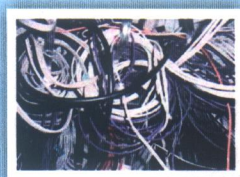




Ж.Ж. Ниеталина, Г.М. Байкенова

МОДЕЛИРОВАНИЕ СИСТЕМ И СЕТЕЙ ЭЛЕКТРОСВЯЗИ



**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РЕСПУБЛИКИ КАЗАХСТАН**

**КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ имени К. И. САТПАЕВА**

**Ж. Ж. Ниеталина
Г. М. Байкенова**

**МОДЕЛИРОВАНИЕ СИСТЕМ И СЕТЕЙ
ЭЛЕКТРОСВЯЗИ**

**Рекомендовано Научно-методическим советом
Университета в качестве учебного пособия**

Алматы 2011

УДК 621.39(075.8)

ББК 32.88я73

Н 55

Н 55 **Ниеталина Ж.Ж., Байкенова Г.М.** Моделирование систем и сетей электросвязи: Учеб. пособие. – Алматы: КазНТУ, 2011 – 124 с. Ил.15. Табл.3. Библиогр. – 15 назв.

ISBN 978–601–228–299–3

В учебном пособии дается систематизированное изложение вопросов, отражающие основные положения теории моделирования в телекоммуникационных системах и сетях электросвязи, ориентированные на практическое использование.

Первая глава посвящена определению моделирования и его видов, а также основным принципам моделирования потоков событий. Основы статистического моделирования изложены во второй главе. В третьей главе рассматривается теория телетрафика как раздел теории массового обслуживания. Приводятся примеры различных моделей СМО, содержится материал по анализу систем массового обслуживания классическими методами теории вероятности. В четвертой главе рассмотрены основные вопросы по измерению и вычислению трафика.

Предназначено для бакалавров, обучающихся по специальности 050719 – «Радиотехника, электроника и телекоммуникация».

УДК 621.39(075.8)

ББК 32.88я73

Рецензенты:

Н. С. Заурбеков, зав. каф. «Прикладная информатика»

КазЭУ им. Т. Рыскулова, канд. физ.- мат. наук, доц.

А. А. Айдосов, КазАТК им. М.Тынышпаева, д-р техн. наук, проф.

А.С.Ермаков, зав. каф. «Вычислительная техника»

КазНТУ им. К.И. Сатпаева, канд. техн. наук, проф.

Печатается по плану издания Министерства образования и науки Республики Казахстан на 2011г.

ISBN 978–601–228–299–3

© Ж. Ж. Ниеталина, 2011

Г. М. Байкенова, 2011

© КазНТУ, 2011

ВВЕДЕНИЕ

Основным методом исследований телекоммуникационных систем является моделирование. При проектировании, модернизации, развитии и эксплуатации телекоммуникационной сети возникают задачи оптимизации структуры сети и применяемых алгоритмов управления, а также прогноза производительности сети и оценки качества обслуживания для услуг разных видов. Для количественных ответов на эти вопросы применяются методы теории массового обслуживания. В данной дисциплине изложены теоретические основы количественного описания многих понятий, связанные с качеством телекоммуникационных услуг.

Основой измерения и прогнозирования качества работы телекоммуникационной сети является теория телетрафика – движения информационных потоков (сообщений, заявок). Основоположителем этой научной дисциплины был датский ученый Агнер Краруп Эрланг. Им была впервые выведена формула для расчета доли вызовов, получающих обслуживание на сельской телефонной станции, и кому придется ожидать, пока делаются внешние вызовы.

Теория телетрафика является разделом теории систем массового обслуживания, или теории очередей (Queueing Theory). И действительно, огромное количество информации, сообщений, заявок, поступающих на вход системы, должны быть обслужены, однако ресурсов для одновременного обслуживания всех запросов недостаточно. Возникает конфликт, следствием чего являются задержки или отказы в обслуживании некоторых запросов. Стремление понять объективные причины этих задержек или отказов и по возможности уменьшить их воздействие является побудительным мотивом развития теории массового обслуживания.

Методами моделирования решаются задачи анализа и синтеза телекоммуникационных систем. Задачей анализа является оценка эффективности функционирования СМО

(действующей ТС), сравнение реальных характеристик с проектными, предоставление объективных оценок качества работы системы. Синтез позволяет находить оптимальные параметры системы.

Основными методами решения задач в теории телетрафика являются аналитические, численные и метод статистического моделирования.

Математическая модель телекоммуникационной системы представляет собой математическое описание процессов функционирования системы. Методологически математическое моделирование делится на аналитическое и имитационное. В данном учебном пособии мы будем рассматривать аналитическое моделирование, которое реализуется с помощью алгебраических, дифференциальных и других уравнений, а также предусматривающие осуществление однозначной вычислительной процедуры, приводящей к их точному решению. Аналитические модели носят обычно вероятностный характер и строятся на основе понятий аппарата теорий массового обслуживания, вероятностей и марковских процессов, а также методов диффузной аппроксимации. В первой главе учебного пособия можно ознакомиться с моделями данных для описания телетрафика, основными понятиями модели потоков событий. Вторая глава посвящена методам статистического моделирования. В третьей главе рассмотрены вопросы аналитического моделирования ТС методами теории систем массового обслуживания. Четвертая глава посвящена моделированию телетрафика, определению параметров нагрузки сети, методам оценки параметров трафика мультисервисной сети.

Данное учебное пособие составлено для ознакомления студентов с основными положениями теории моделирования и использованием их для решения задач анализа и синтеза разнообразных сложных систем.

1. МОДЕЛИРОВАНИЕ. МОДЕЛИ ДАННЫХ ДЛЯ ОПИСАНИЯ ТЕЛЕТРАФИКА

1.1. Моделирование. Виды моделирования

При проектировании систем и сетей электросвязи особую значимость занимает моделирование систем и сетей.

Модель – это прототип, копия объекта, системы, какой-либо технологии, не всегда точная до мелочей, но благодаря которой мы можем иметь представление о нашем исследуемом предмете. Модель может воспроизводить, имитировать «моделируемое» устройство для каких-либо научных, образовательных целей. Модель системы и сети электросвязи может представить собой количественное воплощение определенной функции телекоммуникационной системы и сети связи: количество абонентов и заявок, протяженность маршрута, стоимость и объем трафика, пропускная способность сети. Модель может математическими уравнениями, графом переходов состояний и табличным способом описать процесс передачи информации, а также процесс управления перегрузками сети и задержкой пакетов в сети.

Что же такое моделирование?

Моделирование – построение и изучение моделей реально существующих предметов, процессов или явлений с целью получения объяснений этих явлений, а также для предсказания явлений, интересующих исследователя.

Существуют различные виды моделирования, но их выбор зависит от характера моделируемого объекта, его предназначения, сферы приложения. Например, можно выделить следующие виды моделирования:

- компьютерное моделирование;
- математическое моделирование;
- имитационное моделирование;
- статистическое моделирование;
- физическое моделирование;
- экономико-математическое моделирование и т. д..

Процесс моделирования включает три элемента:

- субъект (исследователь);
- объект исследования;
- модель, определяющую (отражающую) отношения познающего субъекта и познаваемого объекта.

Первый этап построения модели предполагает наличие некоторых знаний об объекте-оригинале. Не всегда есть возможность полностью получить копию объекта, а особенно какого-либо сложного процесса, поэтому для одного объекта может быть построено несколько «специализированных» моделей, концентрирующих внимание на определенных сторонах исследуемого объекта или же характеризующих объект с разной степенью детализации.

На втором этапе модель выступает как самостоятельный объект исследования. На данном этапе можно проводить различные эксперименты, при которых сознательно изменяются условия функционирования модели и систематизируются данные об ее «поведении». Конечным результатом этого этапа является множество (совокупность) знаний о модели.

На третьем этапе осуществляется перенос знаний с модели на оригинал – формирование множества знаний. Одновременно происходит переход с «языка» модели на «язык» оригинала.

Четвертый этап – практическая проверка получаемых с помощью моделей знаний и их использование для построения обобщающей теории объекта, его преобразования или управления им.

Моделирование – циклический процесс. Это означает, что за первым четырехэтапным циклом может последовать второй, третий и т.д. При этом знания об исследуемом объекте расширяются и уточняются, а исходная модель постепенно совершенствуется. Недостатки, обнаруженные после первого цикла моделирования, можно исправить в последующих циклах.

Рассмотрим коротко некоторые виды моделирования.

Компьютерное моделирование является одним из эффективных методов изучения сложных систем. Часто компьютерные модели проще и удобнее исследовать, они позволяют проводить вычислительные эксперименты, реальная

постановка которых затруднена или может дать непредсказуемый результат.

Математическое моделирование – это упрощенное описание реальности с помощью математических понятий.

Физическое моделирование – метод экспериментального изучения различных физических явлений, основанный на их физическом подобии.

Аналитическое моделирование – аналитическими называются модели реального объекта, использующие алгебраические, дифференциальные и другие уравнения, а также предусматривающие осуществление однозначной вычислительной процедуры, приводящей к их точному решению. Аналитические модели носят обычно вероятностный характер и строятся на основе понятий аппарата теорий массового обслуживания, вероятностей и марковских процессов, а также методов диффузной аппроксимации. Могут также применяться дифференциальные и алгебраические уравнения. В то же время аналитические модели имеют ряд существенных недостатков, к числу которых следует отнести:

- значительные упрощения, свойственные большинству аналитических моделей (представление потоков заявок как простейших, предположение об экспоненциальном распределении длительностей обслуживания заявок, невозможность обслуживания заявок одновременно несколькими приборами, например процессором и оперативной памятью, и др.);

- громоздкость вычислений для сложных моделей, например, использование для представления процесса функционирования сети дифференциальных уравнений Колмогорова требует (для установившегося режима) решения сложной системы алгебраических уравнений;

- сложность аналитического описания вычислительных процессов в сети.

Указанные особенности позволяют заключить, что аналитические методы имеют самостоятельное значение лишь при исследовании процессов функционирования вычислительных сетей в первом приближении и в частных, достаточно специфичных задачах.

Имитационное моделирование – это метод, позволяющий строить модели, описывающие процессы таким образом, если бы они проходили в действительности. Математическая модель реализуется в виде программы для компьютера. Проводятся различные эксперименты с моделью. По результатам этих экспериментов можно получить достаточно устойчивую статистику, позволяющую определять изменения в процессе функционирования реального объекта.

Имитационное и аналитическое моделирование – это частный случай математического моделирования.

Существуют явления и ситуации в сети и системах электросвязи, когда модель нельзя задать количественно или математическим соотношением. Тогда используются экспериментальные данные, а способ обработки данных в моделировании называется статистическим.

Статистическое моделирование – численный метод решения математических задач, при котором искомые величины представляют вероятностными характеристиками какого-либо случайного явления, это явление моделируется, после чего нужные характеристики приближенно определяют путем статистической обработки «наблюдений» модели.

Все методы моделирования помогают найти оптимальное управление систем и сетей электросвязи, а также их объектов с помощью применения компьютерного моделирования.

Целью моделирования систем электросвязи является создание систем и сетей высокой надежности с необходимым уровнем качества предоставляемых услуг, оптимальной взаимосвязи различного оборудования и программного обеспечения. В настоящее время при проектировании сетей электросвязи применяется многоуровневый подход, когда все множество модулей системы разбивают на функциональные уровни, а принцип организации интерфейсов называется открытым. При передаче информационных сообщений участники обмена должны согласовать уровни и форму сигналов; способ определения длины сообщений (пакетов); методы контроля

достоверности передачи данных. Соглашение принимается от низшего уровня модели взаимодействия систем – передачи битов, до высшего – сервиса для пользователей сети. Особое место при моделировании систем и сетей электросвязи занимают протоколы связи, которые реализуют аппаратно и программно через концентраторы, маршрутизаторы, коммутаторы и др. Они обеспечивают уровень качества обслуживания при обмене и передаче данных (обычная и электронная почта, факс или нарочный способ). На практике стремятся использовать стандартные протоколы связи и интерфейсы: фирменные, национальные и международные стандарты ISO.

При проектировании сетей электросвязи используют модульность построения систем. По виду обрабатываемых сигналов на входе модель ориентирована на цифровые и аналоговые системы. По характеру временной организации работы моделируют синхронные и асинхронные системы. По количеству выполняемых функций применяют однофазные и многофазные системы моделирования. По способу обслуживания моделируют одноканальные и многоканальные системы. Основным признаком построения систем и сетей электросвязи является наличие нескольких центров сбора и обработки информации, т. е. по архитектуре технической организации они – распределенные.

