. Обозначим через q_{ij} условную вероятность того, что в момент

времени t_{k+1} случайный процесс перейдет в состояние $\mathbf{E}_j$ при условии, что в момент t_k процесс находился в состоянии $\mathbf{E}_i$. Если переход из состояния $\mathbf{E}_i$ в $\mathbf{E}_j$ зависит только от этих двух состояний, то есть условная вероятность q_{ij} не изменяется при дополнительной информации о поведении процесса до момента t_k, получим цепь Маркова.

Цепь Маркова называется ***однородной***, если вероятности переходов не зависят от момента времени t_k, и ***неоднородной***, если вероятности переходов являются функциями t_k, то есть $q_{ij} = q_{ij}(k)$.

Вероятности переходов задаются в виде квадратной ***матрицы вероятностей переходов*** $\mathbf{Q} = [q_{ij} \mid i, j = \overline{1,n}]$, элементы которой удовлетворяют условиям:

$$0 \le q_{ij} \le 1; \quad \sum_{j=1}^{n} q_{ij} = 1 \qquad (i, j = \overline{1,n}). \tag{5.2}$$

Матрица, элементы которой удовлетворяют указанным условиям, называется ***стохастической***.

Последнее условие в виде суммы элементов каждой строки матрицы вероятностей переходов, равной единице, означает, что в момент времени t_k случайный процесс с вероятностью единица выполнит переход в одно из n возможных состояний, включая то же самое состояние, из которого этот переход осуществляется, то есть процесс может остаться в том же состоянии.

Для случайных процессов с непрерывным временем время между переходами из одного состояния в другое случайно. Это означает, что вероятность перехода из одного состояния в другое не может быть задана, поскольку вероятность такого перехода точно в произвольный момент времени t равна нулю. Для описания переходов между состояниями случайного процесса с непрерывным временем вместо вероятностей переходов вводится параметр, называемый *интенсивностью перехода*.

Интенсивность перехода g_{ij} из состояния $\mathbf{E}_i$ в состояние $\mathbf{E}_j$ определяется как предел отношения вероятности перехода $P_{ij}(\Delta\tau)$ системы за промежуток времени $\Delta\tau$ из $\mathbf{E}_i$ в $\mathbf{E}_j$ к длине этого промежутка:

$$g_{ij} = \lim_{\Delta\tau \to 0} \frac{P_{ij}(\Delta\tau)}{\Delta\tau} \qquad (i, j = \overline{1,n};\ i \ne j). \tag{5.3}$$

Отсюда следует, что вероятность перехода за бесконечно малый промежуток времени $\Delta\tau$ равна: $g_{ij}\Delta\tau \quad (i \ne j)$. Вероятность двух и более переходов за время $\Delta\tau$ имеет порядок $(\Delta\tau)^2$ и выше и предполагается бесконечно малой величиной.

Если интенсивности переходов постоянны и не зависят от времени t, то есть от того, в какой момент начинается промежуток $\Delta\tau$, то марковский процесс называется ***однородным***. Если интенсивности g_{ij} представ-

ляют собой функции времени *t*, процесс называется ***неоднородным***.

В дальнейшем будем рассматривать только однородные марковские процессы.

Интенсивности переходов задаются в виде квадратной матрицы $\mathbf{G} = [g_{ij} \mid i, j = \overline{1,n}]$, называемой ***матрицей интенсивностей переходов***, диагональные элементы которой определяются из условия:

$$\sum_{j=1}^{n} g_{ij} = 0 \quad (i = \overline{1,n}),$$

откуда

$$g_{ii} = -\sum_{\substack{j=1 \\ j \neq i}}^{n} g_{ij} \quad (i, j = \overline{1,n}). \tag{5.4}$$

Матрица, в которой сумма элементов в каждой строке равна нулю, называется ***дифференциальной***.

Выше было показано, что в случае экспоненциального закона распределения времени нахождения случайного процесса в некотором состоянии вероятность перехода из одного состояния в другое за бесконечно малый интервал времени определяется выражением (5.1) и равно $P_{ij}(\Delta\tau) = \alpha_{ij}\Delta\tau$. Отсюда следует, что *интенсивность перехода представляет собой параметр экспоненциального распределения*:

$$g_{ij} = \lim_{\Delta\tau \to 0} \frac{P_{ij}(\Delta\tau)}{\Delta\tau} = \alpha_{ij}.$$

Начальные вероятности $p_1(0), \ldots, p_n(0)$, где $p_i(0)$ – вероятность того, что в момент времени $t = 0$ система находится в состоянии $\mathbf{E}_i$ $(i = 1, \ldots n)$, задают состояние системы в начальный момент времени $t = 0$.

Начальные вероятности необходимы при изучении переходных процессов, когда исследуемая система работает в нестационарном режиме. Если марковский процесс обладает эргодическим свойством, что означает работу моделируемой системы в установившемся режиме, то, как будет показано ниже, стационарные характеристики (вероятности) не зависят от начальных вероятностей и, следовательно, могут быть не заданы.

5.2.2. Характеристики марковского случайного процесса

Изучение случайных процессов заключается в определении вероятностей того, что в момент времени t система находится в том или ином состоянии. Совокупность таких вероятностей, описывающих состояния системы в различные моменты времени, дают достаточно полную информацию о протекающем в системе случайном процессе.

Рассмотрим систему с конечным числом состояний: $\mathbf{E}_1$, …, $\mathbf{E}_n$. Обозначим через $p_i(t)$ вероятность того, что в момент времени *t* система

находится в состоянии $\mathbf{E}_i$: $p_i(t) = \Pr\{Z(t) = E_i\}$.

В любой момент времени t система может находиться в одном из n возможных состояний, то есть для любого момента времени t выполняется условие:

$$\sum_{i=1}^{n} p_i(t) = 1, \qquad (5.5)$$

которое называется ***нормировочным***.

Совокупность вероятностей $p_i(t)$ может быть представлена вектором с числом координат, равным числу возможных состояний системы:

$$P(t) = \{p_1(t),..., p_n(t)\},$$

причем

$$0 \le p_i(t) \le 1; \quad \sum_{i=1}^{n} p_i(t) = 1. \qquad (5.6)$$

Вектор, обладающий свойствами (5.6), называется ***стохастическим***.

Стохастический вектор называется ***вектором состояний***, если его компоненты представляют собой вероятности состояний системы.

Вектор состояний $P(t) = \{p_1(t),..., p_n(t)\}$ является основной характеристикой марковского случайного процесса. На основе полученных значений вероятностей состояний случайного процесса, протекающего в исследуемой системе, могут быть рассчитаны представляющие интерес реальные характеристики системы, например для системы массового обслуживания могут быть рассчитаны длины очередей заявок.

5.3. Методы расчета марковских моделей

5.3.1. Эргодическое свойство случайных процессов

Если по истечении достаточно большого промежутка времени вероятности состояний стремятся к предельным значениям $p_1,\ldots, p_n$, не зависящим от начальных вероятностей $p_1(0),\ldots, p_n(0)$ и от текущего момента времени t, то говорят, что случайный процесс обладает ***эргодическим свойством***. Таким образом, для процессов, обладающих эргодическим свойством:

$$\lim_{t\to\infty} P(t) = P(\infty) = \mathbf{P},$$

где $\mathbf{P} = (p_1,\ldots, p_n)$ – вектор вероятностей состояний системы, называемых ***стационарными вероятностями***.

В системе, описываемой марковским случайным процессом, обладающим эргодическим свойством, при $t \to \infty$ устанавливается некоторый предельный режим, при котором характеристики функционирования системы не зависят от времени. В этом случае говорят, что система

работает в **установившемся** или **стационарном режиме**. Если характеристики функционирования системы зависят от времени, то имеем **неустановившийся режим**.

Отметим, что для стационарных вероятностей p_i должно выполняться нормировочное условие (5.5).

При рассмотрении случайных процессов возникает вполне резонный вопрос: *когда случайный процесс обладает эргодическим свойством*?

Случайный процесс *с дискретным временем* обладает *эргодическим свойством*, если матрица вероятностей переходов **не является** *периодической* или *разложимой*.

Матрица является **разложимой**, если она может быть приведена к одному из следующих видов:

1) $\begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix}$, 2) $\begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$, 3) $\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{D} \end{bmatrix}$,

где **A, B, C, D**– ненулевые квадратные подматрицы; **0** – нулевая квадратная подматрица.

В первом случае состояния, соответствующие подмножествам **A** и **D**, называются **замкнутыми**, так как система, находясь в каком-то состоянии одного из этих подмножеств, никогда не сможет перейти в какое-либо состояние другого подмножества. Состояния, соответствующие подмножеству **D** во втором случае и подмножеству **A** в третьем случае, называются **невозвратными**, поскольку после того, как процесс покинет эти состояния, невозможен обратный переход в эти состояния из состояний, соответствующих другим подмножествам.

Матрица является **периодической**, если она может быть приведена к виду:

$$\begin{bmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{C} & \mathbf{0} \end{bmatrix}.$$

Случайный процесс в этом случае будет по очереди переходить из состояний, соответствующих **B**, в состояния, соответствующие **C**.

Итак, если матрица вероятностей переходов $\mathbf{Q} = [q_{ij} \mid i, j = \overline{1,n}]$, случайного процесса с дискретным временем не является периодической или разложимой, то процесс обладает эргодическим свойством:

$$\lim_{k \to \infty} p_i(k) = p_i \qquad (i = \overline{1,n}). \tag{5.7}$$

Транзитивный случайный процесс с *непрерывным временем* и *конечным* числом состояний, среди которых нет невозвратных и поглощающих состояний, всегда обладает эргодическим свойством:

$$\lim_{t \to \infty} p_i(t) = p_i \qquad (i = \overline{1,n}). \tag{5.8}$$

5.3.2. Марковские процессы с дискретным временем

Для однородного марковского процесса с дискретным временем вероятности состояний на момент времени t_k определяются на основе следующего рекуррентного выражения:

$$p_j(k) = \sum_{i=1}^{n} p_i(k-1)\, q_{ij} \quad (j = \overline{1,n};\ k = 1,\ 2,\ldots). \tag{5.9}$$

Если рассматриваемый марковский процесс обладает эргодическим свойством, то, согласно (5.7), при $k \to \infty$ вероятности состояний $p_i(k)$ стремятся к стационарным значениям p_i, не зависящим от момента времени t_k и начальных вероятностей $p_i(0)$. С учётом этого, выражение (5.9) может быть преобразовано к виду:

$$p_j = \sum_{i=1}^{n} p_i q_{ij} \qquad (j = \overline{1,n}) \tag{5.10}$$

а нормировочное условие (5.5) примет вид:

$$\sum_{i=1}^{n} p_i = 1. \tag{5.11}$$

Уравнения (5.10) с условием (5.11) образуют систему линейных алгебраических уравнений для расчёта стационарных вероятностей состояний марковского процесса, которая обладает единственным решением, если **Q** – эргодическая матрица.

Доказательство выражения (5.9).

Рассмотрим однородный марковский процесс с дискретным временем, который может находиться в одном из *n* возможных состояний: $\mathbf{E}_1$, …, $\mathbf{E}_n$. Вероятности переходов q_{ij} заданы в виде матрицы переходов $\mathbf{Q} = [q_{ij} \mid i, j = \overline{1,n}]$, а начальные вероятности на момент времени $t_0 = 0$ в виде вектора $P = \{p_1(0), \ldots, p_n(0)\}$.

Найдем вероятности состояний марковского процесса после первого шага, то есть на момент времени t_1. По формуле полной вероятности получим:

$$\begin{cases} p_1(1) = p_1(0)q_{11} + p_2(0)q_{21} + \ldots + p_n(0)q_{n1}; \\ p_2(1) = p_1(0)q_{12} + p_2(0)q_{22} + \ldots + p_n(0)q_{n2}; \\ \qquad . \quad . \quad . \\ p_n(1) = p_1(0)q_{1n} + p_2(0)q_{2n} + \ldots + p_n(0)q_{nn}, \end{cases}$$

или в компактной форме:

$$p_j(1) = \sum_{i=1}^{n} p_i(0)\, q_{ij} \qquad (j = \overline{1,n}).$$

Вероятности состояний после второго шага на момент времени t_2 определяются аналогично:

$$p_j(2) = \sum_{i=1}^{n} p_i(1)\, q_{ij} \qquad (j = \overline{1,n}).$$

После k-го шага на момент времени t_k $(k = 1, 2, ...)$ вероятности состояний будут определяться как

$$p_j(k) = \sum_{i=1}^{n} p_i(k-1)\, q_{ij} \qquad (j = \overline{1,n}),$$

что и требовалось доказать.

Пример. Рассмотрим систему, которая состоит из двух устройств $У_1$ и $У_2$, каждое из которых может находиться в одном из двух состояний: **0** – выключено и **1** – включено. В определённые моменты времени устройства могут включаться или выключаться. Выделим возможные состояния системы:

$\mathbf{E}_i$	$\mathbf{E}_0$	$\mathbf{E}_1$	$\mathbf{E}_2$	$\mathbf{E}_3$
$У_1$	0	1	0	1
$У_2$	0	0	1	1

.

Состояние $\mathbf{E}_0$ соответствует простою системы, когда оба устройства выключены, а состояние $\mathbf{E}_3$ соответствует случаю, когда оба устройства включены.

Положим, что заданы вероятности переходов в виде матрицы

$\mathbf{Q} =$	$\mathbf{E}_0$	$\mathbf{E}_1$	$\mathbf{E}_2$	$\mathbf{E}_3$
$\mathbf{E}_0$	0	0,2	0,5	0,3
$\mathbf{E}_1$	0,5	0	0,1	0,4
$\mathbf{E}_2$	0,5	0	0	0,5
$\mathbf{E}_3$	0	0,4	0,6	0

и начальные вероятности $p_0(0) = 0{,}8;\ p_1(0) = 0{,}2;\ p_2(0) = 0;\ p_3(0) = 0$.

Определим вероятности нахождения системы в том или ином состоянии на различные моменты времени.

Согласно выражению (5.9) вероятности состояний системы:

- на момент времени t_1:

$$p_0(1) = p_0(0)q_{00} + p_1(0)q_{10} + p_2(0)q_{20} + p_3(0)q_{30} = 0{,}1;$$
$$p_1(1) = p_0(0)q_{01} + p_1(0)q_{11} + p_2(0)q_{21} + p_3(0)q_{31} = 0{,}16;$$
$$p_2(1) = p_0(0)q_{02} + p_1(0)q_{12} + p_2(0)q_{22} + p_3(0)q_{32} = 0{,}42;$$
$$p_3(1) = p_0(0)q_{03} + p_1(0)q_{13} + p_2(0)q_{23} + p_3(0)q_{33} = 0{,}32;$$

- на момент времени t_2:

$$p_0(2) = p_0(1)q_{00} + p_1(1)q_{10} + p_2(1)q_{20} + p_3(1)q_{30} = 0{,}29;$$

$$p_1(2) = p_0(1)q_{01} + p_1(1)q_{11} + p_2(1)q_{21} + p_3(1)q_{31} = 0{,}148;$$
$$p_2(2) = p_0(1)q_{02} + p_1(1)q_{12} + p_2(1)q_{22} + p_3(1)q_{32} = 0{,}258;$$
$$p_3(2) = p_0(1)q_{03} + p_1(1)q_{13} + p_2(1)q_{23} + p_3(1)q_{33} = 0{,}304.$$

Аналогично вероятности состояний системы могут быть рассчитаны на моменты времени $t_3, t_4, \ldots$.

Нетрудно убедиться, что сумма вероятностей состояний системы на каждый момент времени равна единице: $p_0(k) + p_1(k) + p_2(k) + p_3(k) = 1$ для $k = 1, 2, \ldots$.

Матрица вероятностей переходов рассматриваемой системы – неразложимая и непериодическая, следовательно, случайный процесс обладает эргодическим свойством, и вероятности состояний системы для стационарного режима (стационарные вероятности) p_0, p_1, p_2, p_3 могут быть найдены из системы линейных алгебраических уравнений (5.10) с учётом нормировочного условия (5.11):

$$\begin{cases} p_0 = p_0q_{00} + p_1q_{10} + p_2q_{20} + p_3q_{30} = 0{,}5p_1 + 0{,}5p_2; \\ p_1 = p_0q_{01} + p_1q_{11} + p_2q_{21} + p_3q_{31} = 0{,}2p_0 + 0{,}4p_3; \\ p_2 = p_0q_{02} + p_1q_{12} + p_2q_{22} + p_3q_{32} = 0{,}5p_0 + 0{,}1p_1 + 0{,}6p_3; \\ p_3 = p_0q_{03} + p_1q_{13} + p_2q_{23} + p_3q_{33} = 0{,}3p_0 + 0{,}4p_1 + 0{,}5p_2; \\ p_0 + p_1 + p_2 + p_3 = 1. \end{cases}$$

Решая систему уравнений, получим значения стационарных вероятностей: $p_0 = \frac{13}{55} \approx 0{,}236$, $p_1 = \frac{9}{55} \approx 0{,}164$, $p_2 = \frac{17}{55} \approx 0{,}309$, $p_3 = \frac{16}{55} \approx 0{,}291$.

Таким образом, система будет простаивать 23,6% времени, а более 76% времени система будет находиться в рабочем состоянии, причем почти 30% времени (точнее 29,1%) во включённом состоянии будут одновременно находиться оба устройства системы. Среднее число устройств, находящихся одновременно во включённом состоянии, будет равно: $M = p_1 + p_2 + 2p_3 = 1{,}055$, то есть во включённом состоянии находится в среднем одно устройство.

5.3.3. Марковские процессы с непрерывным временем

Для однородного марковского процесса с непрерывным временем вероятности состояний на произвольный момент времени t определяются из системы дифференциальных уравнений:

$$\frac{dp_j(t)}{dt} = \sum_{i=1}^{n} p_i(t)\, g_{ij} \quad (j = \overline{1,n};\ t > 0) \qquad (5.12)$$

с учетом начальных условий $p_1(0), \ldots, p_n(0)$.

Для систем обладающих эргодическим свойством, имеет место стационарный режим, для которого, согласно (5.8), вероятности состояний

$p_1,\ldots, p_n$ при $t \to \infty$ не зависят от начальных вероятностей и текущего момента времени *t*, и система дифференциальных уравнений (5.12) для установившегося режима преобразуется в систему линейных алгебраических уравнений:

$$\sum_{i=1}^{n} p_i g_{ij} = 0 \quad (j = \overline{1,n}), \tag{5.13}$$

которая совместно с нормировочным условием (5.11) образует систему, обладающую единственным решением.

Доказательство выражений (5.12) и (5.13).

Рассмотрим однородный марковский процесс с непрерывным временем, который может находиться в одном из *n* возможных состояний: $\mathbf{E}_1$, …, $\mathbf{E}_n$. Интенсивности переходов q_{ij} заданы в виде матрицы $\mathbf{G} = [g_{ij} \mid i, j = \overline{1,n}]$, в которой диагональные элементы рассчитаны в соответствии с формулой (5.4). Начальные вероятности на момент времени $t = 0$ заданы в виде вектора $P = \{p_1(0),\ldots, p_n(0)\}$.

Определим вероятность $p_j(t)$ того, что в момент времени $t > 0$ случайный процесс находится в состоянии $\mathbf{E}_j$.

Придадим времени t малое приращение Δt и найдем вероятность $p_j(t + \Delta t)$ того, что случайный процесс в момент времени $(t + \Delta t)$ окажется в состоянии $\mathbf{E}_j$.

Случайный процесс может оказаться в состоянии $\mathbf{E}_j$ в момент $(t + \Delta t)$ двумя способами:

1) в момент времени t процесс находился в состоянии $\mathbf{E}_j$ и в течение промежутка времени Δt не перешел в другое состояние;

2) в момент времени t процесс находился в состоянии $\mathbf{E}_i$ $(i \neq j)$ и за время Δt совершил переход в состояние $\mathbf{E}_j$.

Вероятность первого способа $p_j^{(1)}(t + \Delta t)$ найдем как произведение вероятности $p_j(t)$ того, что в момент времени t случайный процесс находился в состоянии $\mathbf{E}_j$, на условную вероятность того, что, будучи в $\mathbf{E}_j$, процесс не перешел в другие состояния $\mathbf{E}_i$ $(i = 1, 2,..., j-1, j+1,..., n)$. Эта условная вероятность равна

$$1 - \sum_{\substack{i=1 \\ i \neq j}}^{n} g_{ji} \Delta t .$$

Последнее выражение становится очевидным, если вспомнить, что произведение $g_{ji} \Delta t$ с точностью до бесконечно малых высших порядков

определяет вероятность перехода случайного процесса из состояния $\mathbf{E}_j$ в состояние $\mathbf{E}_i$ за промежуток времени Δt, а сумма этих вероятностей есть вероятность перехода из состояния $\mathbf{E}_j$ в любое другое состояние, не совпадающее с $\mathbf{E}_j$. Вычитая эту сумму из единицы, получим требуемую вероятность противоположного события.

Таким образом, вероятность первого способа с учетом выражения (5.4) равна

$$p_j^{(1)}(t+\Delta t) = p_j(t)(1+g_{jj}\Delta t) \qquad (j=\overline{1,n}).$$

Аналогично определяется вероятность $p_j^{(2)}(t+\Delta t)$ второго способа оказаться в состоянии $\mathbf{E}_j$ в момент $(t+\Delta t)$: она равна вероятности $p_j(t)$ того, что в момент времени t процесс находился в состоянии $\mathbf{E}_i$, умноженной на вероятность $g_{ij}\Delta t$ перехода за время Δt в состояние $\mathbf{E}_j$: $p_i(t)\,g_{ij}\Delta t$. Суммируя эти вероятности по всем возможным состояниям, исключая состояние $\mathbf{E}_j$, получим искомую вероятность:

$$p_j^{(2)}(t+\Delta t) = \sum_{\substack{i=1\\ i\neq j}}^{n} p_i(t)\,g_{ij}\,\Delta t \qquad (j=\overline{1,n}).$$

Применив правило сложения вероятностей, получим вероятность нахождения случайного процесса в состоянии $\mathbf{E}_j$ в момент времени $(t+\Delta t)$:

$$p_j(t+\Delta t) = p_j^{(1)}(t+\Delta t) + p_j^{(2)}(t+\Delta t) = p_j(t)(1+g_{jj}\Delta t) + \sum_{\substack{i=1\\ i\neq j}}^{n} p_i(t)\,g_{ij}\,\Delta t$$

или $p_j(t+\Delta t) - p_j(t) = \sum_{i=1}^{n} p_i(t)\,g_{ij}\,\Delta t \qquad (j=\overline{1,n}).$

Разделим левую и правую части последнего выражения на Δt и перейдем к пределу при $\Delta t \to 0$:

$$\lim_{\Delta t\to 0} \frac{p_j(t+\Delta t) - p_j(t)}{\Delta t} = \sum_{i=1}^{n} p_i(t)\,g_{ij} \qquad (j=\overline{1,n}).$$

Левая часть полученного выражения представляет собой производную по времени от функции $p_j(t)$:

$$\frac{dp_j(t)}{dt} = \sum_{i=1}^{n} p_i(t)\,g_{ij} \qquad (j=\overline{1,n}). \tag{5.14}$$

Таким образом, получена система дифференциальных уравнений марковского случайного процесса, которая при заданных начальных условиях $P=\{p_1(0),\ldots,p_n(0)\}$ позволяет выполнить исследование нестационарного (переходного) режима работы моделируемой системы

путем расчёта вероятностей состояний марковского процесса в произвольный момент времени $t > 0$.

Для случайных процессов, обладающих эргодическим свойством, имеет место стационарный режим, для которого согласно (5.8) вероятности состояний при $t \to \infty$ не зависят от начальных вероятностей и текущего момента времени *t*. Тогда производные $\frac{dp_j(t)}{dt} = 0$, и система дифференциальных уравнений (5.14) преобразуется в систему линейных алгебраических уравнений (5.13).

Пример. В качестве примера марковского процесса с непрерывным временем рассмотрим модель «гибели и размножения», которая часто встречается в разнообразных практических задачах. Своим названием эта модель обязана биологической задаче об изменении численности популяции и распространении эпидемий, которая формулируется следующим образом.

Рассмотрим развитие некоторой популяции, особи которой могут рождаться и умирать. Положим, что при наличии *i* особей в популяции рождение новых особей происходит с интенсивностью λ_i и с интенсивностью μ_i – особи умирают. Пусть в любой момент времени может происходить рождение или гибель только одной особи, и интервалы времени между двумя моментами рождения и гибели распределены по экспоненциальному закону с параметрами λ_i и μ_i соответственно. Тогда процесс "гибели и размножения" может быть представлен марковским случайным процессом с непрерывным временем (рис.5.1,а), в котором состояние $\mathbf{E}_i$ соответствует наличию *i* особей в популяции (*i*=0, 1, …), причем число состояний может быть конечным или бесконечным. Отметим, что состояние $\mathbf{E}_0$ соответствует вырождению популяции.

Таким образом, марковский процесс называется «**процессом гибели и размножения**», если её граф переходов имеет вид цепочки состояний, в которой каждое состояние (кроме крайних) связано с двумя соседними состояниями, а крайние состояния $\mathbf{E}_0$ и $\mathbf{E}_n$ (в случае конечного числа состояний) или только нулевое состояние $\mathbf{E}_0$ (в случае бесконечного числа состояний) – только с одним соседним состоянием.

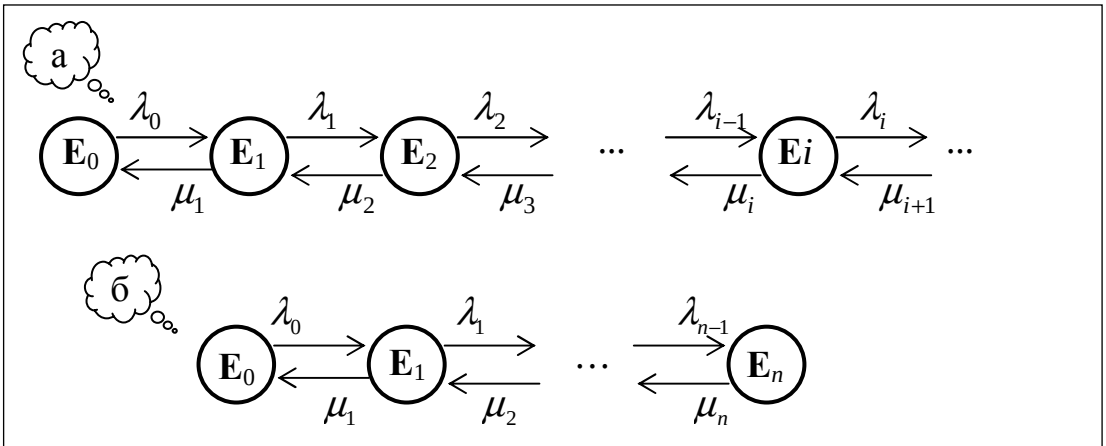


Рис. 5.2. Графы состояний процесса гибели и размножения с бесконечным (а) и конечным (б) числом состояний

Графу переходов процесса гибели и размножения с конечным числом состояний (рис.5.2,б) соответствует матрица интенсивностей переходов:

$$\mathbf{G} = \begin{array}{c|cccccc} \mathbf{E}_i & 0 & 1 & 2 & \dots & n-1 & n \\ \hline 0 & -\lambda_0 & \lambda_0 & 0 & \dots & 0 & 0 \\ 1 & \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & \dots & 0 & 0 \\ 2 & 0 & \mu_2 & -(\lambda_2+\mu_2) & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ n-1 & 0 & 0 & 0 & \dots & -(\lambda_{n-1}+\mu_{n-1}) & \lambda_{n-1} \\ n & 0 & 0 & 0 & \dots & \mu_n & -\mu_n \end{array}.$$

Диагональные элементы матрицы определяются из условия (5.4) – сумма элементов каждой строки должна быть равна нулю.

Система линейных алгебраических уравнений для определения стационарных вероятностей может быть составлена по графу переходов или по матрице интенсивностей переходов.

Сформулируем **правила составления уравнений для стационарных вероятностей состояний** марковского процесса с непрерывным временем по графу переходов и по матрице интенсивностей переходов.

Правило 1 (по графу переходов). В левой части каждого уравнения записывается вероятность рассматриваемого состояния, умноженная на сумму интенсивностей переходов из данного состояния во все другие состояния. Правая часть уравнения представляет собой сумму членов, число которых равно числу входящих в данное состояние дуг, и каждый такой член представляет собой произведение интенсивности перехода, соответствующей данной дуге, на вероятность состояния, из которого исходит эта дуга.

Для нашего примера применение правила 1 дает следующую систему линейных алгебраических уравнений:

$$\begin{cases} \lambda_0 p_0 = \mu_1 p_1 \\ (\lambda_1+\mu_1) p_1 = \lambda_0 p_0 + \mu_2 p_2 \\ \dots \\ (\lambda_k+\mu_k) p_k = \lambda_{k-1} p_{k-1} + \mu_{k+1} p_{k+1} \\ \dots \\ \mu_n p_n = \lambda_{n-1} p_{n-1} \\ p_0 + p_1 + \dots + p_n = 1 \end{cases},$$

где последнее уравнение представляет собой нормировочное условие (5.11).

Правило 2 (по матрице интенсивностей переходов). Для каждого столбца матрицы интенсивностей переходов составляется соответствующее уравнение как сумма произведений интенсивностей переходов на стационарную вероятность состояния с номером соответствующей строки, приравненная нулю.

Применение правила 2 для нашего примера дает следующую систему линейных алгебраических уравнений:

$$\begin{cases} -\lambda_0 p_0 + \mu_1 p_1 = 0 \\ -(\lambda_1 + \mu_1) p_1 + \lambda_0 p_0 + \mu_2 p_2 = 0 \\ \ldots \\ -(\lambda_k + \mu_k) p_k + \lambda_{k-1} p_{k-1} + \mu_{k+1} p_{k+1} = 0 \\ \ldots \\ -\mu_n p_n + \lambda_{n-1} p_{n-1} = 0 \\ p_0 + p_1 + \ldots + p_n = 1 \end{cases}.$$

Легко убедиться, что обе системы уравнений эквивалентны.

Решая полученную систему уравнений аналитически или с применением численных методов, можно определить значения $p_0, p_1, \ldots, p_n$ стационарных вероятностей состояний марковского процесса. Кроме того, могут быть рассчитаны другие характеристики исследуемой системы, в частности, среднее число особей в популяции как математическое ожидание случайной величины:

$$M = \sum_{k=1}^{n} k p_k .$$

5.4. Марковские модели систем массового обслуживания

В данном параграфе в качестве примеров применения случайных процессов для изучения свойств систем с дискретным характером функционирования подробно рассматриваются марковские модели систем массового обслуживания (СМО). Примеры представлены в порядке возрастания их сложности, начиная с простейшей одноканальной СМО с однородным потоком заявок без накопителя и заканчивая СМО с накопителями ограниченной ёмкости и приоритетным обслуживанием неоднородного потока заявок.

В каждом примере приводится *описание* исследуемой системы, а также *предположения и допущения*, принятые при построении математической модели и необходимые для того, чтобы протекающий в системе случайный процесс был марковским. Разработка марковской модели исследуемой системы в терминах случайных процессов предполагает

выполнение следующих этапов:

- *кодирование состояний* случайного процесса;
- построение *размеченного графа переходов*;
- формирование *матрицы интенсивностей переходов*;
- составление *системы линейных алгебраических уравнений* для расчёта стационарных вероятностей состояний марковского процесса.

Матрица интенсивностей переходов может использоваться для задания системы линейных алгебраических уравнений в матричном виде при компьютерном расчёте стационарных вероятностей.

При исследовании различного рода *реальных* систем, моделями которых служат СМО, вряд ли кого-то интересуют вероятности состояний. Гораздо больший интерес представляют такие характеристики СМО, как длина очереди заявок перед обслуживающим прибором, время ожидания и время пребывания заявок в системе, загрузка и коэффициент простоя системы, доля потерянных заявок и т.д., значения которых могут быть рассчитаны по найденным значениям стационарных вероятностей состояний. Поэтому ниже особое внимание уделяется математическим зависимостям, позволяющим рассчитать в каждом конкретном примере наиболее важные характеристики функционирования исследуемых систем. Максимально подробно процесс получения таких зависимостей изложен в нескольких первых рассматриваемых ниже примерах. На основе этих зависимостей в некоторых примерах проводится анализ свойств исследуемой системы. Для остальных примеров подобный анализ рекомендуется читателю выполнить самостоятельно.

В первом примере (п.5.4.1) одноканальной СМО без накопителя представлена диаграмма функционирования исследуемой системы, с помощью которой показано, что протекающий в системе случайный процесс при сформулированных предположениях и допущениях (а это, прежде всего, экспоненциальный характер процессов поступления и обслуживания заявок) является марковским. Аналогично, и для остальных систем можно показать, что протекающие в них случайные процессы при сформулированных предположениях и допущениях являются марковскими.

В некоторых случаях на основе марковских моделей могут быть получены математические выражения для расчёта стационарных вероятностей состояний в явном виде без применения методов численного анализа. В частности, такие результаты представлены ниже для одноканальной и многоканальной СМО без накопителя (с отказами) и одноканальной СМО с накопителем неограниченной и ограниченной ёмкости.

5.4.1. Одноканальная СМО без накопителя (M/M/1/0)

Рассмотрим простейшую одноканальную систему массового обслуживания (СМО) с отказами, в которую поступает случайный поток заявок, задерживаемых в приборе на случайное время (рис.5.3,а). Посколь-

ку перед обслуживающим прибором нет накопителя, то заявка, поступившая в систему и заставшая прибор занятым, получает отказ в обслуживании и теряется. Таким образом, в системе, кроме входящего потока заявок с интенсивностью λ, образуются еще два потока: выходящий поток обслуженных в приборе заявок с интенсивностью λ' и поток необслуженных заявок (получивших отказ в обслуживании) с интенсивностью λ''. Очевидно, что $\lambda' + \lambda'' = \lambda$.

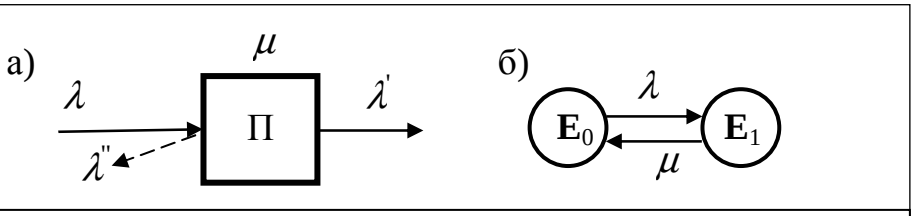


Рис. 5.3. СМО с отказами (а) и её граф переходов (б)

1. Описание системы.

1.1. Система содержит один обслуживающий прибор (П), то есть является *одноканальной*.

1.2. В систему поступает один класс заявок, то есть поток заявок *однородный*.

1.3. В приборе происходит задержка (обслуживание) поступающих в систему заявок на некоторое случайное время.

1.4. Перед прибором не предусмотрены места для ожидания заявок, то есть в системе отсутствует накопитель.

2. Предположения и допущения.

2.1. Поступающие в систему заявки образуют *простейший* поток с интенсивностью λ.

2.2. Длительность обслуживания заявок в приборе распределена по *экспоненциальному* закону с интенсивностью $\mu = 1/b$, где b – средняя длительность обслуживания.

2.3. Дисциплина буферизации – *с отказами*: заявка, поступившая в систему и заставшая прибор занятым обслуживанием другой заявки, теряется.

2.4. Дисциплина обслуживания – *в естественном порядке*: заявка, поступившая в систему и заставшая прибор свободным, принимается на обслуживание.

Очевидно, что в СМО с отказами всегда будет существовать установившийся режим, поскольку даже при больших значениях нагрузки ($y >> 1$) число заявок в системе не может вырасти до бесконечности. Это обусловлено тем, что с ростом нагрузки увеличивается доля заявок, получающих отказ в обслуживании.

3. Кодирование состояний случайного процесса.

В качестве параметра, описывающего состояние случайного процесса, будем рассматривать количество заявок *k*, находящихся в СМО. Очевидно, что система в любой момент времени может находиться в

одном из двух состояний:

$\mathbf{E}_0$: $k = 0$ – в системе нет заявок (прибор простаивает);

$\mathbf{E}_1$: $k = 1$ – в системе (на обслуживании в приборе) находится 1 заявка (прибор работает).

4. Размеченный граф переходов случайного процесса (рис.5.3,б).

В процессе функционирования рассматриваемой системы в один и тот же момент времени может наступить только одно из двух возможных событий, которые приводят к изменению состояния случайного процесса, протекающего в системе.

1. *Поступление заявки в систему* с интенсивностью λ. При этом:

- если случайный процесс находится в состоянии $\mathbf{E}_0$ (прибор простаивает), то произойдет переход в состояние $\mathbf{E}_1$ (начнется обслуживание поступившей заявки), причем *интенсивность перехода совпадает с интенсивностью поступления* заявок в систему λ;
- если же случайный процесс находится в состоянии $\mathbf{E}_1$ (прибор работает), то состояние $\mathbf{E}_1$ случайного процесса не изменится, что будет соответствовать отказу в обслуживании поступившей заявке.

Таким образом, переход из состояния $\mathbf{E}_0$ в состояние $\mathbf{E}_1$ происходит с интенсивностью λ.

2. *Завершение обслуживания заявки*, находящейся в приборе. Это событие может наступить только в том случае, если в приборе на обслуживании находится заявка, то есть случайный процесс находится в состоянии $\mathbf{E}_1$. При этом происходит переход в состояние $\mathbf{E}_0$, причем *интенсивность перехода совпадает с интенсивностью обслуживания* заявки в приборе μ.

5. Диаграммы функционирования системы.

Рассмотрим диаграммы функционирования системы и покажем, что случайный процесс, протекающий в системе, при сформулированных выше предположениях является марковским.

На рис.5.4 показаны диаграммы следующих процессов:

а) **поступления** в СМО заявок, интервалы между которыми в случае простейшего потока распределены по экспоненциальному закону;

б) **перехода** из состояния $\mathbf{E}_0$ в состояние $\mathbf{E}_1$ и обратно, в которых может находиться система, при этом время нахождения случайного процесса в состоянии $\mathbf{E}_1$ равно длительности обслуживания заявки в приборе, которая представляет собой случайную величину, распределенную по экспоненциальному закону;

в) **выхода** из системы обслуженных заявок в моменты времени t_1', t_3', t_4', t_7';

г) **выхода** из системы необслуженных заявок, получивших отказ из-за занятости прибора в моменты времени t_2', t_5', t_6', t_8';

д) **формирования** интервалов времени между соседними переходами случайного процесса.

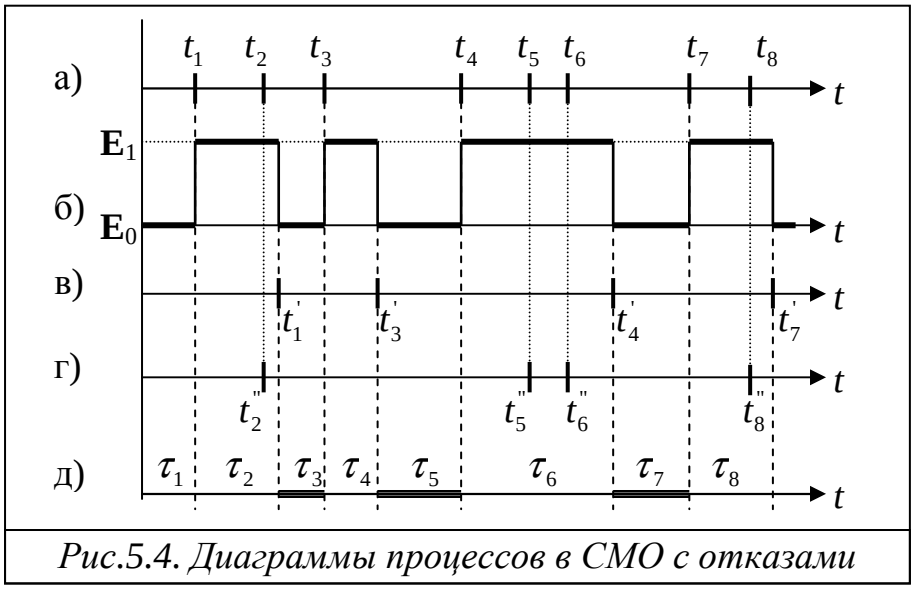


Рис.5.4. Диаграммы процессов в СМО с отказами

Как показано в п. 5.1.2, дискретный случайный процесс с непрерывным временем будет марковским, если интервалы между соседними переходами распределены по *экспоненциальному* закону.

В нашем случае, интервал τ_1 представляет собой интервал между поступающими заявками, который для простейшего потока имеет экспоненциальное распределение. Интервалы $\tau_2, \tau_4, \tau_6, \tau_8$, как видно из диаграммы, представляют собой время нахождения случайного процесса в состоянии $\mathbf{E}_1$, равное длительности обслуживания заявки в приборе, которая распределена по экспоненциальному закону. Таким образом, интервалы $\tau_1, \tau_2, \tau_4, \tau_6, \tau_8$ *распределены по экспоненциальному закону* и, следовательно, удовлетворяют сформулированному выше условию.

Рассмотрим теперь выделенные интервалы τ_3, τ_5, τ_7. Каждый из этих интервалов представляет собой промежуток времени от момента завершения обслуживания некоторой заявки до момента поступления новой заявки, принимаемой на обслуживание в приборе. В п.5.1.2 сформулировано замечательное *свойство экспоненциального распределения*, которое гласит, что в случае экспоненциального распределения интервалов времени между двумя событиями *интервал времени от любого случайного момента до момента наступления очередного события* имеет *такое же экспоненциальное распределение с тем же параметром*. В соответствии с этим свойством интервалы τ_3, τ_5, τ_7 имеют *экспоненциальное распределение* с параметром λ и, следовательно, также удовлетворяют сформулированному выше условию для марковского процесса.

Таким образом, случайный процесс, протекающий в системе с простейшим потоком заявок и экспоненциальным обслуживанием, является марковским.

6. Матрица интенсивностей переходов.

Графу переходов (рис.5.3) соответствует **матрица интенсивностей переходов**:

$$\mathbf{G} = \begin{array}{c|cc} \mathbf{E}_i & 0 & 1 \\ \hline 0 & -\lambda & \lambda \\ 1 & \mu & -\mu \end{array}.$$

Действительно, переход из состояния $\mathbf{E}_0$ в состояние $\mathbf{E}_1$ соответствует поступлению заявки в систему с интенсивностью λ, а переход из состояния $\mathbf{E}_1$ в состояние $\mathbf{E}_0$ соответствует завершению обслуживания заявки в приборе с интенсивностью μ.

Диагональные элементы матрицы определяются из условия (5.4) – сумма элементов каждой строки должна быть равна нулю.

7. Система уравнений.

Система уравнений для определения стационарных вероятностей, составленная по графу переходов с применением *правила 1*, имеет вид:

$$\begin{cases} \lambda p_0 = \mu p_1 \\ \mu p_1 = \lambda p_0 \\ p_0 + p_1 = 1 \end{cases},$$

где последнее уравнение представляет собой нормировочное условие (5.10).

Учитывая, что первое и второе уравнение одинаковы (или, как говорят математики, линейно зависимы) и удаляя одно из них, окончательно получим:

$$\begin{cases} \lambda p_0 = \mu p_1 \\ p_0 + p_1 = 1 \end{cases}.$$

Решая эту систему, получим следующие значения стационарных вероятностей состояний марковского процесса:

$$p_0 = \frac{\mu}{\lambda + \mu} = \frac{1}{1+y}; \qquad p_1 = \frac{\lambda}{\lambda + \mu} = \frac{y}{1+y},$$

где $y = \lambda / \mu$ - нагрузка системы.

8. Расчет характеристик СМО.

Для расчета характеристик СМО можно воспользоваться следующими математическими зависимостями, вытекающими из зависимостей (3.6) – (3.18):

1) нагрузка: $y = \lambda / \mu = \lambda b$ (по определению);

2) загрузка определяется как *вероятность работы прибора*: $\rho = p_1$ и не совпадает с нагрузкой даже в случае $y < 1$, что характерно для систем с отказами и потерями заявок, причём всегда $\rho < y$;

3) коэффициент простоя системы определяется как вероятность отсутствия заявок в системе или, по определению, через загрузку системы: $\eta = p_0 = 1 - \rho$;

4) среднее число заявок в системе: $m = p_1 = \rho$, определяемое как математическое ожидание случайной величины: в системе может

находиться либо ноль заявок с вероятностью p_0, либо одна заявка с вероятностью p_1, тогда среднее число заявок равно $m = 0 \cdot p_0 + 1 \cdot p_1 = p_1$;

5) вероятность потери заявок в результате отказа в обслуживании из-за занятости прибора в соответствии с (3.18) совпадает с вероятностью того, что система занята обслуживанием заявок:

$$\pi = \pi_n = 1 - \frac{\rho}{y} K = 1 - \frac{p_1}{y} = \frac{y}{1+y} = p_1,$$

где учтено, что для рассматриваемой СМО без накопителя $p_1 = \frac{y}{1+y}$;

6) производительность системы: $\lambda' = (1-\pi)\lambda$, определяемая как интенсивность потока обслуженных заявок на выходе системы;

7) интенсивность потока не обслуженных заявок, то есть получивших отказ: $\lambda'' = \pi\lambda$;

8) среднее время пребывания заявок в системе: $u = m / \lambda' = b$ (определяется по формуле Литтла (3.15) и, как следовало ожидать, равно средней длительности обслуживания заявок; отметим, что в формуле Литтла используется интенсивность λ' потока *обслуженных* заявок, а не входящего потока λ).

9. Анализ свойств системы.

Анализ полученных зависимо-стей (рис.5.5) показывает, что с ростом нагрузки коэффициент простоя системы, равный p_0, уменьшается, а загрузка системы, определяемая как вероятность p_1 того, что прибор работает, (а также среднее число зая-вок в системе и вероятность отказа) увеличивается, причем их сумма всегда равна единице. При $y \to \infty$ коэффициент простоя $\eta \to 0$, в то время как загрузка $\rho \to 1$. Заметим также, что нагрузка системы определяется через стационарные вероятности как отношение вероятности работы системы к вероятности простоя: $y = p_1 / p_0$ (что легко может быть получено из выражений для p_0 и p_1) или, что то же самое, через загрузку и коэффициент простоя: $y = \rho / \eta$.

5.4.2. Многоканальная СМО без накопителя (M/M/N/0)

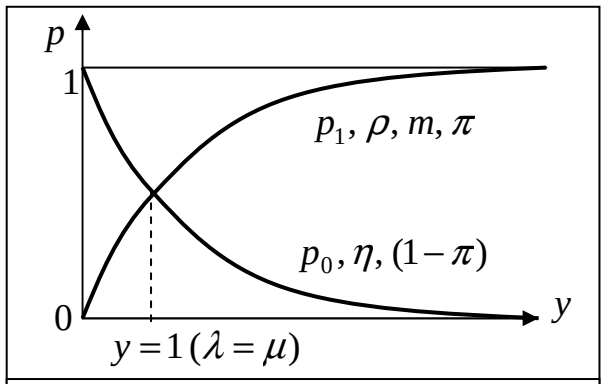


Рис.5.5. Характеристики СМО

Рассмотрим многоканальную систему массового обслуживания (СМО) с отказами, в которую поступает случайный поток заявок, задерживаемых в приборе на случайное время (рис.5.6). Заявка, поступившая в систему и заставшая прибор занятым, получает отказ в обслуживании и теряется. Таким образом, в системе, кроме входящего потока заявок с интенсивностью λ, образуются еще два потока: выходящий поток обслуженных в приборе заявок с интенсив-ностью λ' и поток необслуженных заявок (получивших отказ в обслу-живании) с интенсивностью λ''. Очевидно, что $\lambda' + \lambda'' = \lambda$.

1. Описание системы.

1.1. Система содержит N обслуживающих приборов $П_1,\ldots,П_N$, то есть является *многоканальной*.

1.2. В систему поступает один класс заявок (поток *однородный*).

1.3. Все приборы идентичны, то есть любая заявка может быть обслужена любым прибором за одно и то же случайное время.

1.4. В системе отсутствует накопитель.

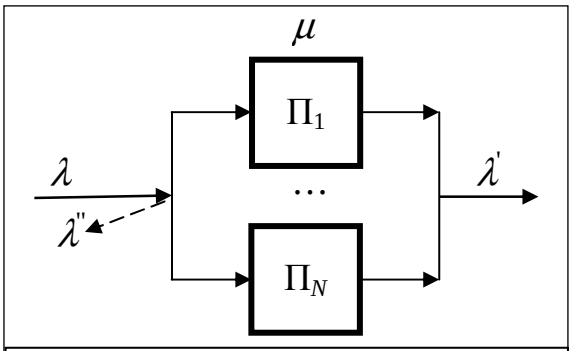


Рис. 5.6. Многоканальная СМО с отказами

2. Предположения и допущения.

2.1. Поступающие в систему заявки образуют *простейший* поток с интенсивностью λ.

2.2. Длительность обслуживания заявок в любом приборе распределена по экспоненциальному закону с интенсивностью $\mu = 1/b$, где b – средняя длительность обслуживания заявок в приборе.

2.3. Дисциплина буферизации – *с отказами*: заявка, поступившая в систему и заставшая все приборы занятыми обслуживанием других заявок, теряется.

2.4. Дисциплина обслуживания – *в естественном порядке*: заявка, поступившая в систему принимается на обслуживание, если есть хотя бы один свободный прибор. Если заявка застала свободными несколько приборов, то она направляется в один из них случайным образом.

3. Кодирование состояний случайного процесса.

В качестве параметра, описывающего состояние случайного процесса, как и ранее, будем рассматривать количество заявок k, находящихся в СМО. При этом система в любой момент времени может находиться в одном из $(N+1)$ состояний:

$\mathbf{E}_0$: $k = 0$ – в системе нет заявок (система простаивает);

$\mathbf{E}_1$: $k = 1$ – в системе находится 1 заявка (один прибор работает, остальные – простаивают);

$\mathbf{E}_2$: $k = 2$ – в системе находится 2 заявки (два прибора работают,

остальные – простаивают);

…

$\mathbf{E}_N$: $k = N$ – в системе находится N заявок (все приборы работают).

4. Размеченный граф переходов случайного процесса представлен на рис.5.7.

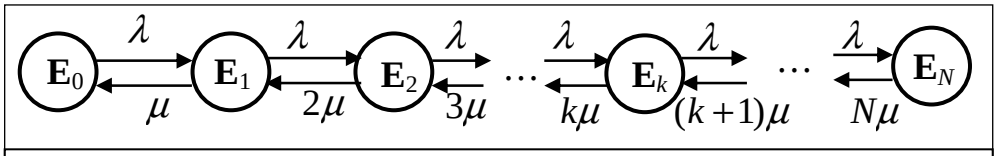


Рис. 5.7. Граф переходов марковского процесса

В один и тот же момент времени в системе может произойти только одно из двух событий, которые приводят к изменению состояния случайного процесса.

1. *Поступление заявки в систему* с интенсивностью λ. При этом:

•если случайный процесс находится в состоянии $\mathbf{E}_k$, причем $k < N$, то произойдет переход в состояние $\mathbf{E}_{k+1}$ (начнется обслуживание поступившей заявки в одном из свободных приборов), причем интенсивность перехода равна интенсивности поступления λ;

•если же случайный процесс находится в состоянии $\mathbf{E}_N$ (все приборы заняты обслуживанием заявок), то состояние $\mathbf{E}_N$ случайного процесса не изменится, что будет соответствовать отказу в обслуживании поступившей заявке.

Таким образом, переход из состояний $\mathbf{E}_k$ в состояние $\mathbf{E}_{k+1}$ (при $k < N$) происходит с интенсивностью λ.

2. *Завершение обслуживания заявки* в одном из приборов с интенсивностью μ.

Это событие может наступить только в том случае, если в системе на обслуживании находится хотя бы одна заявка, то есть случайный процесс находится в состояниях $\mathbf{E}_1$, $\mathbf{E}_2$, …, $\mathbf{E}_N$,. При этом случайный процесс переходит соответственно в состояния $\mathbf{E}_0$, $\mathbf{E}_1$, …, $\mathbf{E}_{N-1}$, причём интенсивности перехода различны. Действительно, если в системе обслуживается только одна заявка (состояние $\mathbf{E}_1$), то интенсивность перехода в состояние $\mathbf{E}_0$ равна μ. Если же в системе обслуживается две заявки (состояние $\mathbf{E}_2$), то есть работают два прибора, то переход случайного процесса в состояние $\mathbf{E}_1$ возможен либо в результате завершения обслуживания заявки в первом приборе с интенсивностью μ, либо в результате завершения обслуживания заявки во втором приборе с такой же интенсивностью μ, причём вероятность завершения обслуживания заявок в обоих приборах в один и тот же момент времени равна нулю. Таким образом, интенсивность перехода из состояния $\mathbf{E}_2$ в состояние $\mathbf{E}_1$ будет равна 2μ (как сумма интенсивностей двух рассмотренных способов).

В общем случае, если в многоканальной системе на обслуживании находится $k = 1,2,..., N$ заявок (случайный процесс находится в состоянии

$\mathbf{E}_k$), то интенсивность перехода в состояние $\mathbf{E}_{k-1}$ будет равна $k\mu$.

По аналогии с предыдущим примером (п.5.4.1) здесь и в последующих примерах можно показать, что случайный процесс, протекающий в системе, при сформулированных предположениях является марковским.

5. Матрица интенсивностей переходов.

Графу переходов (рис.5.7) соответствует матрица интенсивностей переходов:

$$\mathbf{G} = \begin{array}{c|cccccc} \mathbf{E}_i & 0 & 1 & 2 & \dots & N-1 & N \\ \hline 0 & -\lambda & \lambda & 0 & \dots & 0 & 0 \\ 1 & \mu & -(\lambda+\mu) & \lambda & \dots & 0 & 0 \\ 2 & 0 & 2\mu & -(\lambda+2\mu) & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ N-1 & 0 & 0 & 0 & \dots & -(\lambda+(N-1)\mu) & \lambda \\ N & 0 & 0 & 0 & \dots & N\mu & -N\mu \end{array}.$$

Диагональные элементы матрицы определяются из условия (5.4) – сумма элементов каждой строки должна быть равна нулю.

6. Система уравнений.

Система уравнений для определения стационарных вероятностей имеет вид:

$$\begin{cases} \lambda p_0 = \mu p_1 \\ (\lambda+\mu) p_1 = \lambda p_0 + 2\mu p_2 \\ \dots \\ (\lambda+k\mu) p_k = \lambda p_{k-1} + (k+1)\mu p_{k+1} \\ \dots \\ N\mu p_N = \lambda p_{N-1} \\ p_0 + p_1 + \dots + p_N = 1 \end{cases}.$$

Используя метод математической индукции можно показать, что:

$$p_k = \frac{y^k}{k!} p_0 \quad (k = \overline{0, N}),$$

где $y = \lambda b$ – нагрузка системы.