Моделирование учитывает тенденцию слияния информационных технологий:

- локальных информационных сетей с глобальной сетью;
- вычислительных, телефонных и телевизионных информационных технологий для передачи различной информации (звука, текста, дейтаграммы, видеоизображения, рисунков, речи и т. д.).

Методы для оценки производительности сети делятся на две категории: методы измерения и прогнозирующие методы; в свою очередь прогнозирующий метод включает математический анализ и моделирование.

Методы измерения требуют наличия реальной сети для экспериментов. При этом естественно будут учтены абсолютно все детали этой сети, ее характеристики. Однако имеются некоторые ограничения. Сеть, работающая и приносящая доход, не может быть использована в крайней производительности. Экспериментальная сеть может быть ограничена имеющимся количеством и типом источников трафика, таким образом, ограничивая диапазон реальных экспериментальных условий.

Моделирование системы по управлению объектом прогнозирует состояние телекоммуникаций и связи, а также вырабатывает и принимает необходимое решение по достижению заданной цели. Прогнозирование может быть перспективным, годичным, квартальным, декадным и сменно-суточным. Здесь используются программные средства, реализующие экономико-математические методы.

При сравнительном анализе и моделировании, требуется соблюдать основные факторы – точность результатов, время на подготовку выводов и полную стоимость использования метода (включает разработку и использование).

Как правило, средства моделирования сети вычисляют ее производительность на основе показателей ее фактического и оцениваемого трафика, указываемых администратором сети. Другим подходом к моделированию сети является создание вариантов сценария работы, что позволяет программировать уровень трафика на основе действий сетевых приложений. Моделирование позволяет получить лишь ту точность, которую дают базовые данные. Если при измерении трафика не охвачен адекватный диапазон сетевой активности или неверны оценки роста объема трафика, генерируемого новым приложением, получить реалистичное описание производительности невозможно. Средства моделирования способны показать, каким образом изменения могут повлиять на производительность, но интерпретировать данные, разрабатывать план устранения "узких" мест и готовить сценарии для проверки этих планов должен администратор сети.

1.2. Системы и сети электросвязи – объекты моделирования. Основные термины и определения

Современные телекоммуникационные системы и сети представляют собой сложный комплекс технических средств, которые обеспечивают передачу различных видов сообщений на любые расстояния и с заданными параметрами качества. Средства связи (технические и программные) используются для формирования, приема, обработки, хранения, передачи сообщения.

Телекоммуникационные системы и телекоммуникационные сети, взаимодействуя друг с другом, образуют *систему электросвязи* – комплекс технических средств, обеспечивающих электросвязь определенного вида.

Архитектура сети телекоммуникаций является иерархической структурой и включает в себя три уровня:

- 1) системы (службы) электросвязи;
- 2) вторичные сети связи, которые обеспечивают транспортировку, коммутацию, распределение сигналов в службах электросвязи;
- 3) первичные сети, снабжающие вторичные сети каналами передачи и физическими цепями.

Сеть первичных каналов связи является первым (нижним, физическим) уровнем системы передачи данных на котором располагаются первичные каналы связи, оканчивающиеся цифровыми интерфейсами на каналообразующем оборудовании (модеме). Этот уровень представлен каналами связи на базе медных линий связи, волоконно-оптических линий связи и радиотрасс различного частотного диапазона (спутниковые и радиорелейные каналы связи). Назначением каналов связи является передача электрических, оптических и радиосигналов в соответствии с физической природой, используемой в канале связи среды передачи сигналов.

Как известно, сети связи имеют территориальное деление: магистральный участок, внутризоновый и местный.

Магистральная первичная сеть обеспечивает соединение между собой типовых каналов передачи и сетевых трактов

разных внутризоновых первичных сетей общего пользования на всей территории страны.

Внутризональная первичная сеть – технологически сопряженные междугородные сети телекоммуникаций, образуемые в пределах территории одной или нескольких областей.

Местные первичные сети на участке «местный узел – оконечное устройство» являются сетями доступа.

На основе типовых каналов передачи и физических цепей первичной сети с помощью узлов и станций коммутации организуются различные вторичные сети.

Основным требованием к работе единой сети телекоммуникаций является обеспечение показателей надежности сетей и систем связи в соответствии с требованиями МСЭ-Т и нормативно-технических документов.

К основным требованиям также относятся:

1) преемственность наборов услуг, предоставляемых различными системами и сетями (службами) в составе единой сети телекоммуникаций, которая должна быть обеспечена при переходе одного поколения систем к следующему;

2) технологическая совместимость систем и сетей, которая подразумевает согласованность физических и протокольных уровней передачи информации для различных систем, а также обеспечение необходимой надежности и качества связи, включая гарантированный доступ к определенной части сетевых ресурсов;

3) согласованность по адресации и управлению;

4) согласованность организационно-технических решений;

5) экономическая согласованность.

Надежность – свойство объекта сохранять во времени в установленных пределах значения всех параметров, характеризующих способность выполнять требуемые функции в заданных режимах и условиях применения, технического обслуживания, хранения и транспортирования (примечание – надежность является комплексным свойством, которое в зависимости от назначения объекта может включать безотказность, ремонтно-пригодность и долговечность или сочетание этих свойств).

Безотказность – свойство объекта непрерывно сохранять работоспособное состояние в течение некоторого времени или наработки.

Ремонтопригодность – свойство объекта, заключающееся в приспособленности к поддержанию и восстановлению работоспособного состояния путем технического обслуживания и ремонта, при условии соблюдения условий эксплуатации объекта, а также условий проведения процедур технического обслуживания и ремонта.

Долговечность – свойство объекта сохранять работоспособное состояние до наступления предельного состояния при установленной системе технического обслуживания и ремонта.

Показатель надежности – количественная характеристика одного или нескольких свойств, составляющих надежность объекта. Надежность является вероятностной характеристикой (случайной и изменяющейся в пределах $0 \leq P \leq 1$). Математически, надежность – это вероятность безотказной работы системы или сети. Мировой уровень надежности сети $0,98 \div 0,99$. Надежность называется *структурной*, если работоспособность рассматривается в зависимости от способа соединения узлов и элементов системы или структуры, в частности надежность зависит от топологии и структуры сети.

Для описания надежности отдельных устройств используются такие показатели надежности, как *среднее время наработки на отказ*, *вероятность отказа*, *интенсивность отказов*. Эти показатели применяются для оценки надежности простых элементов и устройств, которые при отказе какого-либо своего компонента переходят в неработоспособное состояние. В сложных системах, при отказе какого-нибудь одного из компонентов, может сохраняться работоспособность всей системы.

Интенсивность отказов – условная плотность вероятности отказа невозстанавливаемого объекта, определяемая для рассматриваемого времени при условии, что до этого момента отказ не возник.

(Средняя) наработка на отказ – отношение суммарной наработки восстанавливаемого объекта к математическому

ожидаанию числа его отказов в течение этой наработки (ГОСТ 27.002) или вероятность того, что в пределах заданной наработки отказ объекта не возникнет.

Безотказность – свойство объекта непрерывно сохранять работоспособное состояние в течение некоторого времени или наработки.

Доступность – означает долю времени, в течение которого система находится в работоспособном состоянии. Доступность является долговременной статистической характеристикой, т. к. определяется на большом промежутке времени. Например, доступность коммуникационного оборудования телефонных сетей равна 0,99999, что означает чуть более 5 минут простоя в год.

Еще одной характеристикой надежности является отказоустойчивость.

Отказоустойчивость – способность системы скрывать от пользователя отказ отдельных ее элементов. В отказоустойчивой системе, отказ одного из элементов приводит к некоторому снижению качества ее работы, но не к полной остановке работы.

Для повышения надежности объектов используется резервирование.

Резервирование – метод повышения надежности объекта введением дополнительных элементов и функциональных возможностей сверх минимально необходимых для нормального выполнения объектом заданных функций.

Приведем еще несколько определений, используемых при составлении нормативных документов:

Канал электросвязи – совокупность технических средств и среды распространения, обеспечивающая передачу сообщений соответствующего вида электросвязи. Различают следующие каналы электросвязи: телефонный, телеграфный, телевизионный, факсимильный, звукового вещания, передачи данных.

Цепь (пара) – совокупность тока ведущих жил (проводов), предназначенных для передачи электрических сигналов.

Линия связи – совокупность однотипных и разнотипных систем передачи ЕАСС, имеющие общую среду распространения, линейные сооружения и устройства их обслуживания.

Эксплуатационная готовность – свойство объекта, отражающее как появление отказов, так и время неисправного состояния. Свойство эксплуатационной готовности является сочетанием свойств безотказности и ремонтпригодности объекта.

Средний коэффициент готовности – усредненное на заданном интервале времени значение нестационарного коэффициента готовности.

Отказ – событие, заключающееся в прекращении способности объекта исполнять требуемую функцию (примечание – после отказа объект переходит в неработоспособное состояние).

Кабельные линии связи – совокупность кабелей, кабельной арматуры, кабельных сооружений, обеспечивающих устойчивую передачу электрических сигналов между абонентами сети.

Причина отказа – явления, процессы, события и состояния, вызвавшие возникновение отказа объекта.

Последствия отказа – явления, процессы, события и состояния, обусловленные возникновением отказа объекта.

Восстановление – процесс обнаружения и устранения отказа (повреждения) с целью восстановления его работоспособности (исправления).

Вероятность восстановления – вероятность того, что фактическая продолжительность восстановления работоспособности объекта не превысит заданной.

Интенсивность восстановления – условная плотность вероятности восстановления работоспособности объекта, определенная для рассматриваемого момента времени, при условии, что до этого момента восстановление не было завершено.

Показатель технической эффективности функционирования объекта – мера качества собственно функционирования объекта или целесообразности использования объекта для выполнения заданных функций.

Коэффициент технического использования – отношение средней наработки объекта в единицах времени за некоторый

период эксплуатации к сумме средних значений наработки, времени простоя, обусловленного техническим обслуживанием, и времени ремонтов за тот же период эксплуатации.

Срок службы – календарная продолжительность эксплуатации объекта от начала или возобновления после капитального или среднего ремонта до наступления предельного состояния.

Методы обеспечения качества обслуживания (QoS) занимают сегодня одно из важнейших мест в работе сетей с коммутацией пакетов, т. к. без их применения невозможна работа мультимедийных приложений, таких как IP-телефония, видеоконференцсвязь, дистанционное обучение и т. д. Эти методы используют такие параметры, как скорость передачи данных, задержка пакетов, потеря пакетов. Атрибутом сетей с коммутацией пакетов являются очереди в буфере у каждого коммутатора пакетов. Буферизация пакетов представляет собой основной механизм поддержания пульсирующего трафика, но с другой стороны, это означает определенные задержки при передаче пакетов через сеть, что очень негативно может сказаться для трафика, чувствительного к задержкам. Следовательно, надо искать компромисс между стремлением полностью загрузить сеть и выполнением требований QoS одновременно для всех видов трафика.

При проектировании, модернизации телекоммуникационной сети возникают задачи оптимизации структуры сети и применяемых алгоритмов управления, а также прогноза производительности сети и оценки качества обслуживания для услуг разных видов. Для количественных ответов на эти вопросы применяются методы теории массового обслуживания.

Методами моделирования решаются задачи анализа и синтеза телекоммуникационных систем. Задачей анализа является оценка эффективности функционирования системы, сравнение реальных характеристик с проектными, предоставление объективных оценок качества работы системы. Синтез позволяет находить оптимальные параметры системы.

1.3. Модели потоков событий

Если посмотреть на распечатку работы монитора телефонной сети, то можно увидеть, что моменты времени поступления пакетов или занятия линий представляют некоторый набор случайных чисел. Эти числа образуют случайную неубывающую последовательность, которую принято называть случайным потоком. Для описания таких потоков используют либо понятие распределения количества событий, приходящихся на определенный интервал времени, либо распределение интервала времени между соседними событиями. В первом случае модель потока дается распределением дискретной случайной величины, а во втором – непрерывной случайной величины.

Таким образом *потоком событий* (в данном случае заявок) называется последовательность событий, наступающих одно за другим в какие-то заранее неизвестные, случайные моменты времени. Вид и параметры закона распределения входящего потока определяется характером физических процессов, протекающих в моделируемом объекте. Как показывает опыт, для составления системы моделей в основном используются следующие распределения входящего потока: показательное, Эрланга k -ого порядка, Релея, нормальное, равномерное. Случайный характер потока заявок и длительности их обслуживания порождает в СМО случайный процесс.