Подставляя полученное выражение в последнее уравнение системы линейных алгебраических уравнений, найдем вероятность простоя системы:

$$p_0 = \frac{1}{\sum_{i=0}^{N} \frac{y^i}{i!}}.$$

Тогда стационарные вероятности состояний марковского случайного процесса, протекающего в многоканальной СМО с отказами:

$$p_k = \frac{\frac{y^k}{k!}}{\sum_{i=0}^{N}\frac{y^i}{i!}} \qquad (k = \overline{0, N}).$$

Из последнего выражения при $N = 1$, как частный случай, вытекает результат, полученный в предыдущем примере для одноканальной СМО с отказами.

Задание на самостоятельную работу: *проверить полученные выражения, используя метод математической индукции.*

7. Расчет характеристик СМО.

Для расчета характеристик СМО можно воспользоваться следующими математическими зависимостями:

1) нагрузка: $y = \lambda / \mu = \lambda b$ (по определению);

2) загрузка: $\rho = \frac{1}{N}\sum_{k=0}^{N} k p_k$, учитывающая долю $\left(\frac{k}{N}\right)$ работающих приборов; действительно, система загружена полностью, когда работают все приборы, если же из 10 приборов работает один, то система загружена на 10%, если работают 5 приборов, то система загружена на 50%;

3) коэффициент простоя системы: $\eta = \frac{1}{N}\sum_{k=0}^{N}(N-k)\,p_k = 1-\rho$;

4) среднее число заявок в системе, равное среднему числу работающих приборов: $m = \sum_{k=1}^{N} k p_k = N\rho$;

5) среднее число простаивающих приборов: $\hat{N} = N - m$;

6) вероятность отказа в обслуживании, определяемая как вероятность того, что все приборы заняты обслуживанием заявок: $\pi = p_N = \frac{y^N}{N!} \Big/ \sum_{i=0}^{N}\frac{y^i}{i!}$;

Задание на самостоятельную работу: *доказать последнее выражение для вероятности потери заявок, подставив полученное выше выражение для стационарных вероятностей состояний в формулу (3.18).*

7) производительность системы, определяемая как интенсивность потока обслуженных заявок: $\lambda' = \lambda(1-\pi)$;

8) интенсивность потока не обслуженных заявок, то есть получивших отказ: $\lambda'' = \lambda\pi$;

9) среднее время пребывания заявок в системе: $u = m/\lambda' = b$.

8. Анализ свойств системы.

Анализ свойств многоканальной СМО без накопителя показывает, что с увеличением нагрузки уменьшается вероятность простоя системы и увеличивается загрузка системы, а вместе с ней число работающих

приборов и вероятность отказа.

5.4.3. Одноканальная СМО с накопителем ограниченной емкости (M/M/1/r)

1. Описание системы.

1.1. Система (рис.5.8) содержит один обслуживающий прибор (П), то есть является *одноканальной*.

1.2. Поток поступающих в систему заявок *однородный*.

1.3. Длительность обслуживания заявок в приборе – величина *случайная*.

1.4. Перед прибором имеется r мест для заявок, ожидающих обслуживания и образующих очередь, то есть в системе имеется накопитель *ограниченной* ёмкости: $r < \infty$.

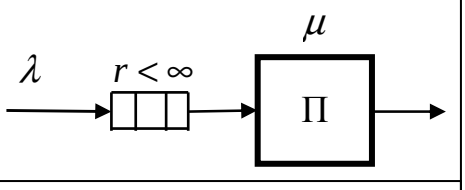


Рис. 5.8. СМО с накопителем ограниченной ёмкости

2. Предположения и допущения.

2.1. Поступающие в систему заявки образуют *простейший* поток с интенсивностью λ.

2.2. Длительность обслуживания заявок в приборе распределена по *экспоненциальному* закону с интенсивностью $\mu = 1/b$, где b – средняя длительность обслуживания заявок в приборе.

2.3. Дисциплина буферизации – *с потерями*: заявка, поступившая в систему и заставшая накопитель заполненным, теряется.

2.4. Дисциплина обслуживания – *в порядке поступления* по правилу «первым пришел – первым обслужен» (FIFO).

В СМО с накопителем ограниченной ёмкости всегда существует установившийся режим, поскольку длина очереди не будет расти до бесконечности даже при больших значениях нагрузки.

3. Кодирование состояний марковского процесса.

В качестве параметра, описывающего состояние марковского процесса, будем рассматривать количество заявок k, находящихся в СМО (в приборе и в накопителе). Тогда марковский процесс в любой момент времени может находиться в одном из следующих $(r+2)$-х состояний:

$\mathbf{E}_0$: $k = 0$ – в системе нет ни одной заявки;

$\mathbf{E}_1$: $k = 2$ – в системе находится 1 заявка на обслуживании в приборе;

$\mathbf{E}_2$: $k = 2$ – в системе находятся 2 заявки: одна – на обслуживании в приборе и вторая ожидает в накопителе;

…

$\mathbf{E}_{r+1}$: $k = r+1$ – в системе находятся $(r+1)$ заявок: одна – на обслуживании в приборе и r – в накопителе.

4. Размеченный граф переходов случайного процесса представлен на рис.5.9.

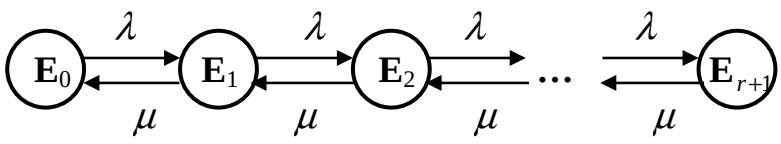


Рис. 5.9. Граф переходов марковского процесса

В один и тот же момент времени в системе может произойти только одно событие:

- поступление заявки с интенсивностью λ, что соответствует увеличению на единицу числа заявок в системе и переходу случайного процесса в состояние с номером на единицу больше;
- завершение обслуживания заявки в приборе с интенсивностью μ, что соответствует уменьшению числа заявок в системе и переходу случайного процесса в состояние с номером на единицу меньше.

Задание на самостоятельную работу: *по графу переходов рис.5.9 построить матрицу интенсивностей переходов.*

5. Система уравнений.

Составим по графу переходов систему уравнений для определения стационарных вероятностей:

$$\begin{cases} \lambda p_0 = \mu p_1 \\ (\lambda+\mu) p_1 = \lambda p_0 + \mu p_2 \\ (\lambda+\mu) p_2 = \lambda p_1 + \mu p_3 \\ \cdots \\ \mu p_{r+1} = \lambda p_r \\ \sum\limits_{k=0}^{r+1} p_k = 1 \end{cases}.$$

Используя метод математической индукции можно показать, что

$p_k = y^k p_0 \quad (k = \overline{0, r+1})$,

где $y = \lambda b$ – нагрузка системы.

Подставляя полученное выражение в последнее уравнение системы линейных алгебраических уравнений, найдем вероятность простоя системы в зависимости от нагрузки:

$$p_0 = \frac{1}{\sum\limits_{k=0}^{r+1} y^k} = \begin{cases} \dfrac{1-y}{1-y^{r+2}}, & y \neq 1 \\ \dfrac{1}{r+2}, & y = 1 \end{cases}.$$

Тогда стационарные вероятности состояний $p_k \quad (k = \overline{0, r+1})$:

$$p_k = \begin{cases} \dfrac{y^k(1-y)}{1-y^{r+2}}, & y \neq 1 \\ \dfrac{y^k}{r+2}, & y = 1 \end{cases} .$$

Задание на самостоятельную работу: *вывести представленные математические зависимости.*

5. Расчет характеристик СМО.

Характеристики СМО при найденных значениях стационарных вероятностей состояний случайного процесса могут быть рассчитаны по следующим формулам:

1) нагрузка $y = \lambda / \mu = \lambda b$;

2) загрузка $\rho = \sum_{k=1}^{r+1} p_k = 1 - p_0$;

3) коэффициент простоя системы $\eta = p_0 = 1 - \rho$;

4) среднее число заявок в очереди $l = \sum_{k=2}^{r+1} (k-1) p_k$;

5) среднее число заявок в системе $m = \sum_{k=1}^{r+1} k p_k = l + \rho$;

6) вероятность потери заявок $\pi = p_{r+1}$;

Задание на самостоятельную работу: *используя выражение (3.18), доказать, что вероятность потери заявок равна вероятности того, что система заполнена, то есть в накопителе нет свободных мест для вновь поступающих заявок.*

7) производительность системы (интенсивность потока обслуженных заявок) $\lambda' = \lambda(1-\pi)$;

8) интенсивность потока потерянных заявок $\lambda'' = \lambda\pi$;

9) среднее время ожидания заявок $w = l / \lambda'$;

10) среднее время пребывания заявок $u = m / \lambda' = w + b$.

5.4.4. Одноканальная СМО с накопителем неограниченной емкости (M/M/1)

1. Описание системы (рис.5.10).

1.1. Система – *одноканальная* – с одним обслуживающим прибором.

1.2. Поток заявок *однородный.*

1.3. В приборе происходит задержка поступающих в систему заявок на некоторое *случайное* время.

1.4. В системе имеется накопитель *неограниченной* ёмкости: $r = \infty$, то есть любая заявка, поступившая в систему, найдет место для ожидания в очереди и

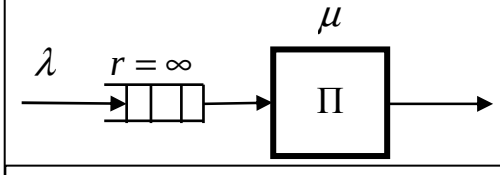


Рис. 5.10. СМО с накопителем неограниченной ёмкости

не будет потеряна.

2. Предположения и допущения.

2.1. Поступающие в систему заявки образуют *простейший* поток с интенсивностью λ.

2.2. Длительность обслуживания заявок в приборе распределена по *экспоненциальному* закону с интенсивностью $\mu = 1/b$, где b – средняя длительность обслуживания заявок в приборе.

2.3. Дисциплина буферизации отсутствует, поскольку накопитель имеет неограниченную ёмкость.

2.4. Дисциплина обслуживания – *в порядке поступления* по правилу «первым пришел – первым обслужен» (FIFO).

2.5. Нагрузка системы совпадает с загрузкой, причём выполняется условие: $y = \rho < 1$, то есть система работает в установившемся режиме без перегрузок. При $y > 1$, в отличие от предыдущих моделей, в СМО устанавливается режим перегрузок.

3. Кодирование состояний марковского процесса.

В качестве параметра, описывающего состояние марковского процесса, как и в предыдущем примере, будем рассматривать количество заявок k, находящихся в СМО (в приборе и в накопителе). Поскольку в системе в произвольный момент времени может находиться любое сколь угодно большое число заявок, то количество состояний марковского процесса равно бесконечности:

$\mathbf{E}_0 : k = 0$ – в системе нет ни одной заявки;

$\mathbf{E}_1$: $k = 1$ – в системе находится 1 заявка (на обслуживании в приборе);

$\mathbf{E}_2$: $k = 2$ – в системе находятся 2 заявки (одна – на обслуживании в приборе и вторая ожидает в накопителе);

…

$\mathbf{E}_k$: k – в системе находятся k заявок (одна – на обслуживании в приборе и $(k-1)$ – в накопителе).

…

4. Размеченный граф переходов случайного процесса представлен на рис.5.11.

В один и тот же момент времени может происходить только одно событие: поступление заявки в систему с интенсивностью λ или завершение обслуживания заявки с интенсивностью μ. Размеченный граф переходов содержит бесконечное число состояний.

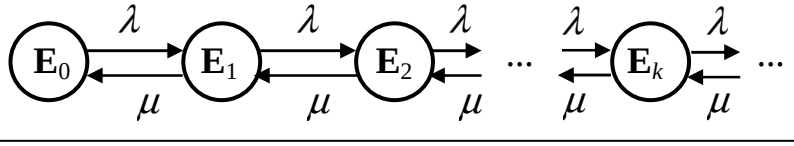


Рис. 5.11. Граф переходов марковского процесса

5. Система уравнений.

Не выписывая матрицу интенсивностей переходов, составим по графу переходов систему уравнений для определения стационарных вероятностей:

$$\begin{cases} \lambda p_0 = \mu p_1 \\ (\lambda+\mu) p_1 = \lambda p_0 + \mu p_2 \\ (\lambda+\mu) p_2 = \lambda p_1 + \mu p_3 \\ \quad \dots \\ (\lambda+\mu) p_k = \lambda p_{k-1} + \mu p_{k+1} \\ \quad \dots \\ p_0 + p_1 + p_2 + \dots + p_k + \dots = 1 \end{cases}$$

Несмотря на то, что система содержит *бесконечное число уравнений* и, соответственно, *бесконечное число неизвестных*, нетрудно методом математической индукции получить аналитическое решение в явном виде для расчета вероятностей состояний одноканальной экспоненциальной СМО с однородным потоком заявок и накопителем неограниченной ёмкости при условии, что нагрузка системы $y < 1$:

$$p_k = y^k (1-y) = \rho^k (1-\rho) \quad (k = 0, 1, 2, \dots),$$

где $\rho = \lambda / \mu$ – загрузка системы, совпадающая с нагрузкой, причём $\rho < 1$, что гарантирует существование установившегося режима в системе.

Таким образом, вероятность нахождения марковского процесса в состоянии $\mathbf{E}_k$ или, что то же самое, вероятность того, что в произвольный момент времени в системе находится k заявок, распределена по геометрическому закону с параметром, равным загрузке (нагрузке) системы.

6. Расчет характеристик СМО.

Для расчета характеристик СМО можно воспользоваться следующими математическими зависимостями:

1) нагрузка $y = \lambda / \mu = \lambda b$;

2) загрузка $\rho = 1 - p_0 = \lambda b$ и совпадает с нагрузкой;

3) коэффициент простоя системы $\eta = p_0 = 1 - \rho$;

4) среднее число заявок в очереди $l = \sum_{k=1}^{\infty} (k-1) p_k = \frac{\rho^2}{1-\rho}$;

5) среднее число заявок в системе $m = \sum_{k=0}^{\infty} k\, p_k = \frac{\rho}{1-\rho}$;

6) вероятность потери заявок $\pi = 0$;

7) производительность системы при отсутствии потерь совпадает с интенсивностью поступления заявок в систему: $\lambda' = \lambda$;

8) интенсивность потерянных заявок $\lambda'' = 0$;

9) среднее время ожидания заявок $w = \frac{l}{\lambda} = \frac{\rho b}{1-\rho}$;

10) среднее время пребывания заявок $u = w + b$ или $u = \frac{m}{\lambda} = \frac{b}{1-\rho}$.

Полученные выражения совпадают с формулами для расчёта характеристик экспоненциальной СМО, представленными в п.4.1.1.

Детальный анализ свойств таких систем выполнен в п.4.1.5.

5.4.5. Многоканальная СМО накопителем ограниченной ёмкости (M/M/2/1)

Рассмотрим многоканальную СМО с отказами, в которую поступает однородный поток заявок. Все приборы системы являются идентичными, и поступившая в систему заявка занимает любой свободный прибор. Заявка, поступившая в систему и заставшая все приборы занятыми, заносится в накопитель ограниченной ёмкости, если он не заполнен до предела. В противном случае, заявка получает отказ и покидает систему не обслуженной. Таким образом, в системе, кроме входящего потока заявок с интенсивностью λ, образуются еще два потока заявок: поток обслуженных заявок с интенсивностью λ' и поток необслуженных заявок (получивших отказ в обслуживании) с интенсивностью λ''. Очевидно, что $\lambda' + \lambda'' = \lambda$.

1. Описание системы (рис.5.12).

1.1. Система *двухканальная* - содержит два обслуживающих прибора.

1.2. В систему поступает *однородный* поток заявок.

1.3. Приборы – *идентичные*, то есть время обслуживания заявок в приборах одинаково.

1.4. Перед прибором имеется накопитель *единичной ёмкости*: $r = 1$.

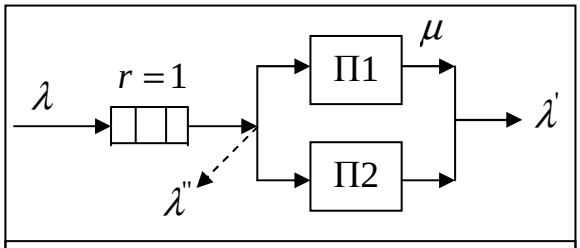


Рис. 5.12. Двухканальная СМО

2. Предположения и допущения.

2.1. Поступающие в систему заявки образуют *простейший* поток с интенсивностью λ.

2.2. Длительность обслуживания заявок в приборе распределена по *экспоненциальному* закону с интенсивностью $\mu = 1/b$, где b – средняя длительность обслуживания заявок в приборе.

2.3. Дисциплина буферизации – *с потерями*: заявка, поступившая в систему и заставшая накопитель заполненным, теряется.

2.4. Дисциплина обслуживания – *в естественном порядке*: заявка, поступившая в систему и заставшая прибор свободным, принимается на обслуживание.

3. Кодирование состояний случайного процесса.

В качестве параметра, описывающего состояние случайного процесса, будем рассматривать количество заявок k, находящихся в СМО.

Система в любой момент времени может находиться в одном из следующих состояний:

$\mathbf{E}_0$: $k = 0$ – в системе нет заявок (оба прибора простаивают);

$\mathbf{E}_1$: $k = 1$ – в системе (на обслуживании в одном из приборов) находится 1 заявка;

$\mathbf{E}_2$: $k = 2$ – в системе (на обслуживании в обоих приборах) находятся 2 заявки;

$\mathbf{E}_3$: $k = 3$ – в системе находятся 3 заявки: две – на обслуживании в приборах и одна – в накопителе.

4. Размеченный граф переходов случайного процесса представлен на рис.5.13.

В один и тот же момент времени может произойти одно из двух событий, которые приводят к изменению состояния случайного процесса, протекающего в системе:

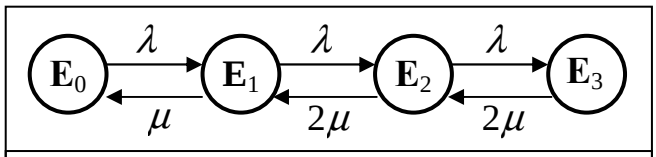


Рис.5.13. Размеченный граф переходов

- поступление заявки в систему с интенсивностью λ, приводящее к переходу в соседнее состояние с бо́льшим номером, причем если случайный процесс находится в состоянии $\mathbf{E}_3$, то его состояние не изменится, что соответствует отказу в обслуживании поступившей заявке;
- завершение обслуживания заявки в одном из приборов с интенсивностью μ, при этом случайный процесс переходит в соседнее состояние с меньшим номером с интенсивностью μ, если работает один прибор, и с интенсивностью 2μ, если работают оба прибора.

5. Расчет характеристик СМО.

Не выписывая матрицу интенсивностей переходов и систему уравнений для определения стационарных вероятностей, запишем формулы для расчёта характеристик СМО:

1) нагрузка: $y = \lambda / \mu = \lambda b$;

2) загрузка: $\rho = (p_1 + 2p_2 + 2p_3)/2$;

3) среднее число работающих приборов: $k' = (p_1 + 2p_2 + 2p_3) = 2\rho$;

4) коэффициент простоя системы: $\eta = 1 - \rho$;

5) среднее число заявок в очереди: $l = p_3$;

6) среднее число заявок в системе: $m = p_1 + 2p_2 + 3p_3 = l + k'$;

7) вероятность потери заявок: $\pi = p_3$;

Задание на самостоятельную работу: *используя выражение (3.18), доказать, что вероятность потери заявок равна вероятности того, что система заполнена, то есть в накопителе нет свободных мест для вновь поступающих заявок.*

8) производительность системы (интенсивность потока обслуженных заявок): $\lambda' = \lambda(1 - \pi)$;

9) интенсивность потока потерянных заявок: $\lambda'' = \lambda\pi$;

10) среднее время ожидания заявок: $w = l / \lambda'$;

11) среднее время пребывания заявок: $u = m / \lambda' = w + b$.

5.4.6. Одноканальная СМО с неоднородным потоком заявок и относительными приоритетами

Рассмотрим одноканальную СМО, в которую поступает неоднородный поток заявок. Ожидающие обслуживания заявки разнесены по разным накопителям ограниченной ёмкости. Между заявками разных классов установлены относительные приоритеты (ОП), означающие, что всякий раз из накопителей на обслуживание выбирается заявка с самым высоким приоритетом. При этом при поступлении в систему высокоприоритетной заявки обслуживание низкоприоритетной не прерывается. При заполненных накопителях поступившая заявка теряется.

1. Описание системы (рис.5.14).

1.1. Система *одноканальная*.

1.2. Входящий поток заявок – *неоднородный:* в систему поступает два класса заявок.

1.3. Накопители для заявок каждого класса – *ограниченной ёмкости*: $r_1 = r_2 = 1$.

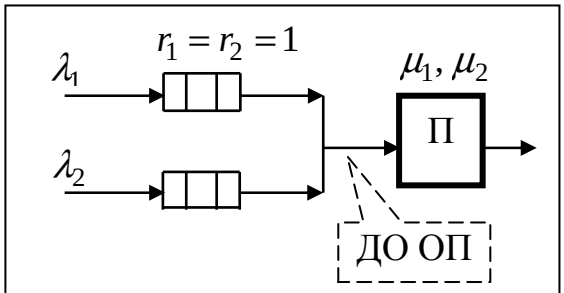


Рис.5.14. СМО с неоднородным потоком заявок

1.4. Дисциплина буферизации – *без вытеснения* заявок: если при поступлении в систему заявки любого класса соответствующий накопитель заполнен до конца, то заявка теряется

1.4. Дисциплина обслуживания – *с относительными приоритетами*: заявки первого класса имеют приоритет по отношению к заявкам второго класса.

2. Предположения и допущения.

2.1. Поступающие в систему заявки двух классов образуют *простейшие* потоки с интенсивностями λ_1 и λ_2 соответственно.

2.2. Длительности обслуживания заявок каждого класса распределены по *экспоненциальному* закону с интенсивностями $\mu_1 = 1/b_1$ и $\mu_2 = 1/b_2$, где b_1 и b_2 – средние длительности обслуживания заявок класса 1 и 2 соответственно.

В СМО всегда существует стационарный режим, так как не может быть бесконечных очередей.

3. Кодирование состояний случайного процесса.

Для описания состояний марковского процесса будем использовать распределение заявок между прибором и накопителями. Закодируем состояния следующим образом: (П/O_1,O_2), где П = {0, 1, 2} – состояние обслуживающего прибора, задаваемое классом заявки, находящейся на

обслуживании («0» – прибор свободен; «1» или «2» – на обслуживании в приборе находится заявка класса 1 или 2 соответственно); O_1, O_2 = {0, 1} – состояние накопителей 1 и 2 соответственно («0» – означает отсутствие заявки в накопителе, «1» –означает наличие одной заявки в накопителе соответствующего класса).

При выбранном способе кодирования система может находиться в следующих состояниях:

E$_0$: (0/0,0) – в системе нет ни одной заявки;

E$_1$: (1/0,0) – на обслуживании в приборе находится заявка класса 1;

E$_2$: (2/0,0) – на обслуживании в приборе находится заявка класса 2;

E$_3$: (1/1,0) – на обслуживании находится заявка класса 1 и одна заявка класса 1 ожидает обслуживания в первом накопителе;

E$_4$: (1/0,1) – на обслуживании находится заявка класса 1 и одна заявка класса 2 ожидает обслуживания соответственно во втором накопителе;

E$_5$: (2/1,0) – на обслуживании находится заявка класса 2 и одна заявка класса 1 ожидает обслуживания в первом накопителе;

E$_6$: (2/0,1) – на обслуживании находится заявка класса 2 и одна заявка класса 2 ожидает обслуживания во втором накопителе;

E$_7$: (1/1,1) – на обслуживании находится заявка класса 1, и по одной заявке каждого класса ожидают обслуживания в соответствующих накопителях;

E$_8$: (2/1,1) – на обслуживании находится заявка класса 2, и по одной заявке каждого класса ожидают обслуживания в соответствующих накопителях.

Отметим, что при кодировании случайных процессов могут быть применены различные способы кодирования.

В рассматриваемом примере состояния случайного процесса вместо представленного выше способа можно закодировать, например, следующим образом: (П,О), где П = {0, 1, 2} – состояние обслуживающего прибора, задаваемое классом заявки, находящейся на обслуживании («0» – прибор свободен; «1» или «2» – на обслуживании в приборе находится заявка класса 1 или 2 соответственно); О = {0, 1, 2, 3} – состояние накопителей 1 и 2 соответственно («0» – означает отсутствие заявок в обоих накопителях; «1» –наличие в первом накопителе заявки класса 1; «2» – наличие во втором накопителе заявки класса 2; «3» – наличие в первом и втором накопителях по одной заявке соответственно класса 1 и 2).

Представленные способы кодирования не применимы, если для заявок обоих классов используется общий накопитель ёмкостью $r = 2$. В этом случае количество состояний случайного процесса увеличится, поскольку в накопителе могут находиться 2 заявки одного и того же класса и состояние накопителя может быть представлено следующим образом: О = {0, 1, 2, 11, 12, 22}, где «0» – означает отсутствие заявок в накопителе; «1» – наличие в накопителе только одной заявки класса 1; «2» – наличие в

накопителе заявки класса 2; «11» – наличие в накопителе двух заявок класса 1; «22» – наличие в накопителе двух заявок класса 2 и «12» – наличие в накопителе одной заявки класса 1 и одной заявки класса 2. Заметим, что состояние «12» не различает, в какой последовательности эти заявки поступили в систему, что обусловлено наличием относительного приоритета между ними – независимо от момента поступления на обслуживание первой всегда будет выбрана заявка класса 1. В случае бесприоритетного обслуживания, когда заявки разных классов выбираются на обслуживание в порядке поступления, следует ввести ещё одно состояние накопителя – «21», означающее, что заявка класса 2 поступила в систему раньше заявки класса 1, в то время как состояние «12» означает, что в систему раньше поступила заявка класса 1.

4. Размеченный граф переходов случайного процесса представлен на рис.5.15.

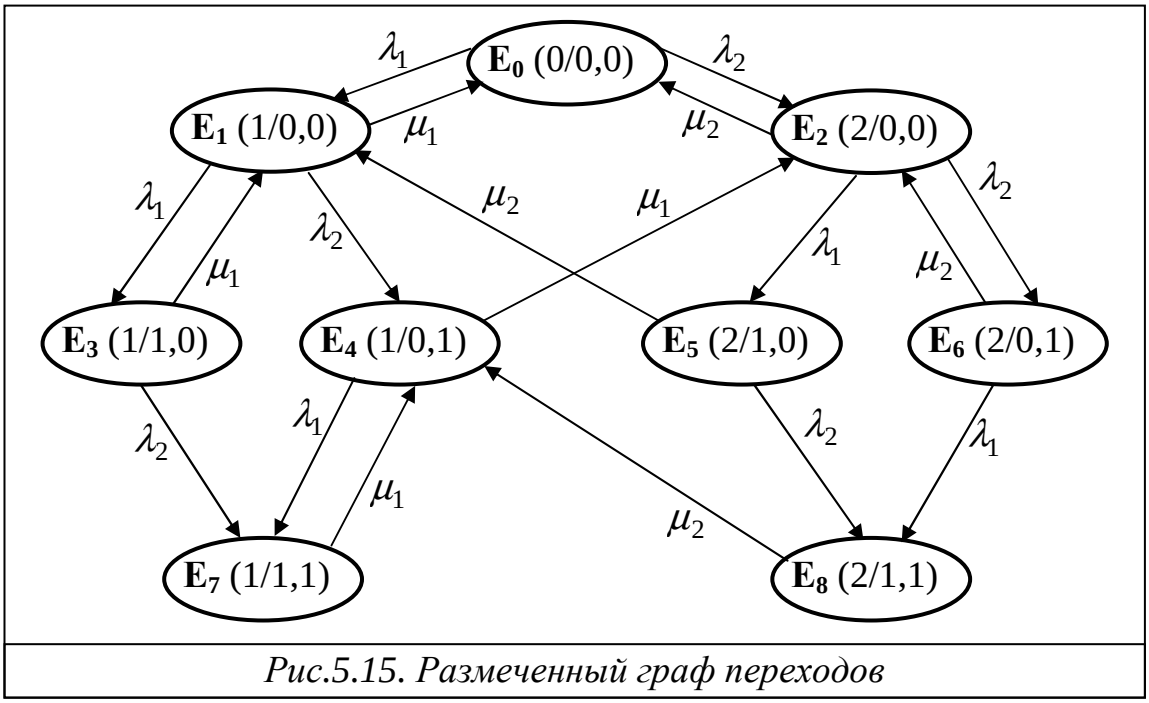


Рис.5.15. Размеченный граф переходов

В каждый момент времени может произойти только одно событие (или поступление заявки какого-либо класса, или завершение обслуживания заявки, находящейся в приборе), поскольку вероятность появления двух и более событий в один и тот же момент времени равна нулю.

При наличии в накопителях заявок первого и второго классов (состояния $\mathbf{E}_7$ и $\mathbf{E}_8$) после завершения обслуживания некоторой заявки в приборе случайный процесс переходит в состояние $\mathbf{E}_4$, означающее, что на обслуживание всегда выбирается высокоприоритетная заявка класса 1.

По графу переходов составим систему уравнений для определения стационарных вероятностей:

$$\begin{cases}(\lambda_1+\lambda_2)\,p_0=\mu_1 p_1+\mu_2 p_2\\ (\lambda_1+\lambda_2+\mu_1)\,p_1=\lambda_1 p_0+\mu_1 p_3+\mu_2 p_5\\ (\lambda_1+\lambda_2+\mu_2)\,p_2=\lambda_2 p_0+\mu_1 p_4+\mu_2 p_6\\ (\lambda_2+\mu_1)\,p_3=\lambda_1 p_1\\ (\lambda_1+\mu_1)\,p_4=\lambda_2 p_1+\mu_1 p_7+\mu_2 p_8\\ (\lambda_2+\mu_2)\,p_5=\lambda_1 p_2\\ (\lambda_1+\mu_2)\,p_6=\lambda_2 p_2\\ \mu_1 p_7=\lambda_2 p_8+\lambda_1 p_4\\ \mu_2\ p_8=\lambda_2 p_5+\lambda_1 p_6\\ p_0+p_1+p_2+p_3+p_4+p_5+p_6+p_7+p_8=1\end{cases}$$

5. Расчет характеристик СМО.

Характеристики обслуживания заявок в СМО с неоднородным потоком заявок делятся на две группы:

- характеристики обслуживания заявок каждого класса;
- характеристики обслуживания заявок суммарного потока.

Расчёт характеристик обслуживания заявок *каждого класса* выполняется по следующим формулам:

1) нагрузка: $y_1=\lambda_1/\mu_1=\lambda_1 b_1;\quad y_2=\lambda_2/\mu_2=\lambda_2 b_2;$

2) загрузка, создаваемая заявками, которая может трактоваться как вероятность того, что на обслуживании в приборе находится заявка класса 1 и 2 соответственно: $\rho_1=p_1+p_3+p_4+p_7;\quad \rho_2=p_2+p_5+p_6+\rho_8;$

3) среднее число заявок в очереди:

$l_1=p_3+p_5+p_7+p_8;\quad l_2=p_4+p_6+p_7+p_8;$

4) среднее число заявок в системе:

$m_1=p_1+2p_3+p_4+p_5+2p_7+p_8=l_1+\rho_1;$

$m_2=p_2+p_4+p_5+2p_6+p_7+2p_8=l_2+\rho_2;$

5) вероятность потери заявок:

$\pi_1=p_3+p_5+p_7+p_8;\quad \pi_2=p_4+p_6+p_7+p_8;$

6) производительность по каждому классу заявок (интенсивность непотерянных заявок):

$\lambda_1^{'}=\lambda_1(1-\pi_1);\quad \lambda_2^{'}=\lambda_2(1-\pi_2);$

7) среднее время ожидания заявок:

$w_1=l_1/\lambda_1^{'};\quad w_2=l_2/\lambda_2^{'}$

8) среднее время пребывания заявок

$u_1=m_1/\lambda_1^{'}=w_1+b;\quad u_2=m_2/\lambda_2^{'}=w_2+b$

Расчёт характеристик обслуживания заявок *суммарного потока* выполняется по следующим формулам:

1) суммарная нагрузка системы: $Y=y_1+y_2;$

2) загрузка системы: $R = \rho_1 + \rho_2$;

3) коэффициент простоя системы: $\eta = 1 - R$;

4) суммарное число заявок во всех очередях: $l = l_1 + l_2$;

5) суммарное число заявок в системе: $m = m_1 + m_2 = l + R$;

6) вероятность потери заявок: $\pi = \pi_1 + \pi_2$;

7) производительность системы (интенсивность суммарного потока обслуженных заявок): $\lambda' = \lambda_1' + \lambda_2' = \lambda(1 - \pi)$;

8) среднее время ожидания заявок суммарного потока:

$w = (\lambda_1' w_1 + \lambda_2' w_2) / \lambda' = l / \lambda'$;

9) среднее время пребывания заявок суммарного потока:

$u = (\lambda_1' u_1 + \lambda_2' u_2) / \lambda' = m / \lambda' = w + b$.

5.5. Марковские модели сетей массового обслуживания

В данном параграфе подробно рассматриваются марковские модели сетей массового обслуживания (СеМО) с однородным потоком заявок. В качестве примеров представлены разомкнутые и замкнутые экспоненциальные СеМО с накопителями ограниченной ёмкости, а также замкнутые неэкспоненциальные СеМО, в которых длительность обслуживания заявок в одном из узлов распределена по закону Эрланга с коэффициентом вариации $\nu < 1$ и гиперэкспоненциальному закону с коэффициентом вариации $\nu > 1$.

Можно показать, что случайный процесс, протекающий в экспоненциальных разомкнутых и замкнутых СеМО при сформулированных предположениях и допущениях является марковским.

Случайный процесс, протекающий в замкнутой неэкспоненциальной сети, не является марковским. Для описания процесса функционирования такой системы в терминах марковских случайных процессов в некоторых случаях можно воспользоваться методом вложенных цепей Маркова, суть которого заключается в том, что функционирование системы рассматривается в определенные моменты времени, образующие цепь Маркова.

Как и для СМО, в каждом примере приводится *описание* исследуемой СеМО и принятые при построении математической модели *предположения и допущения*, необходимые для того, чтобы протекающий в системе случайный процесс мог быть сведён к марковскому. Разработка Марковской модели включает в себя этапы *кодирования состояний* случайного процесса, построения размеченного *графа переходов*, формирования *матрицы интенсивностей переходов* и *системы линейных алгебраических уравнений* для расчёта стационарных вероятностей состояний марковского процесса, на основе которых строятся математические зависимости, позволяющие рассчитать наиболее важные характеристики функционирования исследуемых СеМО.

Применение марковских случайных процессов для расчёта характеристик функционирования и исследования свойств СеМО оказывается наиболее результативным:

- для разомкнутых СеМО с накопителями ограниченной ёмкости, в которых заявки теряются при заполненных накопителях;
- для неэкспоненциальных разомкнутых и замкнутых СеМО, в которых длительности обслуживания заявок в узлах распределены по гипоэкспоненциальному или гиперэкспоненциальному закону.

5.5.1. Разомкнутая экспоненциальная СеМО с накопителями ограниченной емкости

Рассмотрим разомкнутую экспоненциальную СеМО с двумя одноканальными узлами, в которую из внешней среды поступает простейший поток заявок с интенсивностью λ_0 (рис.5.16). Накопители в обоих узлах имеют ограниченную ёмкость, равную единице: $r_1 = r_2 = 1$. Заявка, поступившая в узел и заставшая накопитель заполненным, теряется. Длительности обслуживания в узлах распределены по экспоненциальному закону со средними значениями b_1 и b_2 соответственно. Заявки после обслуживания в узле 2 вероятностью q направляются в узел 1 и вероятностью $(1-q)$ – покидают СеМО.

Отметим, что, поскольку заявки в сети могут теряться, рассматриваемая разомкнутая СеМО является *нелинейной*, то есть интенсивности потоков заявок, поступающих в узлы СеМО, не связаны между собой линейной зависимостью (3.5) и, следовательно, не могут быть рассчитаны путём решения системы линейных алгебраических уравнений (4.16).

1. Описание СеМО (рис.5.16)**.**

1.1. Сеть массового обслуживания – разомкнутая *двухузловая*.

1.2. Узлы 1 и 2 – одноканальные: $K_1 = K_2 = 1$.

1.3. Накопители в узлах ограниченной ёмкости: $r_1 = r_2 = 1$.

1.4. Дисциплины буферизации в узлах – с потерями заявок, если накопители заполнены.

1.5. Поток заявок *однородный с интенсивностью* λ_0.

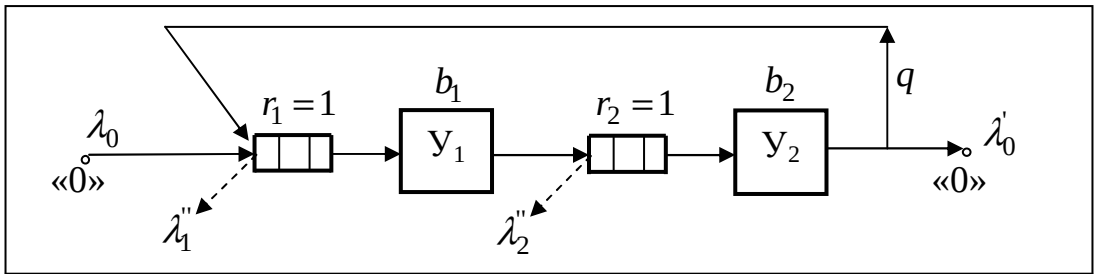


Рис. 5.16. Двухузловая разомкнутая СеМО с потерями

2. Предположения и допущения.

2.1. Поступающие в разомкнутую СеМО заявки образуют

простейший поток с интенсивностью λ_0.

2.2. Длительности обслуживания заявок в узлах СеМО распределены по *экспоненциальному* закону с параметрами, представляющими собой интенсивности обслуживания: $\mu_1 = 1/b_1$ и $\mu_2 = 1/b_2$.

В разомкнутой СеМО при любой нагрузке существует стационарный режим, так как в узлах сети не может быть бесконечных очередей.

3. Кодирование состояний случайного процесса.

Для описания состояний марковского случайного процесса будем использовать распределение заявок между узлами. Закодируем состояния следующим образом: (M_1, M_2), где M_i = {0, 1, 2} – количество заявок в узле i («0» – узел свободен; «1» – на обслуживании в узле находится одна заявка; «2» – в узле находятся две заявки – одна на обслуживания и вторая в накопителе).

При выбранном способе кодирования система может находиться в следующих состояниях:

E$_0$: (0,0) – в СеМО нет ни одной заявки;
E$_1$: (1,0) – в узле 1 находится одна заявка;
E$_2$: (2,0) – в узле 1 находятся две заявки;
E$_3$: (0,1) – в узле 2 находится одна заявка;
E$_4$: (1,1) – в узле 1 и 2 находится по одной заявке;
E$_5$: (2,1) – две заявки находятся в узле 1 и одна – в узле 2;
E$_6$: (0,2) – в узле 2 находятся две заявки;
E$_7$: (1,2) – две заявки находятся в узле 2 и одна – в узле 1;
E$_8$: (2,2) – в узле 1 и 2 находятся по две заявки.

4. Размеченный граф переходов случайного процесса (рис.5.17).

Построим граф переходов, полагая, что в каждый момент времени может произойти только одно событие (поступление заявки в СеМО или завершение обслуживания заявки в одном из узлов), поскольку вероятность появления двух и более событий в один и тот же момент времени равна нулю.

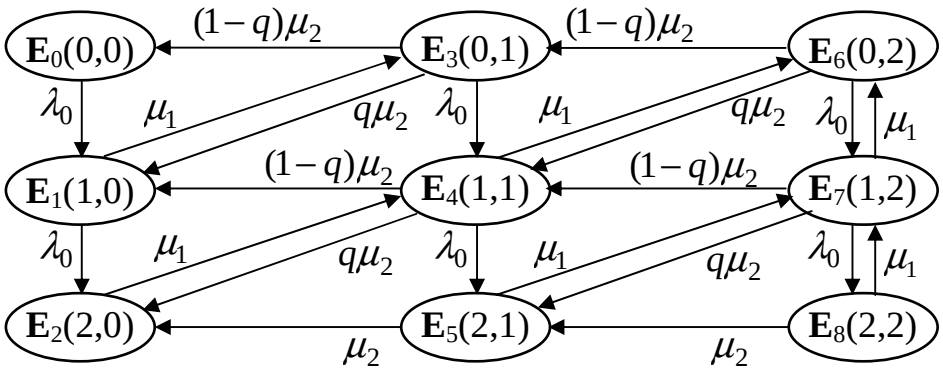


Рис. 5.17. Граф переходов разомкнутой СеМО с потерями

Следует обратить внимание на переходы из состояний **E**$_3$(0,1), **E**$_4$(1,1), **E**$_6$(0,2) и **E**$_7$(1,2), обусловленные завершением обслуживания

заявки в узле 2 с интенсивностью μ_2. В этих случаях с вероятностью q заявка может вернуться в узел 1 и с вероятностью $(1-q)$ – покинуть СеМО, тогда интенсивности соответствующих переходов будут равны $q\mu_2$ и $(1-q)\mu_2$. Если же случайный процесс находится в состояниях $\mathbf{E}_5(2,1)$ и $\mathbf{E}_8(2,2)$, то завершение обслуживания заявки в узле 2 приводит к переходу соответственно в состояния $\mathbf{E}_2(2,0)$ и $\mathbf{E}_5(2,1)$ с интенсивностью μ_2, что соответствует выходу заявки из СеМО с вероятностью $(1-q)$ и потере заявки, которая с вероятностью q будет направлена в узел 1, поскольку в последнем нет места в накопителе. Аналогично, если случайный процесс находится в состояниях $\mathbf{E}_7(1,2)$ и $\mathbf{E}_8(2,2)$, то завершение обслуживания заявки в узле 1 приводит к переходу соответственно в состояния $\mathbf{E}_6(0,2)$ и $\mathbf{E}_7(1,2)$ с интенсивностью μ_1, что соответствует потере заявки, поскольку накопитель узла 1 заполнен.

5. Расчет характеристик СеМО.

Не составляя матрицу интенсивностей переходов и не выписывая систему уравнений для определения стационарных вероятностей, получим математические выражения для определения узловых и сетевых характеристик разомкнутой СеМО с потерями при известных значениях стационарных вероятностей состояний $p_i \quad (i = 0,1,...,8)$.

Заметим, что СеМО с потерями относится к классу *нелинейных* сетевых моделей, расчёт характеристик которых связан с определёнными проблемами, в частности, с необходимостью детального анализа потоков заявок и с невозможностью применения в ряде случаев фундаментальных соотношений для расчёта сетевых характеристик. Кроме того, процесс формирования математических зависимостей для каждой конкретной нелинейной СеМО может существенно отличаться.

В связи с этим, ниже достаточно подробно рассматривается процесс получения математических выражений для расчёта узловых и сетевых характеристик нелинейной разомкнутой СеМО, представленной на рис.5.16.

Узловые характеристики СеМО рассчитываются в такой последовательности:

1) загрузки узлов определяются как суммы вероятностей состояний, в которых соответствующий узел занят обслуживанием заявок:

$\rho_1 = p_1 + p_2 + p_4 + p_5 + p_7 + p_8$; $\rho_2 = p_3 + p_4 + p_5 + p_6 + p_7 + p_8$;

2) коэффициенты простоя узлов: $\eta_1 = 1 - \rho_1$; $\eta_2 = 1 - \rho_2$;

3) среднее число заявок в очередях:

$l_1 = p_2 + p_5 + p_8$; $l_2 = p_6 + p_7 + p_8$;

4) среднее число заявок в узлах:

$m_1 = p_1 + p_4 + p_7 + 2(p_2 + p_5 + p_8) = l_1 + \rho_1$;

$m_2 = p_3 + p_4 + p_5 + 2(p_6 + p_7 + p_8) = l_2 + \rho_2$;

5) производительности узлов (интенсивность обслуженных заявок на

выходе узлов): $\lambda_1^{'} = \frac{\rho_1}{b_1} = \rho_1\mu_1; \quad \lambda_2^{'} = \frac{\rho_2}{b_2} = \rho_2\mu_2;$

6) вероятности потери заявок в узлах СеМО могут быть рассчитаны на основе выражения (3.18) с учётом того, что $K_1 = K_2 = 1$:

$$\pi_1 = 1 - \frac{\rho_1}{y_1}; \qquad \pi_2 = 1 - \frac{\rho_2}{y_2};$$

в этих выражениях: $y_1 = \lambda_1 b_1$ и $y_2 = \lambda_2 b_2$ – создаваемые в узлах нагрузки, где λ_1 и λ_2 – интенсивности поступления заявок в узлы 1 и 2 СеМО, для расчёта которых необходимо выполнить анализ потоков в рассматриваемой СеМО;

интенсивность λ_1 складывается (см. рис.5.16) из интенсивности λ_0 поступления заявок из внешнего источника и интенсивности потока заявок, возвращающихся с вероятностью q в узел 1 после обслуживания в узле 2: $\lambda_1 = \lambda_0 + q\lambda_2^{'}$, где $\lambda_2^{'}$ - рассчитанная ранее интенсивность потока выходящих из узла 2 заявок (производительность узла 2);

аналогично, из рис. 5.16 можно видеть, что интенсивность λ_2 поступающих в узел 2 заявок представляет собой интенсивность $\lambda_1^{'}$ потока выходящих из узла 1 заявок (производительность узла 1): $\lambda_2 = \lambda_1^{'}$;

окончательно, после некоторых преобразований выражения для расчёта вероятностей потери заявок в узлах СеМО примут вид:

$$\pi_1 = 1 - \frac{\lambda_1^{'}}{\lambda_0 + q\lambda_2^{'}}; \qquad \pi_2 = 1 - \frac{\lambda_2^{'}}{\lambda_1^{'}};$$

7) среднее время ожидания заявок в узлах рассчитывается по формулам Литтла с учётом только обслуженных заявок:

$w_1 = l_1 / \lambda_1^{'}; \quad w_2 = l_2 / \lambda_2^{'};$

8) аналогично, среднее время пребывания заявок в узлах:

$u_1 = m_1 / \lambda_1^{'} = w_1 + b; \quad u_2 = m_2 / \lambda_2^{'} = w_2 + b;$

Для расчёта *сетевых характеристик* СеМО могут использоваться следующие формулы:

1) суммарная загрузка узлов СеМО, характеризующая среднее число одновременно работающих узлов в сети: $\rho = \rho_1 + \rho_2$;

2) суммарное число заявок в очередях: $L = l_1 + l_2$;

3) суммарное число заявок в узлах: $M = m_1 + m_2 = L + \rho;$

4) производительность СеМО (интенсивность обслуженных заявок на выходе сети): $\lambda_0^{'} = (1-q)\lambda_2^{'}$;

5) вероятность потери заявок в сети: $\pi = \frac{\lambda_0 - \lambda_0^{'}}{\lambda_0} = 1 - \frac{\lambda_0^{'}}{\lambda_0}$; следует обратить внимание, что вероятность потери заявок в сети определяется как

доля потерянных заявок по отношению к поступившим в СеМО заявкам, в то время как вероятности потери π_1 и π_2 заявок в узлах СеМО определяется как доля потерянных заявок по отношению ко всем заявкам, поступившим в конкретный узел, число которых учитывает и то, что поступившая в СеМО заявка за время нахождения в сети может попасть в данный узел несколько раз.

Математические зависимости для расчёта суммарного времени ожидания заявок и времени пребывания заявок в СеМО не могут быть получены в общем виде в виду нелинейности СеМО с потерями.

5.5.2. Замкнутая экспоненциальная СеМО

1. Описание замкнутой СеМО (рис.5.18).

1.1. Сеть массового обслуживания (СеМО) – замкнутая *двухузловая*.

1.2. Количество приборов в узлах: узел 1 – одноканальный, узел 2 – двухканальный.

1.3. Поток заявок *однородный*.

1.4. В СеМО постоянно циркулируют $M = 3$ заявки.

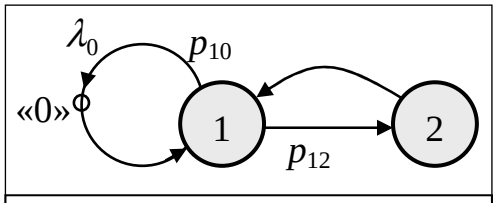


Рис.5.18. Замкнутая СеМО

2. Предположения и допущения.

2.1. Длительности обслуживания заявок в узлах 1 и 2 распределены по экспоненциальному закону с интенсивностями $\mu_1 = 1/b_1$ и $\mu_2 = 1/b_2$ соответственно, где b_1, b_2 – средние длительности обслуживания заявок.

2.2. Приборы в двухканальном узле 2 *идентичны* и любая заявка может обслуживаться в любом приборе.

2.3. Заявка после обслуживания в узле 1 с вероятностью p_{12} переходит в узел 2 и с вероятностью $p_{10} = 1 - p_{12}$ возвращается в этот же узел 1.

2.4. Дуга, выходящая из узла 1 и входящая обратно в этот же узел, рассматривается как внешняя по отношению к СеМО, и на ней выбирается нулевая точка «0».

В замкнутой СеМО всегда существует стационарный режим, так как число заявок в сети ограничено и не может быть бесконечных очередей.

Легко убедиться, что случайный процесс, протекающий в замкнутой экспоненциальной сети, является марковским.

3. Кодирование состояний марковского процесса.

Под состоянием марковского процесса будем понимать распределение заявок по узлам СеМО. Закодируем состояния следующим образом: ($\mathbf{M_1}$, $\mathbf{M_2}$), где $\mathbf{M_1}$ = {0, 1, 2, 3} – количество заявок, находящихся в узле 1 и $\mathbf{M_2}$ = {0, 1, 2, 3} – количество заявок, находящихся в узле 2, причем суммарное число заявок в обоих узлах должно быть равно 3.

При выбранном способе кодирования система может находиться в следующих состояниях:

$\mathbf{E}_0$: (3, 0) – все три заявки находятся в узле 1, причем одна заявка

находятся на обслуживании в приборе и две заявки ожидают в накопителе;

$\mathbf{E}_1$: (2, 1) – две заявки находятся в узле 1 (одна на обслуживании в приборе и одна в накопителе) и одна – на обслуживании в одном из приборов узла 2;

$\mathbf{E}_2$: (1, 2) – одна заявка находится на обслуживании в узле 1 и две – в узле 2 (на обслуживании в обоих приборах);

$\mathbf{E}_3$: (0, 3) – все три заявки находятся в узле 2, причем две заявки находятся на обслуживании в обоих приборах узла 2 и одна заявка ожидает в накопителе.

4. Размеченный граф переходов случайного процесса (рис.5.19).

В один и тот же момент времени в замкнутой СеМО может произойти только одно из двух событий:

1) завершение обслуживания заявки в первом узле с интенсивностью μ_1, при этом заявка с вероятностью p_{12} покинет этот узел и перейдет в узел 2 (интенсивность перехода $p_{12}\mu_1$) или с вероятностью $(1-p_{12})$ останется в этом же узле, то есть состояние случайного процесса не изменится; отметим, что второй случай не отображается на графе переходов в виде дуги, выходящей из узла 1 и снова входящей в узел 1;

2) завершение обслуживания заявки в узле 2 с интенсивностью μ_2, если на обслуживании в этом узле находится одна заявка (работает один прибор), или с интенсивностью $2\mu_2$, если на обслуживании в узле находятся две заявки (работают оба прибора); обслуженная заявка покидает этот узел и с вероятностью 1 переходит в первый узел.

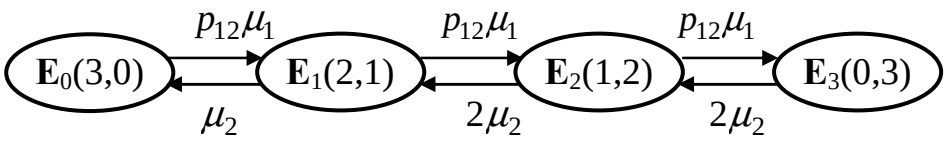


Рис.5.19. Граф экспоненциальной 3СеМО

5. Система уравнений.

Не составляя матрицу интенсивностей переходов, запишем систему уравнений для определения стационарных вероятностей:

$$\begin{cases} p_{12}\mu_1 p_0 = \mu_2 p_1 \\ (p_{12}\mu_1 + \mu_2)p_1 = p_{12}\mu_1 p_0 + 2\mu_2 p_2 \\ (p_{12}\mu_1 + 2\mu_2)p_2 = p_{12}\mu_1 p_1 + 2\mu_2 p_3 \\ 2\mu_2 p_3 = p_{12}\mu_1 p_2 \\ p_0 + p_1 + p_2 + p_3 = 1 \end{cases}$$

6. Расчет характеристик СеМО.

На основе полученных значений стационарных вероятностей рассчитываются узловые и сетевые характеристики СеМО с использованием следующих формул:

1) загрузка узлов:

$\rho_1 = p_0 + p_1 + p_2; \quad \rho_2 = 0{,}5p_1 + p_2 + p_3;$

2) коэффициенты простоя узлов:

$\eta_1 = 1 - \rho_1; \quad \eta_2 = 1 - \rho_2;$

3) средние длины очередей заявок в узлах:

$l_1 = 2p_0 + p_1; \quad l_2 = p_3;$

4) среднее число заявок в узлах:

$m_1 = 3p_0 + 2p_1 + p_2; \quad m_2 = p_1 + 2p_2 + 3p_3;$

5) производительность замкнутой СеМО:

$$\lambda_0 = \frac{\rho_1}{\alpha_1 b_1} = \frac{\rho_2}{\alpha_2 b_2};$$

где α_1 и α_2 – коэффициенты передач соответственно узлов 1 и 2, определяемые путем решения системы уравнений (4.17);

6) среднее время ожидания заявок в узлах СеМО:

$$w_1 = \frac{l_1}{\alpha_1 \lambda_0}; \quad w_2 = \frac{l_2}{\alpha_2 \lambda_0};$$

7) среднее время пребывания заявок в узлах СеМО:

$$u_1 = \frac{l_1}{\alpha_1 \lambda_0}; \quad u_2 = \frac{l_2}{\alpha_2 \lambda_0};$$

8) нагрузка в узлах сети:

$y_1 = \alpha_1 \lambda_0 b_1; \quad y_2 = \alpha_2 \lambda_0 b_2;$

9) среднее число параллельно работающих *узлов* сети, определяемое как суммарная *загрузка* всех узлов СеМО:

$R = \rho_1 + \rho_2;$

10) среднее число параллельно работающих *приборов* во всех узлах сети, определяемое как суммарная *нагрузка* всех узлов СеМО:

$Y = y_1 + y_2;$

11) суммарное число заявок во всех очередях СеМО:

$L = l_1 + l_2;$

12) суммарное (полное) время ожидания заявок в СеМО :

$W = \alpha_1 w_1 + \alpha_2 w_2;$

13) время пребывания заявок в СеМО:

$U = \alpha_1 u_1 + \alpha_2 u_2;$

Суммарное число заявок, циркулирующих в ЗСеМО, рассчитываемое как $M = m_1 + m_2$, должно совпадать с заданным числом заявок $M = 3$.