Случайным процессом (или случайной функцией) называется соответствие, при котором каждому значению аргумента (в данном случае – моменту из промежутка времени проводимого опыта) ставится в соответствие случайная величина (состояние СМО). Случайной величиной называется величина, которая в результате опыта может принять одно, но неизвестное заранее, какое именно, числовое значение из данного числового множества.

Примерами потоков событий могут быть: поток вызовов на телефонной станции, поток сбоев компьютера, поток клиентов в банке и т. д.

Классификация потоков

Входящий поток во многом определяет характеристики производительности функционирования системы. Поэтому правильное описание потока запросов, поступающих в случайные моменты времени, и идентификация его параметров являются весьма важной задачей.

Во входящем случайном потоке запросы поступают в систему в некоторые случайные моменты времени $t_1, t_2, \dots, t_n$.

Интенсивностью потока называют математическое ожидание числа событий в единицу времени в данный момент, т. е. это предел отношения среднего числа событий на интервале Δt к длине этого интервала, стремящейся к нулю:

$$\Lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{\bar{N}(t, t + \Delta t)}{\Delta t}.$$

Параметром потока событий называется предел отношения вероятности поступления хотя бы одного события на интервале $(t, t + \Delta t)$ к длине этого интервала, стремящейся к нулю:

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P_{i \geq 1}(t, t + \Delta t)}{\Delta t}.$$

Если интенсивность потока зависит от t – поток *нестационарный*; если интенсивность потока не зависит от t – поток *стационарный*. Последнее означает на содержательном уровне, что распределение числа запросов, поступивших на некотором интервале времени, зависит от длины этого интервала, но не зависит от расположения этого интервала на временной оси.

Когда вероятность поступления более одной заявки за малый интервал времени t есть величина более высокого порядка малости по сравнению с длиной интервала, говорят об *ординарности* потока, т. е. практическая невозможность одновременного поступления двух и более запросов за время Δt .

Случайный поток является потоком *без последдействия*, если число заявок, поступивших на непересекающихся

интервалах времени, являются независимыми в совокупности случайными величинами.

Случайный поток называется потоком с *ограниченным последствием*, если $\tau_k = t_k - t_{k-1}$, $k \geq 1$ (длина интервала между моментами поступления k -ого и $k-1$ -го запросов) независимы в совокупности.

Случайный поток называется *рекуррентным*, если поток является потоком с ограниченным последствием и величины τ_k одинаково распределены.

Моделирование широкого класса потоков, обладающих свойствами стационарности, ординарности и ограниченного последствия служит для определения моментов появления событий, интервалы между которыми являются случайными величинами. Такими потоками являются простейший поток, потоки Пальма и Эрланга.

Простейший поток называется простейшим, если он обладает свойствами: стационарности, ординарности и отсутствия последствия. Распределение промежутков между вызовами для простейшего потока подчиняется экспоненциальному закону. Вероятность поступления ровно k вызовов за период длительностью t определяется распределением Пуассона:

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

Простейший поток называется также пуассоновским, т. к. интервалы между событиями в этом потоке имеют пуассоновское распределение. Пуассоновские потоки широко применяются в качестве модели реальных потоков благодаря тому, что обладают важным свойством аддитивности. Если взять N источников пуассоновских потоков событий с интенсивностями λ_k , и рассмотреть новый поток как суммарный поток всех событий от этих источников, то результирующий поток окажется также пуассоновским с интенсивностью

$\lambda_{\Sigma} = \sum_{k=1}^N \lambda_k$. Можно также расщеплять пуассоновский поток на

m потоков с интенсивностями λ_k .

Примитивный поток. Этот поток можно считать нестационарным пуассоновским потоком с параметром потока, зависящим от состояния системы, на которую поступает данный поток. Чаще всего модель примитивного потока используется при описании обслуживания нескольких независимых одинаковых пуассоновских источников одной системы. Суммарный пуассоновский поток на входе системы при этом может быть описан суммой потоков каждого из источников. Суммарная интенсивность определяется суммой интенсивностей каждого из потоков. При получении обслуживания события от каждого из таких источников, данный источник исключается из числа создающих нагрузку на систему.

Модель примитивного потока удобна для представления абонентской нагрузки на телефонный коммутатор. Каждый абонент является источником независимого пуассоновского потока заявок. Совокупная нагрузка на коммутатор определяется суммой таких потоков. Однако, когда абонент получает обслуживание своего звонка, его поток исчезает и интенсивность входного потока уменьшается скачком.

Поток Пальма. Однородный поток, обладающий свойствами стационарности, ординарности и ограниченного последствия, называется потоком Пальма. Интервалы между событиями являются независимыми случайными величинами

Если случайные события распределены одинаково, то поток Пальма называется стационарным. Простейший стационарный пуассоновский поток является потоком Пальма, нестационарный пуассоновский поток не является потоком Пальма.

Поток Эрланга. Потоком Эрланга k -го порядка называется поток, полученный из простейшего потока путем операции «разрежения», когда отбрасывают из потока $(k - 1)$ точек

поряд, сохраняя каждое k -ое событие. Простейший поток есть поток Эрланга нулевого порядка.

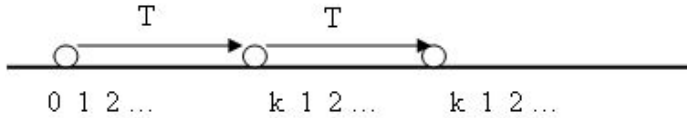


Рис. 1. Поток Эрланга

Промежуток между двумя соседними событиями в потоке Эрланга k -го порядка есть случайная неотрицательная величина T с плотностью распределения ($t > 0$):

$$f(k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}.$$

Таким образом, при моделировании потока Эрланга выбирается заявка только определенного порядка.

Заданные потоки событий, в которых одно событие следует за другим через строго определенные промежутки времени, называются *регулярными* потоками.

Потоки событий с произвольным законом распределения
Рассмотрим ряд распределений случайной величины, которая может описывать интервал времени между поступлениями событий. Каждое такое распределение порождает собственный вид потока событий.

Равномерное распределение. Непрерывная случайная величина x , принимающая значения на отрезке $[a, b]$, распределена равномерно, если ее плотность распределения $P(x)$ и функция распределения $F(x)$ имеют соответственно вид:

$$P(x) = \begin{cases} 0, & x \notin [a, b] \\ \frac{1}{b - a}, & x \in [a, b] \end{cases}.$$

$$F(x) = \begin{cases} 0, & x \leq a \\ x - a & a < x < b \\ b - a & x > b \end{cases}.$$

Математическое ожидание m и дисперсия σ

$$m = \frac{a+b}{2},$$

$$\sigma^2 = \frac{(b-a)^2}{12}.$$

Нормальное распределение (распределение Лапласа-Гаусса).

Случайная величина X нормально распределена с параметрами m и $\sigma > 0$, если плотность и функция распределения имеют вид

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}},$$

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{(z-m)^2}{2\sigma^2}} dz.$$

При $m = 0$ и $\sigma = 1$ случайная величина X имеет стандартное нормальное распределение.

Обычно нормальное распределение используется при больших числах испытаний.

Биномиальное распределение (формула Бернулли). Возникает в тех случаях, когда ставится вопрос: сколько раз происходит некоторое событие в серии из *определенного небольшого числа независимых наблюдений* (опытов), выполняемых в одинаковых условиях.

Для удобства и наглядности будем полагать, что нам известна величина p – вероятность того, что вошедший в магазин посетитель окажется покупателем и $(1 - p) = q$ – вероятность того, что вошедший посетитель не окажется покупателем. Если X – число

покупателей из общего числа n посетителей, то вероятность того, что среди n посетителей оказалось k покупателей, равна

$$P(x = k) = \frac{n!}{k!(n-k)!} \cdot p^k q^{n-k} = C_n^k p^k q^{n-k},$$

где $k = 0, 1, 2, \dots, n$.

Эта формула называется формулой Бернулли. При большом числе испытаний биномиальное распределение стремится к нормальному.

Распределение Пуассона. Как говорилось ранее, простейший поток является пуассоновским. Он играет важную роль в ряде вопросов физики, теории связи, теории надежности, теории массового обслуживания и т. д. Всюду, где в течение определенного времени может происходить случайное число каких-то событий (радиоактивных распадов, телефонных вызовов, отказов оборудования, несчастных случаях и т. п.). Если вероятность появления в единичном испытании интересующего нас события очень мала, то при больших n числах испытаний биномиальное распределение переходит не в нормальное, а в распределение Пуассона.

Случайное число событий, происшедших за время от 0 до t , распределено по закону Пуассона. Вероятность поступления точно k вызовов интервал времени $(0, t)$ составит

$$Pk(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

Вероятность поступления не менее k вызовов за время $(0, t)$ равна

$$P_{i \geq k}(T) = \sum_{i=k}^{\infty} \frac{(\lambda t)^i}{i!} e^{-\lambda t}.$$

Пример 1.1.

На коммутационную систему поступает простейший поток вызовов с интенсивностью $\lambda = 4$. Определить:

1) вероятность поступления точно 0,1,2,3 вызовов за время $t = 1$;

2) вероятность поступления не менее 4-х вызовов за то же время.

Решение:

Вероятность поступления точно k вызовов:

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

$$k = 0, \quad t = 1,$$

$$P_0(t) = \frac{(\lambda t)^0}{0!} \cdot e^{-\lambda t} = \frac{4^0}{0!} e^{-4} = 0.01832,$$

$$k = 1, \quad t = 1,$$

$$P_1(t) = \frac{(\lambda t)^1}{1!} \cdot e^{-\lambda t} = \frac{4^1}{1!} e^{-4} = 0.07326,$$

$$P_2(t) = 0.14652, \quad P_3(t) = 0.19537.$$

Вероятность поступления не менее 4-х вызовов за время $t = 1$:

$$P_{t \geq 4} = \sum_{i=4}^{\infty} \frac{(\lambda t)^i}{i!} e^{-\lambda t} = \sum_{i=4}^{\infty} \frac{4^i}{i!} e^{-4} = 0.05665.$$

1.4. Основные принципы моделирования потоков событий

В основе моделирования потока событий лежит генерирование значений действительной (реже целой) переменной, имеющей смысл текущего модельного времени, которые определяют моменты поступления требований. Обозначим эту переменную *TIME*. Чаще всего генерируются не сами моменты, а интервалы времени между происшедшим событием и

поступлением следующего. Если обозначить последовательность таких межсобытийных интервалов $T_1, T_2, \dots, T_b, \dots$, то на каждом шаге моделирования

$$TIME = TIME + T_i.$$

Величина межсобытийного интервала является случайной величиной и генерируется с помощью специальных программных функций, называемых датчиками случайных чисел.

Наиболее эффективными алгоритмами работы датчиков случайных чисел являются так называемые конгруэнтные генераторы. Алгоритм их работы может быть описан следующими соотношениями:

$$x_{i+1} = (Ax_i + C) \bmod(m)$$

или

$$x_{i+1} = (xx_i) \bmod(m).$$

Здесь A , C , x_b , x – целые числа, а в качестве псевдослучайного числа используется действительное $U_i = x_i/m$. Операции вычисления по модулю ($\bmod$) обеспечиваются работой алгоритма в компьютерах с конечной разрядностью.

Приведем пример работы такого алгоритма при следующих параметрах: пусть

$$x_i = (3x_{i-1}) \bmod(32),$$

т. к. $32 = 2^5$, здесь используются пятиразрядные машинные слова для представления целых чисел. Пока эти числа не превышают 31, результаты операций очевидны. При превышении этого значения будет использоваться остаток от деления на целое, кратное 32.

$x_0 = 1$	00001	$U = 1/32 = 0.03125$
$x_1 = 3$	00011	$U = 3/32 = 0.09375$
$x_2 = 9$	01001	$U = 9/32 = 0.28125$
$x_3 = 27$	11011	$U = 27/32 = 0.84375$

$x_4 = 81 \bmod 32 = 17$	10001	$U = 17/32 = 0.53125$
$x_5 = 51 \bmod 32 = 19$	10011	$U = 19/32 = 0.59375$
$x_6 = 57 \bmod 32 = 25$	11001	$U = 25/32 = 0.78125$
$x_7 = 75 \bmod 32 = 11$	01011	$U = 11/32 = 0.34375$

Если продолжить вычисление, то на следующем шаге будет получена единица и цикл замкнется. Далее последовательность нельзя будет считать псевдослучайной. Ясно, что чем больше период, тем последовательность лучше отражает свойства случайного потока. Нетрудно убедиться, что период будет зависеть как от выбора m , так и от начального значения x .

Генератор случайных чисел запускается один раз за временной интервал, и далее единожды для каждого прибытия ячеек в течение временного интервала. В подходе с продвижением по дискретному событию, модель “ячейка-за-ячейкой” вызывает RNG (Random number generator) единожды на прибытие ячейки, чтобы сгенерировать время между прибытиями для следующей ячейки. На практике применяются различные генераторы с различными параметрами. Например, генераторы с именами функций URN13 (Unified Random Number) и URN38. Для первого период составляет 1 073 741 824 с алгоритмом

$$x_i = 663608941x_{i-1} \bmod 2^{32}.$$

Для второго

$$x_i = 5^{13}x_{i-1} \bmod 32768.$$

Системы моделирования систем массового обслуживания обычно включают такие датчики внутри функций, описывающих потоки событий.

Рассмотрим прикладные аспекты моделирования случайных событий, и в частности, его применение к очереди ячеек и пакетов, а также моделирование дискретного времени. Рассмотрим моделирование дискретной версии M/D/1 очереди, применимой для изучения буферизации АТМ - ячеек. Имеются два основных способа моделирования такой очереди:

- 1) продвижение по дискретному времени;
- 2) продвижение по дискретному событию.

Ранее, модель шла от момента времени i к моменту времени $i+1$ независимо от того, изменилось состояние системы или нет, например, если очередь M/D/1 пуста в момент i , то она может быть пустой и в момент $i+1$, и программа просто перейдет к моменту $i+1$. Эти моменты могут соответствовать тактам ячейки в АТМ. В моделировании с дискретным случайным событием, часы модельного времени производят отсчет всякий раз, когда имеется изменение состояния модели, к примеру, прибытие или отбытие ячеек из очереди M/D/1. Так что при моделировании есть выбор: продвижение по дискретному времени или по дискретному событию. Последний вариант может выполняться быстрее, потому что пропускает передачу тактов, когда очередь пуста, но первый проще для понимания в контексте АТМ, поскольку его проще выполнить, и он моделирует буфер ячеек с точки зрения обслуживающего процесса, т. е. “конвейер” тактов ячеек.

2. МЕТОД СТАТИСТИЧЕСКОГО МОДЕЛИРОВАНИЯ

2.1. Статистическое моделирование. Метод Монте-Карло

Метод статистического моделирования – это способ исследования поведения вероятностных систем (экономических, технических и т. д.) в условиях, когда не известны в полной мере внутренние взаимодействия в этих системах. На практике часто имеют дело со сложными моделями, когда их аналитическое описание громоздко или невозможно. Например, в сетях передачи данных с пакетной коммутацией могут быть такие сетевые протоколы, оценка соответствующих показателей которых может быть осуществлена только при помощи статистического моделирования. Или, если взять сложную сеть связи, информация по которой передается в соответствии с динамическими или адаптивными алгоритмами управления, т. е. направление передачи сообщения из транзитного узла определяется не жесткими правилами, а всем текущим состоянием сети: длинами очередей на соседних узлах связи, времени занятости каналов, характеристиками самого данного сообщения и т. п. В этом случае легче осуществить имитацию реального процесса на статистической модели.

Статистическим моделированием [1] называется численный метод решения математических задач при помощи моделирования структур, процессов с использованием случайных величин и случайных последовательностей величин с последующей статистической оценкой различных характеристик получаемой совокупности реализаций. Этот метод заключается в воспроизведении исследуемого физического процесса при помощи вероятностной математической модели и вычислении характеристик этого процесса. Одно такое воспроизведение функционирования системы называется реализацией или испытанием. После каждого испытания регистрируют совокупность параметров, характеризующих случайный исход реализации. Метод основан на многократных испытаниях построенной модели с последующей статистической обработкой

полученных данных, с целью определения числовых характеристик *рассматриваемого* процесса в виде статистических оценок его параметров. Если для аналитического исследования объекта необходимо записывать формулы или какие-либо рекуррентные соотношения, то при статистическом моделировании допустимо задание в виде взаимосвязей элементов системы, таблицы их отношений друг к другу по входу-выходу и т. д.

Метод статистического моделирования сложных систем на ЭВМ сводится к имитации процесса обслуживания коммутационной системой поступающего потока вызовов, в результате которой можно получить статистические данные для дальнейшего исследования. Случайные величины и их последовательности генерируются определенным образом и представляют собой задачу, имеющую вполне конкретное решение в зависимости от специфики исследуемой задачи.

Статистическое моделирование предполагает следующую схему вычисления (оценивания) искомой величины. Так, искомую величину представляют математическим ожиданием числовой функции f от случайного исхода ω некоторого явления:

$$E(f(\omega)) = \int f(\omega) dP, \quad (2.1.1)$$

т. е. интегралом по вероятностной мере P .

Для того чтобы оценить некоторое значение, необходимо подобрать случайную величину так, чтобы ее математическое ожидание равнялось искомому значению. После этого можно пронаблюдать случайную величину и оценить по выборке ее математическое ожидание. Полученный результат можно считать *оценкой искомого значения*.

Рассмотрим оценку математического ожидания случайной величины

$$\bar{Q} = \overline{E(f(\omega))} = \frac{f(\omega_1) + \dots + f(\omega_N)}{N}, \quad (2.1.2)$$

где $\omega_1, \dots, \omega_N$ – исходы, состоявшие в результате наблюдений.

Оценку (2.1.2) можно рассматривать как квадратурную формулу для указанного интеграла со случайными узлами ω_k и случайной погрешностью R_N .

Таким образом, рассматриваемая схема состоит в проведении серии экспериментов. Каждый i -ый эксперимент представляет собой получение случайного исхода ω_i и вычисление функции $f(\omega_i)$. После этого производятся вычисления по формуле (2.1.2) и полученный результат считается *оценкой искомой величины*.

Случайный выбор на каждом этапе проводится с помощью *случайных чисел*, которые необходимо генерировать тем или иным образом. Так, они могут генерироваться каким-либо физическим датчиком или имитироваться при помощи вычислительной техники по некоторому алгоритму, обеспечивающему заданное распределение (*псевдослучайные числа*).

В реальных условиях функционирования СМО имеются переходные режимы, а входящие и исходящие потоки требований являются далеко не простейшими. В этих условиях для оценки качества функционирования систем обслуживания широко используется метод *статистических испытаний (метод Монте-Карло)*. Метод Монте-Карло – это численный метод решения математических задач при помощи моделирования случайных величин.

Возникновение этого метода стало возможным благодаря появлению ЭВМ, т. к. моделировать случайные величины вручную трудоемко. Согласно этому методу составляется программа для осуществления одного испытания. Затем это испытание повторяется N раз, причем каждый опыт не зависит от всех остальных, и результаты всех опытов соединяются. Поэтому метод Монте-Карло иногда называется методом статистических испытаний, и используется он в основном в тех случаях, когда результат нужен с небольшой точностью (5–10 %).

Моделирование случайных процессов осуществляется с помощью моделирования случайных величин, подчиняющихся различным законам распределения: равномерному, показательному, нормальному и т. д. Для получения таких величин используется случайная величина X , равномерно распределенная на отрезке $(0,1)$, из которой различными преобразованиями получают случайную величину, подчиняющуюся требуемому закону распределения [2].

Случайная величина X называется равномерно распределенной на отрезке $[0,1]$, если ее плотность $f(x)$ на этом отрезке постоянна и равна единице:

$$f(x) = \begin{cases} 1 & 0 \leq x < 1 \\ 0 & 0 > x > 1 \end{cases},$$

а функция распределения $F(x)$ случайной величины X будет:

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}.$$

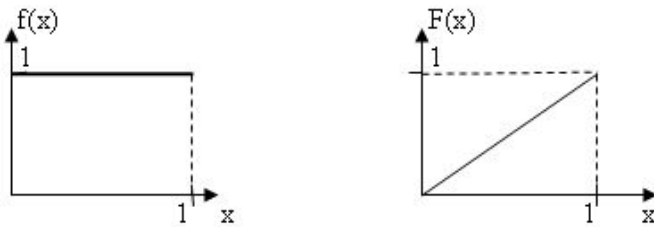


Рис. 2. Плотность и функция распределения случайной величины X

Представим отрезок $[0,1]$ линией, образующей окружность. Тогда случайная величина X окажется равномерно распределенной по длине окружности. Получим аналогию с игрой в рулетку. В связи с этим разнообразные модели, равномерно

распределенной случайной величины часто называются рулеткой, а метод статистических испытаний получил название метода Монте-Карло (название курорта в княжестве Монако).

Основой решения задачи исследования функционирования СМО в реальных условиях является статистическое моделирование входящего потока требований и процесса их обслуживания (исходящего потока требований). Для решения задачи статистического моделирования функционирования СМО должны быть заданы следующие исходные данные:

- описание СМО (тип, параметры, критерии эффективности работы системы);
- параметры закона распределения периодичности поступлений требований в систему;
- параметры закона распределения времени пребывания требования в очереди (для СМО с ожиданием);
- параметры закона распределения времени обслуживания требований в системе.

Решение задачи статистического моделирования функционирования СМО складывается из следующих этапов:

1. Вырабатывают равномерно распределенное случайное число ξ_i .

2. Равномерно распределенные случайные числа преобразуют в величины с заданным законом распределения:

- интервал времени между поступлениями требований в систему (Δt_n);
- время ухода заявки из очереди (для СМО с ограниченной длиной очереди);
- длительность времени обслуживания требования каналами (Δt_w).

3. Определяют моменты наступления событий:

- поступление требования на обслуживание;
- уход требования из очереди;
- окончание обслуживания требования в каналах системы.

4. Моделируют функционирование СМО в целом и накапливают статистические данные о процессе обслуживания.

5. Устанавливают новый момент поступления требования в систему, и вычислительная процедура повторяется в соответствии с изложенным.

6. Определяют показатели качества функционирования СМО путем обработки результатов моделирования методами математической статистики.

Методику решения задачи рассмотрим на примере моделирования СМО с отказами.

Пусть система имеет два однопоточных канала, работающих с отказами, причем моменты времени окончания обслуживания на первом канале обозначим через t_{1i} , на втором канале – через t_{2i} . Закон распределения интервала времени между смежными поступающими требованиями задан плотностью распределения $f_1(t_r)$. Продолжительность обслуживания также является случайной величиной с плотностью распределения $f_1(t_0)$.

Процедура решения задачи будет выглядеть следующим образом:

1. Вырабатывают равномерно распределенное случайное число ξ_i .

2. Равномерно распределенное случайное число преобразуют в величины с заданным законом распределения, используя формулы табл. 1. Определяют реализацию случайного интервала времени (Δt_n) между поступлениями требований в систему.

3. Вычисляют момент поступления заявки на обслуживание: $t_i = t_{i-1} + \Delta t_n$.

4. Сравнивают моменты окончания обслуживания предшествующих заявок на первом $t_{1(i-1)}$ и втором $t_{2(i-1)}$ каналах.

5. Сравнивают момент поступления заявки t_i с минимальным моментом окончания обслуживания (допустим, что $t_{1(i-1)} < t_{2(i-1)}$):

– если $[t_i - t_{1(i-1)}] < 0$, то заявка получает отказ и вырабатывают новый момент поступления заявки описанным способом;

– если $[t_i - t_{2(i-1)}] \geq 0$, то происходит обслуживание.

6. При выполнении условия 5 определяют время обслуживания i -й заявки на первом канале Δt_{1i} путем пре-

образования случайной величины ξ_i в величину (время обслуживания i -й заявки) с заданным законом распределения.

7. Вычисляют момент окончания обслуживания i -й заявки на первом канале $t_{li} = [t_{l(i-1)} + \Delta t_{li}]$.

8. Устанавливают новый момент поступления заявки, и вычислительная процедура повторяется в соответствии с изложенным алгоритмом.

9. В ходе моделирования СМО накапливаются статистические данные о процессе обслуживания.

10. Определяют показатели качества функционирования системы путем обработки накопленных результатов моделирования методами математической статистики.