Следует обратить внимание на то, что временные характеристики обслуживания заявок в узлах СеМО, а, следовательно, и в сети в целом, могут быть рассчитаны только после определения производительности замкнутой СеМО, вычисляемой через найденные значения загрузок узлов.

5.5.3. Замкнутая СеМО с эрланговским обслуживанием

1. Описание СеМО.

1.1. Замкнутая сеть массового обслуживания (ЗСеМО) – *двухузловая*.

1.2. Количество приборов в узлах: $K_1 = K_2 = 1$.

1.3. Поток заявок *однородный*.

1.4. В ЗСеМО постоянно циркулируют $M = 3$ заявки.

Граф рассматриваемой ЗСеМО такой же, как и в предыдущем примере (рис.5.18). Отличие состоит только в том, что в рассматриваемой ЗСеМО узел 2 – одноканальный.

2. Предположения и допущения.

2.1. Длительность обслуживания заявок *в узле 1* распределена по *закону Эрланга 2-го порядка* со средней длительностью обслуживания заявок $b_1 = 1/\mu_1$, а в узле 2 – по экспоненциальному закону со средней длительностью обслуживания заявок $b_2 = 1/\mu_2$, где μ_1, μ_2 – интенсивности обслуживания заявок.

2.2. Заявка после обслуживания в узле 1 с вероятностью p_{12} переходит в узел 2 и с вероятностью $p_{10} = 1 - p_{12}$ возвращается в этот же узел 1.

2.3. Дуга, выходящая из узла 1 и входящая обратно в этот же узел, рассматривается как внешняя по отношению к СеМО, и на ней отмечается нулевая точка «0».

3. Сведение случайного процесса к марковскому.

Случайный процесс, протекающий в замкнутой неэкспоненциальной сети, не является марковским.

Для описания процесса функционирования такой системы в терминах марковских случайных процессов будем рассматривать функционирование системы в определенные моменты времени, в которые случайный процесс обладает марковским свойством. Для этого воспользуемся представлением случайной величины, распределенной по закону Эрланга 2-го порядка, в виде суммы двух экспоненциально распределенных случайных величин (см. раздел 2, п.2.6.1). При этом будем полагать, что обслуживание заявки в первом узле проходит две фазы, длительность каждой из которых распределена по экспоненциальному закону со средним значением $b_1' = b_1 / 2$. Последнее необходимо для того, чтобы полная длительность обслуживания в узле 1 была равна b_1 .

Таким образом, обслуживание заявки в СеМО можно представить как двухфазное обслуживание в первом узле и однофазное – во втором узле (рис.5.20). Длительности обслуживания в фазах Ф1 и Ф2 первого узла ЗСеМО распределены по экспоненциальному закону с одним и тем же параметром $\mu_1' = 1/b_1'$ и с параметром $\mu_2 = 1/b_2$ – в единственной фазе второго узла. Моменты завершения обслуживания в каждой из фаз

образуют цепь Маркова, так как времена нахождения в них распределены по экспоненциальному закону. Такое представление случайного процесса требует другого подхода к кодированию состояний, учитывающего распределение заявок по фазам обслуживания.

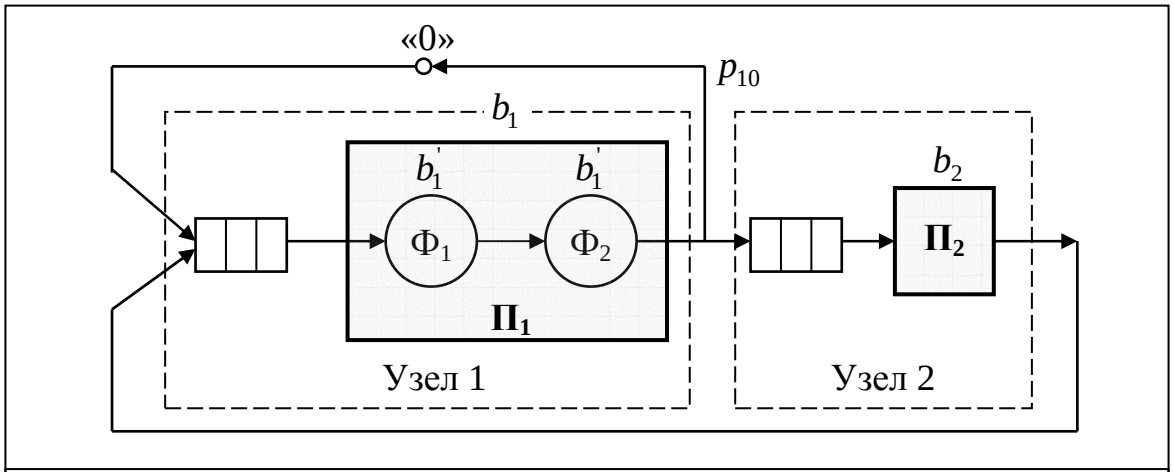


Рис.5.20. 3СеМО с двухфазным представлением распределения Эрланга

4. Кодирование состояний случайного процесса.

Под состоянием марковского процесса будем понимать распределение заявок по узлам СеМО с учетом того, на какой фазе обслуживания распределения Эрланга находится заявка в узле 1.

Для этого закодируем состояния следующим образом: (**M_1, M_2**), где **M_1** = {0, 1_1, 1_2, 2_1, 2_2, 3} – количество заявок, находящихся в узле 1 (индексы отражают нахождение заявки на 1-й или 2-й фазе распределения Эрланга), и **M_2** = {0, 1, 2, 3} – количество заявок, находящихся в узле 2, причем суммарное число заявок в обоих узлах должно быть равно 3.

При выбранном способе кодирования система может находиться в следующих состояниях:

E_1: (3_1, 0) – все три заявки находятся в узле 1, причем одна заявка находятся на обслуживании в приборе на *первой фазе*, и две заявки ожидают в накопителе;

E_2: (3_2, 0) – все три заявки находятся в узле 1, причем одна заявка находятся на обслуживании в приборе на *второй фазе*, и две заявки ожидают в накопителе;

E_3: (2_1, 1) – две заявки находятся в узле 1 (одна на обслуживании в приборе на *первой фазе* и одна в накопителе) и одна – на обслуживании в узле 2;

E_4: (2_2, 1) – две заявки находятся в узле 1 (одна на обслуживании в приборе на *второй фазе* и одна в накопителе) и одна – на обслуживании в узле 2;

E_5: (1_1, 2) – одна заявка находится в узле 1 на обслуживании в приборе на *первой фазе* и две заявки находятся в узле 2, причем одна из них находится на обслуживании в приборе, а вторая заявка ожидает в накопителе;

E$_6$: (1$_2$, 2) – одна заявка находится в узле 1 на обслуживании в приборе на *второй фазе* и две заявки находятся в узле 2, причем одна из них находится на обслуживании в приборе, а вторая заявка ожидает в накопителе;

E$_7$: (0, 3) – все три заявки находятся в узле 2, причем одна заявка находится на обслуживании в приборе, а две другие – ожидают в накопителе.

5. Размеченный граф переходов случайного процесса.

На рис.5.21 представлен граф переходов марковского процесса для рассматриваемой неэкспоненциальной СеМО. Для понимания процесса составления графа переходов вместо номеров состояний в вершинах графа указаны коды состояний.

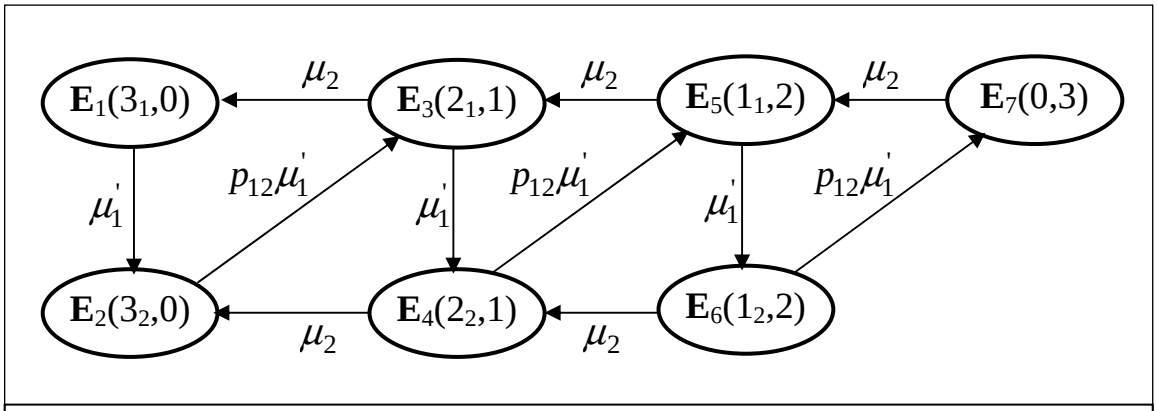


Рис. 5.21. Граф переходов ЗСеМО с эрланговским обслуживанием

Из состояния **E**$_1$=(3$_1$, 0) переход возможен только в одно состояние **E**$_2$=(3$_2$, 0) с интенсивностью μ_1' обслуживания на первой фазе, поскольку все заявки в первом узле обязательно проходят две фазы обслуживания.

Из состояния **E**$_2$=(3$_2$, 0) переход возможен также только в одно состояние **E**$_3$=(2$_1$, 1). Это соответствует завершению обслуживания на второй фазе заявки в узле 1 (с интенсивностью μ_1') и ее передаче в узел 2 (с вероятностью p_{12}). Отсюда интенсивность перехода марковского процесса в состояние **E**$_3$=(2$_1$, 1) будет равна произведению $p_{12}\mu_1'$. Заметим, что с вероятностью $(1-p_{12})$ марковский процесс останется в том же состоянии, что соответствует возврату заявки в узел 1.

Из состояния **E**$_3$=(2$_1$, 1) переход возможен в одно из двух состояний:

- в состояние **E**$_4$=(2$_2$, 1), что соответствует завершению обслуживания заявки на первой фазе в узле 1 (с интенсивностью μ_1') и переходу к обслуживанию на второй фазе в том же узле 1;
- в состояние **E**$_1$=(3$_1$, 0), что соответствует завершению обслуживания заявки в узле 2 (с интенсивностью μ_2) и ее передаче в узел 1 (с вероятностью 1).

Следует помнить, что в любой момент времени может произойти

только одно событие. Вероятность двух и более событий пренебрежимо мала. Поэтому из состояния $\mathbf{E}_3=(2_1, 1)$ не возможен переход в состояние $\mathbf{E}_3=(3_2, 0)$, означающий завершение обслуживания заявки на первой фазе в узле 1 и одновременное завершение обслуживания заявки в узле 2.

Аналогичные рассуждения справедливы для состояний $\mathbf{E}_4=(2_2, 1)$, $\mathbf{E}_5=(1_1, 2)$ и $\mathbf{E}_6=(1_2, 2)$.

Из состояния $\mathbf{E}_7=(0, 3)$ переход возможен только в состояние $\mathbf{E}_5=(1_1, 2)$, означающий завершение обслуживания заявки в узле 2 и ее передачу на обслуживание в узел 1, причем обслуживание новой заявки всегда начинается на первой фазе. По этой причине переход в состояние $\mathbf{E}_6=(1_2, 2)$ не возможен.

6. Система уравнений.

Не составляя матрицу интенсивностей переходов, запишем систему уравнений для определения стационарных вероятностей состояний:

$$\begin{cases} \mu_1' p_1 = \mu_2 p_3 \\ p_{12}\mu_1' p_2 = \mu_1' p_1 + \mu_2 p_4 \\ (\mu_1' + \mu_2) p_3 = p_{12}\mu_1' p_2 + \mu_2 p_5 \\ (p_{12}\mu_1' + \mu_2) p_4 = \mu_1' p_3 + \mu_2 p_6 \\ (\mu_1' + \mu_2) p_5 = p_{12}\mu_1' p_4 + \mu_2 p_7 \\ (p_{12}\mu_1' + \mu_2) p_6 = \mu_1' p_5 \\ \mu_2 p_7 = p_{12}\mu_1' p_6 \\ p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7 = 1 \end{cases}$$

7. Расчет характеристик СеМО.

Характеристики ЗСеМО определяются в такой последовательности:

1) загрузка и коэффициенты простоя узлов:

$\rho_1 = p_1 + p_2 + p_3 + p_4 + p_5 + p_6;\quad \rho_2 = p_3 + p_4 + p_5 + p_6 + p_7;$

$\eta_1 = 1 - \rho_1;\quad \eta_2 = 1 - \rho_2;$

2) среднее число параллельно работающих *узлов* сети, определяемое как суммарная *загрузка* всех узлов СеМО:

$R = \rho_1 + \rho_2;$

3) среднее число заявок в очередях и в узлах СеМО:

$l_1 = 2(p_1 + p_2) + p_3 + p_4;\quad l_2 = p_5 + p_6 + 2p_7;$

$m_1 = 3(p_1 + p_2) + 2(p_3 + p_4) + p_5 + p_6;\quad m_2 = p_3 + p_4 + 2(p_5 + p_6) + 3p_7;$

4) суммарное число заявок во всех очередях СеМО:

$L = l_1 + l_2;$

5) производительность замкнутой СеМО:

$$\lambda_0 = \frac{\rho_1}{\alpha_1 b_1} = \frac{\rho_2}{\alpha_2 b_2};$$

где α_1 и α_2 – коэффициенты передач соответственно узла 1 и узла 2;

6) средние времена ожидания и пребывания заявок в узлах СеМО:

$$w_1 = \frac{l_1}{\alpha_1 \lambda_0}; \quad w_2 = \frac{l_2}{\alpha_2 \lambda_0};$$

$$u_1 = \frac{l_1}{\alpha_1 \lambda_0}; \quad u_2 = \frac{l_2}{\alpha_2 \lambda_0};$$

7) суммарное (полное) время ожидания и время пребывания заявок в СеМО:

$$W = \alpha_1 w_1 + \alpha_2 w_2;$$

$$U = \alpha_1 u_1 + \alpha_2 u_2;$$

8) нагрузка в узлах сети:

$$y_1 = \alpha_1 \lambda_0 b_1; \quad y_2 = \alpha_2 \lambda_0 b_2;$$

9) среднее число параллельно работающих *приборов* во всех узлах сети, определяемое как суммарная *нагрузка* всех узлов СеМО:

$$Y = y_1 + y_2;$$

Суммарное число заявок, циркулирующих в СеМО, рассчитываемое как $M = m_1 + m_2$, должно совпадать с заданным числом заявок в замкнутой сети: $M = 3$.

5.5.4. Замкнутая СеМО с гиперэкспоненциальным обслуживанием

1. Описание СеМО.

1.1. Сеть массового обслуживания (СеМО) – *двухузловая*.

1.2. Количество приборов в узлах: $K_1 = K_2 = 1$.

1.3. Поток заявок *однородный*.

1.4. В СеМО постоянно циркулируют *M*=3 заявки.

2. Предположения и допущения.

2.1. Длительность обслуживания заявок *в узле 1* распределена по *гиперэкспоненциальному закону* со средней длительностью обслуживания заявок $b_1 = 1/\mu_1$ и коэффициентом вариации $\nu_{b_1} = 2$, а в узле 2 – по экспоненциальному закону со средней длительностью обслуживания заявок $b_2 = 1/\mu_2$, где μ_1, μ_2 – интенсивности обслуживания заявок.

2.2. Заявка после обслуживания в узле 1 с вероятностью p_{12} переходит в узел 2 и с вероятностью $p_{10} = 1 - p_{12}$ возвращается в этот же узел 1.

2.3. Дуга, выходящая из узла 1 и входящая обратно в этот же узел, рассматривается как внешняя по отношению к СеМО, и на ней выбирается нулевая точка «0».

В замкнутой СеМО всегда существует стационарный режим.

3. Сведение случайного процесса к марковскому.

Для описания процесса функционирования в замкнутой неэкспоненциальной сети в терминах марковских случайных процессов, как и ранее,

будем рассматривать функционирование системы в определенные моменты времени, в которые случайный процесс обладает марковским свойством. Для этого воспользуемся представлением случайной величины, распределенной по гиперэкспоненциальному закону, в виде композиции двух экспоненциально распределенных случайных величин (см. раздел 2, п.2.6.2), каждая из которых появляется с вероятностями q и $(1-q)$ соответственно. В первом узле ЗСеМО такое представление реализуется в виде двух параллельных экспоненциальных фаз, обслуживающих заявки по следующей схеме (рис.5.22):

- заявка с вероятностью $q = 0{,}1$ попадает на обслуживание в первую фазу, длительность обслуживания в которой распределена по экспоненциальному закону со средним значением $b_1^{'}$, после чего покидает узел;
- заявка с вероятностью $(1-q) = 0{,}9$ попадает на обслуживание во вторую фазу, длительность обслуживания в которой распределена по экспоненциальному закону со средним значением $b_1^{''}$, после чего покидает первый узел.

Значения длительностей обслуживания в этих двух фазах таковы, что выполняется условие: $qb_1^{'} + (1-q)b_1^{''} = b_1$. Последнее необходимо для того, чтобы средняя длительность обслуживания в узле 1 была равна b_1 .

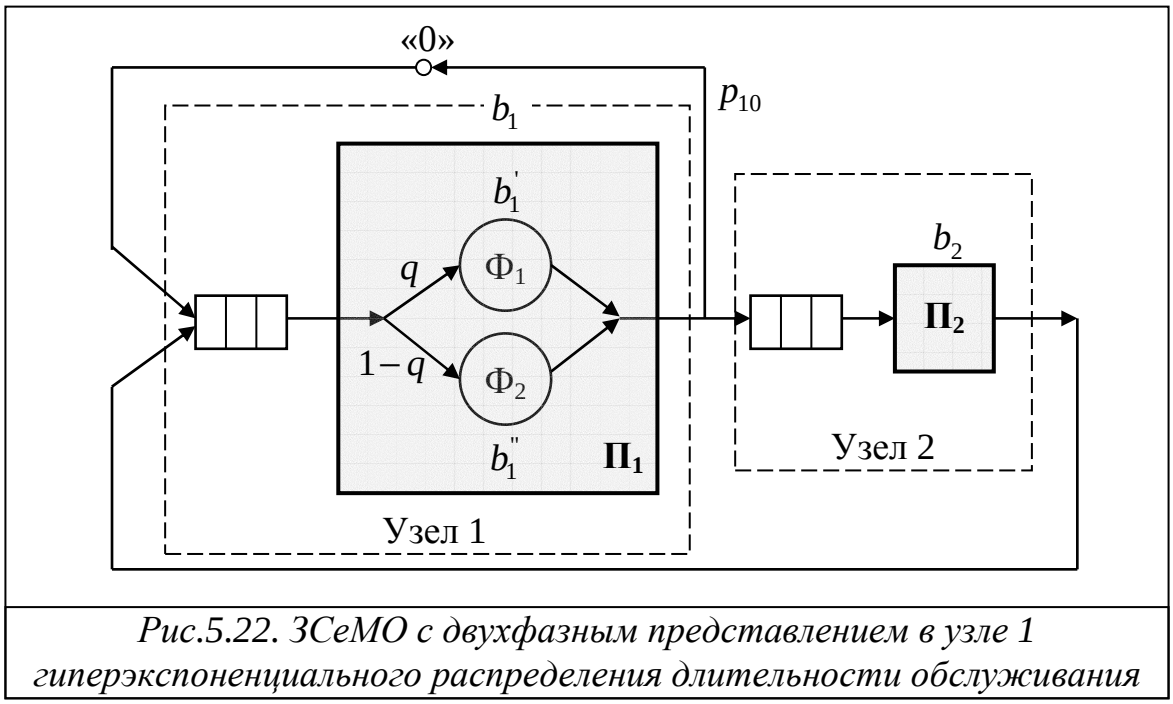


Рис.5.22. ЗСеМО с двухфазным представлением в узле 1 гиперэкспоненциального распределения длительности обслуживания

Моменты завершения обслуживания в каждой из фаз образуют цепь Маркова, так как времена нахождения в них распределены по экспоненциальному закону.

4. Кодирование состояний случайного процесса.

Под состоянием марковского процесса будем понимать распределение заявок по узлам СеМО с учетом того, на какой фазе

обслуживания в узле 1 находится заявка.

Для этого закодируем состояния следующим образом: (**M_1, M_2**), где **M_1** = {0, 1_1, 1_2, 2_1, 2_2, 3} – количество заявок, находящихся в узле 1 (индексы отражают нахождение заявки на 1-й или 2-й фазе гиперэкспоненциального распределения), и **M_2** = {0, 1, 2, 3} – количество заявок, находящихся в узле 2, причем суммарное число заявок в обоих узлах должно быть равно 3.

При выбранном способе кодирования система может находиться, как и в предыдущем примере, в следующих состояниях:

E_1: (3_1, 0) – все три заявки находятся в узле 1, причем одна заявка находятся на обслуживании в приборе на *первой фазе*, и две заявки ожидают в накопителе;

E_2: (3_2, 0) – все три заявки находятся в узле 1, причем одна заявка находятся на обслуживании в приборе на *второй фазе*, и две заявки ожидают в накопителе;

E_3: (2_1, 1) – две заявки находятся в узле 1 (одна на обслуживании в приборе на *первой фазе* и одна в накопителе) и одна – на обслуживании в узле 2;

E_4: (2_2, 1) – две заявки находятся в узле 1 (одна на обслуживании в приборе на *второй фазе* и одна в накопителе) и одна – на обслуживании в узле 2;

E_5: (1_1, 2) – одна заявка находится в узле 1 на обслуживании в приборе на *первой фазе* и две заявки находятся в узле 2, причем одна из них находится на обслуживании в приборе, а вторая заявка ожидает в накопителе;

E_6: (1_2, 2) – одна заявка находится в узле 1 на обслуживании в приборе на *второй фазе* и две заявки находятся в узле 2, причем одна из них находится на обслуживании в приборе, а вторая заявка ожидает в накопителе;

E_7: (0, 3) – три заявки находятся в узле 2, причем одна заявка – на обслуживании в приборе, а две другие – ожидают в накопителе.

5. Размеченный граф переходов случайного процесса.

На рис.5.23 представлен граф переходов марковского процесса для рассматриваемой неэкспоненциальной СеМО с гиперэкспоненциальным распределением длительности обслуживания заявок в первом узле. Для понимания процесса составления графа переходов вместо номеров состояний в вершинах графа указаны коды состояний, а для того чтобы не загромождать рисунок, используются следующие обозначения для интенсивностей переходов: $g_1 = (1-q)(1-p_{12})\mu_1'$; $g_2 = (1-q)p_{12}\mu_1'$; $g_3 = q(1-p_{12})\mu_1''$; $g_4 = qp_{12}\mu_1''$.

Рассмотрим подробно все возможные переходы для каждого состояния $\mathbf{E}_i$ $(i = \overline{1,7})$ марковского случайного процесса.

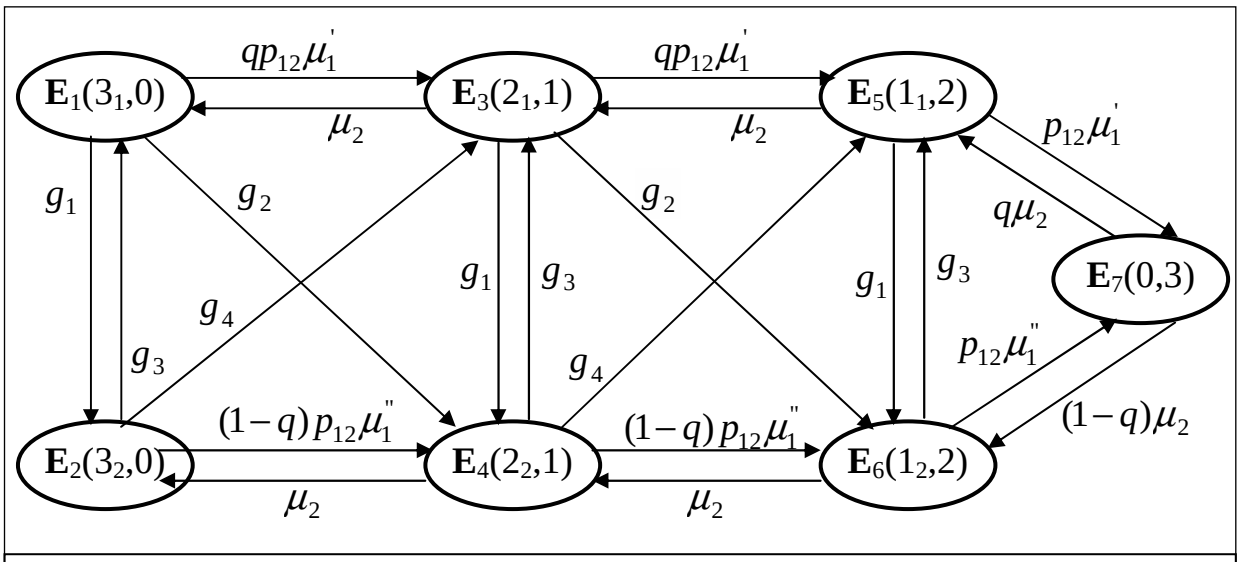


Рис.5.23. Граф переходов ЗСеМО с гиперэкспоненциальным обслуживанием

Состояние $\mathbf{E}_1$. Если случайный процесс находится в состоянии $\mathbf{E}_1=(3_1, 0)$, то по завершению обслуживания заявки случайный процесс может перейти в одно из трёх состояний: $\mathbf{E}_2=(3_2, 0)$, $\mathbf{E}_3=(2_1, 1)$ и $\mathbf{E}_4=(2_2, 1)$ или остаться в том же состоянии. Напомним, что если случайный процесс остаётся в том же состоянии, то это никак не отображается на графе переходов.

Случайный процесс перейдёт из состояния $\mathbf{E}_1=(3_1, 0)$ в состояние $\mathbf{E}_2=(3_2, 1)$ при выполнении следующих условий:

- завершится обслуживание заявки, находящейся на обслуживании в фазе Φ_1; интенсивность этого события $\mu_1' = 1/b_1'$;
- заявка, завершившая обслуживание в узле 1, вернётся в этот же узел и встанет в конец очереди; вероятность этого события равна $p_{10} = 1 - p_{12}$;
- в узле 1 очередная заявка, которая поступит на обслуживание из очереди в прибор Π_1, попадёт на обслуживание в фазу Φ_2; вероятность этого события равна $(1-q)$.

Таким образом, интенсивность перехода из состояния $\mathbf{E}_1=(3_1, 0)$ в состояние $\mathbf{E}_2=(3_2, 0)$ будет равна $g_1 = (1-q)(1-p_{12})\mu_1'$.

Случайный процесс перейдёт из состояния $\mathbf{E}_1=(3_1, 0)$ в состояние $\mathbf{E}_3=(2_1, 1)$ при выполнении следующих условий:

- завершится обслуживание заявки, находящейся на обслуживании в фазе Φ_1; интенсивность этого события $\mu_1' = 1/b_1'$;
- заявка, завершившая обслуживание в узле 1, перейдёт в узел 2; вероятность этого события равна p_{12};
- в узле 1 новая заявка, которая поступит на обслуживание из очереди в прибор Π_1, попадёт на обслуживание в фазу Φ_1; вероятность этого события $- q$.

Таким образом, интенсивность перехода из состояния $\mathbf{E}_1=(3_1, 0)$ в

состояние $\mathbf{E}_3=(2_1, 1)$ будет равна $qp_{12}\mu_1^{'}$.

Случайный процесс перейдёт из состояния $\mathbf{E}_1=(3_1, 0)$ в состояние $\mathbf{E}_4=(2_2, 1)$ при выполнении следующих условий:

- завершится обслуживание заявки, находящейся на обслуживании в фазе Φ_1; интенсивность этого события $\mu_1^{'}=1/b_1^{'}$;
- заявка, завершившая обслуживание в узле 1, перейдёт в узел 2; вероятность этого события равна p_{12};
- в узле 1 новая заявка, которая поступит на обслуживание из очереди в прибор Π_1, попадёт на обслуживание в фазу Φ_2; вероятность этого события – $(1-q)$.

Таким образом, интенсивность перехода из состояния $\mathbf{E}_1=(3_1, 0)$ в состояние $\mathbf{E}_4=(2_2, 1)$ будет равна $g_2=(1-q)p_{12}\mu_1^{'}$.

Состояние $\mathbf{E}_2$. Случайный процесс из состояния $\mathbf{E}_2=(3_2, 0)$ по завершению обслуживания заявки также может перейти в одно из трёх состояний: $\mathbf{E}_1=(3_1, 0)$, $\mathbf{E}_3=(2_1, 1)$ и $\mathbf{E}_4=(2_2, 1)$ или остаться в том же состоянии.

Случайный процесс перейдёт из состояния $\mathbf{E}_2=(3_2, 0)$ в состояние $\mathbf{E}_1=(3_1, 1)$ при выполнении следующих условий:

- с интенсивностью $\mu_1^{''}=1/b_1^{''}$ завершится обслуживание заявки в фазе Φ_2;
- с вероятностью $p_{10}=1-p_{12}$ заявка, завершившая обслуживание в узле 1, вернётся в этот же узел и встанет в конец очереди;
- с вероятностью q в узле 1 очередная заявка, которая поступит из очереди в прибор Π_1, попадёт на обслуживание в фазу Φ_1.

Таким образом, интенсивность перехода из состояния $\mathbf{E}_1=(3_1, 0)$ в состояние $\mathbf{E}_2=(3_2, 0)$ будет равна $g_3=q(1-p_{12})\mu_1^{''}$.

Случайный процесс перейдёт из состояния $\mathbf{E}_2=(3_2, 0)$ в состояние $\mathbf{E}_3=(2_1, 1)$ при выполнении следующих условий:

- с интенсивностью $\mu_1^{''}=1/b_1^{''}$ завершится обслуживание заявки в фазе Φ_2;
- с вероятностью p_{12} заявка, завершившая обслуживание в узле 1, перейдёт в узел 2;
- с вероятностью q в узле 1 очередная заявка, которая поступит из очереди в прибор Π_1, попадёт на обслуживание в фазу Φ_1.

Таким образом, интенсивность перехода из $\mathbf{E}_2=(3_2, 0)$ в $\mathbf{E}_3=(2_1, 1)$ будет равна $g_4=qp_{12}\mu_1^{''}$.

Случайный процесс перейдёт из состояния $\mathbf{E}_2=(3_2, 0)$ в состояние $\mathbf{E}_4=(2_2, 1)$ при выполнении следующих условий:

- с интенсивностью $\mu_1^{''}=1/b_1^{''}$ завершится обслуживание заявки в фазе Φ_2;

•с вероятностью p_{12} заявка, завершившая обслуживание в узле 1, перейдёт в узел 2;

•с вероятностью $(1-q)$ в узле 1 очередная заявка, которая поступит из очереди в прибор $П_1$, попадёт на обслуживание в фазу $Ф_2$.

Таким образом, интенсивность перехода из $\mathbf{E}_2=(3_2, 0)$ в $\mathbf{E}_4=(2_2, 1)$ будет равна $(1-q)p_{12}\mu_1^{''}$.

Состояния $\mathbf{E}_3$ и $\mathbf{E}_4$. Если случайный процесс находится в состоянии $\mathbf{E}_3=(2_1, 1)$ или $\mathbf{E}_4=(2_2, 1)$, то кроме аналогичных переходов, связанных с завершением обслуживания заявки в узле 1, имеется ещё один переход в состояния $\mathbf{E}_1=(3_1, 0)$ и $\mathbf{E}_2=(3_2, 0)$ соответственно, связанный с завершением обслуживания заявки в узле 2. Интенсивность перехода из $\mathbf{E}_3=(2_1, 1)$ в $\mathbf{E}_1=(3_1, 0)$ и из $\mathbf{E}_4=(2_2, 1)$ в $\mathbf{E}_2=(3_2, 0)$ равна интенсивности обслуживания μ_2 в узле 2. Отметим, что переходы из $\mathbf{E}_3=(2_1, 1)$ в $\mathbf{E}_2=(3_2, 0)$ и из $\mathbf{E}_4=(2_2, 1)$ в $\mathbf{E}_1=(3_1, 0)$ отсутствуют, так как заявка, находящаяся на обслуживании в первом узле, остаётся в той же фазе обслуживания, которая была в момент завершения обслуживания заявки в узле 2. Это является следствием того, что в случайных процессах с непрерывным временем вероятность одновременного появления двух событий (завершение обслуживания в узле 1 и в узле 2) равна нулю.

Состояния $\mathbf{E}_5$ и $\mathbf{E}_6$. Переходы из состояний $\mathbf{E}_5=(1_1, 2)$ и $\mathbf{E}_6=(1_2, 2)$ аналогичны переходам из $\mathbf{E}_3=(2_1, 1)$ и $\mathbf{E}_4=(2_2, 1)$ за исключением переходов в состояние $\mathbf{E}_7=(0, 3)$. Интенсивности переходов из $\mathbf{E}_5=(1_1, 2)$ и $\mathbf{E}_6=(1_2, 2)$ в $\mathbf{E}_7=(0, 3)$ определяются как произведение интенсивности обслуживания в соответствующей фазе узла 1 на вероятность того, что заявка, завершившая обслуживание в узле 1, перейдёт в узел 2: $p_{12}\mu_1^{'}$ и $p_{12}\mu_1^{''}$.

Состояние $\mathbf{E}_7$. Переходы из состояния $\mathbf{E}_7=(0, 3)$ связаны с завершением обслуживания с интенсивностью μ_2 заявки в узле 2, которая переходит в узел 1 и с вероятностью q попадает на обслуживание в фазу $Ф_1$ или с вероятностью $(1-q)$ – в фазу $Ф_2$. Соответственно интенсивности переходов будут равны $q\mu_2$ и $(1-q)\mu_2$

6. Расчет характеристик СеМО.

Не составляя матрицу интенсивностей переходов и не выписывая систему линейных алгебраических уравнений для определения стационарных вероятностей состояний, приведём математические зависимости для расчёта характеристик функционирования ЗСеМО:

1) загрузка и коэффициенты простоя узлов:

$\rho_1 = p_1 + p_2 + p_3 + p_4 + p_5 + p_6; \quad \rho_2 = p_3 + p_4 + p_5 + p_6 + p_7;$

$\eta_1 = 1 - \rho_1; \quad \eta_2 = 1 - \rho_2;$

2) среднее число параллельно работающих *узлов* сети, определяемое как суммарная *загрузка* всех узлов СеМО:

$R = \rho_1 + \rho_2;$

3) среднее число заявок в очередях и в узлах СеМО:

$l_1 = 2(p_1 + p_2) + p_3 + p_4; \quad l_2 = p_5 + p_6 + 2p_7;$

$m_1 = 3(p_1 + p_2) + 2(p_3 + p_4) + p_5 + p_6;$

$m_2 = p_3 + p_4 + 2(p_5 + p_6) + 3p_7;$

4) суммарное число заявок во всех очередях СеМО:

$L = l_1 + l_2;$

5) производительность замкнутой СеМО:

$$\lambda_0 = \frac{\rho_1}{\alpha_1 b_1} = \frac{\rho_2}{\alpha_2 b_2};$$

где α_1 и α_2 - коэффициенты передачи соответственно узла 1 и узла 2;

6) средние времена ожидания и пребывания заявок в узлах СеМО:

$$w_1 = \frac{l_1}{\alpha_1 \lambda_0}; \quad w_2 = \frac{l_2}{\alpha_2 \lambda_0};$$

$$u_1 = \frac{l_1}{\alpha_1 \lambda_0}; \quad u_2 = \frac{l_2}{\alpha_2 \lambda_0};$$

7) суммарное (полное) время ожидания и время пребывания заявок в СеМО:

$W = \alpha_1 w_1 + \alpha_2 w_2;$

$U = \alpha_1 u_1 + \alpha_2 u_2;$

8) нагрузка в узлах сети:

$y_1 = \alpha_1 \lambda_0 b_1; \quad y_2 = \alpha_2 \lambda_0 b_2;$

9) среднее число параллельно работающих *приборов* во всех узлах сети, определяемое как суммарная *нагрузка* всех узлов СеМО:

$Y = y_1 + y_2.$

Суммарное число заявок, циркулирующих в СеМО, рассчитываемое как $M = m_1 + m_2$, должно совпадать с заданным числом заявок в замкнутой сети: $M = 3$.

Задание на самостоятельную работу:

1. По графу переходов рис.5.23 построить матрицу интенсивностей переходов и составить систему линейных алгебраических уравнений для расчёта стационарных вероятностей состояний.

2. Выполнить детальный анализ свойств исследуемой системы.

5.6. Резюме

1. Марковские случайные процессы используются в качестве математических моделей систем со стохастическим характером функционирования. Марковская модель представляется в виде систем дифференциальных и алгебраических уравнений, для решения которых обычно применяются численные методы. Поэтому марковские случайные процессы можно отнести к *численным методам моделирования*.

2. Случайный процесс полностью описывается перечнем *состояний*, которые задаются значениями некоторых переменных, и *переходами* между состояниями.

3. Для с*лучайного процесса с дискретными состояниями* характерен скачкообразный переход из состояния в состояние, которые могут быть пронумерованы. При этом число возможных состояний может быть *конечным* или *бесконечным*. Для *случайного процесса с непрерывными состояниями* характерен плавный переход из состояния в состояние.

4. Случайные процессы с дискретными состояниями делятся на процессы с *дискретным временем*, когда переходы из состояния в состояние возможны в строго *определенные заранее фиксированные моменты времени*, которые можно пронумеровать, и с *непрерывным временем*, когда интервал времени между соседними переходами является *случайным*, и переход возможен в любой заранее не известный момент времени.

5. Случайные процессы с дискретными состояниями изображаются в виде *графа переходов (состояний)*. В *размеченном* графе переходов на дугах графа указываются условия перехода в виде *вероятностей переходов* или *интенсивностей переходов*.

Состояния случайного процесса могут быть *невозвратными* и *поглощающими*.

6. Случайный процесс называется *марковским*, если вероятность любого состояния в будущем зависит только от его состояния в настоящем и не зависит от того, когда и каким образом процесс оказался в этом состоянии. Для того чтобы случайный процесс с непрерывным временем был *марковским*, необходимо, чтобы интервалы времени между соседними переходами из состояния в состояние были распределены *по экспоненциальному закону*, который обладает замечательным *свойством*: если время нахождения случайного процесса в некотором состоянии распределено по экспоненциальному закону, то *интервал от любого случайного момента времени до момента перехода* в другое состояние имеет *то же экспоненциальное распределение с тем же параметром*. Эта особенность является следствием *отсутствия последействия*, присущего процессам с экспоненциальным распределением времени нахождения в том или ином состоянии.

7. Для описания марковского случайного процесса используется следующая совокупность параметров:

- *перечень состояний* $\mathbf{E}_1$, ..., $\mathbf{E}_n$;

•*матрица переходов*, в виде *матрицы вероятностей переходов* **Q** для процессов *с дискретным временем* или *матрицы интенсивностей переходов* **G** для процессов *с непрерывным временем*;

•*начальные вероятности* $p_1(0),\ldots, p_n(0)$.

8. Для описания переходов между состояниями случайного процесса *с дискретным временем* используется квадратная *матрица вероятностей переходов* $\mathbf{Q} = [q_{ij} \mid i, j = \overline{1,n}]$, элементы которой удовлетворяют условиям:

$$0 \le q_{ij} \le 1; \quad \sum_{j=1}^{n} q_{ij} = 1 \qquad (i, j = \overline{1,n}).$$

Для описания переходов между состояниями случайного процесса *с непрерывным временем* используется квадратная *матрица интенсивностей переходов* $\mathbf{G} = [g_{ij} \mid i, j = \overline{1,n}]$, в которой *интенсивность перехода* g_{ij} определяется как предел отношения вероятности перехода $P_{ij}(\Delta\tau)$ из состояния $\mathbf{E}_i$ в состояние $\mathbf{E}_j$ за промежуток времени $\Delta\tau$ к длине этого промежутка:

$$g_{ij} = \lim_{\Delta\tau \to 0} \frac{P_{ij}(\Delta\tau)}{\Delta\tau} \qquad (i, j = \overline{1,n};\ i \ne j),$$

а диагональные элементы определяются из условия:

$$\sum_{j=1}^{n} g_{ij} = 0 \quad (i = \overline{1,n}).$$

9. Изучение случайных процессов заключается в определении вероятностей состояний $p_1(t),..., p_n(t)$, которые могут быть представлены *стохастическим* вектором:

$$P(t) = \{p_1(t),..., p_n(t)\},$$

причем

$$0 \le p_i(t) \le 1; \quad \sum_{i=1}^{n} p_i(t) = 1.$$

Вектор состояний $P(t) = \{p_1(t),..., p_n(t)\}$ является *основной* характеристикой марковского случайного процесса.

10. Случайный процесс обладает *эргодическим свойством*, если по истечении достаточно большого промежутка времени вероятности состояний стремятся к предельным (стационарным) значениям $p_1,\ldots, p_n$, не зависящим от начальных вероятностей $p_1(0),\ldots, p_n(0)$ и от самого промежутка времени. В этом случае система, в которой протекает случайный процесс, работает в *установившемся* или *стационарном режиме*. В противном случае система работает в нестационарном режиме.

Случайный процесс *с дискретным временем* обладает *эргодическим свойством*, если матрица вероятностей переходов *не является периодической* или *разложимой*. Случайный процесс с *непрерывным временем* и *конечным* числом состояний всегда обладает эргодическим свойством.

11. Для марковского процесса с дискретным временем, обладающего эргодическим свойством, стационарные вероятности состояний определяются из системы линейных алгебраических уравнений:

$$p_j = \sum_{i=1}^{n} p_i q_{ij} \quad (j = \overline{1,n}),$$

которая совместно с нормировочным условием $\sum_{i=1}^{n} p_i = 1$ образует систему, обладающую единственным решением.

Аналогично, для марковского процесса с непрерывным временем, обладающего эргодическим свойством, стационарные вероятности состояний определяются из системы линейных алгебраических уравнений:

$$\sum_{i=1}^{n} p_i g_{ij} = 0 \quad (j = \overline{1,n}),$$

которая совместно с нормировочным условием образует систему, обладающую единственным решением.

5.7. Практикум: обсуждение и решение задач

Вопрос 1. Существуют ли реальные системы, в которых протекающие в них случайные процессы являются марковскими?

Обсуждение. Марковский процесс является такой же идеализированной моделью реальных систем, как и простейший поток, представляющий собой идеализированную модель случайного потока заявок. Эта идеализация заключается в том, что с вероятностью, отличной от нуля, марковский процесс может находиться в любом из состояний бесконечно долго. Это обусловлено тем, что плотность экспоненциального распределения ограничена слева и не ограничена справа. Очевидно, что в реальных системах это невозможно. В то же время, как показывают многочисленные исследования, такая идеализация часто оказывается оправданной, поскольку при определённых условиях позволяет получить для многих реальных систем вполне приемлемые результаты, погрешность которых лежит в допустимых для практики пределах в 10-20%. Кроме того, в некоторых случаях предположение о марковском характере протекающих в исследуемой системе процессов позволяет получить верхние оценки характеристик функционирования системы.

Вопрос 2. Когда случайный процесс с непрерывным временем не обладает эргодическим свойством?

Обсуждение. Случайный процесс с непрерывным временем не обладает эргодическим свойством, если среди его состояний имеются невозвратные или поглощающие состояния. В первом случае это означает, что по истечении некоторого (иногда достаточно большого) времени случайный процесс никогда не сможет попасть в невозвратные состояния,

а во втором случае – процесс окажется в одном из поглощающих состояний, из которого он никогда не сможет выйти.

Вопрос 3. Обладает ли эргодическим свойством случайный процесс с непрерывным временем, имеющий бесконечное число состояний?

Обсуждение. Случайный процесс с непрерывным временем и бесконечным числом состояний может обладать или не обладать эргодическим свойством. Применительно к случайным процессам, протекающим в системах массового обслуживания, наличие эргодического свойства определяется наличием установившегося режима в моделируемой системе, а точнее отсутствием перегрузок в системе с накопителями неограниченной ёмкости. Если же система перегружена, что со временем приводит к бесконечному увеличению длины очереди заявок в системе, то можно утверждать, что соответствующий случайный процесс не будет обладать эргодическим свойством.

Задача 1. Определить, обладает ли эргодическим свойством случайный процесс с дискретным временем с заданной матрицей вероятностей переходов *P*, сопроводив ответ необходимыми пояснениями.

$$P = \begin{array}{c|cccc} & E_1 & E_2 & E_3 & E_4 \\ E_1 & 0 & 0.2 & 0 & 0.8 \\ E_2 & 0.1 & 0.2 & 0.3 & 0.4 \\ E_3 & 0 & 0.5 & 0 & 0.5 \\ E_4 & 0.2 & 0.4 & 0 & 0.4 \end{array}$$

Решение. Случайный процесс обладает эргодическим свойством, если матрица вероятностей переходов не является разложимой или периодической. Переставляя столбцы и строки матрицы, проверим, является ли заданная матрица *P* разложимой или периодической.

Рассмотрим два варианта перестановок столбцов и строк:

$$P_1 = \begin{array}{c|cccc} & E_1 & E_3 & E_2 & E_4 \\ E_1 & 0 & 0 & 0.2 & 0.8 \\ E_3 & 0 & 0 & 0.5 & 0.5 \\ E_2 & 0.1 & 0.3 & 0.2 & 0.4 \\ E_4 & 0.2 & 0 & 0.4 & 0.4 \end{array} \qquad P_2 = \begin{array}{c|cccc} & E_2 & E_4 & E_1 & E_3 \\ E_2 & 0.3 & 0.4 & 0.1 & 0.3 \\ E_4 & 0.4 & 0.4 & 0.2 & 0 \\ E_1 & 0.2 & 0.8 & 0 & 0 \\ E_3 & 0.5 & 0.5 & 0 & 0 \end{array}$$

Полученные матрицы P_1 и P_2 с нулевыми подматрицами в верхнем левом углу в P_1 и в нижнем правом углу в P_2 не являются разложимыми или периодическими, следовательно, случайный процесс обладает эргодическим свойством

Задача 2. Известны вероятности состояний трехузловой замкнутой экспоненциальной СеМО: *P*(0,0,2)=0,3; *P*(0,1,1)=0,4; *P*(0,2,0)=0,1; *P*(1,0,1)=0,05; *P*(1,1,0)=0,05; P(2,0,0)=0,1. Длительности обслуживания заявок во всех одноканальных узлах одинаковы. Определить значения коэффициентов передач второго и третьего узлов сети, если известно, что

коэффициент передачи первого узла равен 4.

Дано: ЗСеМО: $n=3$; $K_1=K_2=K_3=1$;

$b_1=b_2=b_3=b$; $\alpha_1=4$;

$P(0,0,2)=0,3$; $P(0,1,1)=0,4$; $P(0,2,0)=0.1$;

$P(1,0,1)=0,05$; $P(1,1,0)=0,05$; $P(2,0,0)=0.1$.

Определить: $\alpha_2=?$ и $\alpha_3=?$

Решение.

1) По заданным значениям стационарных вероятностей состояний с учётом того, что все узлы одноканальные, рассчитаем загрузки каждого узла замкнутой СеМО как сумму вероятностей состояний, в которых соответствующий узел занят обслуживанием заявок:

$\rho_1=P(1,0,1)+P(1,1,0)+P(2,0,0)=0.05+0.05+0.1=0.2$;

$\rho_2=P(0,1,1)+P(0,2,0)+P(1,1,0)=0.4+0.1+0.05=0.55$;

$\rho_3=P(0,0,2)+P(0,1,1)+P(1,0,1)=0.3+0.4+0.05=0.75$.

2) Загрузка узлов СеМО (см.п.3.4.3) определяется по формуле:

$$\rho_j=\frac{\alpha_j\lambda_0 b_j}{K_j} \quad (j=\overline{1,3})$$

или с учётом того, что $K_1=K_2=K_3=1$ и $b_1=b_2=b_3=b$, получим:

$$\rho_j=\alpha_j\lambda_0 b \quad (j=\overline{1,3}),$$

где λ_0 - интенсивность потока заявок, проходящих через нулевой узел ЗСеМО, значение которой не известно.

Зная загрузку $\rho_1=0.2$ и коэффициент передачи $\alpha_1=4$ узла 1, найдём:

$\lambda_0 b=\rho_1/\alpha_1=0.2/4=0.05$.

3) Теперь с использованием того же выражения для расчёта загрузок узлов 2 и 3 можно определить значения соответствующих коэффициентов передач:

$$\alpha_2=\frac{\rho_2}{\lambda_0 b}=\frac{0,55}{0,05}=11; \quad \alpha_3=\frac{\rho_3}{\lambda_0 b}=\frac{0,75}{0,05}=25.$$

Задача 3. На автозаправочную станцию (АЗС) с одной колонкой прибывают автомобили со средним интервалом между моментами прибытия *X* минут. Водитель каждого автомобиля сначала заправляет бензином автомобиль в течение случайного времени, распределённого по экспоненциальному закону, со средним значением Y минут, а затем идёт к оператору АЗС и оплачивает бензин, затрачивая на это в среднем ещё Y минут. После этого автомобиль покидает заправку, и к колонке подъезжает следующий ожидающий заправки автомобиль. Ожидающие автомобили образуют очередь перед АЗС.

1) Сформулировать предположения и допущения, при которых процесс функционирования бензозаправочной станции можно рассматри-

вать как марковский.

2) Нарисовать и подробно описать модель в терминах теории массового обслуживания.

3) Выполнить кодирование и нарисовать размеченный граф переходов марковского процесса.

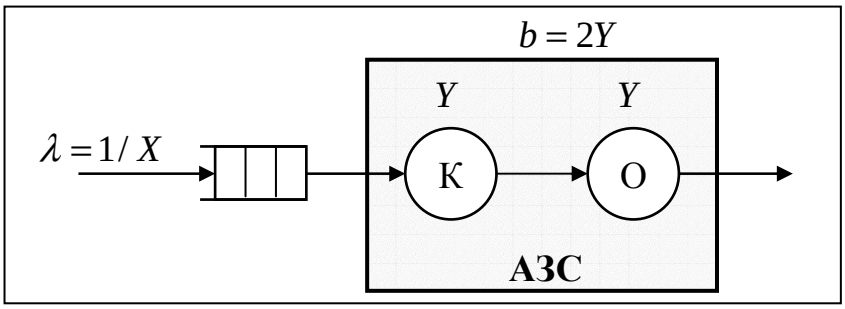
4) Сформулировать требования, при которых марковский процесс обладает эргодическим свойством.

Решение.

1) *Предположения и допущения*, при которых процесс функционирования бензозаправочной станции можно рассматривать как марковский:

- прибывающие на бензозаправочную станцию автомобили образуют *простейший поток*;
- время, затрачиваемое на заправку, и время, затрачиваемое на оплату за бензин, представляют собой случайные величины, распределённые по *экспоненциальному закону*;
- интервал времени от момента отъезда от бензоколонки заправленного автомобиля до момента подъезда к бензоколонке следующего ожидающего автомобиля предполагается много меньшим по сравнению со временем заправки и принимается равным *нулю*;
- в очереди ожидающих заправки автомобилей может находиться любое их количество, то есть имеем накопитель *неограниченной ёмкости*.

2) *Модель* в терминах теории массового обслуживания:



Модель АЗС представляет собой *одноканальную* СМО с накопителем *неограниченной* ёмкости, в которую поступает *простейший* поток заявок (автомобилей) с интенсивностью $\lambda = 1/X$. Обслуживание в приборе складывается из *двух экспоненциальных фаз*: на первой фазе (К) выполняется заправка на колонке автомобиля бензином, а на второй (О) – оплата за бензин. Интенсивность обслуживания на каждой фазе равна $\mu = 1/Y$ заявок в минуту, следовательно, интенсивность обслуживания в приборе (АЗС) составляет $1/(2Y) = \mu/2$. Предположение об экспоненциальном характере обслуживания на каждой фазе обусловливает распределение длительности обслуживания в приборе по *закону Эрланга 2-го порядка*.

3) *Кодирование* и *размеченный граф* переходов марковского процесса.

В качестве параметра, описывающего состояние марковского процесса, будем рассматривать количество заявок *k*, находящихся в СМО

(на обслуживании в приборе и в накопителе), при этом следует различать, на какой экспоненциальной фазе обслуживания в приборе находится заявка. Поскольку в системе в произвольный момент времени может находиться любое сколь угодно большое число заявок, то количество состояний марковского процесса равно бесконечности:

$\mathbf{E}_0: k = 0$ – в системе нет ни одной заявки;

$\mathbf{E}_1$: $k = 1_1$ – в системе находится 1 заявка на обслуживании в фазе 1;

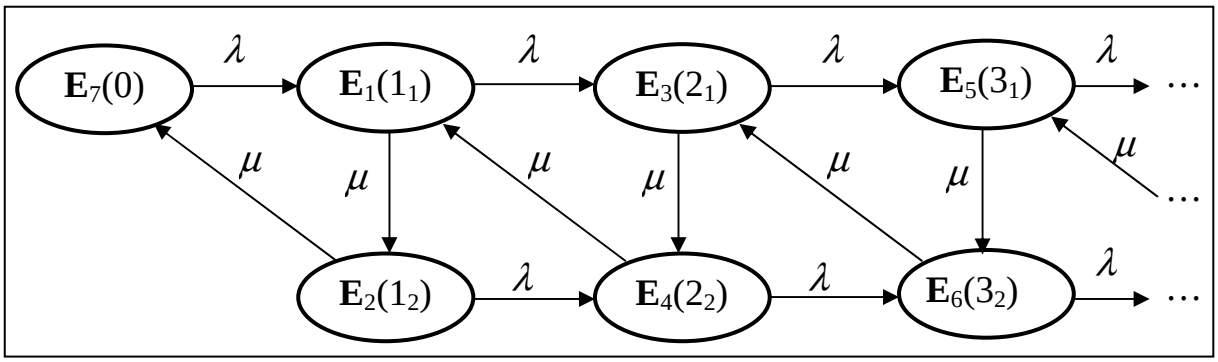
$\mathbf{E}_2$: $k = 1_2$ – в системе находится 1 заявка на обслуживании в фазе 2;

$\mathbf{E}_3$ $k = 2_1$ – в системе находятся 2 заявки (одна – на обслуживании в фазе 1 и вторая ожидает в накопителе);

$\mathbf{E}_4$ $k = 2_2$ – в системе находятся 2 заявки (одна – на обслуживании в фазе 2 и вторая ожидает в накопителе);

…

Размеченный граф переходов имеет следующий вид:



4) *Требования*, при которых марковский процесс обладает эргодическим свойством.