Таблица 2.1

Формулы для моделирования случайных величин

Закон распределения случайной величины	Плотность распределения	Формула для моделирования случайной величины
Экспоненциальный	$f(x) = \lambda \cdot e^{-\lambda \cdot x}$	$x_i = -\frac{1}{\lambda} \cdot \ln \xi_i$
Вейбула	$f(x) = \frac{a}{b} \cdot \left(\frac{x}{a}\right)^{a-1} \exp\left[-\left(\frac{x}{b}\right)^a\right]$	$x_i = -b \cdot (\ln \xi_i)^{1/a}$
Гамма-распределение (η -целые числа)	$f(x) = \frac{\lambda^\eta}{\Gamma(\eta)} \cdot e^{-\lambda \cdot x} x^{\eta-1}$	$x_i = -\frac{1}{\lambda} \sum_{j=1}^{\eta} \ln(1 - \xi_j)$
Нормальное ($\bar{x}$ – математическое ожидание; σ – средне-квадрат. отклонение)	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\bar{x})^2}{2\sigma^2}}$	$x_i = \bar{x} + \sigma \left(\sum_{j=1}^{12} \xi_j - 6 \right)$

2.2. Моделирование случайных чисел. Псевдослучайные числа

При исследовании сложных систем методом статистического моделирования существенное внимание уделяется учетам случайных факторов (случайные величины, случайные события и случайные процессы). Для моделирования случайной величины необходимо знать ее закон распределения.

Наиболее общим способом получения последовательности случайных чисел, распределенных по произвольному закону, является способ, в основе которого лежит их формирование из исходной последовательности случайных чисел, распределенных в интервале $[0,1]$ по равномерному закону. За последние 50 лет обсуждались в основном три принципиально разных варианта получения последовательностей чисел, которые в дальнейшем использовались в методах статистических испытаний:

- использование таблиц случайных чисел;
- применение физических генераторов (датчиков) случайных чисел;
- программным путем с помощью ЭВМ (псевдослучайные числа).

Таблица случайных чисел. К сожалению, объем любой таблицы конечен, поэтому сложные расчеты применить с ее помощью не всегда возможно. Через некоторое время приходится повторяться. При составлении таких таблиц выполняется требование, чтобы каждая из этих цифр от 0; 1;...; 9 встречалась примерно одинаково часто и независимо от других с вероятностью $p_i = 0,1$. Самая большая из опубликованных таблиц случайных чисел содержит 1000 000 цифр. Таблицы случайных чисел составить не так просто. Они требуют тщательной проверки с помощью специальных статистических тестов.

Генераторы (датчики) случайных чисел. Данные генераторы преобразуют результаты случайного физического процесса в двоичные числа.

В качестве случайного физического процесса обычно используются собственные шумы (случайным образом меняющееся

напряжение). Основной недостаток – нестабильность, непредсказуемые отклонения от заданного распределения (обычно – равномерного). Кроме того, трудно проверить качество вырабатываемых чисел, случайные числа не воспроизводимы (если их не запоминать), и, как следствие, нельзя повторить расчет на компьютере для исключения случайного сбоя. Генераторы случайных чисел являются основой метода Монте-Карло.

Реальные физические процессы обычно сложны, не всегда предсказуемы, поэтому использование генераторов случайных чисел на основе физических процессов на практике сильно затруднено. В данном случае используются генераторы псевдослучайных чисел.

Генератор псевдослучайных чисел (ГПСЧ, PRNG) – алгоритм, генерирующий последовательность чисел, элементы которой почти независимы друг от друга и подчиняются заданному распределению. Никакой алгоритм не может генерировать полностью случайные числа, а только лишь аппроксимировать некоторые свойства случайных чисел.

Любой ГПСЧ с ограниченными ресурсами рано или поздно закикливается. Длина циклов ГПСЧ зависит от самого генератора и в среднем составляет около $2^{(n/2)}$, где n – это размер внутреннего состояния в битах. Если ГПСЧ может сходиться к слишком коротким циклам, такой генератор становится предсказуемым и непригодным.

Псевдослучайные числа (п.ч.) – это числа, полученные по какой-либо формуле и имитирующие значения случайной величины, т. е. числа, удовлетворяющие ряду тестов таким образом, если бы они были значениями этой случайной величины. Наиболее распространены алгоритмы, в которых каждое следующее число вычисляется по предыдущему. Получаемые таким образом последовательности псевдослучайных чисел имеют период, что существенно отличает их от последовательностей случайных чисел. При вычислениях по методу статистических испытаний отдаётся предпочтение псевдослучайным числам, т.к. свойства последовательности псевдослучайных чисел можно исследовать путём пробных

вычислений, а экспериментальные устройства дают новые последовательности случайных чисел при каждом их использовании.

Назовем достоинства метода псевдослучайных чисел:

- на получение каждого случайного числа затрачивается несколько простых операций, так что скорость генерирования случайных чисел имеет тот же порядок, что и скорость работы компьютера;

- малый объем памяти для программирования;

- любое из чисел легко воспроизвести;

- качество генерируемых случайных чисел достаточно проверить один раз.

Один из главных недостатков метода – ограниченность «запаса» псевдослучайных чисел (наличие периода, ГПСЧ рано или поздно заикливается). Однако существуют генераторы с очень большим периодом. Подавляющее большинство расчетов по методам Монте-Карло в настоящее время осуществляется с использованием псевдослучайных чисел.

В расчетном методе (методе псевдослучайных чисел) в простейшем случае каждый следующий член последовательности рассчитывается по предыдущему. Например, так:

$$Z_{n+1} \equiv MZ_n \pmod{P},$$

где Z_0 – начальное значение (заданное целое положительное число),

M – параметр алгоритма (заданное целое положительное число), $P = 2^m$, где m – число двоичных разрядов представления чисел, с которыми манипулирует компьютер. Знак $\equiv$ здесь означает теоретико-числовую операцию сравнения, т. е. взятие дробной части от MZ_n/P и отбрасывание целой.

Существует довольно много методов и способов получения, равномерно распределенных в интервале $(0,1)$, псевдослучайных последовательностей.

В них предусматривается сдвиг исходного числа на несколько разрядов влево, затем сдвиг исходного числа на несколько разрядов вправо, сложение двух новых чисел, взятие какой-либо части нового числа и другие арифметические и логические операции для получения следующего случайного числа, равномерно распределенного на отрезке $[0,1]$.

Для увеличения периода в качестве исходных выбирается не одно, а несколько случайных чисел, используется не одно, а несколько различных рекуррентных соотношений.

Приведем несколько примеров. Первый алгоритм для получения псевдослучайных чисел предложил Дж. Нейман. Это так называемый *метод середины квадратов*, который заключается в следующем: пусть задано 4-значное целое число $n_1 = 9876$. Возведем его в квадрат. Получим, вообще говоря, 8-значное число 97 535 376. Выберем четыре средние цифры из этого числа и обозначим $n_2 = 5353$. Затем возведем его в квадрат (28 654 609) и снова извлечем 4 средние цифры. Получим $n_3 = 6546$. Далее, 42 850116, $n_4 = 8501$ и т. д. В качестве значений случайной величины предлагалось использовать значения 0,9876; 0,5353; 0,6546; 0,8501; 0,2670; 0,1289 и т. д.

Но этот алгоритм не оправдал себя: получалось больше чем нужно малых значений. Поэтому разными исследователями были разработаны другие алгоритмы.

Из современных ГПСЧ широкое распространение получил *Mersenne twister*, предложенный в 1997 году Мацумото и Нишимурой. Его достоинствами являются колоссальный период ($2^{19937}-1$), равномерное распределение в 623 измерениях (линейный конгруэнтный метод даёт более или менее равномерное распределение от силы в 5 измерениях), быстрая генерация случайных чисел (в 2–3 раза быстрее, чем стандартные ГПСЧ, использующие линейный конгруэнтный метод). Однако существуют сложные алгоритмы, распознающие последовательность, порождаемую с помощью Mersenne twister, как неслучайную. Это делает Mersenne twister неподходящим для криптографии.

Линейный конгруэнтный метод. Применяется в простых случаях и не обладает криптографической стойкостью.

Этот алгоритм заключается в применении следующей формулы:

$$Z_{n+1} = (q Z_n + c) \pmod{M},$$

где $q > 0$, $c > 0$, $M > 0$ – некоторые целочисленные константы.

Получаемая последовательность зависит от выбора стартового числа Z_0 , и при разных его значениях получаются различные последовательности случайных чисел. Кроме того, многие свойства последовательности не зависят от стартового числа, а зависят от коэффициентов в формуле. Последовательность чисел, генерируемых таким алгоритмом, периодична с периодом, не превышающим M .

При реализации выбирают $M = 2^e$, где e – число бит в машинном слове.

Метод суммирования. Метод суммирования предложен Ван Вейнгарденом и основан на использовании формулы вида:

$$Z_{i+n} = \sum_{k=1}^{n-1} C_k Z_{i+k} \pmod{M}.$$

Этот метод позволяет получить более длительные периоды в последовательности равномерно распределенных псевдослучайных чисел по сравнению с методом вычетов, т. к. здесь все множество чисел должно повторно принять некоторую комбинацию значений. Числа в получающейся последовательности имеют малую корреляцию. Для практической реализации этого метода используется алгоритм Дэвиса, которому удачным выбором двух начальных величин Z_1 и Z_2 удалось получить хорошую последовательность равномерно распределенных псевдослучайных чисел. Алгоритм Дэвиса использует формулу:

$$Z_{i+1} = [Z_i + Z_{i-1}] \pmod{4},$$

где $Z_1 = \pi$; $Z_2 = 5^{17} 2^{-42} = 0,542101887$.

Принцип моделирования непрерывной случайной величины, распределенной по любому закону

Покажем принцип моделирования случайной величины Ξ , равномерно распределенной в интервале $[a, b)$, и случайной величины Ξ , распределенной по показательному закону.

Значения случайной величины Ξ , распределенной по любому закону в интервале $[a, b)$ с плотностью распределения $f(\xi)$ определяются по формуле

$$x = \int_a^{\xi_i} f(\xi) d\xi. \quad (2.2.1)$$

Для каждой реализации величины X решается последнее уравнение относительно ξ_i , т. е. определяется реализация величины Ξ .

Равномерно распределенная в интервале $[a, b)$ случайная величина Ξ имеет плотность

$$f(x) = \begin{cases} \frac{1}{b-a} & a < \xi < b \\ 0 & a > \xi > b \end{cases}, \quad (2.2.2)$$

тогда

$$x = \int_a^{\xi_i} f(\xi) d\xi = \int_a^{\xi_i} \frac{d\xi}{b-a} = \frac{\xi_i - a}{b-a},$$

откуда

$$\xi_i = a + x(b - a). \quad (2.2.3)$$

Таким образом мы получили последовательность случайных величин ξ , равномерно распределенных в интервале $[a, b)$.

Плотность распределения случайной величины Ξ в интервале $[0, \infty)$:

$$f(\xi) = \begin{cases} \lambda e^{-\lambda t} & \xi \geq 0 \\ 0 & \xi < 0 \end{cases}, \quad (2.2.4)$$

$$x = \int_0^{\xi_i} \lambda e^{-\lambda \xi} d\xi, \quad 1 - x = e^{-\lambda \xi_i},$$
$$\xi_i = -\frac{1}{\lambda} \ln(1 - x) = -\frac{1}{\lambda} \ln x. \quad (2.2.5)$$

Мы получили последовательность случайных величин ξ , распределенных по показательному закону с заданным параметром λ .

Моделирование непрерывных случайных величин дает возможность моделировать любой поток вызовов, заданный последовательностью функций распределения промежутков между вызовами.

Математическое ожидание $M\xi$ и дисперсия $D\xi$ являются важными характеристиками случайной величины. Математическое ожидание случайной величины есть среднее значение, т. е. если мы будем наблюдать величину ξ много раз и получим значения $\xi_1, \xi_2, \dots, \xi_N$ (каждое из которых будет равно одному из чисел $x_1, x_2, \dots, x_n$), то среднее арифметическое этих значений будет близко к $M\xi$, а дисперсия $D\xi$ характеризует разброс этих значений около среднего.

В этом учебном пособии мы лишь знакомим с основными принципами статистического моделирования, не углубляясь в программирование метода.

3. ОСНОВЫ ТЕОРИИ СИСТЕМ МАССОВОГО ОБСЛУЖИВАНИЯ

3.1. Основы построения систем массового обслуживания (СМО). Понятие СМО. Классификация моделей систем массового обслуживания

Теория массового обслуживания возникла в начале XX века. Ее основоположником является датский ученый А.К. Эрланг, работавший в шведской телефонной компании и занимавшийся вопросами проектирования телефонных сетей. В дальнейшем теория получила интенсивное развитие в других областях науки, техники, экономики, производства. Это объясняется тем, что эта теория изучает широко распространенные в человеческой практике ситуации, когда имеется некоторый ограниченный ресурс и множество запросов (потоков) на его использование, следствием чего являются задержки или отказы в обслуживании некоторых запросов. Стремление понять объективные причины этих задержек или отказов и по возможности уменьшить их воздействие является побудительным мотивом развития теории массового обслуживания.

Системы, в которых, с одной стороны, возникают массовые запросы на выполнение каких либо видов услуг, а с другой стороны, происходит удовлетворение этих запросов, называются *системами массового обслуживания*. Любая СМО служит для выполнения потока заявок.

Заявка – абонент в канале, неисправность в устройстве, пассажир в билетной кассе, появление частицы в счетчике и т. д.

СМО имеет один или несколько обслуживаемых единиц – каналы обслуживания – число мастеров в парикмахерской, число каналов обслуживания, число колонок на АЗС и т. д.

Задачей СМО является анализ и синтез.

Анализ – оценка эффективности функционирования СМО при постоянных исходных характеристиках: структуре системы, дисциплине обслуживания, потоках требований, законе распределения времени обслуживания.

Синтез – поиск оптимальных параметров СМО.

Основными компонентами СМО любого вида являются:

- входной поток заявок на обслуживание;
- дисциплина очереди;
- механизм обслуживания.

Входной поток требований. Для его описания требуется задать вероятностный закон, определяющий последовательность моментов поступления требований на обслуживание и указать количество таких требований в каждом очередном поступлении. При этом, как правило, оперируют понятием «вероятностное распределение моментов поступления требований». Здесь могут поступать как единичные, так и групповые требования. В последнем случае речь идет о системе с параллельно-групповым обслуживанием. Источник заявок формирует входной поток, задерживая на какой-то отрезок времени поступление заявки в его состав. Заявки поступают на вход очереди, в котором реализуется заданный закон *дисциплины очереди*. Этот закон определяет порядок обслуживания входных заявок, который может быть детерминированным (первой обслуживается заявка, которая первой поступила) или случайным (закон Эрланга). Чаще всего используются дисциплины очереди, определяемые следующими правилами:

- первый пришел – первый обслужился;
- пришел последним – обслуживаешься первым;
- случайный отбор заявок;
- отбор по критерию приоритетности.

Канал обслуживания осуществляет обслуживание каждой заявки в соответствии с заданным детерминированным или случайным законом обслуживания. Выходной поток заявок отличается от входного в зависимости от законов дисциплины и очереди обслуживания.

Механизм обслуживания определяется характеристиками самой процедуры обслуживания и структурой обслуживающей системы.

К характеристикам процедуры обслуживания относятся: продолжительность обслуживания и количество требований, удовлетворяемых в результате выполнения каждой такой процедуры. Структура обслуживающей системы определяется

количеством и взаимным расположением каналов обслуживания (механизмов, приборов). Система обслуживания может иметь не один канал обслуживания, а несколько. Система такого рода способна обслуживать одновременно несколько требований. В этом случае все каналы обслуживания предлагают одни и те же услуги, и, следовательно, можно утверждать, что имеет место параллельное обслуживание.

Существует большое количество различных СМО. Перечислим основные *классы СМО* по разным основаниям:

а) марковские и немарковские. В марковских системах входящий поток требований и выходящий поток обслуженных требований являются пуассоновскими. Пуассоновские потоки позволяют легко описать и построить математическую модель СМО. Данные модели имеют достаточно простые решения, поэтому большинство известных приложений теории массового обслуживания используют марковскую схему. В случае немарковских процессов задачи исследования СМО значительно усложняются и требуется применения статистического моделирования, численных методов с использованием ЭВМ;

б) одноканальные и многоканальные (по числу каналов обслуживания, которые могут одновременно обслуживать входные заявки);

в) с отказами и без отказов (в зависимости от того, разрешается входной заявке ждать в очереди или нет, если разрешается – ограничена очередь по длине или времени либо нет);

г) многофазные и однофазные: при последовательном процессе обслуживания заявки несколькими приборами;

д) открытые и замкнутые: обслуженная заявка либо покидает СМО, либо снова поступает на обслуживание;

е) одиночные и сети СМО: сложные комбинации всех рассматриваемых выше СМО.

СМО, в которых требования, поступающие в момент, когда все приборы обслуживания заняты, получают отказ и теряются, называются системами с потерями или с отказами.

СМО, у которых возможно появление как угодно длинной очереди требований к обслуживающему устройству, называются системами с ожиданием.

Системы, допускающие очередь, но с ограниченным числом мест в ней, называются системами с ограниченной длиной очереди. Например, при попытке дозвониться до справочной «09», если все места для ожидания заполнены, то вежливый голос просит вас повторить вызов позже.

СМО, допускающие очередь, но с ограниченным сроком пребывания каждого требования в ней, называются системами с ограниченным временем ожидания. Обслуживание с ограниченным временем ожидания давно используется в телефонных станциях. Например, если вы слышите акустический сигнал «ответ станции», но не набираете номер в течение определенного времени, то обслуживание будет прервано (акустический сигнал «занято»). Некоторые справочные службы используют оба вида ограничений – по длительности ожидания и числу мест в очереди.

Могут быть СМО с условными потерями. Они подразумевают, что при отсутствии свободного обслуживающего прибора, заявка ожидает освобождения. Обычно считается, что условные потери не приводят к отказу в обслуживании, но слишком долгое ожидание может привести к желанию абонента отказаться от попытки вызова. Другой тип классификации приведен ниже. Он основан на принципе приоритетности обслуживания.

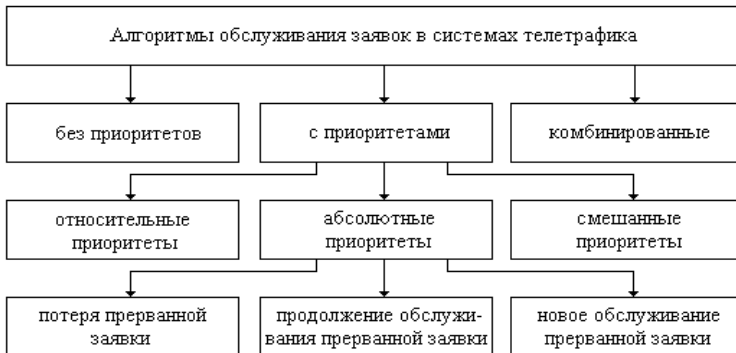


Рис.3. Тип классификации СМО

Заявки могут обрабатываться без приоритета. Комбинированный алгоритм используется при определенных условиях, например, резкий рост трафика, приводящий к снижению качества обслуживания. Для анализа приоритетных стратегий введем простую модель. Все заявки, поступающие в СМО, делятся на группы, которым присваивается приоритет от 1 до N . Заявка с приоритетом j имеет преимущество перед заявками с приоритетами с $j+1$ до N . В СМО с относительными приоритетами обслуживание заявок не прерывается. Допустим, заявка с приоритетом j застала все обслуживающие приборы занятыми. Тогда она встает в очередь перед всеми заявками, имеющими более низкий приоритет. Среди заявок с приоритетом j она будет последней.

СМО с абсолютными приоритетами основаны на прерывании обслуживания заявок. Такая возможность предусмотрена для всех случаев, когда обслуживаются заявки более низкого приоритета. При этом – нижняя часть рисунка – могут использоваться три основных варианта возобновления прерванного процесса обслуживания заявок.

В некоторых СМО используются смешанные приоритеты. Тогда множество $\{J\}$ разбивается на несколько классов – $\{J_1\} \dots \{J_L\}$. Чем меньше индекс у класса $\{J_M\}$, тем выше абсолютный приоритет у обслуживаемых заявок. В пределах каждого класса $\{J_M\}$ заявкам назначаются относительные приоритеты.

Одной из форм классификации СМО является кодовая (символьная) классификация Д. Кендалла. При этой классификации характеристику системы записывают в виде символов, например $A/B/X/Y/Z$, где A – тип распределения входящего потока требований, B – тип распределения времени обслуживания, X – число каналов обслуживания, Y – допустимая длина очереди, Z – порядок отбора требований (приоритет).

В качестве примеров СМО можно привести системы, представляющие собой банки различных типов, страховые организации, станции технического обслуживания, магазины и т. п.

Сеть связи является системой массового обслуживания, которая характеризуется:

- 1) потоком вызовов;
- 2) обслуживающими устройствами;
- 3) дисциплиной обслуживания.

Основные элементы сети связи:

- 1) конечные устройства;
- 2) коммутационные узлы;
- 3) каналы связи (соединительные пути).

Предметом теории массового обслуживания является установление зависимости между факторами, определяющими функциональные возможности системы массового обслуживания, и эффективностью ее функционирования.

При моделировании СМО вначале, если это возможно, составляется аналитическая модель системы в виде системы дифференциальных или алгебраических уравнений. В результате решения систем уравнений получают числовые характеристики системы.

В случае, если аналитическая модель не может быть построена, либо требует большого количества вычислений, используется вычислительная модель: составляется алгоритм работы СМО, переводится на язык программирования, либо язык моделирования и осуществляется машинное моделирование процесса обслуживания. Высокий уровень автоматизации процесса моделирования позволяет путем многократных повторных прогонов с коррекцией выбирать оптимальные режимы работы системы. При этом, по сравнению с концепцией аналитических моделей, здесь прежде всего обеспечивается высокая адекватность исследуемой имитационной модели реальной системе, а это как раз является слабым местом всех аналитических моделей.

Однако за аналитическими моделями остаются преимущества в части дедуктивных методов их исследования, синтеза, а также понимания динамики, внутренних закономерностей их поведения. При машинном моделировании этому противопоставляется возможность быстрого повторного моделирования при новых измененных режимах.

3.2. Марковские системы массового обслуживания

Математическое изучение функционирования СМО значительно упрощается, если протекающий в ней случайный процесс является марковским. В этом случае работа СМО сравнительно легко описывается с помощью линейных дифференциальных уравнений первого порядка, а в предельном режиме (при достаточно длительном функционировании СМО) – с помощью аппарата конечных систем линейных алгебраических уравнений.

В результате можно выразить в явном виде основные характеристики эффективности функционирования СМО через параметры СМО, потока заявок и дисциплины работы СМО.

Случайный процесс, протекающий в СМО, называется *марковским* (или процессом без последствия, или процессом без памяти), если вероятность любого состояния СМО в будущем зависит только от ее состояния в настоящем и не зависит от ее состояний в прошлом.

А.А. Марков (1856 – 1922) – известный русский математик, ординарный академик Петербургской академии наук, профессор Петербургского университета.

Чтобы случайный процесс был марковским, необходимо и достаточно, чтобы все потоки событий, под воздействием которых происходят переходы системы из состояния в состояние, были пуассоновскими.

Поток событий, обладающий свойствами отсутствия последствия и ординарностью (вероятность наступления за элементарно-малый промежуток времени более одного события пренебрежимо мала по сравнению с вероятностью наступления за этот промежуток времени одного события) называется пуассоновским.

Полезность марковских моделей мотивируется тем, что во многих случаях для обоснованных рекомендаций по практическому управлению СМО совсем не требуется знаний точных ее характеристик, а вполне достаточно иметь в своем распоряжении их приближенные значения. В классе марковских

процессов выделяются процессы с дискретными состояниями, называемыми марковскими цепями. Когда множество состояний процесса конечно, марковскую цепь называют конечной. Конечная марковская цепь может быть определена в непрерывном или дискретном времени. В первом случае переходы процесса из одного состояния в другое связываются с произвольными моментами времени и цепь называется непрерывной; а во втором – только в фиксированные моменты времени, обозначаемые порядковыми номерами, и цепь называется дискретной.

Важным частным случаем цепи Маркова с непрерывным временем является так называемый процесс гибели и размножения.

Случайный процесс называется процессом гибели и размножения, если он удовлетворяет условиям:

- пространство состояний процесса есть множество неотрицательных чисел;
- время пребывания процесса в состоянии i имеет показательное распределение и не зависит от предыдущего поведения процесса;
- после завершения пребывания процесса в состоянии i он переходит в состояние $i - 1$ с вероятностью q_i , $0 < q_i < 1$ и в состояние $i + 1$ с вероятностью $p_i = 1 - q_i$.