Марковский процесс с непрерывным временем и бесконечным количеством состояний обладает эргодическим свойством, если в моделируемой системе нет перегрузок. Для этого необходимо, чтобы загрузка системы не превышала единицы:

$$\rho = \lambda b = \frac{2Y}{X} < 1.$$

Отсюда вытекает очевидное требование следующего вида: $X > 2Y$, то есть средний интервал между прибывающими на АЗС автомобилями должен быть больше, чем среднее время их обслуживания, затрачиваемое на заправку и оплату.

Если это условие не выполняется, можно ограничить ёмкость накопителя, построив перед АЗС площадку с ограниченным числом мест для ожидающих автомобилей, полагая, что при отсутствии на этой площадке свободных мест автомобили отправятся на другую АЗС.

5.8. Самоконтроль: перечень вопросов и задач

1. Понятие случайного процесса.
2. Что понимается под состоянием случайного процесса?
3. Классификация случайных процессов.
4. В чём отличие дискретного случайного процесса от непрерывного?
5. Привести примеры систем, в которых процессы непрерывными.
6. Привести примеры систем, в которых процессы дискретными.
7. В чём отличие дискретного случайного процесса с непрерывным временем от процесса с дискретным временем?
8. Понятие марковского случайного процесса.
9. Как называется процесс, в котором переход из одного состояния в другое зависит только от состояния, в котором находится процесс?
10. При каком условии случайный процесс с непрерывным временем является марковским?
11. По какому закону должны быть распределены интервалы времени между соседними переходами, чтобы дискретный случайный процесс был марковским? Ответ обосновать.
12. Дать определение интенсивности перехода для марковского случайного процесса с непрерывным временем.
13. Из какого условия определяются диагональные элементы матрицы интенсивностей переходов?
14. Чему равны диагональные элементы матрицы интенсивностей переходов?
15. Понятие эргодического свойства случайного процесса.
16. В чем различие между случайными процессами, обладающими и не обладающими эргодическим свойством?
17. Что означает понятие "стационарная вероятность состояния случайного процесса"?
18. Перечислить условия, при которых марковский процесс с дискретным временем обладает эргодическим свойством.
19. Объяснить на примере, почему марковский процесс с разложимой и периодической матрицей вероятностей переходов не обладает эргодическим свойством?
20. Определить, обладает ли эргодическим свойством случайный процесс с дискретным временем с заданной матрицей вероятностей переходов P, сопроводив ответ необходимыми пояснениями.

$$P = \begin{array}{c|cccc} & E_1 & E_2 & E_3 & E_4 \\ E_1 & 0.2 & 0 & 0.8 & 0 \\ E_2 & 0 & 0.2 & 0.3 & 0.4 \\ E_3 & 0.4 & 0 & 0.6 & 0 \\ E_4 & 0.2 & 0.3 & 0 & 0.5 \end{array}$$

21. Известны вероятности состояний двухузловой замкнутой СеМО: P(0,4)=0,4; P(1,3)=0,1; P(2,2)=0,2; P(3,1)=0,2; P(4,0)=0,1, где состояние (M_1,M_2) задает число заявок в одноканальном узле 1 и трехканальном узле

2 соответственно. Определить среднее число заявок в СеМО, находящихся в состоянии ожидания.

22. Известны вероятности состояний трехузловой замкнутой СеМО: P(0,0,2)=0,2; P(0,1,1)=0,1; P(0,2,0)=0,15; P(1,0,1)=0,35; P(1,1,0)=0,15; P(2,0,0)=0,05, где состояние (M_1,M_2,M_3) задает число заявок в узле 1, 2, 3 соответственно. Определить среднее число параллельно работающих узлов ЗСеМО.

23. Известны вероятности состояний трехузловой замкнутой СеМО: P(0,0,2)=0,1; P(0,1,1)=0,3; P(0,2,0)=0,4; P(1,0,1)=0,05; P(1,1,0)=0,05; P(2,0,0)=0,1. Длительности обслуживания заявок во всех одноканальных узлах одинаковы. Определить значения коэффициентов передач второго и третьего узлов сети, если известно, что коэффициент передачи первого узла равен 2.

24. Известны вероятности состояний трехузловой ЗСеМО: P(2,0,0)=0,05; P(1,1,0)=0,25; P(0,2,0)=0,1; P(1,0,1)=0,1; P(0,1,1)=0,3; P(0,0,2)=0,2. Определить производительность ЗСеМО, если известно, что коэффициент передачи третьего узла (двухканального) равен 2, а средняя длительность обслуживания заявок в этом узле равна 0,1 с.

25. Система содержит два обслуживающих прибора и накопитель единичной емкости (для одной заявки). В систему поступает простейший поток заявок с интенсивностью λ. Заявки с равной вероятностью попадают в один из них, если оба прибора свободны, и занимают свободный прибор, когда другой прибор занят обслуживанием. Когда оба прибора заняты, заявка заносится в накопитель, если он свободен, или теряется, если накопитель занят. Длительность обслуживания заявок в обоих приборах распределена по гиперэкспоненциальному закону, причем первый прибор работает с двое большей скоростью. Нарисовать модель системы и размеченный граф переходов марковского процесса с необходимыми для понимания комментариями. Составить систему уравнений для стационарных вероятностей .

26. Система содержит два обслуживающих прибора и накопитель единичной емкости (для одной заявки). В систему поступают заявки с интенсивностью λ. Если оба прибора свободны, то поступившая заявка всегда попадает в первый прибор, и занимают свободный прибор, когда другой прибор занят обслуживанием. Когда оба прибора заняты, заявка заносится в накопитель, если он свободен, или теряется, если накопитель занят. Первый прибор работает с вдвое меньшей скоростью. 1) Сформулировать условия (предположения и допущения), при которых случайный процесс, протекающий в системе, будет марковским. 2) Нарисовать модель системы. 3) Выполнить кодирование марковского процесса. 4) Нарисовать размеченный граф переходов марковского процесса. 5) Выписать систему уравнений для определения стационарных вероятностей состояний. 6) Сформулировать условия, при которых марковский процесс обладает эргодическим свойством.

27. На автозаправочной станции (АЗС) имеется две колонки: одна для заправки легковых автомобилей бензином и другая для заправки грузовых автомобилей дизельным топливом. На станцию прибывают автомобили со средним интервалом между моментами прибытия T_0 минут, причём легковые автомобили прибывают в 4 раза чаще, чем грузовые. Время заправки легковых автомобилей в среднем составляет X минут, а грузовых – в два раза больше. Перед АЗС имеется площадка для ожидания прибывающих автомобилей, на которой могут разместиться один грузовой или два легковых автомобиля. Если площадка занята, то автомобили покидают АЗС не заправившись. 1) Сформулировать предположения и допущения, при которых процесс функционирования бензозаправочной станции можно рассматривать как марковский. 2) Нарисовать и подробно описать модель в терминах теории массового обслуживания. 3) Выполнить кодирование и нарисовать размеченный граф переходов марковского процесса. 4) Сформулировать требования, при которых марковский процесс будет обладать эргодическим свойством.

28. В мужской парикмахерской работает один мастер. Средний интервал между моментами прихода клиентов составляет X минут. Каждый клиент просит сначала побрить, а затем постричь. Мастер тратит на каждую из этих операций случайное время со средним значением Y минут. В парикмахерской имеется одно кресло для ожидания. Если кресло занято, то очередной пришедший клиент уходит из парикмахерской не обслуженным. 1) Сформулировать предположения и допущения, при которых процесс функционирования парикмахерской можно рассматривать как марковский. 2) Нарисовать и подробно описать модель в терминах теории массового обслуживания. 3) Выполнить кодирование марковского процесса. 4) Нарисовать размеченный граф переходов марковского процесса. 5) Выписать систему уравнений для определения вероятностей состояний. 6) Сформулировать требования, при которых марковский процесс обладает эргодическим свойством.

29. В парикмахерскую, в которой работают мастер и ученик, приходят клиенты в среднем с интервалом t_1 минут. Пришедший клиент направляется к мастеру, если он свободен, и к ученику, в противном случае. Когда мастер и ученик заняты, клиент располагается в зале на имеющемся там единственном стуле для ожидания, если он свободен. Если стул занят, то пришедший клиент покидает парикмахерскую. Мастер работает вдвое быстрей, чем ученик. 1) Сформулировать условия, при которых процесс функционирования парикмахерской можно представить в виде марковского процесса. 2) Нарисовать детальную модель системы с подробным ее описанием. 3) Выполнить кодирование состояний и нарисовать размеченный граф переходов марковского процесса. 4) Составить систему уравнений для стационарных вероятностей.

Раздел 6. ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ

«Если эксперимент удался, что-то здесь не так…»
(Первый закон Финэйгла)

6.1. Основы имитационного моделирования

6.1.1. Понятие имитационного моделирования

Статистическое моделирование – метод исследования сложных систем, основанный на описании процессов функционирования отдельных элементов в их взаимосвязи с целью получения множества частных результатов, подлежащих обработке методами математической статистики для получения конечных результатов. В основе статистического моделирования лежит метод статистических испытаний – метод Монте-Карло.

Имитационная модель – универсальное средство исследования сложных систем, представляющее собой логико-алгоритмическое описание поведения отдельных элементов системы и правил их взаимодействия, отображающих последовательность событий, возникающих в моделируемой системе.

Если статистическое моделирование выполняется с использованием имитационной модели, то такое моделирование называется *имитационным*.

Понятия «статистическое и имитационное моделирование» часто рассматривают как синонимы. Однако следует иметь в виду, что статистическое моделирование не обязательно является имитационным. Например, вычисление определённого интеграла методом Монте-Карло путем определения подынтегральной площади на основе множества статистических испытаний, относится к статистическому моделированию, но не может называться имитационным.

Наиболее широкое применение имитационное моделирование получило при исследовании сложных систем с дискретным характером функционирования, в том числе моделей массового обслуживания. Для описания процессов функционирования таких систем обычно используются временны*е* диаграммы.

Временная диаграмма – графическое представление последовательности событий, происходящих в системе. Для построения временны*х* диаграмм необходимо достаточно четко представлять взаимосвязь событий внутри системы. Степень детализации при составлении диаграмм зависит от свойств моделируемой системы и от целей моделирования.

Поскольку функционирование любой системы достаточно полно отображается в виде временной диаграммы, *имитационное моделирование можно рассматривать как процесс реализации диаграммы функционирования исследуемой системы на основе сведений о характере функционирования отдельных элементов и их взаимосвязи*.

Имитационное моделирование обычно проводится на ЭВМ в соответствии с программой, реализующей заданное конкретное логико-алгоритмическое описание. При этом несколько часов, недель или лет работы исследуемой системы могут быть промоделированы на ЭВМ за несколько минут. В большинстве случаев модель является не точным аналогом системы, а скорее её символическим отображением. Однако такая модель позволяет производить измерения, которые невозможно произвести каким-либо другим способом.

Имитационное моделирование обеспечивает возможность испытания, оценки и проведения экспериментов с исследуемой системой без каких-либо непосредственных воздействий на нее.

Первым шагом при анализе любой конкретной системы является выделение элементов, и формулирование логических правил, управляющих взаимодействием этих элементов. Полученное в результате этого описание называется **моделью системы**. Модель обычно включает в себя те аспекты системы, которые представляют интерес или нуждаются в исследовании.

Поскольку целью построения любой модели является исследование характеристик моделируемой системы, в имитационную модель должны быть включены средства сбора и обработки статистической информации по всем интересующим характеристикам, основанные на методах математической статистики.

6.1.2. Принципы организации имитационного моделирования

«Даже маленькая практика стоит большой теории» ***(Закон Буккера)***

Рассмотрим принципы имитационного моделирования на примере простейшей базовой модели в виде одноканальной системы массового обслуживания с однородным потоком заявок (рис.6.1), в которую поступает случайный поток заявок с интервалами между соседними заявками, распределёнными по закону $A(\tau)$, а длительность обслуживания заявок в приборе распределена по закону $B(\tau)$.

Процесс функционирования такой системы может быть представлен в виде временных диаграмм, на основе которых могут быть измерены и рассчитаны характеристики обслуживания заявок. Поскольку процессы поступления и обслуживания заявок в системе носят случайный характер, то для построения диаграмм необходимо иметь *генераторы случайных чисел*.

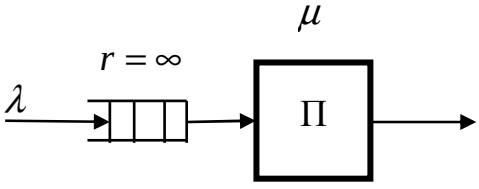


Рис. 6.1. СМО с накопителем неограниченной ёмкости

Положим, что в нашем распоряжении имеются генераторы случай-

ных чисел, формирующие значения соответствующих случайных величин *с заданными законами распределений* $A(\tau)$ и $B(\tau)$. Тогда можно построить временны*е* диаграммы, отображающие процесс функционирования рассматриваемой системы.

На рис.6.2 представлены четыре диаграммы, отображающие:

1) «**процесс поступления заявок**» в виде моментов t_i поступления заявок в систему, формируемых по правилу: $t_i = t_{i-1} + \tau_{a_i}$ $(t_0 = 0)$, где τ_{a_i} $(i = 1, 2, \ldots)$ – интервалы между поступающими в систему заявками, значения которых вырабатываются с помощью генератора случайных величин $A(\tau)$;

2) «**процесс обслуживания в приборе**», представленный в виде длительностей обслуживания τ_{bi}, которых вырабатываются с помощью генератора случайных величин $B(\tau)$, и моментов завершения обслуживания $t_i^{'}$ заявок в приборе, определяемых по следующему правилу:

$t_i^{'} = t_i + \tau_{b_i}$, если на момент поступления *i*-й заявки обслуживающий прибор был свободен;

$t_i^{'} = t_{i-1}^{'} + \tau_{b_i}$, если на момент поступления *i*-й заявки обслуживающий прибор был занят обслуживанием предыдущей заявки ($i = 1, 2, \ldots$; $t_0^{'} = 0$);

3) «**модельное или реальное время**», показывающее дискретное (скачкообразное) изменение времени в реальной системе, каждый момент которого соответствует одному из следующих событий: поступление заявки в систему или завершение обслуживания заявки в приборе; отметим, что в эти моменты времени происходит изменение состояния системы, описываемое числом заявок, находящихся в системе;

4) «**число заявок в системе**», описывающее состояние дискретной системы и изменяющееся по правилу: увеличение на 1 в момент поступления заявки в систему и уменьшение на 1 в момент завершения обслуживания.

При соблюдении выбранного временн*о*го масштаба представленные диаграммы позволяют путем измерения определить значения вероятностно-временны*х* характеристик функционирования моделируемой системы, в частности, как показано на второй диаграмме, время нахождения (пребывания) каждой заявки в системе: τ_{u_i} $(i = 1, 2, \ldots)$.

Очевидно, что время пребывания заявок в системе – величина случайная. В простейшем случае, применяя методы математической статистики, можно рассчитать два первых момента распределения времени пребывания:

- математическое ожидание:

$$u = \frac{1}{N} \sum_{i=1}^{N} \tau_{u_i} ;$$

- второй начальный момент:

$$u^{(2)} = \frac{1}{N-1}\sum_{i=1}^{N}\tau_{u_i}^2 ,$$

где N - количество значений времени пребывания заявок, полученных на диаграмме, то есть количество заявок, отображенных на диаграмме как прошедшие через систему и покинувшие её.

Отсюда легко могут быть получены значения дисперсии, среднеквадратического отклонения и коэффициента вариации времени пребывания заявок в системе.

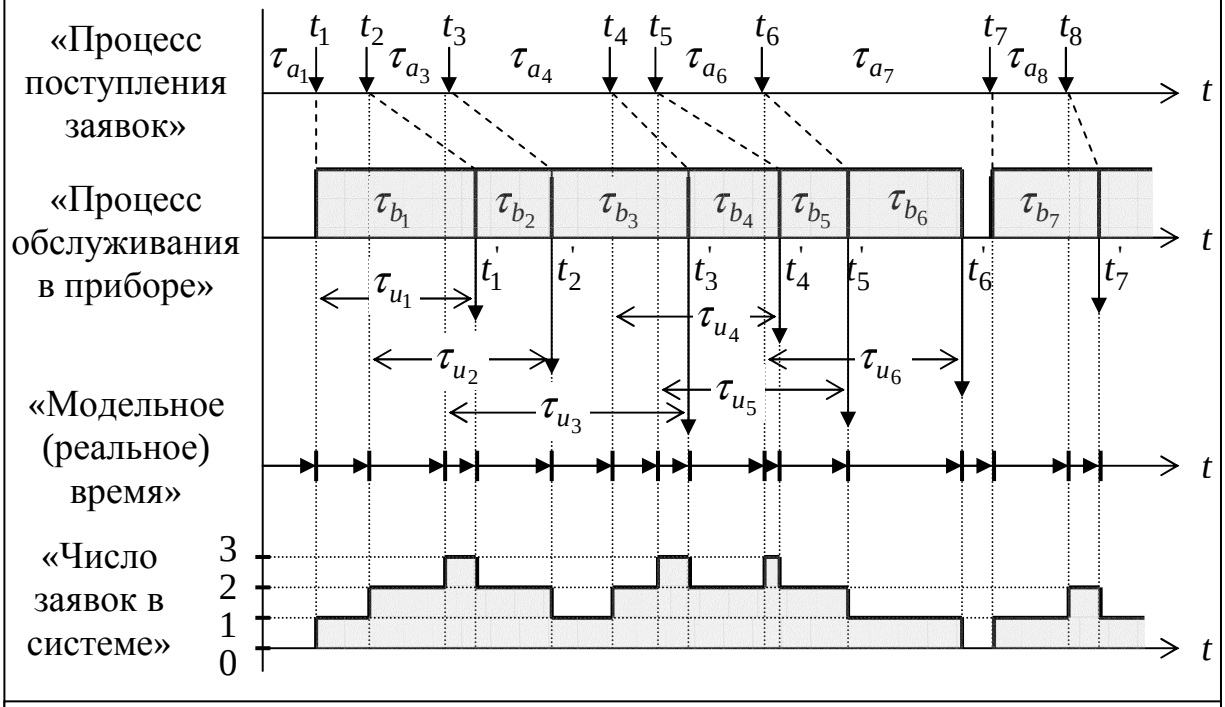


Рис.6.2. Диаграммы функционирования одноканальной СМО

На основе полученных с помощью временны*х* диаграмм значений времени пребывания заявок в системе можно построить гистограмму функции или плотности распределения времени пребывания.

Точность полученных числовых моментов распределения и качество гистограмм существенно зависит от количества значений N времени пребывания заявок, на основе которых они рассчитываются: чем больше N, тем точнее результаты расчета. Значение N может составлять от нескольких тысяч до десятков миллионов. Конкретное значение N зависит от многих факторов, влияющих на скорость сходимости результатов к истинному значению, основными среди которых при моделировании систем и сетей массового обслуживания являются законы распределений интервалов между поступающими заявками и длительностей обслуживания, загрузка системы, сложность модели, количество классов заявок и т.д.

Ясно, что построение вручную таких временных диаграмм с тысячами и более проходящими через систему заявками, нереально. В то же время, использование ЭВМ для реализации временны*х* диаграмм

позволяет существенно ускорить процессы моделирования и получения конечного результата. Поэтому, как сказано выше, имитационное моделирование можно рассматривать как процесс реализации диаграммы функционирования исследуемой системы.

Таким образом, имитационная модель представляет собой алгоритм реализации временной диаграммы функционирования исследуемой системы. Наличие встроенных в большинство алгоритмических языков генераторов случайных чисел значительно упрощает процесс реализации имитационной модели на ЭВМ. Однако при этом остаётся ряд проблем, требующих своего решения. Одна из них заключается в принципе реализации временной диаграммы и, связанной с ней, проблемой организации службы времени в имитационной модели.

В простейшем случае временная диаграмма может быть реализована следующим образом: сначала формируются моменты поступления всех заявок в систему, а затем для каждой заявки определяются длительности обслуживания в приборе и формируются моменты завершения обслуживания (выхода заявок из системы). Очевидно, что такой подход неприемлем, поскольку даже для нашей очень простой системы придётся хранить в памяти ЭВМ одновременно миллионы значений моментов поступления и завершения обслуживания заявок, а также других переменных, причём с увеличением количества классов заявок и количества обслуживающих приборов это число увеличится многократно.

Второй подход, который может быть предложен для реализации временной диаграммы, – пошаговое построение диаграммы. Для этого следует сформировать переменную для модельного времени и выбрать шаг Δt его изменения. В каждый такой момент времени необходимо проверять, какое событие (поступление в систему или завершение обслуживания заявки) произошло в системе за предыдущий интервал Δt.

Этот подход значительно сокращает потребность в памяти, поскольку в этом случае в каждый момент времени необходимо хранить в памяти ЭВМ значения параметров (моментов поступления и завершения обслуживания) только тех заявок, которые находятся в системе на данный момент времени.

Недостатки такого подхода очевидны. Во-первых, проблематичным является выбор длины интервала Δt. С одной стороны, интервал Δt должен быть как можно меньше для уменьшения методической погрешности моделирования, с другой стороны, интервал Δt должен быть как можно больше для уменьшения времени моделирования.

Наиболее эффективным подходом признан подход *с переменным шагом продвижения модельного времени*, который реализуется в соответствии с принципом «до ближайшего события». ***Принцип «продвижения модельного времени до ближайшего события»*** заключается в следующем. По всем процессам, параллельно протекающим в исследуемой системе, в каждый момент времени формируются моменты наступления «бли-

жайшего события в будущем». Затем модельное время продвигается до момента наступления ближайшего из всех возможных событий. В зависимости от того, какое событие оказалось ближайшим, выполняются те или иные действия. Если ближайшим событием является поступление заявки в систему, то выполняются действия, связанные с занятием прибора при условии, что он свободен, и занесение заявки в очередь, если прибор занят. Если же ближайшим событием является завершение обслуживания заявки в приборе, то выполняются действия, связанные с освобождение прибора и выбором на обслуживание новой заявки из очереди, если последняя не пуста. Затем формируется новый момент наступления этого же события. На третьей диаграмме «Модельное (реальное) время» (рис.6.2) продвижение времени в соответствии с этим принципом показано в виде стрелок.

Для того чтобы обеспечить правильную временну*ю* последовательность событий в имитационной модели, используются **системные часы**, хранящие значение текущего *модельного времени*. Изменение значения модельного времени осуществляется в соответствии с принципом «пересчёта времени до ближайшего события». Например, если текущее значение модельного времени равно 25, а очередные события должны наступить в моменты времени 31, 44 и 56, то значение модельного времени увеличивается сразу на 6 единиц и «продвигается» до значения 31. Отметим, что единицы времени в модели не обязательно должны быть конкретными единицами времени, такими как секунда или час. Основной единицей времени в модели можно выбрать любую единицу, которая позволит получить необходимую точность моделирования. Важно помнить, единицы времени выбираются исходя из требований пользователя к точности моделирования. Какая бы единица ни была выбрана, например миллисекунда или одна десятая часа, она должна неизменно использоваться во всей модели.

Кроме рассмотренной службы времени в имитационной модели необходимо реализовать процедуры, связанные с формированием потоков заявок и имитацией обслуживания, с организацией очередей заявок, с организацией сбора и статистической обработки результатов моделирования.

Таким образом, имитационное моделирование дискретных систем со стохастическим характером функционирования, таких как системы и сети массового обслуживания, предполагает использование ряда типовых процедур, обеспечивающих реализацию соответствующих имитационных моделей. К таким процедурам, в первую очередь, относятся следующие процедуры:

1) выработка (генерирование) случайных величин:
   - равномерно распределенных;
   - с заданным законом распределения;
2) формирование потоков заявок и имитация обслуживания;
3) организация очередей заявок;
4) организация службы времени;
5) сбор и статистическая обработка результатов моделирования.

6.2. Методы формирования случайных чисел

«То, что ищешь, найдешь только обыскав все»
(Закон Буба)

Функционирование элементов системы, подверженных случайным воздействиям, задается генераторами (датчиками) случайных чисел: *аппаратными* или *программными*. Генераторы случайных чисел в ЭВМ обычно реализуются программными методами, вырабатывающими псевдослучайные последовательности.

Псевдослучайными последовательностями называются вполне *детерминированные* числа, обладающие:

- *статистическими свойствами случайных чисел*, определяемых путем их проверки специальными тестами,
- *периодичностью*, то есть повторяемостью через определенные промежутки времени.

Количество случайных величин, вырабатываемых между двумя одинаковыми значениями, называется ***длиной периода генератора*** случайных величин.

При моделировании используются интервалы последовательностей псевдослучайных чисел, в которых нет ни одного числа, встречающегося более одного раза.

Для формирования случайных чисел с заданными законами распределений в качестве исходных используют случайные числа, выработанные программными генераторами равномерно распределенных случайных чисел в интервале (0,1), встроенные практически во все языки программирования. Специализированные программные средства, предназначенные для вероятностного моделирования, обычно имеют специальные встроенные процедуры генерирования случайных величин с разными законами распределений.

6.2.1. Формирование равномерно распределённых случайных величин

Для формирования равномерно распределённых случайных чисел в интервале (0; 1) могут использоваться следующие методы:

- метод квадратов;
- метод произведений;
- мультипликативный конгруэнтный метод;
- методы, представляющие модификации перечисленных методов.

Метод квадратов является одним из простейших методов и служит хорошей иллюстрацией принципа алгоритмического формирования равномерно распределённых случайных величин.

Алгоритм формирования равномерно распределённых случайных величин по методу квадратов заключается в выполнении следующих этапов:

1) выбирается некоторое исходное *n*-разрядное целое число, которое должно удовлетворять определённым условиям для получения качественного генератора случайных величин с максимально возможной длиной периода;

2) выбранное *n*-разрядное число возводится в квадрат, в результате чего получается целое число с вдвое большей разрядностью;

3) из полученного 2*n*-разрядного числа выделяются *n* средних разрядов, которые рассматриваются как дробная часть случайного числа, равномерно распределённого в интервале (0; 1);

4) выделенные на предыдущем этапе *n* средних разрядов рассматриваются как новое исходное *n*-разрядное целое число;

5) повторяются этапы 2 – 4.

Проиллюстрируем метод квадратов на следующем примере.

Пример 1. Для простоты будем оперировать десятичными числами, а не двоичными, как это реализуется в программных генераторах.

Пусть выбрано некоторое исходное четырехразрядное целое число, равное 7153. Результаты применения описанного алгоритма представлены в виде следующей таблицы:

Исх.число	Квадрат	Случайное число
7153	51 **1654** 09	0,1654
1654	02 **7357** 16	0,7357
7357	54 **1254** 49	0,1254
1254	01 **5725** 16	0,5725
5725	32 **7756** 25	0,7756
7756	60 **1555** 36	0,1555

Очевидно, что максимальная длина периода генератора, то есть максимальное количество неповторяющихся случайных чисел определяется количеством разрядов в дробной части. В нашем примере максимально возможная длина периода равна 9999 (от 0,0001 до 0,9999). Однако в действительности длина периода меньше максимально возможной и зависит от исходного целого числа. Неудачно выбранное значение исходного числа может привести к двум неприятностям: маленькой длине периода или даже к вырождению генератора, когда значения случайной величины начинают повторяться, как это показано в примере 2.

Пример 2. Исходное четырехразрядное целое число = 1357

Исх.число	Квадрат	Случайное число
1357	01 **8414** 49	0,8414
8414	70 **7953** 96	0,7953
7953	63 **2502** 09	0,2502
2502	06 **2600** 04	0,2600
2600	06 **7600** 00	0,7600
7600	57 **7600** 00	0,7600

Метод произведений аналогичен методу квадратов. Отличие состоит в том, что перемножаются два *n*-разрядных целых числа, одно из которых, называемое *ядром* или *множителем*, не меняется, а второе, называемое *множимым*, формируется из *n* последних (правых) разрядов полученного 2*n*-разрядного числа, представляющего собой произведение ядра и множимого. Естественно, что вначале, как и в методе квадратов, необходимо грамотно выбрать исходные значения ядра и множителя.

Пример 3. Ядро = 5167; множитель = 3729

Множимое	Произведение	Случайное число
3729	19 2677 43	0,2677
7743	40 0080 81	0,0080
8081	41 7545 27	0,7545
4527	23 3910 09	0,3910
1009	05 2135 03	0,2135
3501	….	…

Здесь, в отличие от предыдущего примера, в качестве следующего значения множителя выбираются не средние разряды полученного произведения, а последние *n* разрядов произведения.

Конгруэнтные методы генерирования случайных чисел получили наиболее широкое распространение для формирования на ЭВМ псевдослучайных последовательностей [13].

Два целых числа *a* и *b* называются ***конгруэнтными*** (сравнимыми) по модулю *m*, где *m* – целое число, если разность $(a-b)$ делится на *m* без остатка, а числа *a* и *b* дают одинаковые остатки от деления на *m*. Например, 2568 и 148 (по модулю 10), 1746 и 511 (по модулю 5), 6493 и 2221 (по модулю 2) и т.д.

Конгруэнтные методы описываются в виде рекуррентного соотношения следующего вида:

$$X_{i+1} = \lambda X_i + \mu (\mathrm{mod}\, m) \qquad (i = 0, 1, 2, ...),$$

где X_i, λ, μ, m – неотрицательные целые числа; X_0 – начальное значение псевдослучайной последовательности; λ – множитель; μ – аддитивная константа; m – модуль.

Каждое новое значение X_{i+1} псевдослучайной последовательности представляет собой целочисленный остаток от деления на модуль m суммы произведения предыдущего значения X_i на множитель λ и аддитивной константы μ. Последовательность псевдослучайных чисел в интервале (0; 1) формируется путем деления полученных целочисленных значений X_i на модуль m: $x_i = X_i / m \quad (i = 1, 2, ...)$.

Описанный метод генерирования псевдослучайных чисел получил название **смешанного конгруэнтного метода**.

В некоторых случаях используется более простой метод генерирования псевдослучайных чисел, представляющий собой частный случай смешанного метода, когда $\mu = 0$, и получивший название ***мультипликативного конгруэнтного метода***. В этом случае рекуррентное соотношение имеет вид:

$$X_{i+1} = \lambda X_i (\mathrm{mod}\, m) \qquad (i = 0, 1, 2, ...).$$

На каждом шаге полученное случайное число (множимое) умножается на некоторое постоянное число (множитель) и затем делится на другое постоянное число (делитель). В качестве нового случайного числа принимается остаток от деления, который служит дробной частью случайного числа, равномерно распределённого в интервале (0; 1).

Пример 4. Первое постоянное число (множитель) = 1357; второе постоянное число (делитель) = 5689.

Исходное число	Произведение	Частное, целая часть	Остаток	Случайное число
1357	1 8414 49	323	3902	0,3902
3902	5 2950 14	930	4244	0,4244
4244	5 7591 08	1012	1840	0,1840
1840	…	…	…	…

6.2.2. Проверка генераторов равномерно распределенных псевдослучайных чисел

«Когда не знаешь, что именно ты делаешь, де-лай это тщательно» ***(Правило для лаборантов)***

Достоверность и точность результатов имитационного моделирования в значительной степени определяется качеством используемых в моделях программных генераторов псевдослучайных последовательностей.

Проверка генераторов равномерно распределенных псевдослучайных чисел предполагает формирование большой совокупности или, как говорят, представительной выборки случайных чисел и выполнение множества проверочных тестов, позволяющих оценить качество генераторов.

Различают три вида проверки программных генераторов равномерно распределенных псевдослучайных чисел:

- на периодичность;
- на случайность;
- на равномерность.

Проверка на периодичность требует обязательного определения длины периода, что в значительной степени определяет качество генератора случайных чисел. Чем больше длина периода, тем генератор более качественный.

Проверка на случайность. При проверке на случайность программных генераторов *двоичных* случайных чисел можно использовать

совокупность тестов, а именно тесты проверки:

- частот;
- пар;
- комбинаций;
- серий;
- корреляции.

Тест проверки частот предполагает разбиение диапазона распределения на несколько интервалов и подсчет количества (частот или вероятностей) попаданий случайных чисел в выделенные интервалы.

Тест проверки пар заключается в подсчете количества "1" для каждого *разряда* всей совокупности выработанных генератором двоичных случайных чисел. Очевидно, количество "1" во всех разрядах должно составлять примерно 50% от количества выработанных генератором случайных чисел.

Тест проверки комбинаций сводится к подсчету "1" в случайных числах, количество которых в среднем должно составлять половину от количества разрядов.

Тест проверки серий заключается в подсчете количества различных длин последовательностей одинаковых значений (1 или 0).

Тест проверки корреляции заключается в определении коэффициента корреляции между последовательностями случайных чисел, вырабатываемых двумя разными генераторами.

Проверка на равномерность. При проверке на равномерность можно использовать тест проверки частот, так как гистограмма частот хорошо отражает равномерность распределения случайных чисел по всему диапазону изменения.

6.2.3. Методы формирования псевдослучайных чисел с заданным законом распределения

Методы формирования псевдослучайных чисел с заданным законом распределения основаны на использовании генераторов равномерно распределённых случайных величин. При этом наибольшее распространение получили следующие методы:

- аналитический (метод обратной функции);
- табличный;
- метод композиций, основанный на функциональных особенностях генерируемых распределений.

Аналитический метод заключается в построении математической зависимости, связывающей значения случайной величины с заданным законом распределения со значениями случайной величины, распределённой равномерно в интервале (0; 1).

Суть аналитического метода иллюстрируется на графике (рис.6.3). Пусть задана некоторая функция распределения $F(x)$, значения которой лежат в интервале (0; 1). Положим, что имеется генератор равномерно

распределённых в том же интервале случайных чисел: $S \in (0;1)$. Тогда, генерируя последовательность значений $s_1, s_2, s_3, \ldots$ и откладывая их по оси ординат, можно найти соответствующие значения $x_1, x_2, x_3, \ldots$ случайной величины *X*, распределённой по заданному закону $F(x)$.

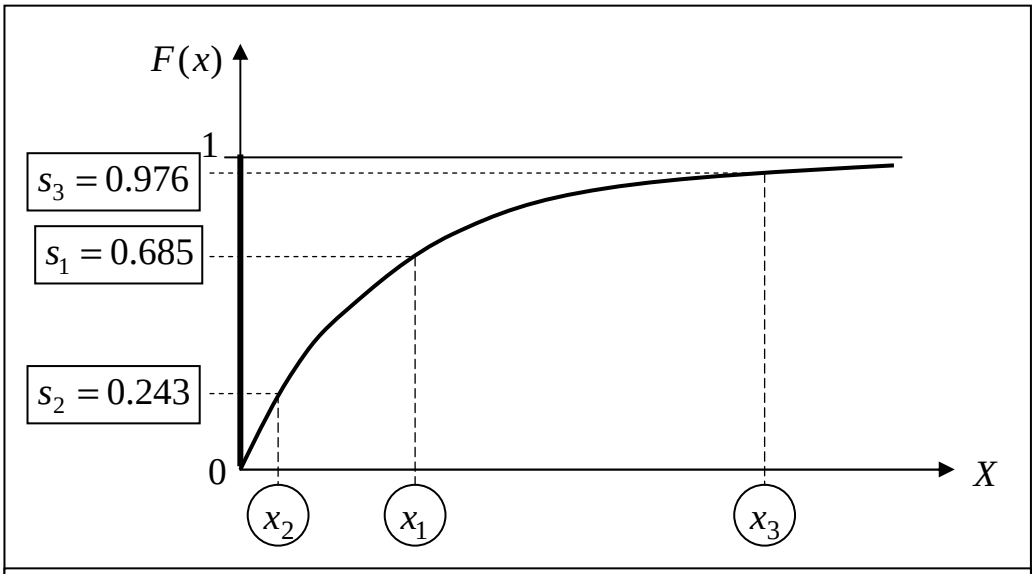


Рис.6.3. Графическая иллюстрация аналитического метода

Выведем аналитическую зависимость для расчета значений случайной величины, распределённой по экспоненциальному закону с функцией: $F(x) = 1 - e^{-\alpha x}$, где $\alpha > 0$ – параметр экспоненциального распределения.

Для этого, в соответствии с выше изложенным, необходимо решить уравнение $F(x) = s$, где s – значение равномерно распределённой в интервале (0; 1) случайной величины. Таким образом, имеем:

$1 - e^{-\alpha x} = s$ или $e^{-\alpha x} = 1 - s$.

Логарифмируя левую и правую части последнего выражения, после некоторых преобразований получим:

$$x = -\frac{1}{\alpha}\ln(1-s) \quad \text{или} \quad \boxed{x = -\frac{1}{\alpha}\ln s}.$$

Отметим, что $\frac{1}{\alpha}$ представляет собой математическое ожидание экспоненциально распределённой случайной величины.

Оба полученных выражения равнозначны, поскольку с позиций теории вероятностей случайные величины s и $(1-s)$ распределены по одному и тому же равномерному закону в интервале (0; 1). В то же время, последнее выражение предпочтительнее, поскольку не требует выполнения «лишней» операции вычитания, что позволяет уменьшить время моделирования с учетом того, что в процессе моделирования генерируются миллионы случайных чисел.

Достоинства аналитического метода:

- высокая точность метода;
- не требуется составления и хранения в памяти таблиц, как в табличном методе.

Недостатки аналитического метода:

- метод распространяется только на те функции, которые позволяют вычислить интеграл от функции плотности аналитически;
- использование численных методов вычисления интегралов приводит к погрешностям и большим затратам машинного времени;
- выражение, используемое для вычислений, содержит в себе функции вычисления логарифмов, возведения в степень, вычисления радикалов, что требует значительных затрат машинного времени.

Табличный метод заключается в формировании таблицы, содержащей пары чисел: значение функции распределения $F(x)$ и соответствующее ему значение x случайной величины. В качестве аргумента при обращении к таблице используется значение $s \in (0;1)$ равномерно распределенной случайной величины S, задающее значение функции распределения $F(x)$, а в качестве функции – значение x случайной величины X с соответствующим законом распределения $F(x)$.

Значение случайного числа, находящегося между узлами табуляции, обычно рассчитывается методом *линейной интерполяции*.

В ранних версиях GPSS для генерирования случайных чисел, распределённых по экспоненциальному закону, использовался табличный генератор со следующими значениями функции распределения $F(x)$ (от 0 до 0.9997) и соответствующими им значениями случайной величины x (от 0 до 8):

$F(x)$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.84	0.88
x	0	0.104	0.222	0.335	0.509	0.69	0.915	1.2	1.38	1.6	1.83	2.12

0.9	0.92	0.94	0.95	0.96	0.97	0.98	0.99	0.995	0.998	0.999	0.9997
2.3	2.52	2.81	2.99	3.2	3.5	3.9	4.6	5.3	6.2	7	8

Представленные в таблице значения $F(x)$ и x соответствуют экспоненциальному распределению *с математическим ожиданием, равным единице*. Если математическое ожидание экспоненциально распределённой случайной величины отличается от 1, то полученное с помощью этой таблицы значение случайной величины умножается на значение математического ожидания.

Заметим, что табулирование функции выполнено с переменным шагом: в начале таблицы шаг изменения аргумента (значений экспоненциальной функции распределения $F(x)$) равен 0.5, а в конце – 0.0007. Это обусловлено необходимостью обеспечить приемлемую методическую погреш-

ность, возникающую в результате линейной интерполяции при вычислении значений случайной величины, находящихся между узлами табуляции.

Достоинства табличного метода:

- существует принципиальная возможность построения таблицы для формирования случайных последовательностей с любым законом распределения, в том числе полученного экспериментальным путём;
- можно обеспечить любую заданную точность генерирования случайных чисел за счет увеличения количества интервалов табуляции (уменьшения шага табуляции);
- для генерирования случайных величин с заданным законом распределения вероятностей требуется только генератор равномерно распределенных случайных чисел и выполнение несложных операций, занимающих мало времени.

Недостатки табличного метода:

- значительные затраты памяти для хранения большого числа таблиц с разными законами распределений;
- наличие методической погрешности, обусловленной применением линейной интерполяции для определения значений случайных чисел, находящихся между узлами табуляции;
- для уменьшения методической погрешности формирования случайных последовательностей при использовании линейной интерполяции следует увеличивать количество точек табуляции, что приводит к увеличению размера таблиц и, как следствие, к дополнительным затратам памяти и времени;
- в связи с неодинаковой скоростью изменения функции распределения для обеспечения высокой точности формирования случайных последовательностей табулирование должно выполняться с переменным шагом, выбор которого связан с определёнными проблемами.

Метод композиций основан на функциональных особенностях вероятностных распределений, таких как распределение Эрланга, гипоэкспоненциальное и гиперэкспоненциальное распределения.

Метод используется, как правило, в тех случаях, когда не удаётся получить аналитическим методом решение в явном виде. Например, значения случайных величин, распределённых по закону Эрланга и гипоэкспоненциальному закону могут быть получены путём сложения нескольких экспоненциально распределённых случайных величин, а значения случайных величин, распределённых по гиперэкспоненциальному закону – путём вероятностного формирования смеси из нескольких экспоненциально распределённых случайных величин с разными математическими ожиданиями.

Для оценки качества случайных последовательностей с заданным законом распределения наиболее часто используют тест проверки частот и метод доверительного интервала для математического ожидания.

6.3. Введение в систему имитационного моделирования GPSS World

«Ошибаться человеку свойственно, но окончательно всё запутать может только компьютер»
(Пятый закон ненадежности)

GPSS (General Purpose Simulation System) – общецелевая система имитационного моделирования (СИМ), предназначенная для разработки моделей сложных систем с дискретным и непрерывным характером функционирования и проведения экспериментов с целью изучения свойств и закономерностей процессов, протекающих в них, а также выбора наилучшего проектного решения среди нескольких возможных вариантов.

Среди множества реализаций GPSS одной из наиболее доступных и популярных является GPSS World для работы на персональных компьютерах под управлением ОС Windows. GPSS World обладает удобным многооконным пользовательским интерфейсом, встроенными средствами визуализации и интерактивного управления процессом моделирования, обширной библиотекой встроенных процедур, включающей, в том числе, генераторы случайных величин для более чем двух десятков вероятностных распределений. Все это делает процесс моделирования эффективным и наглядным.

В GPSS World включены специальные средства для моделирования большого класса дискретных систем со стохастическим характером функционирования, в частности, систем и сетей массового обслуживания, что позволяет сделать модели ясными и лаконичными.

6.3.1. Состав системы имитационного моделирования GPSS World

Система имитационного моделирования GPSS World включает:

- язык GPSS – высокоуровневый язык имитационного моделирования;
- язык PLUS (Programming Language Under Simulation) – встроенный в GPSS язык программирования низкого уровня;
- компилятор – программа для трансляции (перевода) с языка высокого уровня на язык компьютера.

Объектами СИМ GPSS World являются:

- **«Модель»** или **«GPSS-модель»** – программа, написанная на *языке GPSS* и представляющая собой последовательность *операторов*, описывающих логику работы моделируемой системы, каждый из которых реализует некоторую конкретную функцию.
- **«Процесс моделирования»** – непосредственно исполняемый объект, создаваемый в результате трансляции объекта «GPSS-модель»; реализация «процесса моделирования» заключается в перемещении в модели некоторых подвижных объектов, называемых *транзактами*.

- **«Отчёт»** – создается автоматически по завершении процесса моделирования и содержит результаты моделирования.
- **«Текстовый объект»** – текстовые файлы, используемые для упрощения разработки больших моделей и формирования библиотеки исходных текстов.

Три первых объекта являются основными и всегда используются при имитационном моделировании.

6.3.2. Элементы языка GPSS World

Элементами языка GPSS World являются:

- ***алфавитно-цифровые символы***: латинские прописные и строчные буквы от «A» до «Z» и цифры от 0 до 9;
- ***имя*** – совокупность алфавитно-цифровых символов (от 1 до 200), *начинающаяся всегда с алфавитного символа*, причем допускается использование букв только *латинского алфавита*; для того чтобы имя не совпало с зарезервированными ключевыми словами (названиями операторов, системными числовыми атрибутами и т.п.), рекомендуется использование символа «_» (подчеркивание); примеры *правильных* имен: AS_27, R25, Pribor, W5Fix, Object_New1;
- ***метка*** – *имя*, расположенное в поле метки оператора для задания имени объекта GPSS-модели (памяти, таблицы, переменной,…) или для обозначения местоположения блока;
- ***переменная*** пользователя – *имя*, используемое в процессе моделирования для хранения числовых и строковых величин;
- ***числа*** – могут быть трёх типов:
  - *целочисленные* 32-разрядные (при переполнении преобразуются в вещественные);
  - *вещественные* 64-разрядные с плавающей точкой двойной точности (порядок может изменяться от –308 до +308, а точность ограничена примерно 15-ю десятичными разрядами),
  - *строковые* – массив символов произвольной длины, определяемой пользователем;
- ***системные числовые атрибуты (СЧА)*** – переменные, описывающие состояния процесса моделирования, автоматически поддерживаемые в GPSS и доступные в течение всего процесса моделирования;
- ***арифметические операторы*** – задают арифметические операции (перечислены в порядке приоритетности выполнения операций):
  - ^ (возведение в степень);
  - # (умножение), может быть изменено пользователем на *,
    / (деление),
    \ (целочисленное деление);
  - @ (остаток от деления);
  - + (сложение),
    - (вычитание);

- ***операторы отношения*** – задают логические условия (перечислены в порядке приоритетности выполнения операций):
  - > или ’G’ (больше),
    >= или ‘GE’ (больше или равно),
    < или ‘L’ (меньше),
    <= или ‘LE’ (меньше или равно);
  - = или ‘E’ (равно),
    != или ‘NE’ (не равно);
- ***логические операторы*** – задают логические операции (перечислены в порядке приоритетности выполнения операций):
  - & или ‘AND’ (логическое «И»);
  - | или ‘OR’ (логическое «ИЛИ»);
- ***выражения*** – часть языка PLUS: представляют собой совокупность *переменных*, *чисел* и *СЧА*, связанных *арифметическими операторами*, *логическими операторами* и *операторами отношения*; могут использоваться в операндах операторов GPSS и в PLUS-процедурах; всегда заключаются в круглые скобки;
- ***процедуры*** – программы на языке PLUS (PLUS-процедуры), встроенные в GPSS World (*стандартная процедура*) или созданные пользователем (*пользовательская процедура*); обращение к процедуре осуществляется путем задания в качестве операнда GPSS-операторов имени процедуры с её параметрами; библиотека стандартных процедур включает:
  - **обслуживающие процедуры** для управления прогонами процессов моделирования и анализа экспериментов;
  - **математические процедуры**: ABS (абсолютное значение), EXP (степень экспоненты), INT (целая часть), LOG (натуральный логарифм), SQR (квадратный корень), SIN (синус), COS (косинус), TAN (тангенс), ATN (арктангенс);
  - **процедуры запроса** для получения информации о состоянии находящегося в модели транзакта;
  - **строковые процедуры** для операций со строками;
  - **процедуры потоков данных** для управления потоками данных внутри PLUS-процедуры;
  - **процедуры динамического вызова** для вызова функций, хранящихся во внешних исполняемых файлах, включая динамически подключаемые библиотеки DLL;
  - **вероятностные распределения**.

6.3.3. Объекты GPSS-модели

> «Машинная программа выполняет то, что вы ей приказали делать, а не то, что бы вы хотели, чтобы она делала» ***(Третий закон Грида)***

GPSS-модель представляет собой написанную на языке GPSS программу и включает в себя множество объектов, которые могут быть

разбиты на 6 групп (рис.6.4):

- основные объекты;
- оборудование;
- числовые объекты;
- генераторы случайных чисел;
- групповые списки;
- потоки данных.

К *основным объектам* GPSS-модели относятся:

- **операторы (блоки и команды)** – основные объекты GPSS-модели, определяющие совокупность действий, которая должна быть выполнена в модели в соответствии с заданными в операторе параметрами, называемыми ***операндами***;
- **транзакты** – динамические объекты, движущиеся в GPSS-модели от одного оператора (блока) к другому в заданной последовательности.

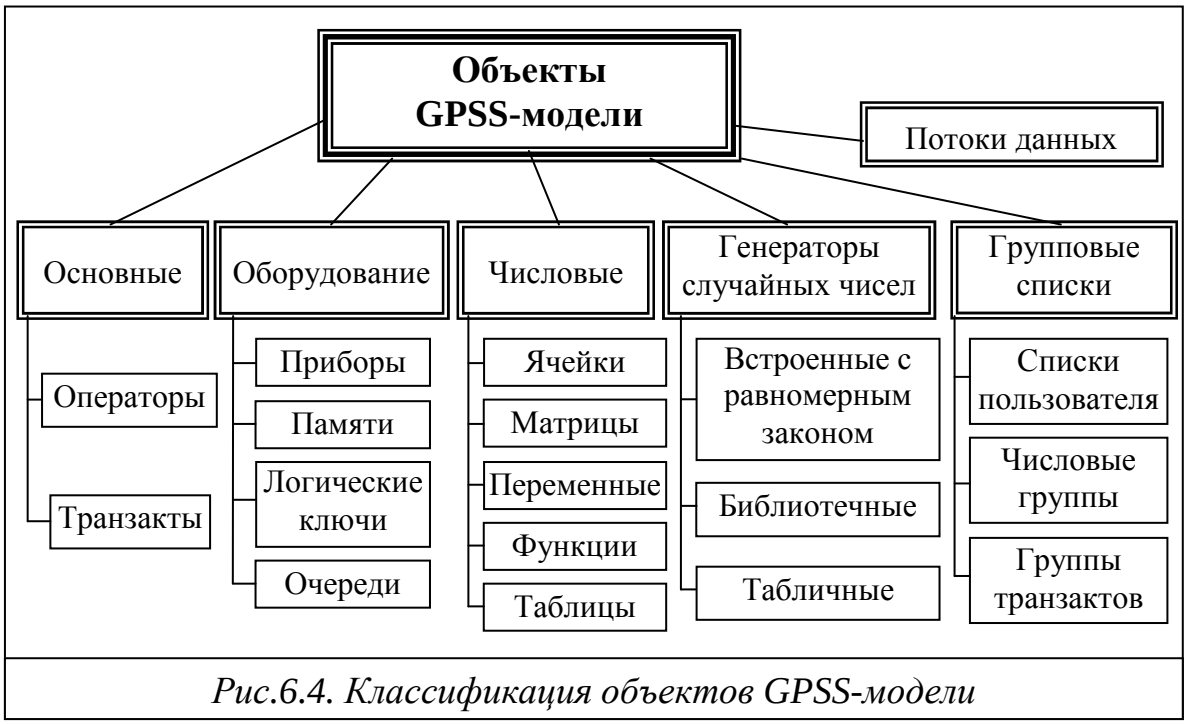


Рис.6.4. Классификация объектов GPSS-модели

Объектами *оборудования* являются:

- **приборы (одноканальные устройства)** – объекты, которые могут находиться в одном из двух состояний: свободном или занятом; при моделировании систем массового обслуживания используются для имитации процессов занятия и освобождения прибора, то есть для моделирования одноканальных СМО; занятие и освобождение прибора транзактом выполняется в GPSS-модели с помощью операторов SEIZE и RELEASE;
- **памяти (многоканальные устройства)** – объекты, состоящие из множества элементов, которые занимаются и освобождаются транзактами, при этом один транзакт может занять один или несколько элементов памяти, но не более чем её *ёмкость*; при моделировании систем массового обслуживания «память» используется для имитации процессов занятия и

освобождения приборов многоканальных СМО; ёмкость памяти задается в области описания GPSS-модели с помощью оператора STORAGE, а занятие и освобождение элементов «памяти» транзактом – с помощью операторов ENTER и LEAVE;

- **очереди** – объекты, используемые для накапливания транзактов, находящихся в состоянии ожидания какого-то события, например освобождения прибора или памяти; при моделировании систем массового обслуживания «очередь» используется для имитации процессов ожидания перед обслуживающими приборами; следует иметь в виду, что понятие «очередь» весьма относительное, поскольку в действительности транзакты, ожидающие освобождения прибора или памяти, заносятся в *«список задержки»* соответствующего прибора или памяти, при этом формирование списков задержки, то есть занесение в очередь и удаление из очереди, происходит автоматически, независимо от наличия операторов QUEUE и DEPART; последние используются только с целью сбора статистики по очередям путем фиксирования моментов поступления транзакта в очередь и удаления его из очереди;
- **логические ключи** – объекты, которые могут находиться только в двух состояниях: «установлен» или «сброшен»; установка, сброс или инвертирование ключа осуществляется с помощью оператора LOGIC.

К *числовым объектам* GPSS-модели относятся:

- **ячейки** – объекты для хранения величин, которым могут быть присвоены некоторые значения;
- **матрицы** – объекты для хранения массивов элементов размерности от 2 до 6;
- **переменные** – объекты для хранения величин, значения которых вычисляются на основе некоторого заданного выражения; переменные описываются с помощью операторов VARIABLE (арифметическая переменная), FVARIABLE (арифметическая переменная с плавающей точкой), BVARIABLE (булева переменная);
- **функции** – объекты, позволяющие вычислять значения в зависимости от некоторого аргумента; функции описываются с помощью оператора FUNCTION;
- **таблицы** – объекты, используемые для построения гистограммы плотности распределения случайной величины и представляющие собой набор чисел, отображающих частоту попадания значений случайной величины в тот или иной частотный диапазон (интервал); таблицы описываются с помощью оператора TABLE.

Генераторы случайных (точнее, псевдослучайных) чисел представляют собой объекты GPSS-модели, которые можно разделить на *три* группы:

- **встроенные** генераторы равномерно распределённых в интервале (0; 1) случайных чисел, основанные на мультипликативном конгруэнтном методе, с длиной периода 2 147 483 646; количество таких генераторов равно 999, причём номер генератора (от 1 до 999) определяет начальное

число для запуска генератора; при обращении к генератору с помощью системного числового атрибута (СЧА) RNj, где j – номер генератора, вырабатываются целочисленные случайные величины в интервале (0; 999);

- **библиотечные** генераторы случайных чисел с конкретными законами распределений, реализованные в виде встроенных библиотечных процедур количеством более 20;
- **табличные** генераторы случайных чисел с произвольными законами распределений, реализуемые пользователем в виде таблиц с помощью оператора описания FUNCTION.

Кроме перечисленных объектов при разработке больших сложных GPSS-моделей дополнительно могут использоваться:

- *групповые списки*, включающие в себя:
  - списки пользователя;
  - числовые группы;
  - группы транзактов,
- *потоки данных*.

Объекты в GPSS-модели могут формироваться автоматически, либо должны объявляться с использованием специальных команд – операторов описания. К объявляемым объектам относятся: памяти, переменные, матрицы, таблицы, функции, а также параметры транзактов.

6.3.4. Состав и структура GPSS-модели

GPSS-модель представляет собой программу, написанную на *языке GPSS* в виде последовательности *операторов*, описывающих логику работы моделируемой системы.

Операторы GPSS-модели делятся на две группы:

- GPSS-операторы;
- PLUS-операторы.

В свою очередь, GPSS-операторы делятся на *команды* и *блоки*.

Команды предназначены:

- для описания (определения) некоторых объектов, таких как памяти, переменные, функции, матрицы, таблицы; эти команды называются также *операторами описания*;
- для управления процессом моделирования (запуск, остановка и продолжение процесса моделирования, сброс статистики, завершение моделирования и т.п.); некоторые из этих команд могут находиться как в GPSS-модели, так и задаваться пользователем в процессе моделирования извне в качестве интерактивных операторов с использованием соответствующих пунктов меню GPSS World; эти команды называются также *операторами управления.*

Все команды делятся на:

- *срочные*, выполнение которых начинается в момент их появления;
- *несрочные*, которые заносятся в специальную очередь команд и последовательно выбираются для выполнения в процессе моделирования.

Операторы блоков или просто ***блоки*** представляют собой *исполняемые операторы* и реализуют в процессе моделирования некоторые действия, предписанные этими операторами. Название «блок» происходит от так называемых *блок-диаграмм*, которые используются для графического представления GPSS-моделей. **Блок-диаграмма** представляет собой набор стандартных блоков, связанных между собой в последовательности, определяемой логикой работы моделируемой системы. Каждому такому блоку в языке GPSS соответствует определенный оператор, реализующий некоторую конкретную функцию. При этом полагается, что реализация оператора связана с выполнением достаточно большой совокупности действий (блока). В GPSS World изображение этих блоков и перемещение транзактов в процессе моделирования можно увидеть в окне «BLOCK ENTITIES», доступ к которому выполняется из главного меню: **Window/Simulation Window/Blocks Window**.

Система имитационного моделирования GPSS, предназначенная для моделирования сложных систем со стохастическим характером функционирования, обладает эффективными встроенными средствами для описания параллельных процессов, протекающих в исследуемой системе. Эти средства позволяют выполнять описание каждого из параллельных процессов независимо друг от друга в виде отдельных программных модулей, что существенно упрощает процесс программирования и создания модели. В то же время, качество модели в значительной степени зависит от того, насколько достоверно и корректно отражены в модели связи между модулями, отображающие логику функционирования моделируемой системы.

Таким образом, укрупнено структуру GPSS-модели можно представить в виде множества модулей (рис.6.5), каждый из которых описывает один из протекающих в исследуемой системе параллельных процессов. Здесь же приведены некоторые операторы описания (команды) и исполняемые операторы (блоки), используемые в соответствующих модулях.

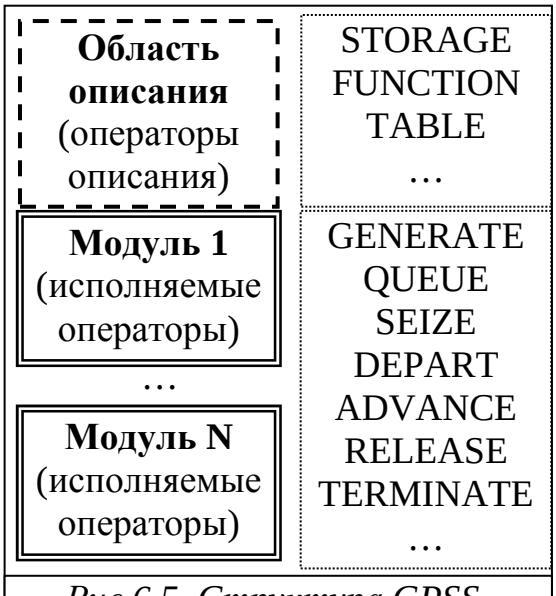


Рис.6.5. Структура GPSS-

Следует отметить, что выделение специальной области описания является желательным, но не обязательным. Такое выделение позволяет создать хорошо структурированную GPSS-модель, наглядную и легко понимаемую. В принципе, операторы описания могут быть в любом месте исполняемой области. При этом в GPSS World даже необязательно, чтобы оператор описания объекта находился до того, как соответствующий объект будет использован.

Оператор GPSS World, в общем случае, содержит 4 поля:

<Метка>	**<Операция>**	**<Операнды>**	**; <Комментарий>**

Поле **<Метка>** содержит *имя*, которое может быть присвоено оператору блока и оператору описания.

Каждый оператор занимает вполне определенное место в GPSS-модели. Для изменения естественного порядка выполнения процесса моделирования *любому исполняемому оператору* может быть присвоена *метка* в виде *имени*. В поле метки *оператора описания* указывается *имя* описываемого объекта (памяти, таблицы, функции и т.д.).

Поле **<Операция>** содержит зарезервированное слово GPSS World, определяющее функциональное назначение блока и задающее совокупность действий, которые должны быть выполнены.

Названия операторов (операции) в GPSS World обычно записываются прописными (но могут и строчными) буквами, не допускают сокращений и не могут использоваться в качестве переменных или имён объектов. Примеры операций: GENERATE (ГЕНЕРИРОВАТЬ), TERMINATE (ЗАВЕРШИТЬ), QUEUE (СТАТЬ В ОЧЕРЕДЬ), DEPART (ПОКИНУТЬ ОЧЕРЕДЬ), SEIZE (ЗАНЯТЬ), RELEASE (ОСВОБОДИТЬ), ENTER (ВОЙТИ), LEAVE (ВЫЙТИ), ADVANCE (ЗАДЕРЖАТЬ), PRIORITY (НАЗНАЧИТЬ ПРИОРИТЕТ), GATE (ВПУСТИТЬ), ASSIGN (НАЗНАЧИТЬ), TEST (ПРОВЕРИТЬ).

В поле **<Операнды>** задаются данные, необходимые для выполнения операции и представляющие собой параметры (операнды) оператора, разделяемые запятыми или пробелами, например: <**A,B,C,D**> или <**A B C D**>. При этом некоторые операнды являются *обязательными*, то есть должны быть всегда заданы, а другие – *необязательными*, то есть могут быть опущены при записи оператора. В последнем случае значения этих операндов определяются транслятором по умолчанию.

Между двумя соседними операндами может находиться только запятая или пробел: <**A,B**> или <**A B**>. Если между операндами **A** и **B** после запятой появится пробел: <**A, B**> или между ними будет находиться два пробела, то это равносильно двум запятым, и операнд **B** будет восприниматься транслятором как третий операнд, а значение второго операнда будет определяться по умолчанию.

Большинство операторов содержат один или два операнда.

В некоторых операторах в качестве операндов могут использоваться операторы отношения, задающие условия, выполнение которых проверяется в процессе выполнения операции.

Поле **«Комментарий»** располагается после операндов, от которых отделяется символом «точка с запятой».

GPSS-модель может содержать *комментарий*, который занимает всю строку. В этом случае признаком комментария служит символ «звездочка»

<*> или «точка с запятой» <;>, располагающийся в первой позиции строки, что говорит о наличии в этой строке только комментария. Если комментарий располагается после оператора в той же самой строке, то в качестве разделителя (признака комментария) используется только символ «точка с запятой». В поле **«Комментарий»** могут использоваться как латинские, так и русские буквы, а также любые другие символы.

Все операторы, кроме оператора описания FUNCTION, записываются в одну строку и могут содержать до 250 символов, включая комментарий.

6.4. Процесс моделирования в среде GPSS World

> «Если отладка - процесс удаления ошибок, то программирование должно быть процессом их внесения» ***(Э.Дейкстра)***

6.4.1. Запуск процесса моделирования

В результате трансляции (компиляции) GPSS-модели с использованием пунктов меню **Command / Create Simulation** создаётся исполняемый объект, реализующий **процесс моделирования**.

Для запуска процесса моделирования используется команда START, которая может находиться в GPSS-модели в качестве последнего оператора или может быть задана интерактивно после трансляции.

Если команда START находится в GPSS-модели, процесс моделирования запускается сразу же после трансляции автоматически. В противном случае, запуск процесса моделирования осуществляется путем задания команды START с использованием пунктов меню **Command / Start** системы имитационного моделирования GPSS World.

6.4.2. Транзакты

Реализация процесса моделирования заключается в перемещении в модели некоторых подвижных объектов, называемых *транзактами*. Транзакты последовательно перемещаются от блока к блоку в заданной алгоритмом моделирования последовательности.

Транзакты создаются и уничтожаются в модели с помощью операторов (блоков): GENERATE и TERMINATE.

В начале моделирования в GPSS-модели нет ни одного транзакта. В процессе моделирования транзакты формируются в модели в определенные моменты времени в соответствии с условиями, заданными с помощью блока GENERATE. Транзакты покидают модель (уничтожаются), попадая в блок TERMINATE. В общем случае, *в модели может находиться множество транзактов*, однако *в один и тот же момент времени продвигается только один транзакт*. Транзакт, попадая в определенный блок, вызывает к исполнению совокупность действий, предписанных соответствующим оператором, и затем пытается войти в следующий по

порядку блок. Такое продвижение транзакта продолжается до тех пор, пока не произойдет одно из следующих событий:

- транзакт входит в блок, функцией которого является задержка транзакта на некоторое заданное время (блок ADVANCE);
- транзакт пытается войти в блок, который "отказывается" принять его до тех пор, пока в модели не изменятся некоторые условия (например, блоки SEIZE, ENTER);
- транзакт входит в блок, функцией которого является удаление транзакта из модели (блок TERMINATE).

При возникновении одного их перечисленных событий транзакт прекращает движение и начинается перемещение в модели *другого* транзакта, то есть моделирование продолжается. Таким образом, моделирование заключается в перемещении транзактов между блоками GPSS-модели и выполнении соответствующих действий.

Для изменения последовательности движения транзактов используются условные и безусловные операторы, такие как TRANSFER, TEST, SELECT.

Транзакт, продвигаемый в модели в данный момент времени, называется ***активным***.

Интервал времени, в течение которого транзакт находится в модели, называется ***резидентным временем транзакта***.

Интервал времени, в течение которого транзакт проходит от одной произвольно выбранной точки модели до другой точки, называется ***транзитным временем*** перехода между двумя этими точками.

Каждому транзакту в модели присваивается порядковый номер, начиная с *единицы*.

6.4.3. Модельное время

Работа реальных систем протекает во времени, для отображения которого в GPSS-модели используется ***таймер модельного времени***. Изменение модельного времени происходит путем его продвижения *до ближайшего события*, связанного с изменением состояния моделируемой системы. При моделировании СМО и СеМО такими событиями являются:

- поступление заявок в систему;
- завершение обслуживания заявок в узле СеМО (обслуживающем приборе).

Таким образом, моделирование заключается в определении ближайшего момента наступления каждого события.

Таймер модельного времени корректируется *автоматически* в соответствии с логикой, предписанной моделью.

Таймер GPSS World может принимать любые значения. Единица времени (секунды, минуты, часы или их доли) для таймера задается разработчиком модели. Так как единица времени не сообщается транслятору, то все данные, связанные со временем, должны быть

выражены разработчиком через эту выбранную единицу.

Рассмотрим более подробно механизм изменения таймера модельного времени и логику процесса моделирования на примере моделирования системы массового обслуживания с неоднородным потоком заявок. Для формирования неоднородного потока заявок GPSS-модель будет содержать несколько операторов GENERATE по числу классов заявок.

В начале моделирования значение таймера модельного времени устанавливается в 0. Для всех классов заявок, поступающих в моделируемую систему, в каждом из блоков GENERATE определяется по одному ближайшему моменту появления транзакта, что соответствует моменту поступления очередной заявки данного класса. Очевидно, что число таких моментов будет равно количеству классов заявок. Среди всех этих моментов определяется момент с наименьшим значением, то есть момент, соответствующий ближайшему событию, и значение таймера модельного времени устанавливается равным значению (продвигается до) этого момента. Такое изменение значения таймера модельного времени приводит к тому, что соответствующий транзакт с моментом поступления, равным значению таймера, начинает движение в модели от блока GENERATE к следующему по порядку блоку. Движение транзакта в модели продолжается до тех пор, пока он не попадет в блок, функцией которого является задержка на некоторое заданное время, или в блок, который "отказывается" принять его до тех пор, пока в модели не изменятся некоторые условия. В последнем случае транзакт остаётся в предыдущем блоке. Если транзакт входит в блок, функцией которого является удаление транзакта из модели, то этот транзакт уничтожается.

Если в модели имеется ещё транзакт с таким же моментом формирования, то начинается его продвижение, причем значение таймера модельного времени не изменяется. Изменение таймера модельного времени происходит только в том случае, если в модели больше нет ни одного транзакта с таким же моментом формирования.

6.4.4. Списки

Описанный принцип продвижения транзактов в GPSS-модели реализуется с помощью так называемых *списков* или *цепей* (Chain).

В каждый момент модельного времени все транзакты, находящиеся в модели, соотносятся с одним из списков. В зависимости от принадлежности транзакта тому или иному списку, он может продвигаться в модели, быть готовым к дальнейшему продвижению, либо ожидать наступления заданного момента модельного времени или выполнения некоторого условия.

Такими списками в GPSS-модели являются:

- **список текущих событий (СТС)** – содержит транзакты, которые могут продвигаться в модели в текущий момент модельного времени;
- **список будущих событий (СБС)** – содержит транзакты, ожидаю-

щие наступления более позднего момента модельного времени;

- **списки повторных попыток (СПП)** – содержат транзакты, не удовлетворяющие условиям входа в блок, причем каждый объект GPSS-модели имеет свой СПП;
- **списки прибора** (одноканального устройства), включающие:
  - *список отложенных прерываний*, в котором находятся транзакты, ожидающие занятия устройства по приоритету с возможностью вытеснения транзакта, ранее занимавшего устройство;
  - *список прерываний*, в котором находятся транзакты, вытесненные из данного устройства, то есть обслуживание которых было прервано более высокоприоритетным транзактом;
  - *список задержки*, в котором находятся транзакты, ожидающие занятия устройства;
  - *список повторных попыток*;
- **списки памяти** (многоканального устройства), включающие:
  - *список задержки*,
  - *список повторных попыток*;
- **списки пользователя**, используемые для построения моделей с разнообразными функциональными возможностями, в частности, для организации моделирования различных алгоритмов формирования очередей (дисциплин буферизации и дисциплин обслуживания заявок в моделях массового обслуживания).

В любой модели всегда формируется *один список текущих* и *один список будущих событий*. Остальные списки формируются по мере необходимости.

В каждый момент модельного времени последовательно один за другим продвигаются только те транзакты, которые находятся в СТС, причем только один транзакт, который продвигается в данный момент реального времени, является активным. Продвижение каждого транзакта осуществляется до тех пор, пока это возможно. Например, если транзакт попадает в блок ADVANCE, функцией которого является задержка на некоторое время, то он переводится в СБС. Транзакт будет находиться в СБС до тех пор, пока модельное время не станет равным моменту, когда он может покинуть блок ADVANCE. В этом случае транзакт будет переведён в СТС.

Транзакты, расположенные в СТС с учётом их приоритетов, выбираются последовательно один за другим. Когда в СТС не остаётся транзактов, которые могут быть продвинуты в текущий момент модельного времени, происходит изменение модельного времени, которое продвигается до ближайшего запланированного момента времени для транзакта, находящегося первым в СБС. Этот, а также все другие транзакты, движение которых может быть возобновлено в тот же момент модельного времени, переносятся из СБС в СТС, где размещаются в порядке убывания приоритетов.

Каждый транзакт может иметь множество параметров, называемых

атрибутами транзакта, которые сопровождают его в течение «жизни» в модели. К ним, в частности, относятся:

- *параметры*, закрепляемые пользователем за каждым транзактом, число которых не ограничено; идентификатором параметра может служить его *номер* (целое положительное число) или *имя*; параметры транзакта должны быть определены, до того как они будут востребованы;
- *приоритет* – преимущественное право на использование общего ресурса, причём более высокому приоритету соответствует большее значение; транзакты с одинаковым приоритетом обычно выбираются в порядке поступления;
- *время входа транзакта в систему* – значение абсолютного времени в момент первого входа транзакта в модель или в блок MARK без операнда A;
- *текущий блок* – номер блока, в котором находится транзакт;
- *следующий блок* – номер следующего блока, в который должен перейти данный транзакт;
- *список*, в котором находится транзакт в некоторый момент времени:
  - ACTIVE – транзакт находится в СТС и является активным;
  - SUSPENDED – транзакт находится в СБС или в СТС и ожидает возможности стать активным;
  - PASSIVE – транзакт находится в состоянии ожидания: в списке пользователя, списке задержки или списке отложенных прерываний;
  - PREEMPTED – обслуживание транзакта в устройстве прервано, и он находится в списке прерываний;
  - TERMINATED – транзакт удаляется из модели и больше не участвует в процессе моделирования.

6.4.5. Завершение моделирования

Достоверность результатов моделирования в значительной степени определяется продолжительностью процесса моделирования, которое устанавливается разработчиком модели или пользователем.

В GPSS World завершение процесса моделирования может быть реализовано:

- принудительно с помощью срочной команды HALT, задаваемой из подменю COMMAND; использование этой команды позволяет принудительно остановить процесс моделирования в любой момент времени;
- по некоторому условию, задаваемому командой STOP, которая может находиться в GPSS-модели;
- по достижению содержимого «счётчика завершений» значения меньше или равного нулю.

Последний способ, используемый наиболее часто при моделировании систем и сетей массового обслуживания, рассмотрим более подробно.

Начальное значение «счётчика завершений» устанавливается с помощью команды START, которая запускает процесс моделирования. В процессе моделирования всякий раз при попадании транзакта в какой-либо блок TERMINATE (таких блоков в модели может быть несколько) из содержимого «счётчика завершений» вычитается значение, указанное в качестве параметра в соответствующем блоке TERMINATE. При достижении нулевого или отрицательного значения «счётчика завершений» процесс моделирования останавливается. Отметим, что при отсутствии параметра в блоке TERMINATE содержимое «счётчика завершений» не изменяется. Если во всех блоках TERMINATE отсутствуют параметры или их значения равны нулю, содержимое «счётчика завершений» не будет изменяться, и процесс моделирования (при отсутствии в модели команды STOP) будет длиться до тех пор, пока не будет введена команда HALT.

В одной и той же модели может быть предусмотрено несколько способов завершения моделирования. Например, в модели может находиться несколько команд STOP, задающих разные условия, и предусмотрено завершение моделирования по достижению содержимого «счётчика завершений» значения равного нулю. В этом случае завершение процесса моделирования происходит при достижении ближайшего по времени наступления условия.

По завершению процесса моделирования формируется и выводится на экран стандартный отчет, содержащий основные результаты моделирования, в том числе характеристики основных объектов – очередей, приборов, многоканальных устройств и т.д. Состав включаемых в отчёт результатов моделирования может быть изменён на вкладке **Reports** меню **EDIT/SETTINGS**.

Кроме отчета, содержащего числовые значения характеристик моделируемых систем, GPSS World предоставляет возможность получения результатов в графическом виде, в частности, путём формирования гистограмм плотностей распределений вероятностных характеристик. Для этого в GPSS-модели используются команды TABLE и QTABLE.

Более подробно результаты моделирования, представленные в отчете и в виде гистограмм, рассматриваются ниже при описании GPSS-моделей массового обслуживания.

6.4.6. Системные числовые атрибуты

Числовые и строковые переменные, используемые в процессе моделирования, называются ***атрибутами***. Атрибуты могут использоваться в операндах операторов GPSS и в выражениях.

Числовые атрибуты, *автоматически поддерживаемые в GPSS* и доступные в течение процесса моделирования, называются ***системными числовыми атрибутами (СЧА)*** (System Numerical Attributes – SNA). Их значения могут изменяться в процессе моделирования и доступны пользователю за счет использования специальных наименований этих атрибутов.

В GPSS используются СЧА трёх типов:

- СЧА *объектов*, описывающие состояние таких объектов GPSS-модели как приборы (одноканальные устройства), памяти (многоканальные устройства), очереди, таблицы и др.;
- СЧА *системы*, описывающие состояние модели в целом;
- СЧА *транзактов*, описывающие их свойства и параметры.

Имя СЧА объектов состоит из двух частей:

- первая часть указывает *групповое имя*, идентифицирующее *тип объекта* (прибор, многоканальное устройство, очередь, таблица) и *тип информации* (количество входов в объект, загрузка объекта, среднее время занятия объекта и т.д.);
- вторая часть (число или имя) идентифицирует *конкретного члена группы*.

Если конкретный член группы задан в виде имени, то вторая часть имени СЧА отделяется от первой (группового имени) символом $. Таким образом, имя СЧА может иметь вид:

<Групповое имя><*Число*> или **<Групповое имя>$**<*Имя*>.

Например: **F***5*, **QT***23*, **FR$***Pribor*, **SR$***New_System*.

Групповые имена (наименования) и значения основных ***СЧА объектов***, таких как приборы, многоканальные устройства, очереди и таблицы приведены ниже в табл.1-3.

Таблица 1

СЧА приборов

Групповое имя	Значение
F	1, если прибор занят; 0, если свободен
FC	Число занятий прибора транзактами
FR	Загрузка прибора, выраженная в долях тысячи
FT	Среднее время занятия прибора транзактом

Таблица 2

СЧА многоканальных устройств

Групповое имя	Значение
R	Количество незанятых приборов (каналов)
S	Количество занятых приборов (каналов)
SA	Среднее количество занятых приборов
SC	Счетчик числа входов в многоканальное устройство (при каждом выполнении блока ENTER значение счетчика увеличивается на величину операнда B блока)
SM	Максимальное количество занятых приборов (максимальное значение Sj или S$ имя)
SR	Загрузка многоканального устройства, выраженная в долях тысячи
ST	Среднее время нахождения транзакта в устройстве.

Таблица 3

СЧА очередей

Групповое имя	Значение
Q	Текущее значение длины очереди
QA	Среднее значение длины очереди
QC	Количество входов в очередь (увеличивается на величину операнда B блока QUEUE)
QM	Максимальное значение длины очереди
QT	Среднее время пребывания в очереди с учетом нулевых входов
QX	Среднее время пребывания в очереди для входов без учета нулевых входов
QZ	Количество нулевых входов в очередь, при которых время ожидания было равно нулю

Кроме того, могут использоваться следующие ***СЧА объектов***:

- СЧА таблиц:
  - **TB$**<*Имя*> – Среднее значение элементов таблицы
  - **TC $**<*Имя*> – Количество учтенных в таблице элементов
  - **TD $**<*Имя*> – Стандартное отклонение элементов таблицы
- СЧА функции:
  - **FN$**<*Имя*> – результат вычисления функции;
- СЧА переменной:
  - **V$**<*Имя*> – результат вычисления переменной.

Примеры СЧА объектов:

FR3 – возвращает значение загрузки прибора с номером **3**;

FT$Auto_Master – возвращает среднее время занятия транзактом прибора с именем **Auto_Master**.

S22 – возвращает количество занятых приборов в многоканальном устройстве с номером **22**;

SM$Kassa_2m – возвращает максимальное количество занятых приборов в многоканальном устройстве с именем **Kassa_2m**.

V$F_5 – возвращает значение переменной **F_5**.

К ***СЧА системы*** относятся такие глобальные переменные как:

- **AC1** – значение абсолютного модельного времени (с момента начала моделирования или последней команды CLEAR);
- **C1** – значение относительного модельного времени (с момента последней команды RESET);
- **TG1** – текущее значение счетчика завершения;
- **Z1** – свободная оперативная память ЭВМ в байтах.

К ***СЧА транзактов*** относятся:

- **MP**<*Число*> или **MP$**<*Имя*> – транзитное время транзакта

(абсолютное модельное время минус значение, содержащееся в параметре <*Число*> или <*Имя*>);

- **P**<*Число*> или **P$**<*Имя*> – значение параметра <*Число*> или <*Имя*>;
- **PR** – приоритет транзакта;
- **M1** – резидентное время транзакта (абсолютное модельное время минус время появления транзакта в модели);
- **XN1**– номер активного транзакта.

6.4.7. Встроенные вероятностные распределения

Встроенная библиотека процедур GPSS World содержит более 20 вероятностных распределений, в том числе:

- равномерное (Uniform);
- экспоненциальное (Exponential);
- геометрическое (Geometric);
- Пуассона (Poisson);
- Бета (Beta);
- Гамма (Gamma);
- биномиальное (Binomial);
- дискретно-равномерное (Discrete Uniform);
- треугольное (Triangular);
- нормальное (Normal);
- Парето (Pareto); …

Для обращения к вероятностному распределению необходимо указать имя библиотечной процедуры и её параметры, заключённые в круглые скобки и отделённые друг от друга запятой:

<Имя процедуры>(G,A,B, …)

Здесь G – номер генератора равномерно распределённых случайных чисел (от 1 до 999) – используется в качестве аргумента для формирования случайных величин с заданным законом распределения. Остальные параметры **A, B, …,** количество которых для разных распределений составляет от 1 до 4, задают непосредственно параметры вероятностного распределения.

Ниже рассматриваются только некоторые из перечисленных распределений, наиболее часто используемые в моделях массового обслуживания.

1. *Равномерное распределение*:

UNIFORM (G,Min,Max),

где Min и Max – соответственно минимальное и максимальное значение равномерно распределённой случайной величины.

2. *Экспоненциальное распределение*:

EXPONENTIAL (G,Min,Mean),

где Mean – математическое ожидание (среднее значение) случайной

величины, распределённой по экспоненциальному закону;
Min – смещение распределения относительно нуля (минимальное значение случайной величины).

3. *Распределение Пуассона*:

POISSON (G,Mean),

где Mean – математическое ожидание (среднее значение) случайной величины.

4. *Геометрическое распределение*:

GEOMETRIC (G,P),

где **P** – параметр распределения, принимающий значения в интервале (0;1).

Библиотечные процедуры вероятностных распределений могут использоваться в выражениях, в том числе арифметических, а также в качестве операнда A в операторах GENERATE и ADVANCE. В последнем случае они рассматриваются как выражения языка PLUS и должны быть заключены в круглые скобки.

6.5. Операторы блоков GPSS World

6.5.1. Общие сведения

В GPSS World используются 53 оператора блоков. Среди операторов блоков имеются так называемые взаимодополняющие операторы, представляющие собой пару операторов, каждый из которых является зеркальным отображением другого оператора, означающим, что совокупность действий, реализуемых одним оператором, является противоположной по отношению к совокупности действий, реализуемых другим оператором. Примерами взаимодополняющих операторов могут служить операторы GENERATE и TERMINATE, SEIZE и RELEASE, QUEUE и DEPART, ENTER и LEAVE, PREEMPT и RETURN.

Для построения имитационных моделей *простейших* систем и сетей массового обслуживания в среде GPSS World оказывается достаточным использование примерно половины из всех операторов блоков, которые по функциональному назначению могут быть разбиты на следующие группы:

1. Операторы генерирования, задержки и удаления транзактов: GENERATE, ADVANCE, TERMINATE.
2. Операторы одноканальных устройств (приборов): SEIZE, RELEASE.
3. Операторы многоканальных устройств (памятей): ENTER, LEAVE.
5. Операторы очередей: QUEUE, DEPART.
4. Условные операторы: TEST, TRANSFER, GATE.
5. Операторы приоритетного обслуживания: PRIORITY, PREEMPT, RETURN.
6. Оператор логических ключей: LOGIC.
7. Прочие операторы: ASSIGN, MARK, TABULATE.

Операторы могут быть без операндов или содержать от 1 до 7 операндов, некоторые из которых могут быть необязательными, то есть могут отсутствовать. В последнем случае значения необязательных операндов принимаются по умолчанию. Если после отсутствующего операнда в операторе имеются другие операнды, то признаком отсутствия необязательного операнда служит лишняя запятая. Например, следующая запись в поле операций: **<„C>** означает, что операнды **A** и **B** не используются, и их значения принимаются по умолчанию.

Ниже представлены краткие описания операторов, наиболее часто используемых при построении имитационных моделей массового обслуживания. При изображении структуры оператора *необязательные операнды заключены в квадратные скобки.*

Отсутствие обязательных операндов приводит к ошибке.

6.5.2. GENERATE (ГЕНЕРИРОВАТЬ)

Назначение оператора: генерирование транзактов в соответствии с заданным правилом формирования интервалов между транзактами.

Формат оператора:

GENERATE [A],[B],[C],[D],[E]

Значения операндов:

A – средний интервал времени между генерируемыми транзактами или вероятностное распределение интервала из встроенной библиотеки процедур, заключённое в круглые скобки;; [по умолчанию – ноль];

B – величина полуинтервала равномерно распределенного интервала или модификатор таблично заданной функции; [по умолчанию – ноль];

C – смещение – момент формирования первого транзакта; [по умолчанию – ноль];

D – ограничитель – число генерируемых данным оператором транзактов; [по умолчанию – не ограничено];

E – уровень приоритета от 0 до 127 (чем больше номер, тем выше приоритет); [по умолчанию – ноль].

Примечание. Несмотря на то, что операнды A и D – необязательные операнды, в операторе **GENERATE** обязательно должен использоваться один из них: либо операнд **A** , либо операнд **D**.

Примеры:

GENERATE 25; интервал времени между генерируемыми транзактами – величина детерминированная равная 25, количество генериру-емых транзактов не ограничено.

GENERATE „„25;операнды **A**, **B** и **C** не используются, и их значения равны нулю по умолчанию; это означает, что в нулевой момент модельного времени будут сгенерированы ровно 25 транзактов.

GENERATE 25,10; интервал времени между транзактами – величина случайная, равномерно распределенная в интервале (25±10), т.е. от 15 до 35.

GENERATE 25,FN$Erlang; интервал времени между транзактами – величина случайная, распределенная по закону, заданному в виде табличной функции Erlang.

GENERATE 25,10,100,250,5; интервал времени между транзактами – равномерно распределенная величина в интервале от 15 до 35; момент формирования первого транзакта равен 100 единицам модельного времени; всего за время моделирования этим оператором будет сгенерировано 250 транзактов, после чего формирование транзактов прекратится; всем сгенерированным транзактам будет присвоен приоритет, равный 5.

GENERATE (Exponential(1,0,50)); интервал времени между транзактами – величина случайная, распределенная по экспоненциальному закону со средним значением 50.

Следует обратить внимание, что в последнем примере имя библиотечной процедуры с параметрами **Exponential(1,0,50)** заключается в круглые скобки. Параметры процедуры **Exponential** имеют следующий смысл: первый параметр – номер встроенного генератора равномерно распределённых в интервале (0; 1) случайных чисел (может иметь значения от 1 до 999); второй и третий параметры – соответственно смещение (минимальное значение) и среднее значение (математическое ожидание) случайной величины, распределённой по экспоненциальному закону.

6.5.3. TERMINATE (ЗАВЕРШИТЬ)

Назначение оператора: удаление транзактов из модели.

Формат оператора:

TERMINATE [A]

Значения операндов:

A – указатель уменьшения счетчика завершений (целое положительное число); [по умолчанию – ноль].

Примеры:

TERMINATE 1; транзакт, поступивший в данный блок, удаляется из модели, и счетчик завершения процесса моделирования, начальное значение которого устанавливается командой START, уменьшается на 1.

TERMINATE; транзакт удаляется из модели, при этом значение счетчика завершения процесса моделирования не изменяется.

6.5.4. ADVANCE (ЗАДЕРЖАТЬ)

Назначение оператора: задержка транзакта на заданное время.

Формат оператора:

ADVANCE [A],[B]

Значения операндов:

A – среднее время задержки или вероятностное распределение из встроенной библиотеки процедур, заключённое в круглые скобки; [по умолчанию – ноль];

B – величина полуинтервала равномерно распределенного интервала задержки или модификатор таблично заданной функции; [по умолчанию – ноль].

Примеры:

ADVANCE 50; поступивший транзакт задерживается в данном блоке на 50 единиц времени.

ADVANCE 50,10; время задержки транзакта – величина случайная, равномерно распределенная в интервале от 40 до 60 (50 ± 10).

ADVANCE 50,FN$Erl_1; время задержки транзакта – величина случайная, распределенная по закону, заданному в виде табличной функции **Erl_1,** со средним значением 50**.**

ADVANCE (Exponential(33,10,50)); время задержки – величина случайная, распределенная по экспоненциальному закону (из встроенной библиотеки процедур) со средним значением 50; номер встроенного генератора равномерно распределённых случайных чисел равен 33; смещение равно 10, то есть случайная величина, распределённая по экспоненциальному закону, принимает значения, начиная от 10.

6.5.5. SEIZE (ЗАНЯТЬ)

Назначение оператора: занятие транзактом прибора.

Формат оператора:

SEIZE A

Значения операндов:

A – идентификатор (число или имя) занимаемого прибора.

Примеры:

SEIZE 4; транзакт пытается занять прибор с номером 4; если прибор занят другим транзактом, то поступивший транзакт помещается в список задержки этого прибора, где находится до момента освобождения прибора, после чего этот транзакт занимает освободившийся прибор и продолжает свое движение к следующему блоку.

SEIZE Pribor_Disk; транзакт пытается занять прибор с именем Pribor_Disk; далее по аналогии с предыдущим примером.

6.5.6. RELEASE (ОСВОБОДИТЬ)

Назначение оператора: удаление транзакта из прибора (освобождение прибора).

Формат оператора:

RELEASE A

Значения операндов:

A – идентификатор (число или имя) освобождаемого прибора.

Примеры:

RELEASE 4; транзакт освобождает прибор с номером 4.

RELEASE Pribor_Disk; транзакт освобождает прибор с именем Pribor_Disk.

6.5.7. QUEUE (СТАТЬ В ОЧЕРЕДЬ)

Назначение оператора: занесение транзакта в очередь (точнее – регистрация статистики очереди, связанная с фиксацией момента поступления транзакта в очередь и увеличением ее длины).

Формат оператора:

QUEUE A,[B]

Значения операндов:

A – идентификатор (число или имя) очереди;

B – количество элементов, на которое должна увеличиться длина очереди; [по умолчанию – один].

Примеры:

QUEUE 3; присоединение транзакта к очереди с номером 3 и увеличение ее длины на 1 (по умолчанию).

QUEUE Jeck,5;присоединение транзакта к очереди с именем Jeck и увеличение ее длины на 5.

6.5.8. DEPART (ПОКИНУТЬ ОЧЕРЕДЬ)

Назначение оператора: удаление транзакта из очереди (точнее – регистрация статистики очереди, связанная с уменьшением ее длины и фиксацией момента удаления транзакта из очереди с целью определения времени ожидания).

Формат оператора:

DEPART A,[B]

Значения операндов:

A – идентификатор (число или имя) очереди;

B – количество элементов, на которое должна уменьшиться длина очереди; [по умолчанию – один].

Примеры:

DEPART 3; удаление транзакта из очереди с номером 3 и уменьшение ее длины на 1 (по умолчанию).

DEPART Jeck,5;удаление транзакта из очереди с именем Jeck и уменьшение ее длины на 5.

6.5.9. ENTER (ВОЙТИ)

Назначение оператора: вход транзакта в многоканальное устройство.

Формат оператора:

ENTER A,[B]

Значения операндов:

A – идентификатор (число или имя) многоканального устройства;

B – количество занимаемых приборов многоканального устройства; [по умолчанию – один].

Примеры:

ENTER 5; транзакт поступает в многоканальное устройство с

номером 3 и занимает один прибор (по умолчанию).

ENTER MANY,4; транзакт, поступая в многоканальное устройство с именем MANY, занимает 4 прибора.

6.5.10. LEAVE (ВЫЙТИ)

Назначение оператора: удаление транзакта из многоканального устройства.

Формат оператора:

LEAVE A,[B]

Значения операндов:

A – идентификатор (число или имя) многоканального устройства;

B – количество освобождаемых приборов многоканального устройства; [по умолчанию – один].

Примеры:

LEAVE 5; транзакт покидает многоканальное устройство с номером 3 и освобождает 1 прибор (по умолчанию).

LEAVE MANY,4; транзакт, покидая многоканальное устройство с именем MANY, освобождает 4 прибора.

6.5.11. TEST (ПРОВЕРИТЬ)

Назначение оператора: поверка значения (обычно СЧА) и передача активного транзакта в блок, отличный от последующего, если указанное условие не выполняется.

Формат оператора:

TEST X A,B,[C],

Значения операндов:

A – проверяемое значение;

B – контрольное значение;

C – имя (метка) блока назначения **C**; [по умолчанию – Режим отказа];

X – оператор отношения, определяющий условие проверки операнда **A** с операндом **B**:

Значения X	Интерпретация в смысле блока TEST
G	**A** больше **B**?
GE	**A** больше или равно **B**?
E	**A** равно **B**?
NE	**A** не равно **B**?
LE	**A** меньше или равно **B**?
L	**A** меньше **B**?

Блок **TEST** может функционировать в двух режимах:

- в режиме альтернативного выхода (если задан операнд **C**);
- в режиме отказа (если операнд **C** не задан).

Когда транзакт пытается войти в блок **TEST** в режиме альтернативного выхода и проверяемое условие не выполняется, транзакт

передается блоку, указанному в операнде **C**.

Когда транзакт пытается войти в блок **TEST** в режиме отказа (при отсутствии операнда **C**) и заданное условие не выполняется, транзакт блокируется до тех пор, пока условие не будет выполнено.

Примеры:

TEST LE Q1,5,Otk_1; если проверяемое условие «длина очереди 1 меньше или равна 5?» выполняется, то активный транзакт передается следующему оператору, в противном случае он направляется к оператору с меткой Otk_1.

TEST G Q1,5; если проверяемое условие «длина очереди 1 больше 5?» выполняется, то активный транзакт передается следующему оператору, в противном случае он блокируется до тех пор, пока условие не будет выполнено.

6.5.12. TRANSFER (ПЕРЕДАТЬ)

Назначение оператора: передача транзакта в блок, отличный от последующего.

Режимы использования оператора TRANSFER:

1) режим безусловной передачи;
2) режим статистической передачи;
3) режим BOTH (ОБА);
4) режим ALL (ВСЕ);
5) режим PICK (выборочный);
6) режим FN (функциональный);
7) режим P (параметрический);
8) режим SBR (подпрограммный);
9) режим SIM (одновременный).

Далее рассматриваются только два первых режима, используемые ниже при построении GPSS-моделей.

1. Режим безусловной передачи

Назначение оператора: безусловная передача транзакта в блок, отличный от последующего.

Формат оператора:

TRANSFER ,B

Значения операндов:

A – НЕ используется, что является признаком режима безусловной передачи;

B – имя блока, к которому направляется активный транзакт.

Пример:

TRANSFER ,UZEL_3; всякий раз активный транзакт будет направляться к блоку с меткой **UZEL_3.**

2. Режим статистической передачи

Назначение оператора: передача транзакта в один из блоков случайным образом.

Формат оператора:

TRANSFER A,[B],C

Значения операндов:

A – частота (вероятность) передачи транзакта в блок **C**;

B – имя блока B; [по умолчанию – Следующий по порядку блок];

C – имя блока C.

Примечание. Частота (вероятность) передачи транзакта в операнде **A** может быть указана двумя способами:

- в виде вероятности – дробного числа с десятичной точкой, принимающего значения строго меньше 1;
- в виде целого положительного числа, принимающего значения от 0 до 1000 и интерпретируемого как доля от тысячи.

Заметим, что значение операнда **A**, равное 1, будет соответствовать вероятности 0,001, а не 1, поскольку транслятор воспринимает любое целое число как долю от 1000.

Следует также отметить, что транслятор не выдаст ошибку, если операнд A будет задан в виде числа с десятичной точкой, имеющего значение больше 1. В этом случае транслятор выделяет целую часть числа и интерпретирует её как долю от тысячи.

Примеры:

TRANSFER 250,UZEL_2,UZEL_3; активный транзакт с вероятностью 250/1000 = 0,25 будет направляться к блоку с меткой **UZEL_3** и с вероятностью 0,75 – к блоку с меткой **UZEL_2** .

TRANSFER 25,UZEL_2,UZEL_3; активный транзакт с вероятностью 25/1000 = 0,025 будет направляться к блоку с меткой **UZEL_3** и с вероятностью 0,975 – к блоку с меткой **UZEL_2** .

TRANSFER .95,,BOX7; активный транзакт с вероятностью 0,95 будет направляться к блоку с меткой **BOX7** и с вероятностью 0,05 – к следующему по порядку блоку (по умолчанию).

6.5.13. PRIORITY (НАЗНАЧИТЬ ПРИОРИТЕТ)

Назначение оператора: изменение уровня приоритета активного транзакта в процессе моделирования.

Формат оператора:

PRIORITY A,[B]

Значения операндов:

A – уровень приоритета, присваиваемый активному транзакту;

B – может принимать только одно значение: **BU** (задает специальный режим, при котором активный транзакт помещается в список текущих событий позади транзактов с таким же приоритетом); [по умолчанию – Транзакт помещается перед транзактами с таким же приоритетом].

6.5.14. PREEMPT (ЗАХВАТИТЬ)

Назначение оператора: захват прибора вновь прибывшим

транзактом.

Формат оператора:

PREEMPT A,[B],[C],[D],[E]

Значения операндов:

A – идентификатор (число или имя) прибора, подлежащего захвату;

B – определяет условие, при котором разрешён захват прибора: **PR** – *приоритетный режим*: захват разрешён, если активный транзакт имеет более высокий приоритет, чем обслуживаемый транзакт; [по умолчанию – *Режим прерывания*: захват разрешён, если обслуживаемый транзакт не является захватчиком];

C – метка блока, в который направляется транзакт, вытесненный из прибора более высокоприоритетным транзактом; [по умолчанию – Транзакт помещается в СБС];

D – номер параметра вытесненного транзакта, в который записывается оставшееся время обслуживания в приборе, если транзакт направляется к блоку **C**; используется совместно с операндом **C**;

E – может принимать только одно значение: **RE**, означающий *режим удаления*: вытесненный транзакт удаляется из состязания за прибор; [по умолчанию – вытесненный транзакт не удаляется из состязания за прибор].

Примечание. Следует обратить внимание, что приоритетный захват возможен только для *прибора*, но невозможен для многоканального устройства.

6.5.15. RETURN (ВЕРНУТЬ)

Назначение оператора: освобождение прибора активным транзактом и выбор нового транзакта.

Формат оператора:

RETURN A

Значения операнда:

A – идентификатор (число или имя) прибора, подлежащего освобождению.

Примечание. В прибор новый транзакт выбирается из списков прибора в строго определённой последовательности: сначала выбирается транзакт из списка отложенных прерываний; если он пуст, то транзакт выбирается из списка прерываний и, наконец, если и список прерываний пуст, то транзакт выбирается из списка задержки.

6.5.16. LOGIC (ИЗМЕНИТЬ)

Назначение оператора: изменение состояния логического ключа.

Формат оператора:

LOGIC X A

Значения операндов:

A – идентификатор (число или имя) логического ключа;

X – логический оператор, указывающий тип операции изменения

состояния: **R** – сбросить (выключить), **S** – установить (включить), **I** – инвертировать.

6.5.17. GATE (ВПУСТИТЬ)

Назначение: изменение маршрута движения транзактов в зависимости от состояния некоторого объекта.

Формат оператора:

GATE X A,[B]

Значения операндов:

A – идентификатор (число или имя) проверяемого объекта;

B – номер блока, к которому переходит транзакт, если объект находится в положении, не отвечающем условию проверки; [по умолчанию – Проверка происходит в режиме отказа];

X – условный оператор, содержащий условие, которому должен удовлетворять объект для успешного завершения теста; может принимать множество значений, в соответствии с которыми проводится проверка состояния некоторого объекта (прибора, многоканального устройства, логического ключа), в том числе:

- **FV** – прибор доступен;
- **FNV** – прибор недоступен;
- **I** – прибор в прерванном состоянии;
- **NI** – прибор в непрерванном состоянии;
- **U** – прибор используется;
- **NU** – прибор не используется;
- **SE** – многоканальное устройство пусто;
- **SNE** – многоканальное устройство не пусто;
- **SF** – многоканальное устройство заполнено;
- **SNF** – многоканальное устройство не заполнено;
- **SV** – многоканальное устройство доступно;
- **SNV** – многоканальное устройство не доступно;
- **LS** – логический ключ установлен (включен);
- **LR** – логический ключ сброшен (выключен).

6.5.18. MARK (ОТМЕТИТЬ)

Назначение оператора: запись значения абсолютного времени в качестве одного из параметров активного транзакта (отметка транзакта).

Формат оператора:

MARK [A]

Значения операндов:

A – номер параметра, в который записывается значение таймера абсолютного времени; [по умолчанию – Значение абсолютного времени помещается на место ранее записанного времени входа транзакта в модель].

6.5.19. ASSIGN (НАЗНАЧИТЬ)

Назначение оператора: назначение и изменение параметра транзакта.

Формат оператора:

ASSIGN A,B,[C]

Значения операндов:

A – номер модифицируемого параметра и вид модификации: присвоение, увеличение (+), уменьшение (-);

B – величина, используемая для модификации;

C – номер функции для модификации.

Примеры:

ASSIGN 4,10.5; параметру с номером 4 присваивается значение 10.5.

ASSIGN 4+,10.5; значение параметра с номером 4 увеличивается на величину 10.5.

ASSIGN 4-,10.5; значение параметра с номером 4 уменьшается на величину 10.5.

ASSIGN 3+,5,7; 1) рассчитывается значение функции 7; 2) это значение умножается на 5; 3) целая часть этого произведения прибавляется к значению параметра 3 вошедшего (активного) транзакта.

ASSIGN 3+,5,FN7; в отличие от предыдущего примера сначала рассчитывается значение функции 7, целая часть которого будет использоваться как **C**, то есть номер функции.

ASSIGN 3+,5,Erl; 1) рассчитывается значение функции с именем Erl; 2) это значение умножается на 5; 3) целая часть этого произведения прибавляется к значению параметра 3 вошедшего (активного) транзакта.

Три последних примера показывают, что в качестве операнда C может использоваться номер функции или её имя (без FN или FN$). Если же операнд C содержит FN или FN$, то это означает косвенное определение номера функции для модификации.

6.5.20. TABULATE (ТАБУЛИРОВАТЬ)

Назначение оператора: занесение значений в таблицу.

Формат оператора:

TABULATE A,[B]

Значения операндов:

A – имя таблицы, в которую заносится соответствующее значение и которая должна быть описана с помощью оператора описания (команды) **TABLE**;

B – весовой коэффициент; [по умолчанию – Коэффициент равен 1].

При попадании активного транзакта в оператор **TABULATE** обновляется статистика, связанная с таблицей, указанной в операнде A (см. пример в п.6.6.4).

6.6. Команды GPSS World

6.6.1. Общие сведения

В GPSS World используются 24 команды (операторов описания и операторов управления).

Для построения и реализации имитационных моделей *простейших* систем и сетей массового обслуживания в среде GPSS World оказывается достаточным использование немногим более половины из всех команд, которые по функциональному назначению могут быть разбиты на две группы:

1. Операторы (команды) описания: FUNCTION, TABLE, QTABLE, STORAGE, VARIABLE.

2. Операторы (команды) управления: CLEAR, CONTINUE, HALT, INCLUDE, REPORT, RESET, SHOW, START, STEP, STOP.

Команды управления используются в процессе моделирования для интерактивного взаимодействия пользователя с GPSS-моделью и управления процессом моделирования.

Команды, как и операторы блоков, могут быть без операндов или содержать от 1 до 5-и операндов, некоторые из которых могут быть необязательными. В последнем случае значения необязательных операндов принимаются по умолчанию. При изображении структуры оператора *необязательные операнды заключены в квадратные скобки.*

Отсутствие обязательных операндов приводит к ошибке.

6.6.2. FUNCTION (ФУНКЦИЯ)

Назначение: описание функции.

Формат:

<Имя> FUNCTION A,B

Здесь: **<Имя>** – имя функции.

Значения операндов:

A – аргумент функции;

B – задаёт тип функции и количество пар данных в виде:

<тип функции><количество пар данных>,

где **<тип функции>** может принимать следующие значения: **C** – непрерывная функция, **D** – дискретная функция, **E** – дискретная атрибутивно-значимая функция, **L** – списковая числовая функция, **M** – списковая атрибутивно-значимая функция;

<количество пар данных> определяет количество пар данных (аргумента и соответствующего ему значения функции) в списке данных функции, который располагается после оператора с первой позиции новой строки и может занимать несколько строк; каждая пара данных определяет значение аргумента X и значение функции Y (или СЧА), разделённые запятой.

Пример. При моделировании систем массового обслуживания функция типа **C** может использоваться для табличного представления

вероятностных законов распределения случайных величин. В частности, в предыдущих версиях GPSS для генерирования случайных чисел, распределённых по экспоненциальному закону, использовался табличный генератор, заданный в виде следующей функции:

EXP1FUNCTION RN100, C24
0,0/.1, .104/.2, .222/.3, .335/.4, .509/.5, .69/.6, .915/.7, 1.2/.75, 1.38/.8, 1.6/.84, 1.83/.88, 2.12/.9, 2.3/.92, 2.52/.94, 2.81/.95, 2.99/.96, 3.2/.97, 3.5/.98, 3.9/.99, 4.6/.995, 5.3/.998, 6.2/.999, 7/.9997, 8

Здесь:

EXP1 – имя табличной функции, которое используется в СЧА класса FN при обращении к функции: FN$EXP1;

RN100 – генератор равномерно распределённых случайных чисел с номером 100, используемый в качестве аргумента функции для вычисления значений экспоненциально распределённых случайных величин; путём изменения номера генератора равномерно распределённых случайных чисел (от 1 до 999) можно создавать множество генераторов экспоненциально распределённых случайных величин;

C24 – тип функции – **C**, означающий, что значения функции для любого значения аргумента определяются с использованием линейной интерполяции; таблица содержит 24 пары значений аргумента и функции, причём каждая пара отделена от другой наклонной чертой.

6.6.3. STORAGE (МНОГОКАНАЛЬНОЕ УСТРОЙСТВО)

Назначение: описание ёмкости многоканального устройства (памяти).

Формат:

<Имя> STORAGE A

Здесь: **<Имя>** – имя многоканального устройства.

Значения операнда:

A – количество приборов (каналов) в многоканальном устройстве.

6.6.4. TABLE (ТАБЛИЦА)

Назначение: описание таблицы, используемой в модели для накопления частоты попадания некоторой случайной величины в заданные частотные интервалы и построения гистограммы плотности распределения.

Формат:

<Имя> TABLE A,B,C,D

Здесь: **<Имя>** – имя таблицы (не более 32-х алфавитно-цифровых символов).

Значения операндов:

A – имя случайной величины (СЧА), значения которой должны учитываться в таблице; операнд **A** игнорируется дисперсионным анализом, но должен быть определен, когда используется блоками **TABULATE**;

B – ширина первого частотного интервала;

C – ширина всех промежуточных частотных интервалов;

D – количество частотных интервалов таблицы, включая левый и правый (целое положительное число).

Пример:

TU_5 TABLE M1,5,10,4

в таблице с именем **TU_5** будет накапливаться частота попаданий значений резидентного времени транзактов в четыре (D=4) частотных интервала шириной 5 единиц времени для первого интервала и 10 – для остальных трёх интервалов: 0-5; 5-15; 15-25; 25-35; когда активный транзакт попадает в блок **TABULATE TU_5**, в соответствии с операндом **A** в команде **TABLE**, заданным в виде СЧА **M1**, вычисляется время нахождения этого транзакта в модели как разница между текущим моментом модельного времени и моментом поступления транзакта в модель; в зависимости от полученного значения резидентного времени прибавляется единица к накапливаемому значению соответствующего частотного интервала; *для всех значений, превышающих правую границу последнего частотного интервала, единица добавляется в последний интервал*.

6.6.5. QTABLE (ТАБЛИЦА ОЧЕРЕДИ)

Назначение: описание таблицы очереди, используемой в модели для накопления частоты попадания времени нахождения транзакта в очереди (времени ожидания) в заданные частотные интервалы и построения гистограммы плотности распределения.

Формат:

<Имя> TABLE A,B,C,D

Здесь: **<Имя>** – имя таблицы (не более 32-х алфавитно-цифровых символов), в которой будут накапливаться значения частот.

Значения операндов:

A – имя очереди, для которой формируется таблица;

B – ширина первого частотного интервала;

C – ширина всех промежуточных частотных интервалов;

D – количество частотных интервалов таблицы, включая левый и правый (целое положительное число).

Пример:

Gis2u TABLE Stell,10,10,40

в таблице с именем **Gis2u** будет накапливаться частота попаданий значений времени нахождения транзактов в очереди с именем **Stell** в сорока (D=40) частотных интервалах шириной по 10 единиц времени, то есть охватывается временной интервал от 0 до 400 единиц времени; значения, превышающие 400 единиц времени, попадут в последний интервал.

6.6.6. VARIABLE (АРИФМЕТИЧЕСКАЯ ПЕРЕМЕННАЯ)

Назначение: описание арифметической переменной.

Формат:

<Имя> VARIABLE X

Здесь: **<Имя>** – имя арифметической переменной.

Значения операнда:

X – арифметическое выражение для вычисления значения переменной **<Имя>**.

Пример:

Vara1 VARIABLE 5#EXP(V$Grad+2)

когда активный транзакт попадает в блок, в котором используется переменная Vara1, (точнее, ссылка на эту переменную в виде СЧА: V$Vara1), например:

ADVANCE V$Vara1,

вычисляется значение переменной **Vara1** в соответствии с заданным арифметическим выражением как $5e^{Grad+2}$, где **V$Grad** – ссылка на другую арифметическую переменную **Grad**, которая тоже должна быть определена с помощью другого оператора **VARIABLE.**

6.6.7. CLEAR (ОЧИСТИТЬ)

Назначение: возврат процесса моделирования в исходное состояние с возможностью сохранения значений некоторых объектов GPSS-модели.

Формат команды:

CLEAR [A]

Операнд **A** может принимать только два значения: **ON** или **OFF**; необязательный операнд [по умолчанию – **ON**].

Когда операнд **A** равен OFF, ячейки, логические ключи и элементы матриц остаются без изменений.

6.6.8. CONTINUE (ПРОДОЛЖИТЬ)

Назначение: возобновление прерванного процесса моделирования.

Формат команды:

CONTINUE

6.6.9. HALT (ОСТАНОВИТЬ)

Назначение: прерывает процесс моделирования и очищает очередь команд. Является *срочной* командой.

Формат команды:

HALT

6.6.10. INCLUDE (ВКЛЮЧИТЬ)

Назначение: вставка в исходную модель и трансляция файла с операторами.