Если процесс, протекающий в системе с дискретными состояниями и непрерывным временем, является марковским, то для вероятностей состояний $P_1(t)$, $P_2(t)$, ... $P_n(t)$ можно составить систему линейных дифференциальных уравнений. Их еще называют уравнениями Колмогорова. При их составлении удобно пользоваться графом состояния системы, на котором против каждой стрелки, ведущей из состояния в состояние, проставлена плотность (интенсивность) потока событий, переводящего систему из состояния в состояние по данной стрелке. Если имеется размеченный граф состояний системы S , то систему дифференциальных уравнений для вероятностей состояний $P_k(t)$, ($k = 1, 2, 3, \dots, n$) можно сразу написать, пользуясь правилом.

В левой части каждого уравнения стоит производная $dP_k(t)/dt$, а в правой части – столько членов, сколько стрелок связано непосредственно с выбранным состоянием; если стрелка ведет в данное состояние, член имеет знак плюс, если ведет из данного состояния, член имеет знак минус. Каждый член равен плотности потока событий, переводящего систему по данной стрелке, умноженной на вероятность того состояния, из которого исходит стрелка. Например, для системы S , размеченный граф состояний которой показан на рис. 4, система дифференциальных уравнений будет:

$$\begin{cases} \frac{dP_0}{dt} - \lambda P_0 + \mu P_1 \\ \frac{dP_1}{dt} - -(\lambda + \mu)P_1 + \lambda P_0 + \mu P_2, \\ \dots \\ \frac{dP_n}{dt} - \lambda P_{n-1} - \mu P_n \end{cases} \quad (3.2.1)$$

где λ – интенсивность входящего потока,
 μ – интенсивность обслуживания.

Интегрируют всю систему при начальных условиях $P_0(0) = 1$, $P_k(0) = 0$, $k > 0$, вначале процесса все каналы исправны и свободны.

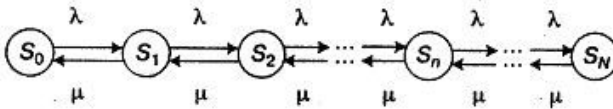


Рис. 4. Размеченный граф состояний системы S

Предельным режимом для системы X называется случайный процесс, устанавливающийся при $t \rightarrow \infty$. Если все потоки событий, переводящие систему из состояния в состояние, стационарны ($\lambda_{i,j} = const$), общее число состояний конечно и состояний без выхода нет, то предельный режим существует и характеризуется предельными вероятностями

X – определяет количество служебных каналов;

Y – определяет пропускную способность системы (длину очереди);

Z – определяет очередность обслуживания.

Когда пропускная способность системы бесконечна и очередность обслуживания подчиняется принципу «первый пришел – первый обслужился», части Y/Z опускают. В первом разряде (A) используются следующие символы:

M – распределение имеет показательный закон;

G – отсутствие каких-либо предположений о процессе обслуживания, либо он отождествляется с символом GI, означающий рекуррентный процесс обслуживания;

D – детерминированный (время обслуживания фиксированное);

E^n – эрланговское n -ого порядка;

HM^n – гиперэрланговское n -ого порядка.

Во втором разряде (B) используются те же символы.

В четвертом разряде (Y) показывается емкость буфера, т. е. максимальное количество мест в очереди.

В пятом разряде (Z) указывается способ выбора из очереди в системе с ожиданием: SP – равновероятный, PR – приоритетный, FF – первый пришел – первый ушел, LF – последний пришел – первый ушел.

3.3. Основные показатели эффективности и характеристики СМО

Цель теории массового обслуживания – выработка рекомендаций по рациональному построению СМО, рациональной организации их работы и регулированию потока заявок для обеспечения высокой эффективности функционирования СМО.

В качестве характеристик эффективности функционирования СО можно выбрать три основные группы показателей:

1. Показатели эффективности использования СМО:

1.1. Абсолютная пропускная способность СМО – среднее число заявок, которое может обслужить СМО в единицу времени.

1.2. Относительная пропускная способность – отношение среднего числа заявок, обслуживаемых СМО в единицу времени, к среднему числу поступивших заявок за это же время.

1.3. Средняя продолжительность периода занятости СМО.

1.4. Коэффициент использования СМО – средняя доля времени, в течение которого СМО занята обслуживанием заявок, и т. д.

2. Показатели качества обслуживания заявок:

2.1. Среднее время ожидания заявки в очереди.

2.2. Среднее время пребывания заявки в СМО.

2.3. Вероятность отказа заявке в обслуживании без ожидания.

2.4. Вероятность того, что поступившая заявка немедленно будет принята к обслуживанию.

2.5. Закон распределения времени ожидания заявки в очереди.

2.6. Закон распределения времени пребывания заявки в СМО.

2.7. Среднее число заявок, находящихся в очереди.

2.8. Среднее число заявок, находящихся в СМО, и т. п.

3. Показатели эффективности функционирования пары «СМО – потребитель», где под «потребителем» понимают всю совокупность заявок или некий их источник и т.п.

Вероятность того, что поступающее в систему требование откажется присоединяться к очереди и теряется, $P_{отк}$. Этот показатель для системы массового обслуживания с отказами равен вероятности того, что в системе находится столько требований, сколько она содержит приборов (каналов) обслуживания:

$$P_{отк} = P_m,$$

где m – число каналов обслуживания.

Для системы с ограниченной длиной очереди $P_{отк}$ равно вероятности того, что в системе находится $(m + l)$ требований:

$$P_{отк} = P_{m+l},$$

где l – допустимая длина очереди.

Противоположным показателем является вероятность обслуживания требования

$$P_{обсл} = 1 - P_{отк}.$$

Среднее количество требований, ожидающих начала обслуживания,

$$M_{ож} = \sum_{n=m+1}^{m+l} (n-m)P_n,$$

где P_n – вероятность того, что в системе находятся n требований.

При условии простейшего потока требований и экспоненциального закона распределения времени обслуживания формулы для $M_{ож}$ принимают следующий вид:

система с ограниченной длиной очереди:

$$M_{ож} = \frac{P_0 \rho^m}{m!} \sum_{n=1}^l n \left(\frac{\rho}{m} \right)^n,$$

где $\rho = \lambda/\mu$ – коэффициент загрузки системы;

λ – интенсивность входящего потока требований (среднее число требований, поступающих в единицу времени);

μ – интенсивность обслуживания (среднее число обслуженных требований в единицу времени);

система с ожиданием:

$$M_{ож} = \frac{P_0 \rho^{m+l}}{m \cdot m!} \cdot \frac{1}{\left(1 - \frac{\rho}{m}\right)^2}.$$

Относительная q и абсолютная A пропускные способности системы. Эти величины находят соответственно по формулам:

$$q = 1 - P_{отк}, \quad A = q \cdot \lambda.$$

Среднее число, занятых обслуживанием приборов в случае экспоненциального характера потока требований и времени обслуживания:

$$m_3 = \rho \cdot q.$$

Для системы массового обслуживания *с отказами*, можно найти по формуле:

$$m_3 = \sum_{n=1}^m n \cdot P_n.$$

Общее количество требований, находящихся в системе M . Эту величину определяют следующим образом:
система массового обслуживания *с отказами*:

$$M = m_3;$$

система массового обслуживания *с ограниченной длиной очереди и ожиданием*:

$$M = m_3 + M_{ож}.$$

Среднее время ожидания требованиям начала обслуживания $T_{ож}$. Если известна функция распределения вероятности времени ожидания требованиям начала обслуживания $F(t) = P(T_{ож} < t)$, то среднее время ожидания находится как математическое ожидание случайной величины $T_{ож}$:

$$T_{ож} = M[T_{ож}] = \int_0^{\infty} t dF.$$

$T_{ож}$ при показательном законе распределения требований во входящем потоке можно определить по формуле:

$$T_{ож} = \frac{M_{ож}}{\lambda}.$$

Показатели, характеризующие экономические особенности, формируют обычно в соответствии с конкретным видом

системы и ее назначением. Одним из общих экономических показателей является экономическая эффективность

$$E = P_{\text{общ}} \lambda c T - G_{\text{п}},$$

где c – средний экономический эффект, полученный при обслуживании одного требования;

T – рассматриваемый интервал времени;

$G_{\text{п}}$ – величина потерь в системе.

Величину потерь можно определить по следующим формулам:

система с отказами

$$G_{\text{п}} = (q_k \cdot m_{\text{з}} + q_y \cdot P_{\text{отк}} \cdot \lambda + q_{\text{пк}} \cdot m_{\text{св}}) \cdot T,$$

где q_k – стоимость эксплуатации одного прибора в единицу времени;

q_y – стоимость убытков в результате ухода требований из системы в единицу времени;

$q_{\text{пк}}$ – стоимость единицы времени простоя прибора системы;

система с ожиданием

$$G_{\text{п}} = (q_{\text{ож}} \cdot M_{\text{сж}} + q_{\text{гк}} \cdot m_{\text{св}} + q_k \cdot m_{\text{з}}) \cdot T,$$

где $q_{\text{ож}}$ – стоимость потерь, связанных с простоем требований в очереди в единицу времени.

3.4. Одноканальная модель с пуассоновским входным потоком с экспоненциальным распределением длительности обслуживания

Рассмотрим одноканальную модель, которая характеризуется показательным распределением как длительностей интервалов между поступлениями требований, так и длительностей обслуживания.

Пусть поток заявок и обслуживания простейшие, т. е. обладают свойствами стационарности (среднее число событий, воз-

действующих на систему, в течение единицы времени, остается постоянным), ординарности (вероятность попадания на элементарный участок времени двух и более событий пренебрежимо мала) и отсутствия последействия (для любых непересекающихся участков времени количество событий, попадающих на одно из них, не зависит от того, сколько событий попало на другие участки времени).

Для простейшего потока интенсивность входящих требований $\lambda = const$.

Так как поток заявок и обслуживания простейшие, то мы имеем дело с марковским случайным процессом с дискретными состояниями и непрерывным временем.

Рассмотрим три случая работы одноканальной модели:

1) система работает с отказами, т. е. заявка, поступившая в момент времени, когда обслуживающий канал занят передачей ранее поступившей заявки, получает отказ передачи;

2) система работает с ожиданием, но с ограниченным накопителем очереди, т. е. те заявки, которые не вмещаются в буфер памяти, получают отказ;

3) система работает с ожиданием, когда очередь неограниченна, т. е. все поступающие заявки будут обслужены.

Пусть система работает с отказами. Этот тип СМО может быть использован при моделировании каналов передачи в локальных сетях. Необходимо определить абсолютную и относительную пропускную способность системы. Данную систему массового обслуживания можно представить в виде графа (рис. 5), у которого имеются два состояния:

S_0 — канал свободен (ожидание);

S_1 — канал занят (идет обслуживание заявки).

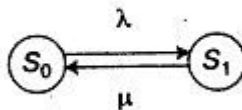


Рис. 5. Граф состояний одноканальной СМО с отказами

Обозначим вероятности состояний: $P_0(t)$ – вероятность состояния «канал свободен»; $P_1(t)$ – вероятность состояния «канал занят». По размеченному графу состояний система дифференциальных уравнений Колмогорова для вероятностей состояний будет:

$$\begin{cases} \frac{dP_0(t)}{dt} = -\lambda P_0(t) + \mu P_1(t) \\ \frac{dP_1(t)}{dt} = -\mu P_1(t) + \lambda P_0(t) \end{cases} \quad (3.4.1)$$

Система линейных дифференциальных уравнений имеет решение с учетом нормировочного условия $P_0(t) + P_1(t) = 1$. Решение данной системы называется неустановившимся, поскольку оно непосредственно зависит от t и выглядит следующим образом:

$$P_0(t) = \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t} + \frac{\mu}{\lambda + \mu}, \quad (3.4.2)$$

$$P_1(t) = 1 - P_0(t). \quad (3.4.3)$$

Нетрудно убедиться, что для одноканальной СМО с отказами вероятность $P_0(t)$ есть не что иное, как *относительная пропускная способность системы* q . Действительно, P_0 – вероятность того, что в момент t канал свободен и заявка, пришедшая к моменту t , будет обслужена, а, следовательно, для данного момента времени t среднее отношение числа обслуженных заявок к числу поступивших также равно $P_0(t)$, т. е. $q = P_0(t)$.