Формат команды:

INCLUDE A

A – полный путь доступа к указанному файлу.

Если **A** – имя файла (без указания пути доступа), то предполагается, что вставляемый файл находится в той же папке, что и исходная модель.

6.6.11. REPORT (СОЗДАТЬ ОТЧЁТ)

Назначение: немедленное создание отчета.

Формат команды:

REPORT

6.6.12. RESET (СБРОСИТЬ)

Назначение: сброс в ноль статистики и атрибутов системы.

Формат команды:

RESET

6.6.13. SHOW (ПОКАЗАТЬ)

Назначение: отображает значение выражения в строке состояния окна «Model».

Формат команды:

SHOW X

Операнд **X** представляет собой выражение (арифметическое или логическое), значение которого необходимо отобразить в строке состояния окна «Model».

6.6.14. START (НАЧАТЬ)

Назначение: запуск процесса моделирования.

Формат команды:

START A,[B],,[D]

A – начальное значение «счетчика завершений»;

B – признак вывода статистики: значение NP (no printout) блокирует вывод стандартной статистики; необязательный операнд;

D – признак вывода списков: значение **1** включает вывод списков будущих и текущих событий в стандартный отчет; необязательный операнд.

Операнд **C** остался от предыдущих версий GPSS и *не используется* в GPSS World.

6.6.15. STEP (ШАГАТЬ)

Назначение: остановка процесса моделирования по определенному количеству входов транзактов в блоки.

Формат команды:

STEP A

A – количество входов в блок (положительное целое число).

Пример: команда

STEP 1

приводит к приостановке процесса моделирования всякий раз, когда транзакт входит в очередной блок.

6.6.16. STOP (ОСТАНОВИТЬ)

Назначение: устанавливает или снимает условие прерывания моделирования.

Формат команды:

STOP [A],[B],[C]

A – номер транзакта (положительное целое число);

B – номер блока (положительное целое число) или метка блока (имя);

C – флаг состояния команды:

- ON – устанавливает условие прерывания;
- OFF – снимает условие прерывания;
- по умолчанию ON.

При отсутствии:

- операнда **A** – любой транзакт, входящий в блок с номером **B**, вызывает условие прерывания;
- операнда **B** – транзакт с номером **A**, при входе в любой блок вызывает условие прерывания;
- операндов **A** и **B** – процесс моделирования немедленно прерывается.

Пример: команда

STOP 100,21

определяет условие прерывания процесса моделирования: при входе транзакта с номером 100 в блок с номером 21. Продолжение моделирования – команда CONTINUE.

6.7. GPSS-модели массового обслуживания

«Если программа полностью отлажена, ее нужно скорректировать» ***(Законы программирования)***

Рассмотрим принципы построения GPSS-моделей на примерах моделей систем (СМО) и сетей (СеМО) массового обслуживания с однородным и неоднородным потоком заявок. GPSS-модели представлены в порядке возрастания сложности. Вначале рассматриваются и подробно комментируются простейшие GPSS-модели, имитирующие работу СМО с однородным потоком заявок и позволяющие получить представление об основных операторах GPSS World. По мере усложнения моделей вводятся новые операторы, необходимые для построения более сложных GPSS-моделей.

Для каждой модели представлено подробное описание моделируемой системы с указанием конкретных значений параметров. Далее приводится текст GPSS-модели и детально рассматривается каждый оператор. Все операторы GPSS-моделей сопровождаются комментариями. Для некоторых моделей приводятся и подробно описываются стандартные отчеты, формируемые автоматически по завершению моделирования и содержащие результаты моделирования.

6.7.1. Модель 1: одноканальная СМО с детерминированным потоком заявок и равномерно распределенной длительностью обслуживания (D/U/1)

Положим, что система содержит один обслуживающий прибор (рис.6.6). В СМО поступает детерминированный поток заявок с интервалом 10 секунд. Заявки выбираются на обслуживание из накопителя неограниченной ёмкости в порядке поступления, то есть по правилу «первым пришел – первым обслужен» (дисциплина обслуживания FIFO – First In First Out).

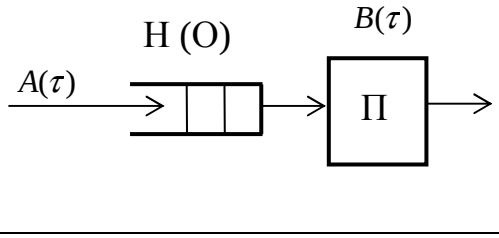


Рис.6.6. Одноканальная СМО

Длительность обслуживания заявок в приборе – величина случайная, распределенная по равномерному закону в интервале от 4 до 12 секунд (8±4 секунды) со средним значением 8 секунд.

Краткое описание рассматриваемой СМО имеет следующий вид:

- количество обслуживающих приборов: 1;
- количество потоков (классов) заявок: 1;
- ёмкость накопителя: *не ограничена* (равна бесконечности);
- интервалы между заявками в потоке: 10 секунд;
- поток заявок: *детерминированный*;
- значение длительности обслуживания заявок в приборе: 8±4 секунд;
- закон распределения длительности обслуживания заявок в приборе: *равномерный*.

Текст GPSS-модели:

```
**********************************************************************
      GENERATE    10; формирование детерминированного потока заявок
      QUEUE       1; отметка момента поступления заявки в очередь 1
      SEIZE       uzel; занятия прибора с именем uzel
      DEPART      1; отметка момента покидания заявкой очереди 1
      ADVANCE     8,4; задержка на время 8±4 единицы времени
      RELEASE     uzel; освобождение прибора с именем uzel
      TERMINATE   1; удаление заявки из модели
**********************************************************************
      START       100000
**********************************************************************
```

Рассмотрим подробно представленную модель и прокомментируем каждый оператор GPSS-модели, сопоставив их с реально протекающими в системе процессами.

Первый оператор **GENERATE** формирует в модели через каждые 10 единиц модельного времени транзакты. Множество формируемых таким образом транзактов моделируют процесс поступления заявок в

систему, образующих детерминированный поток с интервалом 10 секунд.

Когда модельное время становится равным моменту формирования очередного транзакта, последний начинает движение в модели к следующему по порядку оператору **QUEUE**, который заносит транзакт (заявку) в очередь с именем «1». (В действительности же, все транзакты сохраняются в очереди даже при отсутствии оператора QUEUE. Оператор QUEUE отмечает момент поступления транзакта в очередь с целью сбора статистики по очередям).

Далее транзакт продолжает движение к следующему оператору **SEIZE**, в соответствии с которым выполняет попытку занять одноканальное устройство (прибор) с именем «uzel». При этом проверяется занятость устройства. Если прибор занят обслуживанием ранее поступившего транзакта, то рассматриваемый транзакт приостанавливает свое движение и остается в очереди до тех пор, пока не освободится прибор. Если прибор свободен, то рассматриваемый транзакт продвигается к следующему оператору DEPART.

Оператор **DEPART** отмечает момент покидания транзактом очереди с именем (номером) «1» с целью сбора статистики по очередям (определяется время нахождения транзакта в очереди, то есть время ожидания заявки). Двигаясь дальше, транзакт попадает в оператор ADVANCE. Оператор **ADVANCE** задерживает транзакт на случайную величину, формируемую по равномерному закону распределения из интервала 8±4, моделируя, таким образом, процесс обслуживания заявок в приборе. Дальнейшее движение транзакта в модели возможно только тогда, когда значение модельного времени достигнет момента завершения обслуживания заявки в приборе.

При попадании транзакта в операторе **RELEASE** выполняется совокупность действий по освобождению прибора с именем «uzel».

Затем транзакт попадает в последний оператор **TERMINATE**, который выводит транзакт из модели (уничтожает транзакт), при этом из «**Счетчика завершений»** вычитается значение, указанное в качестве операнда A оператора TERMINATE и равное 1 в нашем примере.

Процесс моделирования продолжается до тех пор, пока значение «**Счетчика завершений»** не станет равным нулю.

Начальное значение «**Счетчика завершений»**, указываемое в качестве операнда A, устанавливается с помощью команды **START**, которая одновременно запускает процесс моделирования. Таким образом, моделирование в данном примере завершится после прохождения через модель 100 тысяч транзактов (после обслуживания в моделируемой системе 100 тысяч заявок).

Команда START может находиться *непосредственно в модели* или же может быть задана отдельно, *после трансляции* модели. В первом случае, после трансляции модели сразу же начинается ее выполнение. Во втором случае, выполнение модели начинается только после запуска

команды START.

По завершению моделирования результаты формируются автоматически в виде стандартного отчета, представленного на рис.6.7.

```
                GPSS World Simulation Report - CMO_DU1.2.1

                    Wednesday, January 25, 2006 11:58:53

          START TIME      END TIME  BLOCKS  FACILITIES  STORAGES
              0.000   1000005.010     7        1          0

                        NAME                VALUE
                       UZEL               10000.000

     LABEL          LOC  BLOCK TYPE     ENTRY COUNT CURRENT COUNT RETRY
                1    GENERATE         100000             0       0
                2    QUEUE            100000             0       0
                3    SEIZE            100000             0       0
                4    DEPART           100000             0       0
                5    ADVANCE          100000             0       0
                6    RELEASE          100000             0       0
                7    TERMINATE        100000             0       0

FACILITY    ENTRIES  UTIL.   AVE. TIME AVAIL. OWNER PEND INTER RETRY DELAY
UZEL        100000    0.801       8.008  1        0    0     0     0      0

QUEUE         MAX CONT. ENTRY ENTRY(0) AVE.CONT. AVE.TIME   AVE.(-0) RETRY
 1               1    0 100000  69780     0.040      0.405      1.339   0

   FEC XN    PRI      BDT      ASSEM  CURRENT  NEXT  PARAMETER    VALUE
             100001    0     1000010.000   100001      0      1
```

Рис.6.7. Стандартный отчет Модели 1

Стандартный отчет рассматриваемой модели содержит следующую информацию.

1. Заголовок с именем GPSS-модели:

```
GPSS World Simulation Report - CMO_DU1.2.1
```

2. Дату и время проведения имитационного моделирования (эксперимента):

```
Wednesday, January 25, 2006 11:58:53
```

3. Время старта и завершения моделирования, количество блоков (операторов), одноканальных устройств (приборов) и многоканальных устройств (памятей) в GPSS-модели:

```
START TIME          END TIME  BLOCKS  FACILITIES  STORAGES
    0.000        1000005.010     7        1          0
```

4. Перечень заданных в модели символических имен (блоков, устройств, памятей) и присвоенные им числовые значения (начиная с 10000):

```
NAME                          VALUE
UZEL                          10000.000
```

5. Перечень (BLOCK TYPE) пронумерованных (LOC) блоков с присвоенными им в модели метками (LABEL):

```
LABEL              LOC  BLOCK TYPE     ENTRY COUNT CURRENT COUNT RETRY
                    1    GENERATE         100000             0       0
                    2    QUEUE            100000             0       0
```

```
                    3     SEIZE          100000          0       0
                    4     DEPART         100000          0       0
                    5     ADVANCE        100000          0       0
                    6     RELEASE        100000          0       0
                    7     TERMINATE      100000          0       0
```

Кроме того, для каждого блока указывается:

ENTRY COUNT – количество транзактов, вошедших в данный блок за время моделирования;

CURRENT COUNT – количество транзактов, в данном блоке на момент завершения моделирования;

RETRY – количество транзактов, ожидающих выполнения некоторого специфического условия.

6. Результаты моделирования и дополнительная информация по устройствам:

```
FACILITY   ENTRIES   UTIL.    AVE.TIME  AVAIL. OWNER PEND INTER RETRY DELAY
 UZEL       100000   0.801       8.008  1        0    0    0     0      0
```

Здесь:

FACILITY – символическое имя или номер устройства;

ENTRIES – количество транзактов, вошедших в данное устройство за время моделирования;

UTIL. – коэффициент использования (загрузка) устройства;

AVE.TIME – среднее время занятия устройства одним транзактом (средняя длительность обслуживания заявок);

AVAIL. – состояние устройства на момент завершения моделирования: 1 – устройство доступно (не занято), 0 - устройство недоступно (занято);

OWNER – номер транзакта, находящегося в устройстве на момент завершения моделирования;

PEND – количество транзактов, ожидающих выполнения с прерыванием других транзактов;

INTER – количество прерванных транзактов на момент завершения моделирования (в списке прерываний);

RETRY – количество транзактов, ожидающих выполнения некоторого специфического условия;

DELAY – количество транзактов, ожидающих занятия устройства.

7. Результаты моделирования и дополнительная информация по очередям:

```
QUEUE          MAX CONT. ENTRY ENTRY(0) AVE.CONT. AVE.TIME   AVE.(-0) RETRY
 1               1    0 100000  69780     0.040      0.405      1.339   0
```

Здесь:

QUEUE – имя или номер очереди;

MAX – максимальное количество транзактов в очереди за время моделирования;

CONT. – текущее количество транзактов в очереди на момент завершения моделирования;

ENTRY – количество транзактов, прошедших через очередь за время моделирования;

ENTRY(0) – количество транзактов, прошедших через очередь за время моделирования с нулевым временем ожидания;
AVE.CONT. – средняя длина очереди за время моделирования;
AVE.TIME – среднее время нахождения транзакта в очереди (среднее время ожидания заявок);
AVE.(-0) – среднее время нахождения транзакта в очереди без учета транзактов с нулевым временем ожидания;
RETRY – количество транзактов, ожидающих выполнения некоторого специфического условия;

8. Список будущих событий (FEC):

```
FEC XN   PRI          BDT      ASSEM  CURRENT  NEXT  PARAMETER    VALUE
100001    0     1000010.000   100001       0      1
```

Здесь:
FEC – Future Events Chain;
XN – номера всех транзактов, находящихся в списке будущих событий (в данном примере это единственный транзакт с номером 100001);
PRI – приоритет транзакта;
BDT – момент времени, когда транзакт должен покинуть блок, а, следовательно, и список будущих событий);
ASSEM – номер семейства данного транзакта;
CURRENT – номер блока, в котором находился транзакт на момент завершения моделирования;
NEXT – номер следующего блока, в который будет передан транзакт;
PARAMETER – имя или номер параметра транзакта;
VALUE – значение параметра.

6.7.2. Модель 1.A: одноканальная СМО с простейшим потоком заявок (M/U/1)

Положим теперь, что в рассмотренную выше одноканальную СМО поступает простейший поток заявок, интервалы между которыми распределены по экспоненциальному закону со средним значением 10 секунд.

Текст GPSS-модели:

```
**************************************************************
      GENERATE     (Exponential(5,0,10))
      QUEUE        1
      SEIZE        uzel
      DEPART       1
      ADVANCE      (Uniform(25,4,12))
      RELEASE      uzel
      TERMINATE    2
**************************************************************
      START        100000
**************************************************************
```

Рассмотрим изменения, внесенные в предыдущую модель и выделенные жирным шрифтом.

Первое изменение – в операторе GENERATE, в первом операнде которого указан закон распределения интервалов между генерируемыми транзактами в виде библиотечной процедуры **Exponential(5,0,10),** обеспечивающей формирование случайных величин с экспоненциальным законом распределения. Три параметра процедуры **Exponential(5,0,10)** задают соответственно: 5 – номер исходного стандартного (встроенного) генератора равномерно распределенных случайных величин; 0 – смещение вырабатываемой случайной величины; 10 – среднее значение интервала.

Второе изменение в операторе ADVANCE, в котором используется другая форма задания равномерно распределенной случайной величины – в виде библиотечной процедуры **(Uniform(25,4,12))** из встроенной в GPSS библиотеки вероятностных распределений. Первый параметр процедуры **Uniform** определяет номер встроенного генератора равномерно распределённых в интервале (0; 1) случайных величин, а второй и третий параметры – границы интервала (соответственно нижняя и верхняя) формируемой равномерно распределенной случайной величины.

Третье изменение в операторе TERMINATE, в котором параметр A задан равным 2. Это значит, что при каждом попадании транзакта в этот оператор из счетчика завершений будет вычитаться не 1, как в предыдущем примере, а 2. Следовательно, моделирование завершится после прохождения через модель не 100 000 транзактов, а только 50 000.

6.7.3. Модель 2: многоканальная СМО с накопителем ограниченной ёмкости и обслуживанием заявок по закону Эрланга ($M/E_2/1/r$)

Положим теперь, что в предыдущую модель внесены следующие изменения (рис.6.8):

1) система содержит K=4 идентичных обслуживающих приборов, причём заявка может занять любой свободный прибор;
2) накопитель имеет ограниченную ёмкость $r = 10$, при этом заявка, заставшая накопитель заполненным, получает отказ в обслуживании и теряется;
3) длительность обслуживания заявок в одном приборе распределена по закону Эрланга 2-го порядка со средним значением 40 секунд.

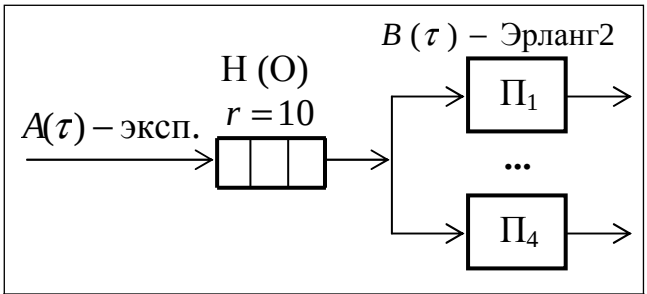


Рис.6.8. Многоканальная СМО с накопителем ограниченной ёмкости

Текст GPSS-модели с *комментариями* (*выделены курсивом*):

```
Uzel    STORAGE     4; задание числа приборов в устройстве с именем Uzel
**************************************
*Область исполняемых блоков (Основной модуль)
        GENERATE    (Exponential(11,0,10)); формирование простейшего потока
        TEST  L     Q$ch_1,10,Otkaz; проверка длины очереди
        QUEUE       ch_1; регистрация момента поступления заявки в очередь ch_1
        ENTER       Uzel; попытка занять один из приборов устройства Uzel
        DEPART      ch_1; регистрация момента покидания заявки очереди ch_1
        ADVANCE     (Exponential(21,0,20)+Exponential(31,0,20)); задержка заявки
*                                  в среднем на 40 единиц модельного времени
        LEAVE       Uzel; освобождение одного прибора многоканального
*                   устройства Uzel
        TERMINATE   1; удаление обслуженной заявки из модели и уменьшение
*                  счетчика завершений
Otkaz   TERMINATE   1; удаление заявки, получившей отказ
```

Рассмотрим изменения, внесенные в предыдущую GPSS-модель и выделенные жирным шрифтом.

Первое изменение заключается в появлении в GPSS-модели «Области описания», которая содержит оператор **STORAGE**, задающий имя (**Uzel**) многоканального устройства (памяти) и количество обслуживающих приборов (ёмкость памяти), равное 4.

Второе изменение заключается в появлении в GPSS-модели нового оператора (блока) **TEST**, позволяющего смоделировать накопитель с ограниченной ёмкостью перед многоканальным устройством.

Рассмотрим оператор **TEST** более подробно в контексте данного примера. Для этого сопоставим оператор **TEST**, записанный в общем виде, с оператором **TEST** в нашей модели:

TEST X A, B, C
TEST L Q$ch_1, 10, Otkaz

Здесь:

X – условный оператор (в нашем примере **L** означает «меньше»);

A – СЧА, значение которого проверяется в соответствии с заданным условным оператором (в нашем примере **Q$ch_1** означает проверку длины очереди с именем **ch_1**);

B – контрольное значение, с которым сравнивается значение числового атрибута, указанного в параметре A (в нашем примере длина очереди **ch_1** сравнивается во значением 10);

C – имя альтернативного оператора, которому передается транзакт, если указанное условие не выполняется (в нашем примере транзакт будет передан оператору **TERMINATE** с именем **Otkaz**).

Таким образом, транзакт, попав в указанный опреатор **TEST**, перейдет к следующему по порядку оператору при условии, что длина очереди **ch_1** меньше 10, и к оператору **TERMINATE** с меткой **Otkaz**, если в очереди **ch_1** уже находятся 10 заявок.

Третье изменение состоит в использовании операторов **ENTER** и **LEAVE,** моделирующих занятие и освобождение многоканального устройства, вместо операторов SEIZE и RELEASE, использующихся для одноканального устройства. Заметим, что в операторах **ENTER** и **LEAVE,** в отличие от SEIZE и RELEASE, могут использоваться два операнда A и B, где второй операнд B определяет количество занимаемых или освобождаемых приборов (каналов), причем при отсутствии операнда B его значение по умолчанию принимается равным 1.

В операторе ADVANCE реализуется случайная задержка заявки в соответствии с законом распределения Эрланга 2-го порядка в виде суммы двух экспоненциально распределенных случайных величин со средними значениями в 20 секунд (одна единица модельного времени равна одной секунде) так, что средняя задержка заявки в приборе составляет 40 секунд.

Еще одной особенностью данной модели является наличие двух операторов **TERMINATE**. Первый оператор удаляет из модели *обслуженные* заявки (транзакты), при этом из «**Счетчика завершений»** вычитается единица. Второй оператор удаляет из модели *необслуженны*е заявки, то есть заявки, заставшие при поступлении в систему накопитель заполненным и получившие отказ в обслуживании, при этом из «**Счетчика завершений»** также вычитается единица. Возникает вопрос: «Можно ли для вывода из модели обслуженных и необслуженных заявок использовать только один оператор **TERMINATE**?». Ответ: «Да, можно!». Зачем же тогда надо было использовать 2 оператора **TERMINATE**? Ответ достаточно простой. Второй оператор нужен только для того, чтобы получить информацию о доле обслуженных и доле потерянных (не обслуженных) заявок. Из стандартного отчета (рис.6.9) видно, что число обслуженных транзактов, прошедших через первый оператор **TERMINATE,** равно 938291, а число необслуженных (потерянных) транзактов, прошедших через второй оператор **TERMINATE,** равно 61709. Таким образом, вероятность потери заявки в моделируемой системе составляет 61709/(938291+61709)=0,061709, то есть 6,2% от общего числа поступивших в систему заявок. Отметим, что наличие в обоих операторах **TERMINATE** операнда, равного 1, означает, что моделирование завершится при достижении суммарного числа обслуженных и необслуженных заявок, покинувших систему, значения, указанного в операнде A команды **START** (в данной модели это значение равно 500000). Если в *первом* операторе **TERMINATE** операнд будет отсутствовать, что по умолчанию соответствует значению 0, то моделирование завершится, когда число *необслуженных (потерянных*) заявок достигнет указанного в команде **START** значения. И наоборот, если операнд будет отсутствовать во *втором* операторе **TERMINATE**, то моделирование завершится, когда число *обслуженных* заявок достигнет указанного в команде **START** значения.

На рис.6.9 представлен стандартный отчет, полученный для рассмотренной модели при задании команды

START 1000000,

означающем, что моделирование завершается после прохождения через систему *миллиона* заявок (транзактов).

```
               GPSS World Simulation Report - Untitled Model 2.1

                    Thursday, September 21, 2006 20:48:27

           START TIME          END TIME  BLOCKS  FACILITIES  STORAGES
                0.000      10007445.339     9         0          1

              NAME                       VALUE
          CH_1                         10001.000
          OTKAZ                            9.000
          UZEL                         10000.000

 LABEL             LOC  BLOCK TYPE     ENTRY COUNT CURRENT COUNT RETRY
                   1    GENERATE       1000006             0       0
                   2    TEST           1000006             0       0
                   3    QUEUE           938297             2       0
                   4    ENTER           938295             1       0
                   5    DEPART          938294             0       0
                   6    ADVANCE         938294             3       0
                   7    LEAVE           938291             0       0
                   8    TERMINATE       938291             0       0
OTKAZ              9    TERMINATE        61709             0       0

QUEUE          MAX CONT. ENTRY ENTRY(0) AVE.CONT. AVE.TIME   AVE.(-0) RETRY
 CH_1           10    3 938297 138746     4.231     45.125     52.955   0

STORAGE        CAP. REM. MIN. MAX.  ENTRIES AVL.  AVE.C. UTIL. RETRY DELAY
 UZEL            4    0   0     4   938295   1    3.753  0.938    0    2

CEC XN   PRI        M1        ASSEM  CURRENT  NEXT  PARAMETER    VALUE
1000004    0  10007414.480   1000004      4      5

FEC XN   PRI         BDT      ASSEM  CURRENT  NEXT  PARAMETER    VALUE
1000007   0    10007447.972   1000007     0      1
999995    0    10007482.391   999995      6      7
1000003   0    10007495.279   1000003     6      7
1000001   0    10007496.268   1000001     6      7
```

Рис.6.9. Стандартный отчет к Модели 2

Следует обратить внимание на то, что завершение процесса моделирования происходит по числу транзактов, прошедших через операторы TERMINATE, а не по числу транзактов, сформированных оператором GENERATE. В нашей модели через операторы TERMINATE с номерами (LOC) 8 и 9 прошли соответственно 938291 и 61709 транзактов (см. раздел LABEL отчета), что в сумме составляет ровно 1000000 транзактов, как указано в команде START. В то же время, количество транзактов, сформированных в операторе GENERATE равно 1000006, то есть на 6 транзактов больше, чем покинуло модель. Эти шесть транзактов на момент завершения моделирования остались в модели и, как видно в том же разделе отчета, они находятся в блоках QUEUE (2 транзакта), ENTER (1 транзакт), ADVANCE (3 транзакта).

6.7.4. Модель 2.A: дополнительная статистика в виде гистограмм

Стандартный отчёт, формируемый автоматически после завершения моделирования, если в команде START не был задан параметр NP, содержит основные результаты моделирования, состав которых может быть задан перед моделированием. Для этого в главном меню нужно выбрать пункт «Edit/Settings …» и на странице «Reports» («Отчёты») журнала настроек модели с помощью набора флажков задать состав результатов, включаемых в отчёт.

В некоторых случаях требуется получить результаты моделирования не только в виде средних значений вероятностных характеристик, но и в виде гистограмм, отображающих законы (плотности) распределений случайных величин.

Положим, что в рассмотренной модели 2 результаты моделирования должны быть представлены в виде гистограмм плотностей распределений времени ожидания и времени пребывания заявок.

Ниже представлен **текст GPSS-модели** с соответствующими добавлениями:

```
**************************************************************************
* Область описания
Uzel     STORAGE   4; задание числа приборов в устройстве с именем Uzel
T_w      QTABLE    ch_1,15,15,10
T_u      TABLE     M1,30,30,10
  ****************************************
*  Область исполняемых блоков (Основной модуль)
         GENERATE  (Exponential(11,0,10)); формирование простейшего потока
         TEST L    Q$ch_1,10,Otkaz; проверка длины очереди  ch_1
         QUEUE     ch_1; регистрация момента поступления заявки в очередь
         ENTER     Uzel; попытка занять один из приборов устройства Uzel
         DEPART    ch_1;  регистрация момента покидания очереди ch_1
         ADVANCE   (Exponential(21,0,20)+Exponential(31,0,20)); задержка
*                  заявки в среднем на 40 единиц модельного времени
         LEAVE     Uzel; освобождение одного прибора многоканального
*                  устройства Uzel
         TABULATE  T_u
         TERMINATE 1; удаление обслуженной заявки из модели и уменьшение
*                  счетчика завершений
Otkaz    TERMINATE 1; удаление заявки, получившей отказ
**************************************************************************
```

Рассмотрим изменения, внесенные в предыдущую модель и выделенные жирным шрифтом.

Во-первых, в области описания появились два новых оператора – команды: **QTABLE** и **TABLE**.

Сопоставим эти операторы, записанные в общем виде, с операторами в нашей модели:

<Имя> QTABLE A, B, C, D
T_w QTABLE ch_1, 15, 15, 10

<Имя> TABLE A, B, C, D
T_u TABLE M1, 30, 30, 10

Первый оператор (команда) **QTABLE** формирует таблицу для гистограммы плотности распределения времени ожидания заявок в очереди, имя которой указано в операнде A.

Имя **T_w** задаёт имя таблицы (гистограммы), а операнды A, B, C и D задают соответственно:

A=**ch_1** – имя очереди, для которой формируется гистограмма;

B=**15** – верхнюю (правую) границу первого частотного интервала гистограммы;

C=**15** – величину всех остальных частотных интервалов;

D=**10** – количество частотных интервалов.

Второй оператор **TABLE** формирует таблицу для гистограммы плотности распределения времени пребывания заявок в системе.

Имя **T_u**, как и в предыдущем случае, задает имя таблицы (гистограммы), а операнды A, B, C и D задают соответственно:

A=**M1** – величину, для которой формируется гистограмма; в нашем примере **M1** представляет собой СЧА, определяющее резидентное время, вычисляемое как разность между текущим значением модельного времени, определяемым в момент вхождения транзакта в блок **TABULATE**, и временем появления транзакта в модели, то есть временем поступления заявки в систему, являющимся одним из параметров транзакта;

B=**30** – верхнюю границу первого частотного интервала;

C=**30 –** величину всех остальных частотных интервалов;

D=**10** – количество частотных интервалов.

Таким образом, команда **TABLE** используется совместно с блоком **TABULATE**, который регистрирует момент прохождения транзактом (заявкой) определенного места в модели. Соответственно блок **TABULATE** должен находиться в модели в том месте, относительно которого измеряется искомое время. Таким местом при измерении времени пребывания заявки в моделируемой системе является точка выхода заявки из системы, когда транзакт покидает прибор многоканальной системы. В качестве параметра A оператора **TABULATE** выступает имя соответствующей таблицы (гистограммы). В нашем случае эта таблица и соответствующая ей гистограмма имеет имя **T_u.**

Оператор **TABLE** так же, как и **QTABLE**, позволяет сформировать гистограмму плотности распределения случайной величины и имеет аналогичную структуру. Основное отличие **TABLE** от **QTABLE** состоит в том, что оператор **TABLE** позволяет формировать гистограмму плотности распределения случайной величины между двумя, в общем случае, *произвольными моментами времени*, в то время как **QTABLE** всегда формирует гистограмму плотности распределения *времени ожидания в очереди*.

На рис.6.10 представлен фрагмент стандартного отчета, полученного для рассмотренной модели при задании команды

START 100000,

означающем, что моделирование завершено после прохождения через систему *ста тысяч* заявок (транзактов).

```
                                  ...
 LABEL            LOC   BLOCK TYPE     ENTRY COUNT CURRENT COUNT RETRY
                  1     GENERATE         100006             0       0
                  2     TEST             100006             0       0
                  3     QUEUE             93548             2       0
                  4     ENTER             93546             1       0
                  5     DEPART            93545             0       0
                  6     ADVANCE           93545             3       0
                  7     LEAVE             93542             0       0
                  8     TABULATE          93542             0       0
                  9     TERMINATE         93542             0       0
OTKAZ            10    TERMINATE          6458              0       0

QUEUE         MAX CONT. ENTRY ENTRY(0) AVE.CONT. AVE.TIME   AVE.(-0) RETRY
 CH_1          10    3  93548  13027     4.365     46.404     53.912   0

STORAGE       CAP. REM. MIN. MAX.  ENTRIES AVL.  AVE.C. UTIL. RETRY DELAY
 UZEL           4    0   0     4    93546   1    3.766  0.941    0    2

TABLE          MEAN    STD.DEV.       RANGE            RETRY FREQUENCY CUM.%
 T_W          46.405   35.680                            0
                                    _ -       15.000          23766    25.41
                              15.000 -       30.000          11890    38.12
                              30.000 -       45.000          11926    50.87
                              45.000 -       60.000          12229    63.94
                              60.000 -       75.000          11917    76.68
                              75.000 -       90.000           9946    87.31
                              90.000 -      105.000           6324    94.07
                             105.000 -      120.000           3365    97.67
                             120.000 -      135.000           1428    99.19
                             135.000 -  _                      754   100.00
 T_U          86.443   45.393                              0
                                    _ -       30.000           9684    10.35
                              30.000 -       60.000          19501    31.20
                              60.000 -       90.000          22921    55.70
                              90.000 -      120.000          20461    77.58
                             120.000 -      150.000          12584    91.03
                             150.000 -      180.000           5605    97.02
                             180.000 -      210.000           1954    99.11
                             210.000 -      240.000            594    99.75
                             240.000 -      270.000            168    99.93
                             270.000 -  _                       70   100.00
                                  ...
```

Рис.6.10. Фрагмент отчета к модели 2.А

Жирным шрифтом в отчёте выделены результаты формирования двух таблиц для построения гистограмм плотности распределения:

T_W – *времени ожидания* заявок в очереди;

T_U – *времени пребыв*ания заявок в системе.

На рис. 6.11 показаны гистограммы плотностей распределений времени пребывания **T_U** и времени ожидания **T_W** заявок, полученные для рассматриваемой модели.

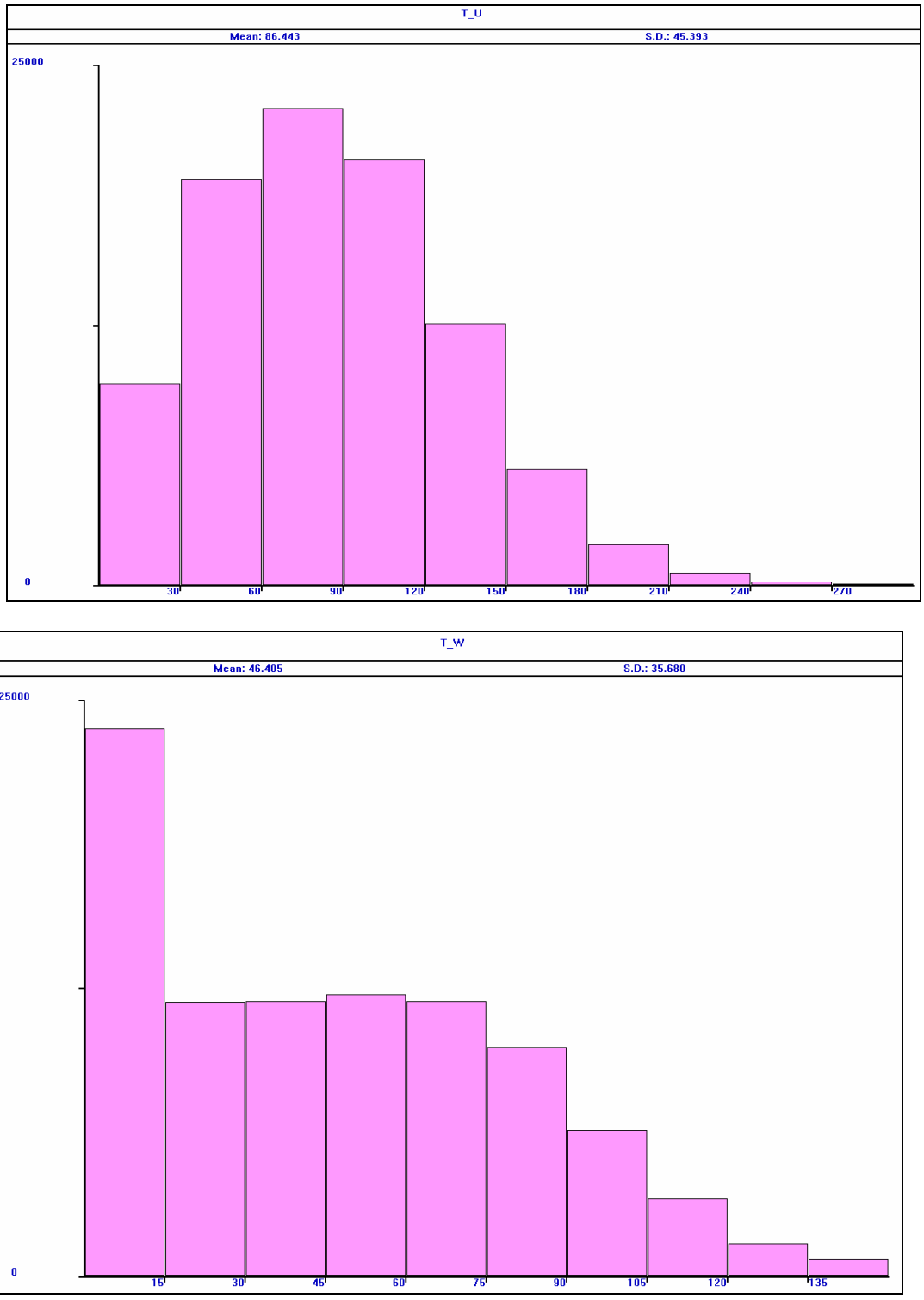


Рис.6.11. Гистограммы плотностей распределений времени пребывания T_U и времени ожидания T_W заявок

Для каждой из таблиц в отчёте приведены следующие данные:

MEAN – среднее значение соответствующей случайной величины;

STD.DEV. – стандартное отклонение случайной величины;

RANGE – нижние и верхние границы частотного класса (интервала);

RETRY – количество транзактов, ожидающих выполнения специфического условия, зависящего от состояния данной таблицы;

FREQUENCY – количество случайных значений, попавших в данный интервал; всякий раз увеличивается на единицу, если значение случайной величины больше нижней границы и меньше или равно верхней границе данного интервала; нижняя граница последнего интервала принимается равной бесконечности, то есть все случайные величины, значения которых больше нижней границы последнего частотного интервала, "попадают" в последний интервал;

CUM.% - накопленная частота, выраженная в процентах от общего количества случайных значений.

Следует отметить наличие определенных проблем, возникающих при задании длин и количества частотных интервалов, задаваемых в качестве операндов команд QTABLE и TABLE. Очевидно, что наглядность и, вытекающая отсюда, информативность гистограмм распределения случайных величин существенно зависит от количества частотных интервалов. Естественно, что для наглядности желательно иметь большое количества частотных интервалов. Однако, чем больше частотных интервалов, тем большую выборку случайных величин необходимо иметь, для того чтобы получить объективную картину, что не всегда возможно и целесообразно. В то же время небольшое количество частотных интервалов (в пределе только 1 интервал) не даёт объективной картины, позволяющей судить о законе распределения анализируемой случайной величины. Таким образом, задание длин и количества частотных интервалов является непростой задачей. Обычно их значения подбираются экспериментальным путем в процессе нескольких реализаций имитационной модели или же на основе предполагаемых значений математического ожидания и среднеквадратического отклонения соответствующей случайной величины.

По значениям среднеквадратического отклонения (S.D.) и математического ожидания (Mean) можно рассчитать коэффициенты вариации времени пребывания ν_u и времени ожидания ν_w заявок: $\nu_u = \frac{45{,}393}{86{,}443} \cong 0{,}525$ и $\nu_w = \frac{35{,}680}{46{,}405} \cong 0{,}769$, значения которых свидетельствуют о близости соответствующих законов распределений к распределению Эрланга 4-го и 2-го порядка соответственно ($k = 1/\nu^2$).

6.7.5. Модель 3: многоканальная СМО с неоднородным потоком заявок и накопителем ограниченной емкости

Внесем теперь в предыдущую модель четырехканальной СМО следующие изменения (рис.6.12):

1) ёмкость накопителя ограничена и равна 4;

2) в систему поступают два класса заявок:

- заявки 1-го класса образуют простейший поток со средним значением интервалов между заявками 20 секунд, и

длительность их обслуживания в приборе постоянна и равна 50 секундам;

- заявки 2-го класса образуют случайный равномерный поток с интервалами между заявками 18±10 секунд и длительностью их обслуживания в приборе, распределённой по экспоненциальному закону со средним значением 40 секунд.

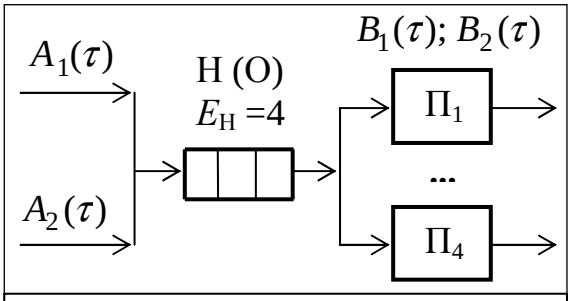


Рис.6.12. Многоканальная СМО с неоднородным потоком заявок

Заявки обоих классов поступают в один и тот же накопитель и выбираются на обслуживание в порядке поступления, то есть в соответствии с дисциплиной обслуживания FIFO.

Текст GPSS-модели с *комментариями* (*выделены курсивом*):

```
**********Модель 3**********
* Область описания
Uzel    STORAGE   4; задание числа приборов в устройстве с именем Uzel
Tw      QTABLE    1,2,2,40
Tu_1    TABLE     M1,50,4,40
Tu_2    TABLE     M1,7,7,40
**************************************
*Модуль 1: моделирование процессов поступления и обслуживания заявок 1-го класса
        GENERATE  (Exponential(1,0,20)); формирование простейшего потока
        TEST  L   Q1,4,Otk_1; проверка длины очереди
        QUEUE     1; регистрация момента поступления заявки в очередь 1
        ENTER     Uzel; попытка занять один из приборов устройства Uzel
        DEPART    1; регистрация момента покидания заявки очереди 1
        ADVANCE   50; задержка заявки на 50 единиц модельного времени
        LEAVE     Uzel; освобождение прибора Uzel
        TABULATE  Tu_1
        TERMINATE 1; удаление из модели обслуженной заявки 1-го класса
Otk_1   ERMINATE  1; удаление не обслуженной заявки 1-го класса
*Модуль 2: моделирование процессов поступления и обслуживания заявок 2-го класса
        GENERATE  18,10; формирование равномерно распределенного потока
        TEST  L   Q1,4,Otk_2; проверка длины очереди
        QUEUE     1; регистрация момента поступления заявки в очередь 1
        ENTER     Uzel; попытка занять один из приборов устройства Uzel
        DEPART    1; регистрация момента покидания заявки очереди 1
        ADVANCE   (Exponential(25,0,40)); задержка заявки 2-го класса
        LEAVE     Uzel; освобождение прибора Uzel
        TABULATE   Tu_2
        TERMINATE 1; удаление из модели обслуженной заявки 2-го класса
Otk_2   TERMINATE 1; удаление не обслуженной заявки 2-го класса
**************************************
        START     500000; запуск модели
```

Краткое описание рассматриваемой СМО:

- количество обслуживающих приборов – 4;
- емкость накопителя – 4;
- количество потоков (классов) заявок – 2;
- закон распределения интервалов между заявками 1-го класса – простейший со средним значением 10 секунд;
- длительность обслуживания заявок 1-го класса – детерминированная и равна 50 секундам;
- закон распределения интервалов между заявками 2-го класса – равномерный с интервалами между заявками 18±10 секунд;
- закон распределения длительности обслуживания заявок 2-го класса – экспоненциальный со средним значением 40 секунд;
- дисциплина буферизации – бесприоритетная с потерей заявки, заставшей в момент поступления накопитель заполненным;
- дисциплина обслуживания – бесприоритетная в порядке поступления (FIFO).

Рассмотрим подробнее представленную GPSS-модель.

В области описания заданы 3 таблицы для построения гистограмм плотностей распределений:

Tw – времени ожидания заявок обоих классов в общей очереди;

Tu_1 – времени пребывания в системе заявок 1-го класса;

Tu_2 – времени пребывания в системе заявок 2-го класса.

Отметим, что таблица **Tw** содержит информацию об усредненном времени ожидания заявок обоих классов.

Исполняемая область модели состоит из двух модулей, каждый из которых моделирует процессы поступления и обслуживания заявок 1-го и 2-го классов.

Последним оператором модели является команда START, задающая значение счетчика завершений равным 500000. Поскольку во всех четырех операторах TERMINATE задано значение операнда A равным 1, то моделирование завершится после прохождения через систему 500 тысяч заявок обоих классов, включая как обслуженные в системе заявки, так и потерянные (не обслуженные) заявки, которые в момент поступления в систему застали накопитель заполненным до конца.

Если в операторах TERMINATE с метками Otk_1 и Otk_2 операнд A не будет указан, то моделирование завершится после прохождения через систему 500 тысяч обслуженных заявок обоих классов, то есть без учета потерянных заявок.

Поскольку команда START включена в состав модели, то после создания модели (трансляции с помощью команды меню «Command/Create Simulation») процесс моделирования начнется автоматически сразу же после завершения трансляции.

6.7.6. Модель 3.A: многоканальная СМО с раздельными накопителями для заявок разных классов

В предыдущей модели время ожидания заявок определяется безотносительно к какому-либо классу, то есть полученное значение является усредненным временем ожидания заявок 1-го и 2-го класса. При этом отсутствует возможность оценки времени ожидания заявок каждого класса в отдельности.

Для определения времени ожидания заявок каждого класса в отдельности можно воспользоваться двумя способами:

1) собирать информацию о времени ожидания заявок 1-го и 2-го классов с помощью двух разных таблиц с использованием операторов TABLE и TABULATE, причем последний должен располагаться перед оператором ADVANCE в обоих исполняемых модулях, отображающих процесс прохождения заявок каждого класса; если при этом сохраняется оператор QTABLE, то в соответствующей таблице будет накапливаться информация об усредненном значении времени ожидания заявок обоих классов;
2) использовать для ожидания заявок 1-го и 2-го классов разные накопители.

Рассмотрим, какие изменения необходимо внести в предыдущую GPSS-модель 3 для реализации второго способа, когда заявки разных классов ожидают в разных накопителях. При этом будем полагать, что ёмкости обоих накопителей одинаковы: $E_{H1}=E_{H2}=2$, а их суммарная ёмкость осталась прежней, равной 4 (рис.6.13).

Текст GPSS-модели с изменениями, выделенными жирным шрифтом, приведён на следующей странице. По сравнению с предыдущей моделью в эту GPSS-модель внесены такие изменения.

В области описания появился второй оператор **QTABLE** с именем таблицы **Tw_2**, в котором в качестве операнда **A** указано имя (номер) накопителя **2**.

В операторах **TEST** модулей 1 и 2 в качестве операндов **A** используются СЧА **Q1** и **Q2**, означающие проверку длин очередей 1 и 2, а в качестве операндов **B** заданы ёмкости соответствующих накопителей, равные в обоих случаях 2. Таким образом, в момент поступления в систему заявки первого или второго класса текущие длины очередей сравниваются с заданными ёмкостями соответствующих накопителей.

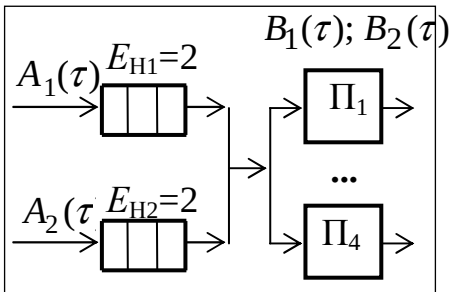


Рис.6.13. СМО с раздельными накопителями

Кроме того, в модуле 2 операнды **A** в операторах **QUEUE** и **DEPART** задают теперь номер очереди 2 для хранения заявок 2-го класса.

```
**********************************************************************
* Область описания
Uzel     STORAGE    4; задание числа приборов в устройстве с именем Uzel
Tw_1     QTABLE     1,2,2,40
Tw_2     QTABLE     2,2,2,40
Tu_1     TABLE      M1,50,4,40
Tu_2     TABLE      M1,7,7,40
**********************************************************************
*Модуль 1: моделирование процессов поступления и обслуживания заявок 1-го класса
         GENERATE   (Exponential(1,0,20)); формирование простейшего потока
         TEST  L    Q1,2,Otk_1; проверка длины очереди
         QUEUE      1; регистрация момента поступления заявки в очередь 1
         ENTER      Uzel; попытка занять один из приборов устройства Uzel
         DEPART     1; регистрация момента покидания заявки очереди 1
         ADVANCE    50; задержка заявки на 50 единиц модельного времени
         LEAVE      Uzel; освобождение прибора Uzel
         TABULATE   Tu_1
         TERMINATE  1; удаление из модели обслуженной заявки 1-го класса
Otk_1    TERMINATE  1; удаление не обслуженной заявки 1-го класса
**********************************************************************
*Модуль 2: моделирование процессов поступления и обслуживания заявок 2-го класса
         GENERATE   18,10; формирование равномерно распределенного потока
         TEST  L    Q2,2,Otk_2; проверка длины очереди 2
         QUEUE      2; регистрация момента поступления заявки в очередь 2
         ENTER      Uzel; попытка занять один из приборов устройства Uzel
         DEPART     2; регистрация момента покидания заявки очереди 2
         ADVANCE    (Exponential(25,0,40)); задержка заявки 2-го класса
         LEAVE      Uzel; освобождение прибора Uzel
         TABULATE   Tu_2
         TERMINATE  1; удаление из модели обслуженной заявки 2-го класса
Otk_2    TERMINATE  1; удаление не обслуженной заявки 2-го класса
**********************************************************************
         START      500000; запуск модели
```

Следует отметить, что результаты моделирования, полученные для СМО с общим накопителем ограниченной ёмкости (модель 3) и с раздельными накопителями ограниченной ёмкости (модель 3.А) для заявок разных классов, будут различны. В то же время, в случае накопителей *с неограниченной ёмкостью* модели 3 и 3.А дадут одинаковые результаты. Для того чтобы убедиться в этом, достаточно в представленных GPSS-моделях закомментировать операторы TEST (поставив в первой позиции символ *), используемые для проверки длины очереди и изменения направления движения транзактов при заполненном накопителе.

6.7.7. Модель 4: одноканальная СМО с относительными приоритетами

Рассмотрим одноканальную СМО с накопителем неограниченной ёмкости с неоднородным потоком заявок и приоритетным обслуживанием заявок разных классов (рис.6.14).

Положим, что в систему поступают 2 класса заявок. Заявки 1-го класса образуют детерминированный поток с интервалом между заявками 30 минут, заявки 2-го класса образуют равномерный поток с интервалами между заявками 15±5,5 минут.

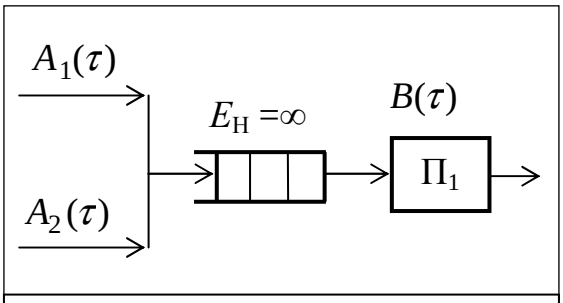


Рис.6.14. Одноканальная СМО с приоритетным обслуживанием

Длительность обслуживания в приборе заявок 1-го и 2-го классов является случайной величиной со средним значением 7 минут 30 секунд и среднеквадратическим отклонением 4 минуты 20 секунд.

Заявки обоих классов поступают в один и тот же накопитель, ёмкость которого не ограничена, и выбираются на обслуживание из накопителя в соответствии с дисциплиной обслуживания с относительными приоритетами, причём заявки 1-го класса имеют более высокий приоритет.

Для формирования в GPSS-модели закона распределения длительности обслуживания заявок воспользуемся аппроксимацией по двум моментам распределения: среднему значению $b = 7{,}5$ и среднеквадратическому отклонению $\sigma_b = 4{,}33$. Для выбора аппроксимирующего распределения рассчитаем коэффициент вариации длительности обслуживания:

$$\nu_b = \frac{\sigma_b}{b} = \frac{4{,}33}{7{,}5} \approx 0{,}577 .$$

В качестве аппроксимирующего распределения случайной величины с коэффициентом вариации, принимающим значения в интервале от 0 до 1, можно воспользоваться распределением Эрланга, коэффициент вариации которого определяется как $\nu = 1/\sqrt{k}$, где k – порядок распределения Эрланга. Тогда:

$$k = \frac{1}{\nu^2} = \frac{1}{0{,}577^2} \approx 3 .$$

Таким образом, в качестве закона распределения длительности обслуживания заявок будем использовать распределение Эрланга 3-го порядка, в соответствии с которым случайная величина формируется как сумма 3-х экспоненциально распределённых случайных величин с математическим ожиданием равным 2,5.

Краткое описание моделируемой СМО:

- количество обслуживающих приборов – 1;
- ёмкость накопителя – не ограничена;
- количество потоков (классов) заявок – 2;
- поток заявок 1-го класса – детерминированный с интервалами между заявками 30 минут;
- поток заявок 2-го класса – случайный с равномерно распределёнными интервалами между поступающими заявками в пределах от 9,5 до 20,5 минут;

- длительность обслуживания заявок 1-го и 2-го класса – случайная, распределённая по закону Эрланга 3-го порядка со средним значением 7,5 минут;
- дисциплина обслуживания заявок – с **относительными** приоритетами.

Текст GPSS-модели с *комментариями* (*выделены курсивом*):

```
* Модуль 1: моделирование процессов поступления и обслуживания заявок 1-го класса
      GENERATE  30,,,,2; детерминированный поток и приоритет, равный 2
      QUEUE     QUzel_1;  поступление заявки в очередь QUzel_1
      SEIZE     Uzel; занятие прибора Uzel
      DEPART    QUzel_1; покидание очереди QUzel_1
      ADVANCE   (Exponential(1,0,2.5)+Exponential(2,0,2.5)+Exponential(3,0,2.5));
      RELEASE   Uzel; освобождение прибора Uzel
      TERMINATE 1; удаление из модели обслуженной заявки 1-го класса
**************************************
* Модуль 2: моделирование процессов поступления и обслуживания заявок 2-го класса
      GENERATE  15,5.5; равномерный поток и приоритет, равный 0
      QUEUE     QUzel_2;  поступление заявки в очередь QUzel_2
      SEIZE     Uzel;  попытка занять один из приборов устройства Uzel
      ADVANCE   (Exponential(4,0,2.5)+Exponential(5,0,2.5)+Exponential(6,0,2.5));
      DEPART    QUzel_2;  покидание очереди QUzel_2
      RELEASE   Uzel; освобождение прибора Uzel
      TERMINATE 1; удаление из модели обслуженной заявки 2-го класса
```

Рассмотрим подробнее представленную GPSS-модель.

Модель включает в себя два модуля, каждый из которых моделирует процессы поступления и обслуживания заявок 1-го и 2-го классов и содержит стандартный набор операторов, используемых для моделирования простейшей одноканальной СМО с однородным потоком заявок.

Особенности данной модели, на которые следует обратить внимание, заключаются в следующем.

В первом операторе **GENERATE** в качестве пятого операнда, используемого для задания уровня приоритета формируемых транзактов (заявок первого класса), задано значение 2. Во втором операторе **GENERATE** значение пятого операнда не задано, что по умолчанию означает нулевой уровень приоритета формируемых транзактов (заявок второго класса). Таким образом, заявки 1-го класса имеют более высокий уровень приоритета по сравнению с заявками 2-го класса, что предоставляет им преимущественное право на занятие того или иного объекта модели: прибора (в нашей модели) или многоканального устройства.

При необходимости предоставить более высокий приоритет заявкам 2-го класса достаточно во втором операторе **GENERATE** задать значение пятого операнда большее, чем 2.

Следует также обратить внимание на второй оператор **GENERATE**, где в поле операндов присутствует запятая, разделяющая операнды A и B, и точка, отделяющая дробную часть десятичного числа от целой части.

Необходимо помнить, что запятая в поле операндов используется только в качестве разделителя операндов, а точка – для отделения дробной части десятичного числа от целой части.

Оператор **ADVANCE** в обоих модулях GPSS-модели в качестве операнда A содержит выражение, реализующее формирование случайной задержки транзактов (заявок) в соответствии с распределением Эрланга 3-го порядка путём сложения трёх экспоненциально распределённых случайных величин. Для уменьшения корреляции между случайными значениями, вырабатываемыми при обращении к библиотечной процедуре **Exponential**, используются разные генераторы равномерно распределённых величин, номера которых указаны в качестве первого параметра библиотечной процедуры **Exponential**.

Следует обратить внимание на местоположение оператора **DEPART** во втором модуле GPSS-модели, который, в отличие от первого модуля, располагается после оператора **ADVANCE**. Таким способом можно в стандартном отчёте получить информацию не о времени ожидания заявок в очереди, как это сделано для заявок 1-го класса, а о времени пребывания заявок 2-го класса в системе (в очереди и на обслуживании). Это легко понять, если вспомнить, что функцией оператора **DEPART** является отметка времени и сбор статистики по времени между моментами прохождения транзакта через операторы QUEUE и DEPART. Описанная ситуация соответствует случаю, когда заявка, находясь на обслуживании в приборе, остаётся в накопителе, который она покидает только после завершения обслуживания. Примером такой ситуации может служить модель передачи пакетов по каналу связи в сети передачи данных. Пакет в процессе передачи по каналу связи остаётся в выходной буферной памяти передающего узла до тех пор, пока не будет получено подтверждение от принимающего узла о безошибочном приёме пакета.

Для одновременного сбора статистики по времени пребывания заявок в системе и по времени ожидания в очереди можно использовать две пары операторов QUEUE и DEPART, как это показано ниже для второго модуля рассматриваемой GPSS-модели.

```
                                        . . .
Gis_U   QTABLE      Q_U,10,10,20
Gis_W   QTABLE      Q_W,10,10,20
                                        . . .
* Модуль 2: моделирование процессов поступления и обслуживания заявок 2-го класса
        GENERATE    15,5.5; равномерный поток и приоритет, равный 0
        QUEUE       Q_U;  регистрация момента занесения в очередь Q_U
        QUEUE       Q_W;  регистрация момента занесения в очередь Q_W
        SEIZE       Uzel;  попытка занять один из приборов устройства Uzel
        DEPART      Q_W;  регистрация момента покидания очереди Q_W
        ADVANCE     (Exponential(4,0,2.5)+Exponential(5,0,2.5)+Exponential(6,0,2.5));
        DEPART      Q_U;  регистрация момента покидания очереди Q_U
        RELEASE     Uzel; освобождение прибора Uzel
        TERMINATE   1; удаление из модели обслуженной заявки 2-го класса
```

При этом для получения соответствующих гистограмм плотностей распределений достаточно в модели описать две таблицы с помощью команды QTABLE.

Здесь пара операторов **QUEUE Q_U** и **DEPART Q_U** используется для сбора статистики по времени пребывания транзактов (заявок), а пара операторов **QUEUE Q_W** и **DEPART Q_W** – для сбора статистики по времени ожидания заявок.

Команда START не включена в состав модели и для запуска процесса моделирования должна быть задана с использованием меню GPSS World.

Подводя итог, следует отметить, что для реализации относительных приоритетов в GPSS не предусмотрены специальные операторы. Единственным средством для отображения относительных приоритетов служит присвоение определённого уровня приоритета разным транзактам. Однако для реализации абсолютных приоритетов в GPSS используются специальные операторы, которые рассматриваются в следующей модели.

6.7.8. Модель 4.А: одноканальная СМО с абсолютными приоритетами

Положим, что в той же одноканальной СМО (рис.6.14), обслуживание неоднородного потока заявок осуществляется на основе *абсолютных* приоритетов, допускающих прерывание обслуживания низкоприоритетной заявки при поступлении в систему высокоприоритетной заявки.

Текст GPSS-модели с *комментариями* (*выделены курсивом*):

```
* Модуль 1: моделирование процессов поступления и обслуживания заявок 1-го класса
      GENERATE  30,,,,2; детерминированный поток и приоритет, равный 2
      QUEUE     QUzel_1;  поступление заявки в очередь QUzel_1
      PREEMPT   Uzel;  занятие прибора с вытеснением низкоприоритетной заявки
      DEPART    QUzel_1; покидание очереди QUzel_1
      ADVANCE   (Exponential(1,0,2.5)+Exponential(2,0,2.5)+Exponential(3,0,2.5));
      RETURN    Uzel;  освобождение прибора и возврат прерванной заявки
      TERMINATE 1; удаление из модели обслуженной заявки 1-го класса
 ****************************************
* Модуль 2: моделирование процессов поступления и обслуживания заявок 2-го класса
      GENERATE  15,5.5; равномерный поток и приоритет, равный 0
      QUEUE     QUzel_2;  поступление заявки в очередь QUzel_2
      SEIZE     Uzel;  попытка занять один из приборов устройства Uzel
      ADVANCE   (Exponential(4,0,2.5)+Exponential(5,0,2.5)+Exponential(6,0,2.5));
      DEPART    QUzel_2;  покидание очереди QUzel_2
      RELEASE   Uzel; освобождение прибора Uzel
      TERMINATE 1; удаление из модели обслуженной заявки 2-го класса
```

Краткое описание моделируемой СМО:

- количество обслуживающих приборов – 1;
- емкость накопителя – не ограничена;
- количество потоков (классов) заявок – 2;

- поток заявок 1-го класса – детерминированный с интервалами между заявками 30 минут;
- поток заявок 2-го класса – случайный с равномерно распределёнными интервалами между заявками от 9,5 до 20,5 минут;
- длительность обслуживания заявок 1-го и 2-го класса – случайная, распределённая по закону Эрланга 3-го порядка со средним значением 7,5 минут;
- дисциплина обслуживания – с **абсолютными** приоритетами.

Для моделирования абсолютных приоритетов в GPSS World используются специальные операторы: PREEMPT (ЗАХВАТИТЬ) и RETURN (ВЕРНУТЬ), позволяющие реализовать прерывание обслуживания в приборе низкоприоритетного транзакта с последующим его восстановлением, а также оператор PRIORITY, позволяющий изменять приоритет транзакта в модели.

Оператор PREEMPT реализует захват прибора с вытеснением находящегося в приборе транзакта с более низким приоритетом (прерывание обслуживания низкоприоритетной заявки).

Если при поступлении в этот оператор транзакта прибор с именем, указанным в операнде A, свободен, то оператор PREEMPT инициирует такие же действия, что и оператор SEIZE, то есть транзакт занимает прибор независимо от уровня его приоритета. Если же прибор занят, то сравниваются уровни приоритета поступившего транзакта и транзакта, находящегося в приборе. Если уровень приоритета поступившего транзакта выше, чем находящегося в приборе, то в *приоритетном режиме*, признаком которого служит операнд B, имеющий значение PR, выполняется вытеснение из прибора (прерывание обслуживания) низкоприоритетного транзакта, и поступивший высокоприоритетный транзакт захватывает прибор. Отсутствие операнда по умолчанию означает *режим прерывания*, при котором захват прибора высокоприоритетным транзактом возможен только в том случае, если обслуживаемый транзакт не является захватчиком. В этом случае поступивший транзакт помещается в список отложенных прерываний. Транзактам из списка отложенных прерываний предоставляется возможность занять устройство раньше, чем вытесненным транзактам или транзактам из списка задержки. Если уровень приоритета поступившего транзакта ниже, чем находящегося в приборе, то он помещается в список задержки прибора с учётом приоритета.

Таким образом, наличие двух режимов оператора PREEMPT предоставляет возможность смоделировать разные стратегии реализации абсолютных приоритетов.

Оператор RETURN реализует удаление из прибора обслуженного транзакта (освобождение прибора) и, при необходимости, возврат низкоприоритетного транзакта в прибор для продолжения обслуживания. При этом очередной транзакт на обслуживание выбирается из списков в следующем порядке: сначала из списка отложенных прерываний, затем из

списка прерываний и, наконец, из списка задержки.

Заметим, что во втором модуле занятие прибора и его освобождение реализуется с помощью операторов SEIZE и RELEASE, поскольку формируемые там транзакты имеют самый низкий приоритет и не могут вытеснять никакие другие транзакты.

Использование вместо операторов SEIZE и RELEASE операторов PREEMPT и RETURN, а также операндов C, D и E в операторе PREEMPT позволяет смоделировать более сложные стратегии реализации абсолютных приоритетов, при которых, например, вытесненные транзакты могут быть направлены в другие блоки модели или исключены из состязания за прибор.

Следует помнить, что приоритетный захват возможен только для прибора, но невозможен для многоканального устройства!

Как и в предыдущей модели, команда START должна быть задана с использованием меню GPSS World.

6.7.9. Модель 5: двухузловая разомкнутая СеМО с однородным потоком заявок

Положим, что линейная разомкнутая СеМО с однородным потоком заявок содержит два узла (рис.6.15).

В РСеМО из внешней среды, обозначенной на рисунке как «0», в узел 1 поступает простейший поток заявок со средним интервалом 100 секунд. После обслуживания в узле 1 заявки с вероятностью $p_{12} = 0{,}8$ переходят на обслуживание в узел 2 и с вероятностью $p_{10} = 0{,}2$ покидают СеМО, возвращаясь во внешнюю среду. Из условия линейности СеМО следует, что $p_{10} + p_{12} = 1$, поскольку заявки в сети не теряются и не размножаются.

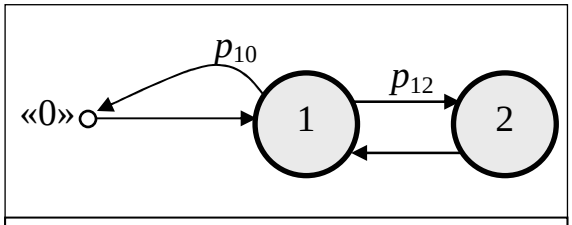


Рис.6.15. Разомкнутая СеМО

Длительность обслуживания заявок в узле 1, представляющем собой двухканальную СМО, – величина случайная, распределенная по равномерному закону в интервале от 10 до 20 секунд, то есть в интервале 15±5 с (со средним значением 15 с).

Длительность обслуживания заявок в узле 2, представляющем собой одноканальную СМО, – величина случайная, распределенная по экспоненциальному закону со средним значением 20 с.

Краткое описание рассматриваемой РСеМО:

- количество потоков (классов) заявок: $H = 1$;
- количество узлов в сети: $n = 2$;
- количество обслуживающих приборов в узле 1: $K_1 = 2$;
- количество обслуживающих приборов в узле 2: $K_2 = 1$;

- ёмкость накопителей в узлах сети – не ограничена, то есть в сети не может быть потерь заявок, что обусловливает линейность сети;
- поток заявок – простейший;
- средний интервал между поступающими в сеть заявками: $a = 100\,\text{с}$;
- длительность обслуживания заявок в узле 1 распределена равномерно в интервале от 10 до 20 с (15±5 с);
- длительность обслуживания заявок в узле 2 распределена по экспоненциальному закону со средним значением 20 с.

Рассмотрим **GPSS-модель** разомкнутой СеМО и прокомментируем некоторые операторы модели.

```
**************************************************************************
* Модель 5: разомкнутая однородная СеМО с двумя узлами
**************************************************************************
* Модуль 0:   область описания
Uz_1    STORAGE    2; задание числа приборов в узле 1
Tw_1    QTABLE     1,0,1,20; время ожидания в узле 1
Tw_2    QTABLE     2,50,50,20; время ожидания в узле 2
T_U     TABLE      M1,150,150,20; время пребывания в сети
**************************************************************************
* Модуль 1: моделирование процессов поступления и обслуживания заявок в узле 1
        GENERATE   (Exponential(10,0,100)); формирование простейшего потока
Met_1   QUEUE      1;  регистрация момента поступления заявки в очередь узла 1
        ENTER      Uz_1; попытка занять один из приборов узла 1
        DEPART     1;  регистрация момента покидания заявки очереди узла 1
        ADVANCE    15,5; задержка (обслуживание) заявки в узле 1
        LEAVE      Uz_1; выход обслуженной заявки из узла 1
        TRANSFER   .8,,Met_2; передача транзакта с вероятностью 0,8 в узел 2
        TABULATE   T_U
        TERMINATE  1; удаление из модели (СеМО) обслуженной заявки
**************************************************************************
* Модуль 2:  моделирование процесса обслуживания заявок в узле 2
Met_2   QUEUE      2;  регистрация момента поступления заявки в очередь узла 2
        SEIZE      2;  попытка занять прибор узла 2
        DEPART     2;  регистрация момента покидания заявки очереди узла 2
        ADVANCE    (Exponential(50,0,20));  обслуживание заявки в узле 2
        RELEASE    2;  освобождение прибора и выход заявки из узла 2
        TRANSFER   ,Met_1; безусловная передача транзакта в узел 1
**************************************************************************
        START      100000; запуск модели
```

Модуль 0 содержит описание двухканального устройства первого узла СеМО с именем Uz_1 и трех таблиц для формирования гистограмм плотностей распределений следующих случайных величин:

Tw_1 – времени ожидания заявок в узле 1 СеМО;

Tw_2 – времени ожидания заявок в узле 2 СеМО;

T_U – полного времени пребывания заявок в СеМО.

Отметим, что для каждого узла в модели определяется так называе-

мое ***единичное время ожидания*** заявок в узле СеМО, то есть время, соответствующее одному попаданию заявки в узел. В отличие от единичного, ***полное время ожидания*** заявки в узле СеМО учитывает, сколько раз заявка попала в данный узел за время ее нахождения в СеМО.

Как уже отмечалось, выбор длин и числа частотных интервалов, задаваемых в качестве операндов операторов QTABLE и TABLE, является непростой задачей, если мы хотим получить гистограмму приемлемого вида, дающего достаточно хорошее представление о законе (плотности) распределения. Обычно их значения подбираются экспериментальным путем в процессе нескольких реализаций имитационной модели. В частности, для рассматриваемой модели таким способом были подобраны операнды B=0, C=1 и D=20 в операторе:

Tw_1 QTABLE 1,0,1,20 .

Значение операнда B, определяющего длину первого частотного класса, было выбрано равным 0, поскольку в узле 1 почти 80% заявок имели нулевое время ожидания.

Модули 1 и 2 моделируют функционирование соответственно узлов 1 и 2 СеМО.

В модулях 1 и 2 используется новый оператор – оператор передачи **TRANSFER,** который реализует передачу активного транзакта к другому блоку (оператору) и может работать в 9 режимах. Активным является транзакт, поступивший в рассматриваемый момент времени в оператор **TRANSFER**.

В нашей модели оператор **TRANSFER** используется в двух режимах: *статистическом* и *безусловном*.

В общем случае оператор **TRANSFER** в *статистическом режиме* имеет вид:

TRANSFER A,[B],C ,

где **A** – частота (которая может трактоваться как вероятность) передачи активного транзакта к оператору с именем (меткой), указанной в качестве операнда **C**; может задаваться двумя способами: в виде *дробной величины* в интервале (0; 1) либо в виде неотрицательного *целого числа*, которое интерпретируется как доля от тысячи;

B – метка альтернативного оператора, которому будет передан активный транзакт (с вероятностью p=1-**A)**; если операнд не указан, то транзакт передается следующему по порядку оператору;

C – метка оператора, к которому передается активный транзакт с вероятностью, заданной в операнде **A**.

В нашем случае оператор **TRANSFER** в статистическом режиме имеет вид:

TRANSFER .8,,Met_2

Активный транзакт с частотой (вероятностью) 0,8 направляется к оператору с меткой **Met_2** и с вероятностью 0,2 – к следующему по порядку оператору.

Оператор **TRANSFER 800,,Met_2** эквивалентен предыдущему оператору и отличается только тем, что операнд **A** задан в виде целого числа, которое интерпретируется как доля от тысячи. Заметим, что оператор **TRANSFER 8,,Met_2** не эквивалентен двум предыдущим, поскольку вероятность перехода в данном случае будет равна 0,008.

Если в операторе **TRANSFER** операнд **A** отсутствует, то это означает, что оператор применяется в *безусловном режиме*: активный транзакт всякий раз будет направляться к оператору с меткой, указанной в качестве операнда **B.**

В нашем случае оператор **TRANSFER** в безусловном режиме имеет вид:

TRANSFER ,Met_1 .

Активный транзакт всякий раз направляется к оператору с меткой **Met_1**.

Команда **START**, включённая в состав модели и содержащая в качестве операнда **A** значение 100000, обеспечивает автоматический запуск процесса моделирования сразу же после завершения трансляции GPSS-модели. Процесс моделирования завершится после прохождения через систему 100 тысяч заявок (транзактов).

На рис. 6.16 представлен фрагмент отчёта, из которого могут быть получены основные характеристики функционирования РСеМО (наиболее интересные и важные результаты выделены жирным шрифтом).

Время моделирования (**END TIME**), в течение которого через РСеМО прошло $N_0 = 100000$ заявок, равно 9 994 872.283 секунд, что составляет без малого почти 2 777 часов или более 115 суток работы моделируемой системы. Все 100 тысяч обслуженных заявок, в конечном счете, попали в блок TERMINATE и были удалены из модели. В то же время количество сгенерированных транзактов в блоке GENERATE составляет 100 016. Возникает вопрос: почему было сгенерировано транзактов больше 100 000 и где «лишние» 16 транзактов? Ответ достаточно прост: поскольку процесс моделирования был остановлен по числу покинувших сеть ста тысяч транзактов, то на момент завершения моделирования в РСеМО остались транзакты, которые поступили в сеть, но не успели полностью пройти все этапы обслуживания. Где же эти транзакты? Из представленного отчёта видно, что на момент завершения моделирования в 10-м блоке QUEUE находится 15 транзактов и в 13-м блоке ADVANCE – один транзакт, всего 16 «лишних» транзактов на момент завершения процесса моделирования оказались в узле 2 СеМО: один в приборе и 15 – в очереди.

Через первый узел (блок **ENTER**) транзакты прошли $N_1 = 501310$ раз, через второй (блок **SEIZE**) – $N_2 = 401295$ раз. Эти значения позволяют рассчитать коэффициенты передач для обоих узлов РСеМО:

$$\alpha_1 = \frac{N_1}{N_0} = \frac{501310}{100000} \cong 5 \qquad \text{и} \quad \alpha_2 = \frac{N_2}{N_0} = \frac{401295}{100000} \cong 4,$$

что соответствует теоретическим значениям, которые могут быть найдены путём решения системы уравнений (4.16), как это описано в п.4.4.2.

Загрузки узлов (**UTIL.**) соответственно равны: $\rho_1 = 0{,}376$ и $\rho_2 = 0{,}800$.

```
START TIME              END TIME  BLOCKS  FACILITIES  STORAGES
             0.000       9994872.283    15        1          1

LABEL            LOC  BLOCK TYPE     ENTRY COUNT CURRENT COUNT RETRY
                   1    GENERATE         100016             0       0
MET_1              2    QUEUE            501310             0       0
                   3    ENTER            501310             0       0
                   4    DEPART           501310             0       0
                   5    ADVANCE          501310             0       0
                   6    LEAVE            501310             0       0
                   7    TRANSFER         501310             0       0
                   8    TABULATE         100000             0       0
                   9    TERMINATE        100000             0       0
MET_2             10    QUEUE            401310            15       0
                  11    SEIZE            401295             0       0
                  12    DEPART           401295             0       0
                  13    ADVANCE          401295             1       0
                  14    RELEASE          401294             0       0
                  15    TRANSFER         401294             0       0

FACILITY    ENTRIES    UTIL.   AVE. TIME AVAIL. OWNER PEND INTER RETRY
DELAY
 2           401295    0.800      19.936   1    99992    0    0     0
15

QUEUE          MAX CONT. ENTRY ENTRY(0) AVE.CONT. AVE.TIME   AVE.(-0) RETRY
 1               8    0 501310 398681     0.073      1.457      7.115   0
 2              35   15 401310  81329     3.243     80.776    101.307   0

STORAGE       CAP. REM. MIN. MAX.  ENTRIES AVL.  AVE.C. UTIL. RETRY DELAY
 UZ_1           2    2   0     2   501310   1    0.752  0.376    0    0

TABLE         MEAN     STD.DEV.        RANGE            RETRY FREQUENCY CUM.%
 TW_1         1.457    3.795                              0
                                  _  -        0.000          398681    79.53
                              0.000  -        1.000            8939    81.31

                                     ...
                             18.000  -  _                      4203   100.00
 TW_2        80.772   98.691                                  0
                                  _  -       50.000          207020    51.59
                             50.000  -      100.000           75108    70.30

                                     ...
                            900.000  -      950.000               5   100.00
 T_U        486.446  736.978                                  0
                                  _  -      150.000           38624    38.62
                            150.000  -      300.000           17615    56.24

                                        ...
                           2850.000  -  _                      1819   100.00


  FEC XN   PRI         BDT      ASSEM  CURRENT  NEXT  PARAMETER    VALUE
  99992    0     9994893.688    99992     13     14
 100017    0     9994946.528   100017      0      1
```

Рис.6.16. Фрагмент отчета модели разомкнутой СеМО

Средние длины очередей (**AVE.CONT.**) в узлах СеМО составляют: $l_1 = 0{,}073$ и $l_2 = 3{,}243$.

Использование в модели таблиц для построения гистограмм плотностей распределений времён ожидания заявок в узлах СеМО и времени пребывания заявок в сети, кроме средних значений временных характеристик, позволяет получить их среднеквадратические отклонения:

$w_1 = 1{,}457$ с; $\sigma_{w_1} = 3{,}795$ с;

$w_2 = 80{,}772$ с; $\sigma_{w_2} = 98{,}691$ с;

$U = 486{,}446$ с; $\sigma_U = 736{,}978$ с.

6.7.10. Модель 6: многоузловая разомкнутая СеМО с однородным потоком заявок

В данном пункте рассматривается модель, в которой заявки, покидающие некоторый узел, должны перемещаться с заданными вероятностями по 3-м и более направлениям. Реализация соответствующей имитациионной модели связана с определёнными проблемами, обусловленными особенностями имитационного моделирования.

Положим, что линейная разомкнутая СеМО с однородным потоком заявок содержит три узла (рис.6.17).

Пусть, как и в предыдущей модели, в РСеМО из внешней среды «0» в узел 1 поступает простейший поток заявок со средним интервалом 100 секунд. После обслуживания в узле 1 заявки с вероятностью $p_{12} = 0{,}1$ переходят на обслуживание в узел 2, с вероятностью $p_{13} = 0{,}3$ – на обслуживание в узел 3 и с вероятностью $p_{10} = 0{,}6$ покидают сеть, причём $p_{10} + p_{12} + p_{13} = 1$.

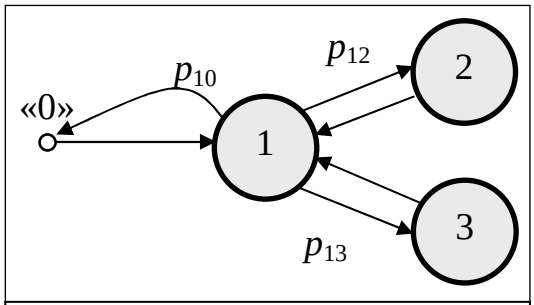


Рис.6.17. Трёхузловая РСеМО

Положим, что все узлы СеМО – одноканальные, а длительности обслуживания заявок в узлах 1, 2 и 3 представляют собой детерминированные величины, равные 10, 20 и 30 секундам соответственно.

Краткое описание рассматриваемой СеМО:

- количество потоков (классов) заявок: $H = 1$;
- количество узлов в сети: $n = 3$;
- все узлы одноканальные: $K_1 = K_2 = K_3 = 1$;
- емкость накопителей в узлах сети – не ограничена;
- поток заявок – простейший;
- средний интервал между поступающими в сеть заявками: $a = 100$ с;
- длительности обслуживания заявок в узлах 1, 2 и 3 детерминированные величины, равные соответственно 10 с, 20 с и 30 с.

Текст GPSS-модели с *комментариями* (*выделены курсивом*):

```
**************************************************************************
* Модель 6 линейной разомкнутой однородной СеМО с тремя узлами
**************************************************************************
*Модуль 1: моделирование процессов поступления и обслуживания заявок в узле 1
        GENERATE    (Exponential(10,0,100)); формирование простейшего потока
Met_1   SEIZE       1;  попытка занять прибор узла 1
        ADVANCE     10;  обслуживание заявки в узле 1
        RELEASE     1;  выход обслуженной заявки из узла 1
        TRANSFER    .6,,Met_0;  передача заявки с вероятностью 0,6 в узел «0»
        TRANSFER    .75,,Met_3;  передача транзакта с вероятностью 0,75 в узел 3
**************************************************************************
* Модуль 2: моделирование процесса обслуживания заявок в узле 2
Met_2   SEIZE       2;  попытка занять прибор узла 2
        ADVANCE     20;  обслуживание заявки в узле 2
        RELEASE     2;  освобождение прибора и выход заявки из узла 2
        TRANSFER    ,Met_1;  безусловная передача транзакта в узел 1
**************************************************************************
* Модуль 3:  моделирование процесса обслуживания заявок в узле 3
Met_3   SEIZE       3;  попытка занять прибор узла 3
        ADVANCE     30;  обслуживание заявки в узле 3
        RELEASE     3;  освобождение прибора и выход заявки из узла 3
        TRANSFER    ,Met_1; безусловная передача транзакта в узел 1
Met_0   TERMINATE   1; удаление из модели (СеМО) обслуженной заявки
```

Рассмотрим представленную GPSS-модель, уделив основное внимание операторам TRANSFER.

В модуле 1, реализующем процесс обслуживания заявки в узле 1 СеМО, используются два оператора TRANSFER. Первый оператор:

TRANSFER .6,,Met_0

направляет активный транзакт с заданной вероятностью 0,6 к оператору TERMINATE, что соответствует удалению обслуженной заявки из СеМО, и с вероятностью 0,4 – к следующему по порядку оператору:

TRANSFER .75,,Met_3 .

Назначение этого оператора – реализация процесса выбора узла, в который будет направлена заявка после завершения обслуживания в узле 1. Активный транзакт с вероятностью 0,75 будет направлен в модуль 3, реализующий процесс обслуживания заявок в узле 3, и соответственно с вероятностью 0,25 – в модуль 2, реализующий процесс обслуживания заявок в узле 2. Естественно, возникает вопрос: почему вероятности равны 0,25 и 0,75, если вероятности передачи в узлы 2 и 3 СеМО заданы равными 0,1 и 0,3? Ответ становится очевидным, если иметь в виду следующее: транзакт, попадающий в статистический TRANSFER, с заданной в операнде A вероятностью p направляется к оператору с меткой, указанной в операнде C, и с вероятностью $(1-p)$, являющейся дополнением до 1, к оператору с меткой, указанной в операнде B, или, по умолчанию, к следующему по порядку оператору, если операнд B опущен. Таким

образом, если в рассматриваемом операторе TRANSFER в качестве операнда A указать вероятность 0,3, то активный транзакт с вероятностью 0,3 будет направлен в модуль 3, и с вероятностью $(1-0{,}3)=0{,}7$ – в модуль 2. Это приведёт к тому, что большая часть заявок после обслуживания в узле 1 будет направляться в узел 2, в то время как в соответствии с исходными данными должно быть наоборот – в узел 3 должно направляться в 3 раза больше заявок, чем в узел 2.

Итак, каким же образом должны рассчитываться новые значения вероятностей? Необходимо, чтобы соотношение между вероятностями сохранялось. Этого можно достичь путём *нормирования вероятностей* так, чтобы сумма вероятностей была равна 1.

Например, пусть имеется N направлений передачи транзактов с вероятностями $p_1, p_2, \ldots, p_N$, причём $p_1 + p_2 + \ldots + p_N = 1$. Указанное вероятностное разветвление потока заявок может быть реализовано в GPSS-модели с помощью последовательности из $(N-1)$ операторов TRANSFER:

TRANSFER p_1,,Napravlenie_1

TRANSFER p_2^*,,Napravlenie_2

TRANSFER p_3^*,,Napravlenie_3

. . .

TRANSFER p_{N-1}^*,,Napravlenie_N-1

В первом операторе указывается заданное значение вероятности p_1 для передачи транзакта к оператору с меткой Napravlenie_1.

Во втором операторе для передачи транзакта к оператору с меткой Napravlenie_2 указывается нормированное значение вероятности, рассчитываемое как $p_2^* = \dfrac{p_2}{p_2 + p_3 + \ldots p_N}$ или, что то же самое, как $p_2^* = \dfrac{p_2}{1-p_1}$.

Аналогично, значение вероятности для третьего оператора:

$$p_3^* = \frac{p_3}{p_3 + p_4 + \ldots p_N} = \frac{p_3}{1-p_1-p_2}$$ и т.д.

И, наконец, для последнего оператора:

$$p_{N-1}^* = \frac{p_{N-1}}{p_{N-1} + p_N} = \frac{p_{N-1}}{1-p_1-\ldots-p_{N-2}}.$$

6.7.11. Модель 7: замкнутая СеМО с однородным потоком заявок

Положим, что рассмотренная выше линейная разомкнутая СеМО с однородным потоком заявок и двумя узлами преобразована в замкнутую

СеМО (рис.6.18), в которой циркулирует постоянное число заявок: $M = 5$.

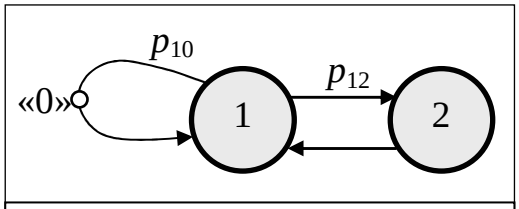


Рис.6.18. Замкнутая СеМО

Как и в предыдущей модели, после обслуживания в узле 1 заявки с вероятностью $p_{12} = 0{,}8$ переходят на обслуживание в узел 2 и с вероятностью $p_{10} = 0{,}2$ возвращаются в узел 1, причем $p_{10} + p_{12} = 1$. Пусть нулевая точка выбрана на дуге, выходящей из узла 1 и входящей снова в узел 1 так, как это показано на рис.6.18. Относительно этой точки будут измеряться такие характеристики сети, как производительность ЗСеМО и время пребывания заявок в сети. Длительность обслуживания заявок в двухканальном узле 1 распределена по равномерному закону в интервале от 10 до 20 секунд, а длительность обслуживания заявок в одноканальном узле 2 распределена по экспоненциальному закону со средним значением 20 с.

Таким образом, краткое описание рассматриваемой замкнутой СеМО имеет следующий вид:

- количество потоков (классов) заявок: $H = 1$;
- количество узлов в сети: $n = 2$;
- количество заявок, циркулирующих в замкнутой сети: $M = 5$;
- количество обслуживающих приборов в узле 1: $K_1 = 2$;
- длительность обслуживания заявок в узле 1 распределена равномерно в интервале от 10 до 20 с (15±5 с);
- количество обслуживающих приборов в узле 2: $K_2 = 1$;
- длительность обслуживания заявок в узле 2 распределена по экспоненциальному закону со средним значением 20 с.
- ёмкость накопителей в узлах сети достаточна для того, чтобы в сети не было потерь заявок, что обусловливает линейность сети; в нашем случае можно считать, что ёмкость каждого накопителя совпадает с числом циркулирующих в сети заявок.

Для того чтобы упростить процесс разработки GPSS-модели замкнутой СеМО, воспользуемся представленной выше GPSS-моделью разомкнутой СеМО. Основное отличие замкнутой СеМО от разомкнутой состоит в отсутствии внешнего источника заявок, при этом в GPSS-модели замкнутой СеМО необходимо реализовать циркуляцию в сети постоянного числа заявок (в нашей модели – пяти заявок).

Рассмотрим представленную ниже **GPSS-модель** и прокомментируем только выделенные жирным шрифтом изменения (операторы), которые были внесены в модель разомкнутой СеМО для ее преобразования в модель замкнутой СеМО.

1) Модель содержит на один модуль больше, чем модель разомкнутой СеМО. Дополнительный третий модуль реализует завершение процесса моделирования путем задания временного интервала функционирования реальной (исследуемой) системы.

```
*************************************************************************
* Модель 7 линейной замкнутой однородной СеМО с двумя узлами и М=5
*************************************************************************
* Модуль 0:   область описания
Uz_1    STORAGE     2; задание числа приборов в узле 1
Tw_1    QTABLE      1,0,0.5,30; время ожидания в узле 1
Tw_2    QTABLE      2,10,10,30; время ожидания в узле 2
T_U     TABLE       M1,40,40,30; время пребывания в сети
****************************************
* Модуль 1:  моделирование процесса обслуживания заявок в узле 1
        GENERATE    ,,,5;  формирование в нулевой момент времени пяти заявок
Met_1   MARK            ; отметка момента времени поступления заявки в сеть
Met_3   QUEUE       1; регистрация момента поступления заявки в очередь узла 1
        ENTER       Uz_1; попытка занять один из приборов узла 1
        DEPART      1;  регистрация момента покидания заявки очереди узла 1
        ADVANCE     15,5; задержка (обслуживание) заявки в узле 1
        LEAVE       Uz_1; выход обслуженной заявки из узла 1
        TRANSFER    .8,,Met_2; передача транзакта с вероятностью 0,8 в узел 2
        TABULATE    T_U
        TRANSFER    ,Met_1; безусловная передача транзакта в узел 1
****************************************
* Модуль 2:  моделирование процесса обслуживания заявок в узле 2
Met_2   QUEUE       2; регистрация момента поступления заявки в очередь узла 2
        SEIZE       2;  попытка занять прибор узла 2
        DEPART      2;  регистрация момента покидания заявки очереди узла 2
        ADVANCE     (Exponential(50,0,20));  обслуживание заявки в узле 2
        RELEASE     2;  освобождение прибора и выход заявки из узла 2
        TRANSFER    ,Met_3; безусловная передача транзакта в узел 1
*************************************************************************
* Модуль 3: завершение процесса моделирования по длительности моделирования
        GENERATE    10000000; задание единичной длительности моделирования
        TERMINATE   1; уменьшение счетчика завершения на 1
*************************************************************************
```

2) Для получения более наглядных временных гистограмм в операторах QTABLE и TABLE модуля 0 изменены параметры, задающие длину и число частотных интервалов, значения которых были подобраны экспериментальным путем.
3) В модуль 1 внесены 3 изменения.
   - Во-первых, изменился оператор **GENERATE**, который принял вид:

     GENERATE ,,,5

     В этом операторе указан только четвертый операнд, определяющий число генерируемых им транзактов за все время моделирования. Поскольку первый параметр отсутствует, то в нулевой момент модельного времени будут сформированы 5 транзактов, которые поступят в очередь первого узла. Таким образом, в моделируемой СеМО появятся 5 заявок.
   - Во-вторых, появился новый оператор с меткой:

Met_1 MARK .

Оператор **MARK (ОТМЕТИТЬ)** предназначен для записи значения абсолютного времени в качестве одного из параметров транзакта (отметка транзакта) и, в общем случае, имеет вид:

MARK A .

Единственный операнд **A** задает *имя* или *номер параметра* активного транзакта, в который записывается значение таймера абсолютного времени, причём при его отсутствии значение абсолютного времени по умолчанию помещается на место ранее записанного времени входа транзакта в модель.

Этот оператор в рассматриваемой GPSS-модели используется для отметки момента прохождения заявкой «нулевой точки», относительно которой измеряется время пребывания заявок в замкнутой СеМО.

- В-третьих, в конце модуля 1 вместо оператора TERMINATE вставлен оператор

TRANSFER ,Met_1 ,

реализующий безусловную передачу транзакта к блоку с меткой **Met_1**, что соответствует в модели возврату заявки в узел 1.

4) Дополнительный модуль 3 состоит только из двух операторов:

```
GENERATE    10000000
TERMINATE   1
```

Такой модуль применяется в GPSS-моделях для реализации завершения процесса моделирования *по времени*, прошедшему в моделируемой системе, а не по числу обслуженных в системе заявок (прошедших через модель транзактов).

В соответствии с этими операторами в момент модельного времени (времени, который наступит в реальной исследуемой системе), равный 10000000, в блоке **GENERATE** модели появится транзакт, который сразу же попадет в блок **TERMINATE** и будет уничтожен. При этом значение счётчика завершений будет уменьшено на единицу. Если модель была запущена командой **START 1**, установившей начальное значение счётчика завершений в 1, то после вычитания 1 значение счётчика завершений станет равным 0 и процесс моделирования завершится. Таким образом, если в предыдущей модели завершение моделирования осуществлялось по числу заявок, покинувших СеМО, то в данной модели использовалось другое условие завершения моделирования – по времени, прошедшему в моделируемой системе.

На рис.6.19 представлен фрагмент отчёта, из которого могут быть получены все основные характеристики функционирования замкнутой СеМО (наиболее интересные и важные результаты моделирования выделены жирным шрифтом).

```
            GPSS World Simulation Report - Untitled Model 1.1.1

         START TIME       END TIME    BLOCKS  FACILITIES  STORAGES
            0.000      10000000.000     18        1          1

                                  ...
LABEL          LOC   BLOCK TYPE     ENTRY COUNT   CURRENT COUNT  RETRY
                 1    GENERATE            5             0        0
MET_1            2    MARK           124309             0        0
MET_3            3    QUEUE          622086             0        0
                 4    ENTER          622086             0        0
                 5    DEPART         622086             0        0
                 6    ADVANCE        622086             1        0
                 7    LEAVE          622085             0        0
                 8    TRANSFER       622085             0        0
                 9    TABULATE       124304             0        0
                10    TRANSFER       124304             0        0
MET_2           11    QUEUE          497781             3        0
                12    SEIZE          497778             0        0
                13    DEPART         497778             0        0
                14    ADVANCE        497778             1        0
                15    RELEASE        497777             0        0
                16    TRANSFER       497777             0        0
                17    GENERATE            1             0        0
                18    TERMINATE           1             0        0

FACILITY  ENTRIES   UTIL.   AVE.TIME  AVAIL. OWNER PEND INTER RETRY DELAY
 2         497778  0.993     19.949     1      4    0     0     0    3

QUEUE     MAX CONT. ENTRY  ENTRY(0) AVE.CONT. AVE.TIME   AVE.(-0) RETRY
 1         4    0  622086  444845     0.145      2.330      8.178   0
 2         4    3  497781   12528     2.929     58.836     60.355   0

STORAGE   CAP. REM. MIN. MAX.  ENTRIES AVL.  AVE.C. UTIL. RETRY DELAY
 UZ_1       2    1   0     2   622086   1    0.933  0.467    0    0

TABLE     MEAN    STD.DEV.       RANGE        RETRY FREQUENCY CUM.%
TW_1      2.330    4.796                        0
                              _  -     0.000        444845    71.51
                          0.000  -     0.500          6254    72.51
                          0.500  -     1.000          6350    73.53
                                       . . .
                         14.000  -  _                28460   100.00
TW_2     58.836   39.993                            0
                              _  -    10.000         38373     7.71
                         10.000  -    20.000         39655    15.68
                         20.000  -    30.000         49849    25.69
                                       . . .
                        290.000  -  _                   71   100.00
T_U     402.214  438.431                            0
                              _  -    40.000         24771    19.93
                         40.000  -    80.000          4846    23.83
                         80.000  -   120.000          8414    30.60
                                       . . .
                       1160.000  -  _                8158   100.00
```

Рис.6.19. Фрагмент отчета к модели замкнутой СеМО

Видно, что время завершения моделирования (**END TIME**) в точности совпадает с временем, заданным в модуле 3 GPSS-модели.

Оператор **GENERATE** сгенерировал за время моделирования только

5 транзактов, которые постоянно циркулировали в модели. При этом через первый узел (блок **ENTER**) транзакты прошли $N_1 = 622086$ раз, через второй (блок **SEIZE**) – $N_2 = 497778$ раз, а через нулевую точку (блок **TABULATE**) – $N_0 = 124304$ раз. Последнее значение позволяет рассчитать одну из основных сетевых характеристик замкнутой СеМО – производительность сети, как отношение числа заявок (транзактов), прошедших через нулевую точку СеМО за время моделирования $T = 10000000$, к этому времени:

$$\lambda_0 = \frac{N_0}{T} = \frac{124304}{10000000} = 0{,}0124304\ \text{с}^{-1} \approx 44{,}75\ \text{ч}^{-1},$$

то есть примерно 45 заявок в час.

Коэффициенты передач для каждого из узлов могут быть рассчитаны следующим образом:

$$\alpha_1 = \frac{N_1}{N_0} = \frac{622086}{124304} \cong 5 \quad \text{и} \quad \alpha_2 = \frac{N_2}{N_0} = \frac{497778}{124304} \cong 4.$$

Загрузки узлов (**UTIL.**) соответственно равны: $\rho_1 = 0{,}467$ и $\rho_2 = 0{,}993$.

Средние длины очередей (**AVE.CONT.**) в узлах СеМО составляют: $l_1 = 0{,}144$ и $l_2 = 2{,}929$.

Использование в модели таблиц для построения гистограмм плотностей распределений времён ожидания заявок в узлах СеМО и времени пребывания заявок в сети, кроме средних значений временных характеристик, позволяет получить их среднеквадратические отклонения:

$w_1 = 2{,}33$ с; $\sigma_{w_1} = 4{,}796$ с;

$w_2 = 58{,}836$ с; $\sigma_{w_2} = 39{,}993$ с;

$U = 402{,}214$ с; $\sigma_U = 438{,}431$ с.

6.7.12. Модель 8: разомкнутая СеМО с неоднородным потоком заявок

Положим, что в линейную разомкнутую СеМО с двумя узлами поступает неоднородный поток заявок двух классов (рис.6.20). Заявки класса 1 (сплошная линия) и класса 2 (пунктирная линия) поступают в узел 1 и образуют простейшие потоки со средними интервалами 100 и 50 секунд соответственно. После обслуживания в узле 1 заявки класса 1 с вероятностью $p_{12} = 0{,}8$ переходят на обслуживание в узел 2 и с вероятностью $p_{10} = 0{,}2$ покидают СеМО. Заявки класса 2 обслуживаются только в узле 1, после чего покидают СеМО.

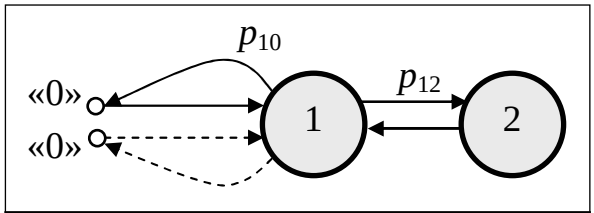


Рис.6.20. Разомкнутая СеМО с неоднородным потоком заявок

Длительности обслуживания заявок класса 1 и 2 в двухканальном узле 1 представляют собой равномерно распределённые случайные величины в интервалах (15±5) и (10 ±5) секунд соответственно.

Длительность обслуживания заявок класса 2 в одноканальном узле 2 – величина случайная, распределенная по экспоненциальному закону со средним значением 20 секунд.

Краткое описание рассматриваемой СеМО:

- количество потоков (классов) заявок: $H = 2$;
- количество узлов в сети: $n = 2$;
- количество обслуживающих приборов в узле 1: $K_1 = 2$;
- количество обслуживающих приборов в узле 2: $K_2 = 1$;
- емкость накопителей в узлах сети – не ограничена, то есть в сети не может быть потерь заявок, что обусловливает линейность сети;
- потоки заявок класса 1 и класса 2 – простейшие;
- средний интервал между поступающими заявками класса 1: $a_0(1) = 100\,\text{с}$;
- средний интервал между поступающими заявками класса 2: $a_0(2) = 50\,\text{с}$;
- длительность обслуживания заявок класса 1 в узле 1 распределена равномерно в интервале от 10 до 20 с: $b_1(1) = 15 \pm 5\,\text{с}$;
- длительность обслуживания заявок класса 2 в узле 1 распределена равномерно в интервале от 5 до 15 с: $b_1(2) = 10 \pm 5\,\text{с}$;
- длительность обслуживания заявок класса 1 в узле 2 распределена по экспоненциальному закону со средним значением 20 с: $b_2(1) = 20\,\text{с}$.

Текст GPSS-модели разомкнутой СеМО с неоднородным потоком заявок представлен на следующей странице.

В рассматриваемой GPSS-модели, в отличие от модели 5 двухузловой РСеМО с однородным потоком заявок, появился третий модуль, моделирующий процессы поступления и обслуживания заявок класса 2 в узле 1.

Таким образом, при моделировании СеМО с неоднородным потоком заявок количество исполняемых модулей GPSS-модели определяется как произведение количества классов заявок на количество узлов моделируемой СеМО.

На рис 6.21 представлен отчет с результатами имитационного моделирования разомкнутой СеМО с двумя классами заявок для значения 1000000 операнда А в команде START, заданного при запуске процесса моделирования.

Анализ представленного отчета позволяет получить основные характеристики функционирования разомкнутой СеМО с неоднородным потоком заявок (наиболее интересные и важные результаты моделирования выделены жирным шрифтом).

```
*************************************************************************
* Модель 8 разомкнутой СеМО с неоднородным потоком заявок
*************************************************************************
* Модуль 0:          область описания
Uzel_1      STORAGE    2; задание числа приборов в узле 1
****************************************
* Модуль 1:  поступление и обслуживание заявок класса 1 в узле 1
            GENERATE     (Exponential(10,0,100)); формирование потока заявок класса 1
Met_1       QUEUE        Quz1_k1;  момент поступления в очередь узла 1
            ENTER        Uzel_1; попытка занять один из приборов узла 1
            DEPART       QUz1_k1;  момент покидания очереди узла 1
            ADVANCE      15,5; задержка (обслуживание) в узле 1
            LEAVE        Uzel_1; выход обслуженной заявки из узла 1
            TRANSFER     .8,,Met_2; передача заявки с вероятностью 0,8 в узел 2
            TERMINATE    1; удаление из модели (СеМО) обслуженной заявки класса 1
****************************************
* Модуль 2:  моделирование процесса обслуживания заявок класса 1 в узле 2
Met_2       QUEUE        QUz2_k1;  момент поступления в очередь узла 2
            SEIZE        Uzel_2;  попытка занять прибор узла 2
            DEPART       QUz2_k1;  момент покидания очереди узла 2
            ADVANCE      (Exponential(50,0,20));  обслуживание в узле 2
            RELEASE      Uzel_2;  выход обслуженной заявки из узла 2
            TRANSFER     ,Met_1; безусловная передача транзакта в узел 1
****************************************
* Модуль 3:  поступление и обслуживание заявок класса 2 в узле 1
            GENERATE     (Exponential(10,0,50)); формирование потока заявок класса 2
            QUEUE        QUz1_k2;  момент поступления в очередь узла 1
            ENTER        Uzel_1; попытка занять один из приборов узла 1
            DEPART       QUz1_k2;  момент покидания очереди узла 1
            ADVANCE      10,5; задержка (обслуживание) в узле 1
            LEAVE        Uzel_1; выход обслуженной заявки из узла 1
            TERMINATE    1; удаление из модели (СеМО) обслуженной заявки класса 2
```

В процессе моделирования через разомкнутую СеМО прошло $N_0 = 1000000$ заявок обоих классов. Все обслуженные заявки попали в два блока TERMINATE и были удалены из модели.

По количеству транзактов каждого класса, прошедших через соответствующие блоки ENTER, SEIZE и TERMINATE можно рассчитать коэффициенты передач для заявок класса 1 ($\alpha_1(1), \alpha_2(1)$) и 2 ($\alpha_1(2), \alpha_2(2)$) в узлах 1 и 2 разомкнутой СеМО соответственно:

$$\alpha_1(1) = \frac{1671938}{333403} \cong 5 \quad \text{и} \quad \alpha_2(1) = \frac{1338533}{333403} \cong 4,$$

$$\alpha_1(2) = \frac{666597}{666597} = 1 \quad \text{и} \quad \alpha_2(2) = 0,$$

что соответствует теоретическим значениям, которые могут быть рассчитаны путём решения системы линейных алгебраических уравнений (4.16), как это описано в п.4.4.2.

Загрузки узлов СеМО (**UTIL.**) равны: $\rho_1 = 0{,}476$ и $\rho_2 = 0{,}802$.

```
           START TIME          END TIME  BLOCKS  FACILITIES  STORAGES
                0.000      33345310.868    21        1          1

            NAME                          VALUE
         MET_1                             2.000
         MET_2                             9.000
         QUZ1_1                        10002.000
         QUZ1_2                        10001.000
         QUZ2                          10003.000
         UZEL_1                        10000.000
         UZEL_2                        10004.000

LABEL              LOC  BLOCK TYPE     ENTRY COUNT CURRENT COUNT RETRY
                    1    GENERATE        333406             0       0
MET_1               2    QUEUE          1671938             0       0
                    3    ENTER          1671938             1       0
                    4    DEPART         1671937             0       0
                    5    ADVANCE        1671937             1       0
                    6    LEAVE          1671936             0       0
                    7    TRANSFER       1671936             0       0
                    8    TERMINATE       333403             0       0
MET_2               9    QUEUE          1338533             0       0
                   10    SEIZE          1338533             0       0
                   11    DEPART         1338533             0       0
                   12    ADVANCE        1338533             1       0
                   13    RELEASE        1338532             0       0
                   14    TRANSFER       1338532             0       0
                   15    GENERATE        666597             0       0
                   16    QUEUE           666597             0       0
                   17    ENTER           666597             0       0
                   18    DEPART          666597             0       0
                   19    ADVANCE         666597             0       0
                   20    LEAVE           666597             0       0
                   21    TERMINATE       666597             0       0

FACILITY  ENTRIES  UTIL.   AVE. TIME AVAIL. OWNER PEND INTER RETRY DELAY
UZEL_2    1338533  0.802      19.984  1    999991   0    0     0     0

QUEUE     MAX CONT.  ENTRY ENTRY(0) AVE.CONT. AVE.TIME   AVE.(-0) RETRY
 QUZ1_2     6   0   666597  465509     0.047     2.332      7.732   0
 QUZ1_1    10   1  1671938 1158921     0.120     2.396      7.809   0
 QUZ2      40   0  1338533  272711     3.232    80.521    101.124   0

STORAGE   CAP. REM. MIN. MAX.  ENTRIES  AVL.  AVE.C. UTIL. RETRY DELAY
 UZEL_1     2    0   0    2    2338535   1    0.952  0.476    0    0

CEC XN  PRI         M1      ASSEM  CURRENT  NEXT  PARAMETER    VALUE
999980    0    33344706.455   999980      3      4

FEC XN   PRI         BDT      ASSEM  CURRENT  NEXT  PARAMETER    VALUE
999993    0    33345317.746   999993      5      6
999991    0    33345318.791   999991     12     13
1000005   0    33345330.450   1000005     0     15
1000003   0    33345459.221   1000003     0      1
```

Рис.6.21. Отчет к модели разомкнутой СеМО с неоднородным

Среднее число заявок класса 1 в очереди (**AVE.CONT.**) в узлах 1 и 2 СеМО: $l_1(1) = 0{,}120$ и $l_2(1) = l_2 = 3{,}232$. Среднее число заявок класса 2 в

очереди узла 1 СеМО: $l_1(2) = 0{,}047$. Заметим, что в узле 2 очередь образуют только заявки первого класса. Суммарная длина очереди заявок в узле 1 равна $l_1 = l_1(1) + l_1(2) = 0{,}167$. Суммарное число заявок, находящихся в состоянии ожидания в СеМО: $L = l_1 + l_2 \cong 3{,}4$.

Средние времена ожидания (**AVE.TIME**) заявок класса 1 в узлах 1 и 2 СеМО соответственно равны: $w_1(1) = 2{,}4\,\text{с}$ и $w_2(1) = 80{,}5\,\text{с}$. Среднее время ожидания заявок класса 2 в узле 1 СеМО: $w_1(2) = 2{,}33\,\text{с}$.

Следует заметить, что стандартный GPSS-отчёт по результатам моделирования содержит информацию не по всем характеристикам, которые могут представлять интерес для исследователя. В частности, представленный выше отчёт не содержит информацию о временах пребывания заявок в узлах и в СеМО в целом.

Эти характеристики могут быть рассчитаны на основе имеющихся в отчёте данных с использованием фундаментальных соотношений, представленных в п.3.4.3. Так, например, легко могут быть рассчитаны средние времена пребывания заявок каждого класса в узлах СеМО:

$u_1(1) = w_1(1) + b_1(1) = 17{,}4\,\text{с}$; $\quad u_2(1) = w_2(1) + b_2(1) = 100{,}5\,\text{с}$;

$u_1(2) = w_1(2) + b_1(2) = 12{,}3\,\text{с}$.

С учётом того, что за время нахождения в сети заявки класса 1 в среднем пройдут через узел 1 $\alpha_1(1) = 5$ раз, а через узел 2 – $\alpha_2(1) = 4$ раза, можно найти среднее время пребывания заявок класса 1 в сети:

$U(1) = \alpha_1(1)u_1(1) + \alpha_2(1)u_2(1) = 489\,\text{с}$.

Аналогично, среднее время пребывания в сети заявок класса 2:

$U(2) = \alpha_1(2)u_1(2) + \alpha_2(2)u_2(2) = 12{,}3\,\text{с}$.

Среднее число заявок каждого класса в СеМО может быть найдено по формулам Литтла, связывающим безразмерные и временны*ы*е сетевые характеристики:

$$M(1) = \lambda_0(1)U(1) = \frac{U(1)}{a_0(1)} = 4{,}89 \quad \text{и} \quad M(2) = \lambda_0(2)U(2) = \frac{U(2)}{a_0(2)} \cong 0{,}25.$$

Для получения дополнительных результатов, например в виде гистограмм плотностей распределений времён ожидания и пребывания заявок в СеМО с целью детального анализа свойств исследуемой системы, в область описания GPSS-модели следует включить команды TABLE и QTABLE:

Tw1_k1	QTABLE	QUz1_k1,0,.25,40; *время ожидания в узле 1 заявок класса 1*
Tw1_k2	QTABLE	QUz1_k2,0,.25,40; *время ожидания в узле 1 заявок класса 2*
Tw2_k1	QTABLE	QUz2_k1,0,10,40; *время ожидания в узле 2 заявок класса 1*
TU_k1	TABLE	M1,50,50,40; *время пребывания в сети заявок класса 1*
TU_k2	TABLE	M1,1,1,40; *время пребывания в сети заявок класса 2*

Для двух последних таблиц TU_k1 и TU_k2, в которых накапливается статистика по временам пребывания заявок обоих классов в СеМО, дополнительно в GPSS-модель необходимо вставить два оператора:

```
TABULATE    TU_k1; удаление из модели (СеМО) обслуженной заявки класса 1
TABULATE    TU_k2; удаление из модели (СеМО) обслуженной заявки класса 2
```

Первый оператор должен быть вставлен в модуль 1 перед оператором TERMINATE для отметки времени выхода из СеМО заявки класса 1, а второй – в модуль 3 перед оператором TERMINATE для отметки времени выхода из СеМО заявки класса 2.

В этом случае кроме средних значений временных характеристик могут быть получены значения среднеквадратических отклонений соответствующих характеристик:

$\sigma_{w_1(1)} = 5{,}02$ с; $\sigma_{w_2(1)} = 98{,}90$ с; $\sigma_{w_2(1)} = 4{,}94$ с;

$\sigma_{U(1)} = 737{,}3$ с; $\sigma_{U(2)} = 5{,}72$ с .

6.8. Резюме

1. Универсальным и наиболее эффективным методом исследования сложных систем со стохастическим характером функционирования является имитационное моделирование, предоставляющее возможность исследования систем *любой сложности с любой степенью детализации и получения наиболее полных результатов*.

Имитационная модель представляет собой программу для ЭВМ, реализующую заданное логико-алгоритмическое описание исследуемой системы.

Имитационное моделирование часто называют статистическим, поскольку сбор и обработка результатов имитационного моделирования реализуется методами математической статистики, позволяющими получить результаты в любом объёме – от средних значений и нескольких первых начальных или центральных моментов *до законов распределений*.

К основным процедурам имитационного моделирования относятся:

- организация службы времени;
- сбор и статистическая обработка результатов моделирования;
- выработка (генерирование) случайных величин с заданными законами распределений.

Одна из основных проблем имитационного моделирования – организация службы времени, определяющей способ изменения (продвижения) модельного времени, протекающего в моделируемой системе. В настоящее время наиболее эффективным признан способ «*продвижения модельного времени с переменным шагом до ближайшего события*», реализуемый в большинстве систем имитационного моделирования.

2. Случайные величины в имитационных моделях формируются *программными генераторами* (датчиками), вырабатывающими *псевдослучайные последовательности*, представляющие собой детерминированные числа и обладающие *статистическими свойствами случайных чисел*.

Основными методами формирования равномерно распределённых в интервале (0; 1) псевдослучайных последовательностей являются:

- метод квадратов;
- метод произведений;
- мультипликативный конгруэнтный метод.

Оценка качества генераторов равномерно распределенных псевдослучайных последовательностей проводится с использованием трёх видов проверки:

- на периодичность;
- на случайность;
- на равномерность.

При проверке на случайность программных генераторов *двоичных* псевдослучайных последовательностей используются тесты проверки частот, пар, комбинаций, серий, корреляций.

Псевдослучайные последовательности с заданным законом распределения формируются в ЭВМ на основе программных генераторов равномерно распределённых случайных величин одним из следующих методов:

- аналитическим (обратной функции);
- табличным;
- методом композиций.

3. Общецелевая система имитационного моделирования GPSS World является одной из наиболее доступных и популярных для работы на персональных компьютерах под управлением ОС Windows. GPSS World обладает специальными средствами, которые делают процесс моделирования эффективным и наглядным. GPSS World включает в себя языки программирования GPSS и PLUS и компилятор.

Основными объектами системы имитационного моделирования GPSS World, которые всегда используются при моделировании, являются:

- программа, написанная на языке *GPSS* (*GPSS-модель*),
- исполняемый объект, создаваемый в результате трансляции GPSS-модели и реализующий *процесс моделирования* на ЭВМ,
- *отчёт* с результатами моделирования.

Элементами языка GPSS World являются алфавитно-цифровые символы, имена, метки, переменные, числа, системные числовые атрибуты (СЧА), арифметические операторы, операторы отношения, логические операторы, выражения, процедуры.

Объекты GPSS-модели могут быть разбиты на следующие группы:

- основные объекты (операторы и транзакты);
- оборудование (приборы или одноканальные устройства, памяти или многоканальные устройства, очереди, логические ключи);
- числовые объекты (ячейки, матрицы, переменные, функции, таблицы);
- генераторы случайных чисел (встроенные, библиотечные, табличные);
- групповые списки (списки пользователя, числовые группы,

группы транзактов);

- потоки данных.

Объекты в GPSS-модели могут формироваться автоматически, либо должны объявляться с использованием специальных команд – операторов описания. К объявляемым объектам относятся: памяти, переменные, матрицы, таблицы, функции, а также параметры транзактов.

4. GPSS-модель представляет собой последовательность операторов, описывающих логику работы моделируемой системы, которые могут быть разбиты на две группы: GPSS-операторы и PLUS-операторы.

GPSS-операторы делятся на исполняемые операторы, операторы описания и операторы управления.

Исполняемые операторы, называемые также операторами блоков или просто *блоками*, непосредственно реализуют процесс моделирования.

Операторы описания и *операторы управления*, называемые в GPSS World *командами*, используются соответственно для *описания* многоканальных устройств, переменных, функций, матриц, таблиц и для *управления* процессом моделирования (запуск, остановка и продолжение процесса моделирования, сброс статистики, завершение моделирования и т.п.).

Команды могут быть *срочными и несрочными*.

Оператор GPSS World, в общем случае, содержит 4 поля: поле *метки*, поле *операции*, поле *операндов* и поле *комментариев*.

Все операторы, кроме оператора описания FUNCTION, записываются в одну строку и могут содержать до 250 символов, включая комментарий.

5. Для запуска процесса моделирования используется команда START, которая может находиться в GPSS-модели в качестве последнего оператора или может быть задана после трансляции из меню GPSS World.

Реализация процесса моделирования заключается в перемещении в модели *транзактов, которые* последовательно переходят от блока к блоку в заданной алгоритмом моделирования последовательности.

Транзакты создаются и уничтожаются в модели с помощью специальных операторов: GENERATE и TERMINATE.

В общем случае, *в модели может находиться множество транзактов*, однако *в один и тот же момент времени двигается только один транзакт*.

Транзакт, продвигаемый в модели в данный момент времени, называется *активным*.

Интервал времени, в течение которого транзакт находится в модели, называется *резидентным временем* транзакта.

Интервал времени, в течение которого транзакт проходит между двумя произвольно выбранными точками модели, называется *транзитным временем*.

Каждому транзакту в модели присваивается порядковый номер по мере появления их в модели, начиная с *единицы*.

Работа реальных систем протекает во времени, для отображения которого в GPSS-модели используется *таймер модельного времени*, содержание которого корректируется *автоматически* в соответствии с логикой, предписанной моделью. Единица времени (секунды, минуты, часы или их доли) задается разработчиком модели.

6. Для реализации перемещения транзактов в GPSS-модели используются следующие *списки* (*цепи*):

- *список текущих событий (СТС)*;
- *список будущих событий (СБС)*;
- *списки повторных попыток (СПП)*;
- списки одноканального устройства, включающие *список отложенных прерываний*, *список прерываний*, *список задержки*, *список повторных попыток*;
- списки многоканального устройства, включающие *список задержки*, *список повторных попыток*;
- *списки пользователя*.

В любой модели всегда формируется *один список текущих* и *один список будущих событий*. Остальные списки формируются по мере необходимости.

7. Каждый транзакт имеет множество параметров, называемых *атрибутами транзакта*, к которым относятся:

- *параметры*, число которых не ограничено;
- *приоритет* транзакта;
- *время входа транзакта в систему*;
- *текущий блок*, в котором находится транзакт;
- *следующий блок*, в который должен перейти данный транзакт;
- *список*, в котором находится транзакт в некоторый момент времени.

8. В GPSS World завершение процесса моделирования может быть реализовано:

- принудительно с помощью срочной команды HALT;
- по некоторому условию, задаваемому командой STOP;
- по достижению содержимого «счётчика завершений» значения меньше или равного нулю.

Последний способ используется наиболее часто при моделировании систем и сетей массового обслуживания.

По завершению моделирования формируется и выводится на экран стандартный отчет, содержащий основные результаты моделирования.

9. Переменные, используемые в операндах операторов GPSS и в выражениях, называются *атрибутами*.

Числовые атрибуты, *автоматически поддерживаемые в GPSS* и доступные в течение всего процесса моделирования, называются *системными числовыми атрибутами (СЧА)*. Их значения в любой момент

в процессе моделирования доступны пользователю за счет использования специальных наименований этих атрибутов.

В GPSS используются СЧА трёх типов: СЧА *объектов*, СЧА *системы*, СЧА *транзактов*.

Имя СЧА объектов состоит из двух частей: *группового имени* (идентифицирующего тип объекта и тип информации) и имени или номера *конкретного члена группы*.

10. Встроенная библиотека процедур GPSS World содержит более 20 вероятностных распределений, в том числе равномерное (Uniform), экспоненциальное (Exponential) и др.

Для обращения к вероятностному распределению необходимо указать имя библиотечной процедуры и её параметры, заключённые в круглые скобки и отделённые друг от друга запятой.

Библиотечные процедуры вероятностных распределений могут использоваться в выражениях, а также в качестве операнда A в операторах GENERATE и ADVANCE.

11. В GPSS World имеется 53 операторов блоков, из которых примерно половина используется для построения имитационных моделей простейших систем и сетей массового обслуживания.

Операторы могут быть без операндов или содержать от 1 до 7 операндов, некоторые из которых могут быть необязательными. При отсутствии операндов их значения принимаются по умолчанию. *Отсутствие обязательных операндов приводит к ошибке*.

К операторам *генерирования*, *задержки* и *удаления* транзактов относятся:

- GENERATE (генерирование транзактов);
- ADVANCE (задержка транзакта на заданное время);
- TERMINATE (удаление транзактов из модели).

Операторы *одноканальных устройств* (приборов):

- SEIZE (занятие транзактом прибора);
- RELEASE (удаление транзакта из прибора).

Операторы *многоканальных устройств* (памятей):

- ENTER (вход транзакта в многоканальное устройство);
- LEAVE (удаление транзакта из многоканального устройства).

Операторы *очередей*:

- QUEUE (фиксация момента поступления транзакта в очередь);
- DEPART (фиксация момента удаления транзакта из очереди).

Условные операторы:

- TEST (поверка значения СЧА и передача активного транзакта в блок, отличный от последующего);
- TRANSFER (передача транзакта в блок, отличный от последующего);
- GATE (изменение маршрута движения транзактов в зависимости от состояния некоторого объекта).

Операторы *приоритетного обслуживания*:

- PRIORITY (изменение уровня приоритета активного транзакта);
- PREEMPT (захват прибора поступившим транзактом);
- RETURN (освобождение прибора активным транзактом).

Оператор *логических ключей*:

- LOGIC (изменение состояния логического ключа).

К *прочим* операторам относятся:

- ASSIGN (назначение и изменение параметра транзакта),
- MARK (запись значения абсолютного времени в качестве одного из параметров активного транзакта),
- TABULATE (занесение значений в таблицу – обновление статистики).

12. В GPSS World используются 24 команды (описания и управления), из которых для построения и реализации имитационных моделей *простейших* систем и сетей массового обслуживания оказывается достаточным использование немногим более половины. Команды, как и операторы блоков, могут быть без операндов или содержать от 1 до 5-и операндов, некоторые из которых могут быть необязательными. Значения необязательных операндов при их отсутствии принимаются по умолчанию. *Отсутствие обязательных операндов приводит к ошибке.*

К командам описания относятся:

- FUNCTION (описание функции);
- TABLE (описание таблицы);
- QTABLE (описание таблицы очереди);
- STORAGE (описание ёмкости многоканального устройства);
- VARIABLE (описание арифметической переменной).

К командам управления относятся:

- CLEAR (сброс процесса моделирования в исходное состояние);
- CONTINUE (возобновление прерванного процесса моделирования);
- HALT (прерывает процесс моделирования и очищает очередь команд);
- INCLUDE (вставка в исходную модель и трансляция файла с операторами);
- REPORT (немедленное создание отчета);
- RESET (сброс в ноль статистики и атрибутов системы);
- SHOW (отображает значение выражения в строке состояния окна «Model»);
- START (запуск процесса моделирования);
- STEP (остановка процесса моделирования по определенному количеству входов транзактов в блоки);
- STOP (устанавливает или снимает условие прерывания моделирования).

6.9. Практикум: обсуждение и решение задач

Вопрос 1. Каково соотношение между терминами «имитационное» и «статистическое» моделирование? Эквивалентны ли эти термины?

Обсуждение. Чаще всего подразумевается, что термины «имитационное» и «статистическое» моделирование эквивалентны, и используются как равнозначные термины. Однако, исходя из смыслового содержания этих терминов, всё-таки следует различать, что «имитационное» моделирование означает, что моделирование основано на подражании исследуемому объекту, а «статистическое» моделирование указывает на то, что результаты моделирования накапливаются и обрабатываются методами математической статистики. Эти же результаты могли бы быть получены точно так же на реальной системе путём многократных измерений и их статистической обработки. Таким образом, можно считать, что имитационное моделирование всегда является статистическим. Однако статистическое моделирование не всегда является имитационным. Как показано выше, вычисление определённого интеграла методом Монте-Карло относится к статистическому моделированию, но не является имитационным.

Вопрос 2. Какими достоинствами обладает имитационное моделирование по сравнению с другими методами моделирования?

Обсуждение. Основным достоинством имитационного моделирования является возможность всестороннего детального исследования системы любой сложности и с любой степенью детализации, что невозможно при аналитическом и численном моделировании. Имитационное моделирование позволяет построить математическую модель, максимально приближенную к оригиналу (реальной системе), и, фактически, заменить измерения на реальной системе измерениями на модели. Степень соответствия имитационной модели оригиналу ограничивается только возможностями ЭВМ (производительностью, ёмкостью памяти), на которой проводится моделирование. Естественно, чем сложнее модель, тем более мощной должна быть ЭВМ. Одна из причин, по которой разрабатываются всё более мощные суперЭВМ, – повышенные требования к производительности при решении задач моделирования реальных систем и процессов. Таким образом, можно считать, что имитационное моделирование является универсальным инструментом исследования реальных систем и процессов.

Вопрос 3. Имеют ли результаты имитационного моделирования методическую погрешность и, если да, то чему она равна и как её оценить?

Обсуждение. Имитационному моделированию присущ статистический разброс результатов, который означает, что получаемые значения характеристик имеют методическую погрешность, обусловленную, прежде всего, такими факторами, как длительность моделирования и качество генераторов случайных величин. Наличие методической погрешности проявляется в том, что значения одних и тех же характеристик могут разли-

чаться при использовании в одной и той же модели разных генераторов случайных чисел, а также при различной длительности имитационного эксперимента, причём с увеличением длительности моделирования методическая погрешность уменьшается и лежит обычно в пределах 1-3%.

Следует отметить, что *методическая погрешность различна для разных характеристик*, в чём легко убедиться на следующем гипотетическом примере.

Положим, что в одноканальной СМО с однородным потоком заявок измеряются две характеристики: время ожидания w и время пребывания $u = w + b$ заявок в системе, где b – детерминированная длительность обслуживания заявок в приборе.

Пусть в результате одного эксперимента было получено следующее значение времени ожидания: $w_1 = 10$. Если длительность обслуживания равна $b = 10$, то время пребывания окажется равным $u_1 = 10 + 10 = 20$.

Пусть в результате другого эксперимента было получено значение времени ожидания: $w_2 = 20$. Тогда время пребывания окажется равным $u_2 = 20 + 10 = 30$.

Разница между полученными значениями, представляющая собой погрешность имитационного моделирования, будет составлять:

$$\delta_w = \frac{|w_2 - w_1|}{w_1} 100\% = 100\%, \qquad \delta_u = \frac{|u_2 - u_1|}{u_1} 100\% = 50\%.$$

Итак, погрешность времени ожидания оказалась существенно больше погрешности времени пребывания заявок в системе, что, если подумать, выглядит вполне логично.

Уменьшение методической погрешности имитационного моделирования при использовании качественных генераторов случайных величин достигается за счёт увеличения длительности имитационного эксперимента. При этом некоторые характеристики имеют минимальную погрешность даже при небольшой длительности моделирования (быстрая сходимость результатов к своему истинному значению), в то время как другие характеристики требуют гораздо большей длительности моделирования (медленная сходимость). При моделировании систем и сетей массового обслуживания быстрой сходимостью обычно обладает загрузка, а для времени ожидания характерна медленная сходимость.

Оценить методическую погрешность характеристик моделируемой системы можно «методом срезов», который заключается в следующем. Проводится моделирование длительностью T, и фиксируются полученные на первом срезе значения $\{h_1^{(1)},...,h_N^{(1)}\}$ характеристик (обычно обладающих медленной сходимостью). Затем моделирование продолжается в течение того же времени T, и фиксируются новые полученные на втором срезе значения $\{h_1^{(2)},...,h_N^{(2)}\}$ тех же характеристик. Рассчитывается относительная разность между значениями одноимённых характеристик:

$\delta_i = \frac{\left|h_i^{(2)} - h_i^{(1)}\right|}{h_i} 100\% \quad (i = 1,..., N)$, где в качестве h_i принимается минимальное из двух значений: $h_i = \min(h_i^{(1)}, h_i^{(2)})$ или их среднее значение: $h_i = (h_i^{(1)} + h_i^{(2)})/2$. Значение $\delta = \max(\delta_1,..., \delta_N)$ может рассматриваться как максимальная погрешность имитационного моделирования. Для получения достоверной оценки погрешности рекомендуется выполнить не менее трёх срезов и, если максимальные относительные разности между первым и вторым срезами и между вторым и третьим срезами значительно отличаются, следует продолжить моделирование до тех пор, пока, как минимум, два (а ещё лучше три) соседних среза не дадут приемлемую и примерно одинаковую погрешность.

Вопрос 4. Для чего и каким образом формируются предположения и допущения при разработке модели?

Обсуждение. Предположения и допущения, формируемые в процессе разработки модели, преследуют две цели. Во-первых, это позволяет, во многих случаях, упростить модель и уменьшить её размерность за счёт отбрасывания несущественных факторов и параметров, оказывающих незначительное влияние на процесс функционирования исследуемой системы и, соответственно, на конечные результаты. Во-вторых, при отсутствии каких-либо исходных данных или недостаточно полных сведений о некоторых из них могут и должны вводиться предположения и допущения, позволяющие решить (пусть и упрощённо) поставленную задачу. Действительно, на практике при разработке моделей и исследовании реальных систем зачастую известны только средние значения нагрузочных параметров, представляющих собой случайные величины, и, возможно, их дисперсии. Закон распределения этих параметров обычно не известен. В этом случае для первоначальных оценочных расчётов можно ввести некоторые предположения о законах распределений, позволяющие получить конечные результаты аналитическими методами. При необходимости, дополнительные исследования влияния закона распределения на характеристики функционирования системы могут быть выполнены с использованием имитационного моделирования. Например, если в задаче не оговаривается характер потока заявок и длительности обслуживания заявок в системе массового обслуживания, то может быть введено предположение о том, что поток заявок – простейший, а длительность их обслуживания распределена по экспоненциальному закону. Если не указано количество приборов в узлах сетевой модели, то может быть введено предположение о том, что оно равно 1. В то же время, при решении задач, как это часто бывает на практике, могут иметься «избыточные» исходные данные, которые, вполне возможно, и не нужны для получения результата, поскольку не влияют на характеристики функционирования системы.

Вопрос 5. Если имитационное моделирование является универсальным инструментом исследования, то не значит ли это, что другие методы моделирования не нужны? Или же имитационное моделирование имеет какие-то недостатки?

Обсуждение. Несмотря на универсальность, имитационное моделирование не может полностью заменить другие методы моделирования – аналитические или численные, что обусловлено присущими ему недостатками.

Первый очевидный недостаток связан с высокими требованиями к производительности ЭВМ, на которой проводится моделирование реальных систем, обладающих сложностью и большой размерностью. Естественно, что имитационные модели таких систем представляют собой большие программные комплексы, разработка которых под силу только высококвалифицированным специалистам, владеющим не только приёмами программирования, но и имеющим опыт разработки имитационных моделей, позволяющий разрабатывать модели, учитывающие все существенные особенности структурно-функциональной организации, не перегружая её незначительными подробностями, не влияющими на конечный результат. Всё это делает имитационное моделирование *дорогостоящим* и требующим *значительных временных затрат* как на разработку модели, так и на её реализацию на ЭВМ с учётом того, что для детального исследования свойств системы и получения достоверных результатов необходимо провести большое множество экспериментов, число которых может достигать десятков тысяч.

Последнее связано со вторым недостатком, который заключается в том, что всякий раз при каждом эксперименте результат моделирования получается в точке, то есть справедлив только для заданных в данном эксперименте структурно-функциональных и нагрузочных параметров системы. Поэтому для получения зависимости характеристик системы только от одного параметра требуется провести несколько экспериментов, а для получения доверительного интервала этой зависимости количество экспериментов возрастает многократно. К тому же количество таких параметров может быть значительным.

Третий существенный недостаток имитационного моделирования состоит в невозможности получить приемлемые результаты для систем и сетей массового обслуживания, работающих в области малых (близких к нулю) и больших (близких к единице) загрузок. Действительно, при загрузке системы менее 0,01 вероятность появления очереди очень мала, и в процессе имитационного моделирования даже после прогона достаточно большого числа заявок через систему может оказаться, что ни одна заявка не ждала в очереди, то есть время ожидания будет строго равно нулю. В то же время, точный аналитический расчёт показывает, что время ожидания, хотя и очень маленькое, но не равно нулю. Для того чтобы имитационная модель выдала результат отличный от нуля, возможно потребуется

провести достаточно длительное моделирование, при котором хотя бы одна заявка окажется в состоянии ожидания.

Покажем это на примере простейшей СМО типа M/M/1. Положим, что интенсивность поступления заявок в СМО $\lambda = 0{,}0001\,\text{с}^{-1}$, а длительность обслуживания $b = 100\,\text{с}$. Тогда загрузка системы $\rho = 0{,}001$. В п.5.4.5 показано, что вероятность нахождения в системе *k* заявок определяется по формуле: $p_k = \rho^k(1-\rho)$ $(k = 0, 1, 2, \ldots)$. Тогда вероятность образования очереди (того, что в системе будет две и более заявок) $p_{k>2} = 1 - p_0 - p_1 = 1 - 0{,}99 - 0{,}0099 = 0{,}0001$. Таким образом, для того чтобы в очереди оказалась хотя бы одна заявка, необходимо при имитационном моделировании пропустить через систему более 10 тысяч заявок.

Ещё больше проблем возникает при имитационном моделировании систем, загрузка которых близка к единице. В этом случае практически невозможно получить результат (например, время ожидания заявок в системе), близкий к реальному. Это связано с характером зависимости характеристик СМО от загрузки системы, которая при загрузке, близкой к 1, резко возрастает и стремится к бесконечности (см. рис.4.2).

Действительно, если загрузка той же СМО M/M/1 будет равна $\rho = 0{,}99$ (например за счёт интенсивности $\lambda = 0{,}0099\,\text{с}^{-1}$), среднее время ожидания в соответствии с (4.1) будет равно $w = \dfrac{\rho b}{1-\rho} = \dfrac{0{,}99 * 100}{0{,}01} = 9900\,\text{с}$.

Однако, как сказано выше, имитационному моделированию присущ статистический разброс результатов, в результате которого может оказаться, что в момент завершения процесса моделирования значение загрузки системы будет равно $\rho = 0{,}98$. Тогда среднее время ожидания $w' = \dfrac{\rho b}{1-\rho} = \dfrac{0{,}98 * 100}{0{,}02} = 4900\,\text{с}$, то есть значение времени ожидания будет отличаться более чем в два раза от действительного значения.

Ещё один существенный недостаток, присущий имитационному моделированию, состоит в том, что для сложных систем с большим количеством структурно-функциональных параметров практически невозможно решать задачи оптимального синтеза. Имитационное моделирование позволяет выбрать наилучший вариант структурно-функциональной организации проектируемой системы из нескольких вариантов, но не предоставляет возможностей для решения оптимизационных задач. Для решения этих задач обычно используется аналитическое моделирование.

Задача 1. Для заданной *GPSS-модели*:

а) нарисовать и подробно описать модель исследуемой системы с указанием всех параметров и законов распределений;

б) пояснить, когда (по какому условию) завершится моделирование;

в) определить, существует ли стационарный режим в системе (с

необходимыми обоснованиями, расчетами и пояснениями).

GPSS-модель:

(начало модели)

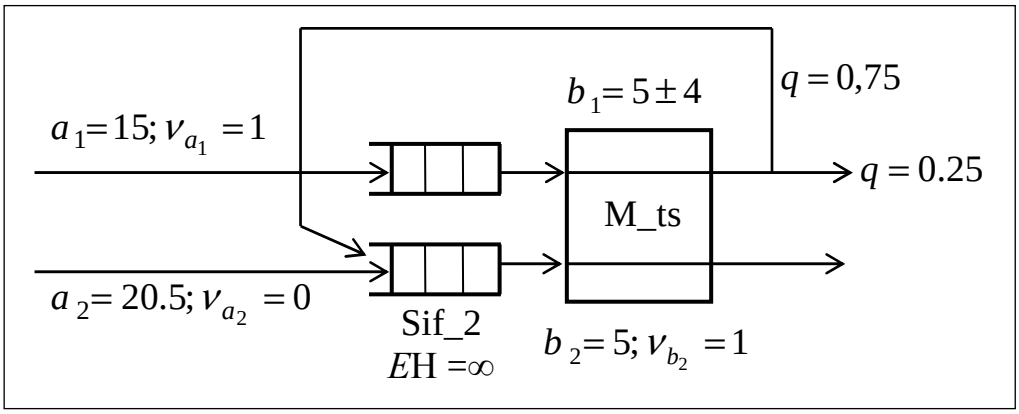
```
GENERATE    (Exponential(2,0,15))
QUEUE       Sif_1
SEIZE       M_ts
DEPART      Sif_1
ADVANCE     5,4
RELEASE     M_ts
TRANSFER    0.25,Noh_1,Noh_2
```

(продолжение модели)

```
            GENERATE    20.5
Noh_1       QUEUE       Sif_2
            SEIZE       M_ts
            DEPART      Sif_2
            ADVANCE     (Exponential(1,0,5))
            RELEASE     M_ts
Noh_2       TERMINATE   2
            START       100000
```

Решение.

а) Наличие двух операторов GENERATE свидетельствует о том, что в моделируемой системе формируется два потока (класса) заявок. Заявки первого класса образуют простейший поток со средним интервалом между заявками 15 единиц времени, а заявки второго класса – детерминированный поток с интервалом 20,5 единиц времени. Формируемые заявки поступают в разные накопители неограниченной ёмкости с именами Sif_1 и Sif_2 соответственно и далее в один и тот же прибор с именем M_ts, где задерживаются на случайное время: заявки класса 1 – на время, равномерно распределённое в интервале (5±4), а заявки класса 2 – на время, распределённое по экспоненциальному закону со средним значением 5 единиц времени. После обслуживания в приборе заявки класса 1 с вероятностью 0,25 направляются к блоку TERMINATE с меткой Noh_2 (удаляются из модели) и с вероятностью 0,75 – к блоку QUEUE с меткой Noh_1 (в накопитель с именем Sif_2) и далее снова попадают в прибор M_ts, где задерживаются на экспоненциально распределённое время со средним значением 5 единиц, то есть обслуживаются уже как заявки класса 2. Таким образом, моделируемая система, показанная на рисунке, представляет собой одноканальную СМО с двумя классами заявок, причём после обслуживания в приборе 75% заявок первого класса переходит во второй класс.



б) Завершение моделирования реализуется оператором TERMINATE и командой START. В момент запуска процесса моделирования в счётчик

завершений заносится значение 100000, указанное в команде START. Всякий раз, когда транзакт (заявка) покидает модель, из счётчика завершений вычитается значение 2, указанное в качестве параметра A оператора TERMINATE. Таким образом, моделирование завершится после обслуживания 50 тысяч заявок.

в) Для того чтобы определить, существует ли стационарный режим в системе, рассчитаем загрузку системы как сумму загрузок, создаваемых заявками классов 1 и 2: $R = \rho_1 + \rho_1$.

Загрузка системы заявками класса 2 рассчитывается как $\rho_2 = \lambda_2 b_2 \approx 0{,}25$, где $\lambda_2 = 1/a_2 = 1/20{,}5 \approx 0{,}05$. Для заявок класса 1 при расчёте загрузки следует учесть, что после первого обслуживания 75% заявок класса 1 остаётся в системе, которые обслуживаются как заявки класса 2 со средним значением $b_2 = 5$ единиц времени. Тогда загрузка системы заявками класса 1 может быть рассчитана как сумма загрузок, создаваемых при первом и втором обслуживании в приборе:

$$\rho_1 = \rho_1^{(1)} + \rho_1^{(2)} = \lambda_1 b_1 + 0{,}75\lambda_1 b_2 = 5/15 + 0{,}75 * 5/15 \approx 0{,}58.$$

Таким образом, загрузка системы $R \approx 0{,}58 + 0{,}05 < 1$, следовательно, система работает без перегрузок, то есть стационарный режим существует.

6.10. Самоконтроль: перечень вопросов и задач

1. Понятия статистического и имитационного моделирования.
2. Основное достоинство имитационного моделирования.
3. Недостатки имитационного моделирования.
4. Основные процедуры имитационного моделирования.
5. По какому принципу осуществляется продвижение модельного времени в имитационной модели.
6. Классификация генераторов случайных величин в зависимости от способа их реализации.
7. Почему величины, вырабатываемые программными генераторами случайных величин, являются псевдослучайными?
8. Перечислить методы генерирования равномерно распределенных случайных величин.
9. Пояснить суть метода квадратов (произведений) для генерирования равномерно распределенных случайных величин.
10. Пояснить суть мультипликативного конгруэнтного метода генерирования равномерно распределенных случайных величин.
11. Понятие длины периода генератора случайных величин.
12. Типы проверок генераторов случайных величин.
13. В чем заключается проверка на периодичность (случайность, равномерность) генераторов случайных величин.
14. Перечислить тесты проверки на случайность генераторов случайных величин.
15. В чем заключается тест проверки частот, пар, комбинаций, серий,

корреляций генераторов равномерно распределённых случайных величин.

16. Перечислить методы генерирования случайных величин с заданным законом распределения.
17. В чем суть аналитического метода (метода обратных функций) генерирования случайных величин с заданным законом распределения?
18. Проиллюстрировать на графике идею аналитического метода генерирования случайных величин с заданным законом распределения?
19. Проиллюстрировать идею аналитического метода (метода обратных функций) генерирования случайных величин на примере экспоненциального закона распределения?
20. Достоинства и недостатки аналитического метода (метода обратных функций) генерирования случайных величин с заданным законом распределения?
21. В чем суть табличного метода генерирования случайных величин с заданным законом распределения?
22. Достоинства и недостатки табличного метода генерирования случайных величин с заданным законом распределения?
23. В чем суть метода генерирования случайных величин с заданным законом распределения, основанного на функциональных особенностях распределений (метод композиций)?
24. Привести примеры генерирования случайных величин с заданным законом распределения, основанного на функциональных особенностях распределений (метод композиций)?
25. Достоинства и недостатки метода генерирования случайных величин с заданным законом распределения, основанного на функциональных особенностях распределений (метод композиций)?
26. Состав системы имитационного моделирования GPSS World.
27. Перечислить элементы языка GPSS World.
28. Классификация объектов системы имитационного моделирования GPSS World.
29. Понятие транзакта.
30. Сколько транзактов может находиться в GPSS-модели одновременно?
31. Сколько транзактов может двигаться в GPSS-модели в один и тот же момент времени?
32. В каких случаях прекращается движение транзакта в GPSS-модели?
33. Какие события в GPSS-моделях массового обслуживания приводят к изменению модельного времени?
34. Какая статистика отражается в стандартном отчёте GPSS-модели?
35. Структура оператора GPSS.
36. Типы GPSS-операторов.
37. В чём отличие операторов блоков от команд?

38. Назначение операторов GENERATE, TERMINATE, ADVANCE, SEIZE, RELEASE, QUEUE, DEPART, ENTER, LEAVE, TEST, TRANSFER, PRIORITY, PREEMPT, RETURN, LOGIC, GATE, MARK, ASSIGN, TABULATE.

39. С помощью какого оператора создаются (уничтожаются) транзакты в GPSS-модели?

40. С помощью каких операторов осуществляется задержка транзакта на определенный период времени (сбор статистики по очередям, изменение маршрута движения транзакта, …) в GPSS-модели?

41. Назначение команд FUNCTION, STORAGE, TABLE, QTABLE, VARIABLE, CLEAR, CONTINUE, HALT, INCLUDE, REPORT, RESET, SHOW, START, STEP, STOP.

42. Задан фрагмент GPSS-модели:

```
GENERATE    20, 10
SEIZE       DIC
ADVANCE     10.5
RELEASE     DIC
TERMINATE
GENERATE    100000
TERMINATE   1
START       10
```

А) Нарисовать и подробно описать модель исследуемой системы (с указанием всех параметров). Б) Пояснить, когда (по какому условию) завершится моделирование. В) Определить, существует ли стационарный режим в системе (с необходимыми обоснованиями, расчетами и пояснениями). Г) Рассчитать среднее число заявок, которые пройдут через систему за время моделирования.

43. Задан фрагмент GPSS-модели:

(начало модели)

```
Met_kom  STORAGE   5
         GENERATE  4.3,1.3
Div_1    ENTER     Met_kom
         ADVANCE   0.5
         LEAVE     Met_kom
         TRANSFER  750, ,Div _2
         TERMINATE
```

(продолжение модели)

```
Div_2    SEIZE     1
         ADVANCE   (Exponential(12,0,4))
         RELEASE   1
         TRANSFER  ,Div _1
         GENERATE  100000
         TERMINATE 2
         START     10
```

А) Нарисовать и подробно описать модель исследуемой системы с указанием всех параметров. Б) Пояснить, когда (по какому условию) завершится моделирование. В) Определить, существует ли стационарный режим в системе (с необходимыми обоснованиями, расчетами и пояснениями). Г) Рассчитать среднее число заявок, которые пройдут через систему за время моделирования.

Заключительный раздел

> «Если кажется, что работу сделать легко, это непременно будет трудно. Если на вид она трудна, значит выполнить ее абсолютно невозможно» ***(Теорема Стакмайера)***

В основе исследования сложных систем с использованием математического моделирования лежит системный подход, конечной целью которого является системное проектирование, направленное на построение системы с заданным качеством. В свою очередь системное проектирование базируется на результатах системного анализа, позволяющего выявить причинно-следственные связи между параметрами и характеристиками исследуемой системы и реализуемого с использованием математических моделей, которые позволяют прогнозировать эффект, достигаемый при изменении структурно-функциональных параметров системы и параметров нагрузки.

Одним из основных требований, предъявляемых к модели, является ее *адекватность* реальной системе, которая достигается за счет использования моделей с различным уровнем детализации, зависящим от особенностей структурно-функциональной организации системы и целей исследования.

Процессы функционирования реальных систем практически невозможно описать полно и детально, что обусловлено существенной сложностью таких систем. Основная проблема при разработке модели состоит в нахождении компромисса между простотой ее описания, что необходимо для её исследования математическими методами, и необходимостью учета многочисленных особенностей, присущих реальной системе. Попытка построить единую универсальную модель сложной системы, несомненно, обречена на неудачу ввиду ее необозримости и невозможности расчета.

Моделирование технических систем в общем случае предполагает выполнение следующих основных этапов:

- формулировка целей моделирования;
- разработка концептуальной модели;
- разработка математической модели;
- параметризация модели;
- выбор методов моделирования;
- выбор средств моделирования;
- проверка адекватности модели (верификация модели);
- проведение экспериментов на модели (расчет характеристик);
- анализ результатов моделирования.

На этапе *определения и формулирования целей моделирования* определяется объект моделирования, формулируются задачи анализа и синтеза, выявляются наиболее важные характеристики, подлежащие

исследованию, формулируются требования к качеству функционирования в виде ограничений, налагаемых на характеристики системы, и формулируется критерий эффективности, определяются требования к точности результатов моделирования и форме их представления.

Основное *назначение концептуальной модели* – выявление наиболее существенных аспектов структурно-функциональной организации, учет которых необходим для получения требуемых результатов. В концептуальной модели обычно в словесной форме приводятся сведения о природе и параметрах элементарных явлений исследуемой системы, о степени их взаимодействия, выявляются параметры, оказывающие наиболее существенное влияние на исследуемые характеристики системы. Одна и та же система может представляться различными концептуальными моделями, которые строятся в зависимости от целей исследования, сформулированных на предыдущем этапе. Например, одна концептуальная модель может отображать временные аспекты функционирования системы, другая – надежностные, третья – масса-габаритные аспекты построения системы.

Концептуальная модель служит основой для разработки математической модели в терминах конкретного математического аппарата.

Создание *математической модели* преследует две основные цели:

1) дать формализованное описание структуры и процесса функционирования системы для однозначности их понимания;

2) попытаться представить процесс функционирования системы в виде, допускающем аналитическое исследование системы с использованием методов, разработанных в рамках данного математического аппарата.

В связи с тем, что состав и номенклатура системных и модельных параметров и характеристик, в общем случае, различается, возникает необходимость установления соответствия между значениями системных и модельных параметров и характеристик, которое выполняется на этапе *параметризации модели*.

Выбор метода моделирования зависит от многих факторов, в том числе от целей моделирования, сложности исследуемой системы, требований к номенклатуре исследуемых характеристик, требований к точности получаемых результатов и т.д. При исследовании и проектировании технических систем, таких как вычислительные системы и сети, наиболее эффективным оказывается использование комбинированного подхода, предполагающего совместное применение аналитических и имитационных методов, что позволяет во многих случаях гарантировать достоверность получаемых результатов. С использованием аналитических методов, применяемых на этапах анализа свойств и синтеза оптимальной системы, решаются задачи, связанные с формированием требований к структурным и функциональным параметрам, обеспечивающим заданное качество функционирования системы, однако получаемые при этом результаты могут иметь значительную погрешность. Имитационные методы, основан-

ные на использовании специализированных языков моделирования, таких как GPSS, позволяют выполнять исследование систем практически любой сложности с любой степенью детализации и должны применяться на заключительном этапе детального анализа спроектированной системы.

Технические и программные *средства моделирования* выбираются с учетом ряда факторов, к которым относятся достаточность и полнота средств для реализации концептуальной и математической модели, доступность средств, простота и легкость освоения технических и программных средств моделирования, наличие методики применения средств для моделирования систем определенного класса. После выбора средств моделирования разрабатывается программная модель.

Проверка адекватности модели исследуемой системе (верификация модели) заключается в анализе ее соответствия исследуемой системе, проявляющегося в близости значений модельных и системных характеристик. Отличие модели от исследуемой системы связана с тем, что обычно модель является упрощенным и идеализированным отображением системы, которое обусловлено идеализацией внешних условий и режимов функционирования, не учитывающей в модели несущественных, по мнению исследователя, факторов и параметров, отсутствием точных сведений о внешних воздействиях и о некоторых конкретных нюансах организации системы, введением ряда упрощающих предположений и допущений. На практике верификация модели обычно проводится путем экспертного анализа разумности результатов моделирования. В случае выявления неадекватности модели исследуемой системе необходимо выполнить корректировку модели.

В процессе проверки адекватности модели необходимо определить область применения модели, то есть оценить диапазон изменения параметров, при котором точность результатов моделирования находится в допустимых пределах.

Исследования на моделях заключаются в *проведении экспериментов*, в процессе которых определяются характеристики системы при разных значениях структурно-функциональных параметров и параметров нагрузки. Большая номенклатура исходных параметров и широкий диапазон их изменения требует предварительного планирования выполняемых на модели экспериментов (расчетов). Планирование направлено на уменьшение количества и длительности экспериментов при условии обеспечения достоверности и полноты результатов моделирования. В случае большой размерности исследуемой системы и многочисленной номенклатуры структурно-функциональных и нагрузочных параметров, изменяющихся в больших пределах, количество экспериментов и соответственно время, затраченное на моделирование, могут оказаться настолько большими, что полученные в конечном счёте результаты потеряют свою актуальность.

Особую значимость планирование экспериментов приобретает при использовании методов имитационного моделирования, характеризующихся большими затратами ресурсов ЭВМ в процессе моделирования.

Одной из основных проблем имитационного моделирования является нахождение компромисса между временем моделирования и затратами памяти ЭВМ, на которой проводится моделирование. Это связано с тем, что имитационное моделирование предъявляет повышенные требования как к производительности, так и к памяти ЭВМ для проведения имитационных экспериментов. Время, затрачиваемое на проведение одного эксперимента с моделью средней сложности даже на высокопроизводительных ЭВМ может достигать нескольких десятков минут и, в некоторых случаях, нескольких часов, а потребность в оперативной памяти ЭВМ – десятков и сотен гигабайт. Причём с увеличением числа проводимых имитационных экспериментов соответственно возрастает время моделирования. Все это обусловливает высокую стоимость имитационного моделирования и требует тщательного планирования имитационных экспериментов с целью сокращения затрат на моделирование.

Анализ результатов моделирования направлен на выявление свойств, присущих исследуемой системе, и включает в себя следующие этапы:

1) обработка результатов для последующего анализа и использования; на этом этапе выделяются наиболее важные, с точки зрения исследователя, результаты, которые представляются в форме, наиболее удобной для изучения свойств исследуемой системы;

2) определение зависимостей характеристик от параметров системы путем варьирования исходных параметров структурно-функциональной организации и нагрузки с целью выявления и формулирования свойств исследуемой системы;

3) принятие решения о работоспособности исследуемой системы и выработка рекомендаций по наиболее эффективной и рациональной организации проектируемой или модернизируемой системы, которые могут быть использованы в дальнейшем при решении задач синтеза в процессе системотехнического проектирования.

Синтез оптимальной системы направлен на построение системы, наилучшим образом соответствующей своему назначению. Решение задачи синтеза связано с определением зависимостей характеристик функционирования системы от параметров, которые представляются сложными математическими конструкциями. При этом возможность получения приемлемых результатов в процессе решения задач синтеза из-за их сложности и большой трудоемкости с учетом специфических особенностей реальных систем превосходит возможности математических методов оптимизации, и задача синтеза в общем виде оказывается математически неразрешимой. Для того, чтобы снизить сложность задачи синтеза, процесс проектирования разделяют на последовательность этапов, на каждом из которых решаются частные задачи синтеза – определяются параметры, связанные с отдельными аспектами структурно-функциональной организации системы, с использованием тех или иных моделей.

Приложение 1

ИСПОЛЬЗУЕМЫЕ АББРЕВИАТУРЫ

АП	абсолютный приоритет
БМ	базовая модель
БП	бесприоритетное обслуживание
ВС	вычислительная система
ДБ	дисциплина буферизации
ДО	дисциплина обслуживания
ЗСВО	закон сохранения времени ожидания
ЗСВП	закон сохранения времени пребывания
ЗСеМО	замкнутая сеть массового обслуживания
ИМ	имитационная модель
ИММ	иерархическое многоуровневое моделирование
КВ	коэффициент вариации
МК СМО	многоканальная система массового обслуживания
ММО	модель массового обслуживания
МП	матрица приоритетов
ОК СМО	одноканальная система массового обслуживания
ОП	относительный приоритет
ОПП	обслуживание в порядке поступления
ООП	обслуживание в обратном порядке
ОР	обслуживание по расписанию
ОСП	обслуживание в случайном порядке
ОЦП	обслуживание в циклическом порядке
РСеМО	разомкнутая сеть массового обслуживания
СеМО	сеть массового обслуживания
СМО	система массового обслуживания
СП	смешанный приоритет
ЧП	чередующийся приоритет
СИМ	общецелевая система имитационного моделирования
СЧА	системные числовые атрибуты
СТС	список текущих событий
СБС	список будущих событий
СПП	списки повторных попыток
ЭВМ	электронная вычислительная машина
FIFO	First In First Out
GPSS	General Purpose Simulation System
LIFO	Last In First Out
PLUS	Programming Language Under Simulation
SNA	System Numerical Attributes

Приложение 2

ОСНОВНЫЕ ОБОЗНАЧЕНИЯ

t_i – i-й момент времени

τ_f – интервал времени f ($f \subset \{a, b, c, w, x, z, v, u\}$)

a – время между заявками во входящем потоке

b – длительность обслуживания заявки

c – время между заявками в выходящем потоке

x – время ожидания начала обслуживания

z – время ожидания в прерванном состоянии

w – полное время ожидания: $w=x+z$

v – время нахождения заявки на обработке: $v=z+b$

u – время пребывания заявки в системе: $u=w+b$

$F(\tau)$ – функция распределения случайной величины τ_f

$f(\tau)$ – плотность распределения случайной величины τ_f:

$$f(\tau) = F'(\tau)$$

$F^*(s)$ – преобразование Лапласа плотности распределения $f(\tau)$

$X^*(z)$ – производящая функция распределения $p_k = \mathrm{P}(X = k)$

$f^{(n)}$ – n-й начальный момент распределения $F(\tau)$

$\hat{f}^{(n)}$ – n-й центральный момент распределения $F(\tau)$

$f = M[\tau_f]$ – математическое ожидание случайной величины τ_f:

$$f = M[\tau_f] = f^{(1)}$$

D_f – дисперсия случайной величины τ_f: $D_f = \hat{f}^{(2)}$

σ_f – среднее квадратическое отклонение случайной величины τ_f:

$$\sigma_f = \sqrt{D_f}$$

ν_f – коэффициент вариации случайной величины τ_f: $\nu_f = \sigma_f / f$

ν_a – коэффициент вариации интервалов τ_a во входящем потоке

ν_b – коэффициент вариации длительности обслуживания τ_b

ν_c – коэффициент вариации интервалов τ_c в выходящем потоке

λ – интенсивность потока заявок

λ' – интенсивность потока обслуженных заявок

λ'' – интенсивность потока потерянных заявок

Λ – суммарная интенсивность объединённого потока заявок

λ_0 – производительность замкнутой СеМО

π_n – вероятность потери заявки

π_0 – вероятность обслуживания заявки

l – средняя длина очереди заявок в СМО

L – суммарная длина очереди заявок в СеМО

m – средняя число заявок в СМО

M – суммарное число заявок в СеМО

H – количество классов заявок

n – число узлов в СеМО

K – число обслуживающих приборов в СМО

Q – матрица приоритетов: $Q = [q_{ij}\,(i, j = \overline{1,H})]$

q_{ij} – элемент матрицы приоритетов, принимающий значения: 0 – нет приоритета), 1 – приоритет относительный и 2 – приоритет абсолютный;

$r_g(i,k)$ – коэффициенты, позволяющие выделить классы заявок с одинаковым соотношением приоритетов

μ – интенсивность обслуживания: $\mu = 1/b$

θ – средняя ресурсоемкость обслуживания заявки

y – нагрузка: $y = \lambda b$

Y – суммарная нагрузка

ρ – коэффициент загрузки (загрузка): $\rho = \min(y/K;1)$

R – суммарная загрузка

α_j – коэффициент передачи j-го узла СеМО

p_{ij} – вероятность передачи заявки из узла i в узел j СеМО

q_{ij} – вероятность перехода марковского процесса с дискретным временем из состояния i в состояние j

g_{ij} – интенсивность перехода марковского процесса с непрерывным временем из состояния i в состояние j

$\mathbf{Q} = [q_{ij} \mid i, j = \overline{1,n}]$ – матрица вероятностей перехода марковского случайного процесса с дискретным временем

$\mathbf{G} = [g_{ij} \mid i, j = \overline{1,n}]$ – матрица интенсивностей перехода марковского случайного процесса с непрерывным временем

$p_i(t)$ –вероятность того, что марковский случайный процесс в момент времени t находится в состоянии i

$p_i(0)$ – начальная вероятность – вероятность того, что марковский случайный процесс в момент времени $t = 0$ находится в состоянии i

$\mathbf{P} = (p_1, \ldots, p_n)$ – вектор стационарных вероятностей состояний

S – стоимость

Обозначения СМО в символике Кендалла:

A/B/N/L–обозначение СМО,

где:

A – закон распределения интервалов времени между моментами поступления заявок в систему;

B – закон распределения длительности обслуживания заявок в приборе;

N – число обслуживающих приборов в системе ($N = 1, 2, ..., \infty$);

L – число мест в накопителе ($L = 0, 1, 2, ...$) (отсутствие **L** означает, что накопитель – неограниченной ёмкости).

Для задания законов распределений **A** и **B** используются следующие обозначения:

G (General) – произвольное распределение общего вида

M (Markovian) – экспоненциальное (показательное) распределение

D (Deterministik) – детерминированное распределение

U (Uniform) – равномерное распределение

$\mathbf{E_k}$ (Erlangian) – распределение Эрланга *k*-го порядка (с *k* последовательными одинаковыми экспоненциальными фазами)

$\mathbf{h_k}$ (hipoexponential) – гипоэкспоненциальное распределение *k*-го порядка (с *k* последовательными разными экспоненциальными фазами)

$\mathbf{H_r}$ (Hiperexponential) – гиперэкпоненциальное распределение порядка *r* (с *r* параллельными экспоненциальными фазами)

g (gamma) – гамма-распределение

P (Pareto) – распределение Парето

Приложение 3

ВОПРОСЫ ДЛЯ ОБСУЖДЕНИЯ

Ниже представлен перечень вопросов, обсуждение которых можно найти в указанных разделах пособия.

Раздел 1:

1. Можно ли персональный компьютер рассматривать как систему, элементами которого являются системный блок и связанные с ним внешние устройства – монитор, принтер и сканер?

2. Насколько велико различие между «параметрами» и «характеристиками» системы? Могут ли характеристики быть параметрами и наоборот?

3. Являются ли синонимами термины «показатель эффективности» и «характеристика»?

4. Сколько критериев эффективности используется при синтезе оптимальной системы?

5. В литературе часто встречается такое понятие как «многокритериальная задача». Означает ли это, что задача оптимального синтеза может решаться с использованием сразу нескольких критериев эффективности?

6. Можно ли систему, работающую в неустановившемся режиме, исследовать методами, разработанными для установившегося режима?

7. Каким способом достигается разумный компромисс между простотой и адекватностью модели?

8. Каково значение параметризации модели в процессе исследования реальной системы?

9. Насколько необходим детальный анализ спроектированной системы?

10. Если, как сказано выше, статистические (имитационные) методы исследования сложных систем являются универсальными, то насколько актуально применение аналитических методов?

11. В некоторых литературных источниках вместо понятия «оптимальная система» используется понятие «рациональная система». Каково соотношение между этими двумя понятиями?

12. В чем различие между понятиями «синтез» и «проектирование»?

Раздел 3:

1. Почему математическая модель называется абстрактной?

2. Насколько предположение о простейшем характере потока заявок соответствует реальности?

3. Когда оправдано использование предположения о простейшем характере потока заявок?

4. Почему в СМО с накопителем неограниченной емкости, работающей без перегрузок, возникают очереди? В каких случаях они не возникают?

5. Что в реальной системе может служить основанием для того, чтобы в соответствующей математической модели заявки были разделены на разные классы?

Раздел 5:

1. Существуют ли реальные системы, в которых протекающие в них случайные процессы являются марковскими?

2. Когда случайный процесс с непрерывным временем не обладает эргодическим свойством?

3. Обладает ли эргодическим свойством случайный процесс с непрерывным временем, имеющий бесконечное число состояний?

Раздел 6:

1. Каково соотношение между терминами «имитационное» и «статистическое» моделирование? Эквивалентны ли эти термины?

2. Какими достоинствами обладает имитационное моделирование по сравнению с другими методами моделирования?

3. Имеют ли результаты имитационного моделирования методическую погрешность и, если да, то чему она равна и как её оценить?

4. Для чего и каким образом формируются предположения и допущения при разработке модели?

5. Если имитационное моделирование является универсальным инструментом исследования, то не значит ли это, что другие методы моделирования не нужны? Или же имитационное моделирование имеет какие-то недостатки?

Список литературы

«При печатании книги в нее всегда вкрадывается несколько ошибок, которые никто не заметит» (***Закон публикаций Джоунса***)

1. Авен О.И., Гурин Н.Н.. Коган Я.А. Оценка качества и оптимизация вычислительных систем. – М.: Наука, 1982. – 464 с.

2. Алиев Т.И. Математические методы теории вычислительных систем. – Л.: ЛИТМО, 1979. – 92 с.

3. Алиев Т.И. Исследование методов диспетчеризации в цифровых управляющих системах. Уч. пособие. – Л.: ЛИТМО, 1986. – 82 с.

4. Бражник А.Н. Имитационное моделирование: Возможности GPSS World. – СПб.: Реноме, 2006. – 439 с.

5. Венцель Е.С. Исследование операций: задачи, принципы, методология. – М.: Наука, 1980. – 408 с.

6. Жожикашвили В.А., Вишневский В.М. Сети массового обслуживания. Теория и применение к сетям ЭВМ. – М.: Радио и связь, 1988. – 192 с.: ил.

7. Кельтон В., Лоу А. Имитационное моделирование. Классика CS. 3-е изд. – СПб.: Питер; Киев: Издательская группа BHV, 2004. – 847 с.: ил.

8. Клейнрок Л. Теория массового обслуживания. Пер. с англ. – М.: Машиностроение, 1979.

9. Клейнрок Л. Вычислительные системы с очередями. – М.: Мир, 1979. – 600 с.

10. Липаев В.В., Колин К.К., Серебровский Л.А. Математическое обеспечение управляющих ЭВМ. – М.: Советское радио, 1972. – 528 с.

11. Основы теории вычислительных систем / С.А.Майоров, Г.И.Новиков, Т.И.Алиев, Э.И.Махарев, Б.Д.Тимченко. – М.: Высшая школа, 1978. – 408 с.

12. Рыжиков Ю.И. Теория очередей и управление запасами: Учебник для вузов. – СПб.: Питер, 2001 год. – 384 с.

13. Советов Б.Я., Яковлев С.А. Моделирование систем: Учебник для вузов. – 4-е изд., стер. – М.: Высшая школа, 2005. – 343 с.: ил.

14. Столингс В. Современные компьютерные сети. – СПб.: Питер, 2003. – 783 с.: ил.

15. Феррари Д. Оценка производительности вычислительных систем. – М.: Мир, 1981.

16. Шварц М. Сети ЭВМ. Анализ и проектирование: Пер. с англ./ Под ред. В.А.Жожикашвили. – М.: Радио и связь, 1981. – 336 с.: ил.

17.Шнепс М.А. Системы распределения информации. Методы расчета. - М.: Связь, 1979.

18. Шрайбер Т.Дж. Моделирование на GPSS. – М.: Машиностроение, 1980.

АЛФАВИТНЫЙ УКАЗАТЕЛЬ

«Если вам непонятно какое-то слово в техническом тексте, не обращайте на него внимания. Текст полностью сохраняет смысл и без него» ***(Закон Купера)***

СОДЕРЖАНИЕ