По истечении большого интервала времени (при $t \rightarrow \infty$) достигается стационарный (установившийся) режим:

$$q = P_0 = \frac{\mu}{\lambda + \mu}.$$

Зная относительную пропускную способность, легко найти абсолютную. *Абсолютная пропускная способность* (A) – среднее число заявок, которое может обслужить система массового обслуживания в единицу времени:

$$A = \lambda q = \frac{\lambda \mu}{\lambda + \mu}.$$

Вероятность отказа в обслуживании заявки будет равна вероятности состояния «канал занят» P_1 , т. е. канал не может обслужить новую заявку, пока не пропустит предыдущую заявку:

$$P_{отк} = P_1 = 1 - P_0 = \frac{\lambda}{\lambda + \mu}.$$

Данная величина $P_{отк}$ может быть интерпретирована как средняя доля необслуженных заявок среди поданных.

Случай второй. Рассмотрим теперь *одноканальную СМО с ожиданием и с ограниченной очередью*. Заявка, поступившая в момент, когда канал занят, становится в очередь и ожидает обслуживания. Данная СМО является наиболее распространенной при моделировании. С той или иной долей приближения с ее помощью можно моделировать практически любой узел локальной вычислительной сети (ЛВС).

Пусть на вход СМО поступает неограниченное число заявок. Предположим, что данная система (очередь + обслуживаемые клиенты) не может вместить более N требований (заявок), т. е. клиенты, не попавшие в ожидание, вынуждены обслуживаться в другом месте. Система М/М/1/Ν. Граф состояний СМО в этом случае имеет вид, показанный на рис. 6.

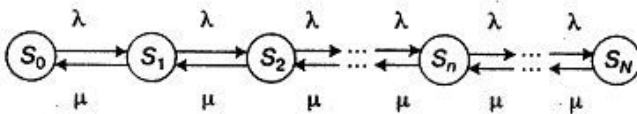


Рис. 6. Граф состояний одноканальной СМО с ожиданием (схема гибели и размножения)

Состояния СМО имеют следующую интерпретацию:

 S_0 – «канал свободен»;

S_1 – «канал занят» (очереди нет);

S_2 – «канал занят» (одна заявка стоит в очереди);

S_n – «канал занят» ($n - l$ заявок стоит в очереди);

S_N – «канал занят» ($N - I$ заявок стоит в очереди).

Стационарный процесс в данной системе будет описываться следующей системой алгебраических уравнений:

$$\begin{cases} -\rho P_0 + P_1 = 0 \\ \dots\dots\dots \\ -(1-\rho)P_n + P_{n+1} + \rho P_{n-1} = 0 , \\ \dots\dots\dots \\ -P_N + \rho P_{N-1} = 0 \end{cases} \quad (3.4.4)$$

где $\rho = \lambda/\mu$ – коэффициент загрузки,

n – номер состояния.

Выполнение условия стационарности для данной системы не обязательно, поскольку число допускаемых в обслуживающую систему заявок контролируется путем введения ограничения на длину очереди, которая не может превышать $(N - I)$, а не соотношением между интенсивностями входного потока, т. е. не отношением $\rho = \lambda / \mu$.

Решение приведенной выше системы уравнений для нашей модели СМО имеет вид:

$$P_n = \begin{cases} \left(\frac{1-\rho}{1-\rho^{N+1}} \right) \cdot \rho^n & \rho \neq 1 \\ \frac{1}{N+1} & \rho = 1 \end{cases} \quad n = 0, 1, 2 \dots N.$$

Начальное значение вероятности:

$$P_0 = \frac{1 - \rho}{1 - \rho^{N+1}}, \quad (3.4.5)$$

тогда

$$P_n = \begin{cases} P_0 \cdot \rho^n & \rho \neq 1, \quad n = 0, 1, 2, \dots, N \\ 1 & \rho = 1. \end{cases} \quad (3.4.6)$$

Определим характеристики одноканальной СМО с ожиданием и ограниченной длиной очереди, равной $(N - 1)$, M/M/1/N:

1) вероятность отказа в обслуживании заявки

$$P_{отк} = P_N = \begin{cases} \left(\frac{1 - \rho}{1 - \rho^{N+1}} \right) \rho^N & \rho \neq 1 \\ \frac{1}{N+1} & \rho = 1 \end{cases}.$$

Одной из важнейших характеристик систем, в которых возможна потеря запросов, является вероятность $P_{отк}$ того, что произвольный запрос будет потерян. При этом вероятность потери запроса совпадает с вероятностью того, что в произвольный момент времени все места для ожидания заняты, т. е. справедлива формула $P_{отк} = P_N$;

2) относительная пропускная способность системы

$$q = 1 - P_{отк} = \begin{cases} 1 - \left(\frac{1 - \rho}{1 - \rho^{N+1}} \right) \rho^N & \rho \neq 1 \\ 1 - \frac{1}{N+1} & \rho = 1; \end{cases}$$

3) абсолютная пропускная способность

$$A = q \cdot \lambda;$$

4) среднее число, находящихся в системе заявок

$$L_S = \sum_{n=0}^N n \cdot P_n = \begin{cases} \frac{\rho \cdot [1 - (N+1) \cdot \rho^N + N \cdot \rho^{N+1}]}{(1-\rho)(1-\rho^{N+1})} & \rho \neq 1 \\ \frac{N}{2} & \rho = 1 \end{cases};$$

5) среднее время пребывания заявки в системе:

$$T_S = \frac{L_S}{\lambda(1 - P_N)};$$

6) средняя продолжительность пребывания клиента (заявки) в очереди

$$T_{оч} = T_S - \frac{1}{\mu};$$

7) среднее число заявок (клиентов) в очереди (длина очереди)

$$L_{оч} = \lambda \cdot (1 - P_N) \cdot T_{оч}.$$

Третий случай. Рассмотрим *одноканальную СМО с ожиданием и с неограниченной очередью*.

Для системы массового обслуживания с неограниченной очередью должно выполняться условие $\rho < 1$, т. е., для того, чтобы в системе не накапливалась бесконечная очередь необходимо, чтобы в среднем запросы в системе обслуживались быстрее, чем они туда поступают.

$$P_0 = 1 - \rho. \quad (3.4.7)$$

Это вытекает из формулы (3.4.5) при $N \rightarrow \infty$.

Основные характеристики:

1) вероятность отказа в обслуживании заявки

$$P_{отк} = 0$$

т. к. все заявки будут обслужены, нет ограничений по очереди;

2) если нет отказа, значит относительная пропускная способность

$$q = 1;$$

3) абсолютная пропускная способность

$$A = \lambda \cdot q = \lambda;$$

4) среднее число, находящихся в системе заявок

$$L_s = \frac{\rho}{1 - \rho};$$

5) среднее время пребывания заявки в системе

$$T_s = \frac{\rho}{\lambda(1 - \rho)};$$

6) средняя продолжительность пребывания клиента (заявки) в очереди

$$T_{оч} = \frac{\rho^2}{\lambda(1 - \rho)} = \frac{L_{оч}}{\lambda};$$

7) среднее число заявок (клиентов) в очереди (длина очереди)

$$L_{оч} = \frac{\rho^2}{1 - \rho}.$$

Сравнивая выражения для среднего времени ожидания в очереди $T_{оч}$ и формулу для средней длины очереди $L_{оч}$, а также среднее время пребывания запросов в системе T_s и среднего числа запросов в системе L_s , видим, что

$$L_{оч} = \lambda \cdot T_{оч}, \quad L_s = \lambda \cdot T_s$$

Отметим, что эти формулы справедливы и для многих более общих, чем рассматриваемая система М/М/1, систем массового обслуживания и называются формулами *Литтла*.

Практическая значимость этих формул состоит в том, что они избавляют от необходимости непосредственного вычисления величин $T_{оч}$ и T_s при известном значении величин $L_{оч}$ и L_s и наоборот.

Пример 3.1.

Межузловая ветвь вторичной сети, имеющая один канал, принимает простейший поток сообщений с интенсивностью $\lambda = 0,04$ сообщений в секунду. Время передачи сообщений по каналу связи распределено по экспоненциальному закону. Среднее время передачи одного сообщения составляет $t = 14$ с. Сообщения, поступающие в моменты времени, когда обслуживающий канал занят передачей ранее поступившего сообщения, получают отказ передачи.

Определить следующие показатели эффективности ветви связи вторичной сети при условии её работы в установившемся режиме:

$P_{отк}$ – вероятность отказа приёма сообщения для передачи по межузловой ветви;

$P_{зан}$ – вероятность занятости канала связи (коэфф. относительной загрузки канала);

Q – относительная пропускная способность межузловой ветви;

A – абсолютная пропускная способность межузловой ветви.

Решение:

Найдём μ и ρ :

$$\mu = \frac{1}{t} = \frac{1}{14} = 0.07 \text{ с}^{-1},$$
$$\rho = \frac{\lambda}{\mu} = 0.04 \cdot 14 = 0.56.$$

Найдём вначале вероятность нулевого состояния анализируемой СМО, при которой канал связи свободен и поступающее сообщение будет принято для передачи по каналу связи:

$$P_0 = \frac{\mu}{\lambda + \mu} = \frac{0,07}{0,04 + 0,07} = 0,64.$$

Вероятность отказа приёма для передачи по каналу связи межузловой ветви

$$P_{\text{отк}} = \frac{\lambda}{\lambda + \mu} = \frac{0,04}{0,04 + 0,07} = 0,36.$$

Для одноканальной СМО с отказами вероятность свободного состояния P_0 численно равна Q – относительной пропускной способности СМО.

$$Q = \frac{\mu}{\lambda + \mu}.$$

Абсолютная пропускная способность межузловой ветви, сооб./с:

$$A = \lambda Q = 0,04 \cdot 0,64 = 0,03.$$

Пример 3.2.

Межузловая ветвь вторичной сети связи, имеющая один канал и неограниченный по объёму накопитель очереди ожидающих сообщений, принимает простейший поток сообщений с интенсивностью $\lambda = 0,04$ сообщений в секунду. Время передачи сообщений распределено по экспоненциальному закону. Среднее время передачи одного сообщения составляет $t = 14$ с. Сообщения, поступающие в моменты времени, когда обслуживающий канал занят передачей ранее поступившего сообщения, принимаются в очередь и не покидают её до момента до начала передачи по каналу связи.

Определить следующие показатели эффективности ветви связи вторичной сети:

$L_{оч}$ – среднее число сообщений в очереди к ветви связи вторичной сети;

$L_{сист}$ – среднее суммарное число сообщений в очереди и передающихся по ветви связи вторичной ветви;

$T_{оч}$ – среднее время пребывания сообщения в очереди до начала передачи;

$T_{\text{сист}}$ – среднее суммарное время пребывания сообщения в системе, складывающееся из среднего времени ожидания в очереди и среднего времени передачи;

$P_{\text{зан}}$ – вероятность занятости канала связи (коэфф. относительной загрузки канала);

Q – относительную пропускную способность межузловой ветви;

A – абсолютную пропускную способность межузловой ветви.

Решение:

$$\mu = \frac{1}{t} = \frac{1}{14} = 0.07 \text{ с}^{-1},$$

$$\rho = \frac{\lambda}{\mu} = \frac{0.04}{0.07} = 0.56.$$

Найдём вначале вероятность нулевого состояния анализируемой СМО, при котором канал связи свободен и в накопителе очереди нет ни одного сообщения:

$$P_0 = 1 - \rho = 1 - 0.56 = 0.44.$$

Вероятность отказа приёма сообщения для передачи по каналу связи межузловой ветви в СМО с бесконечно большим накопителем очереди будет равна нулю при условии $\rho < 1$.

$L_{\text{оч}}$ – среднее число сообщений в очереди к ветви связи вторичной сети:

$$L_{\text{оч}} = \frac{\rho^2}{1 - \rho} = \frac{0.56^2}{1 - 0.56} = 0.71.$$

$L_{\text{сист}}$ – среднее суммарное число сообщений в очереди и передающихся по ветви связи вторичной ветви:

$$L_{\text{сист}} = \frac{\rho}{1 - \rho} = \frac{0.56}{1 - 0.56} = 1.14.$$

$T_{оч}$ – среднее время пребывания сообщения в очереди до начала передачи:

$$T_{оч} = \frac{\rho^2}{\lambda \cdot (1 - \rho)} = \frac{0.56^2}{0.04 \cdot (1 - 0.56)} = 17.42.$$

$T_{сист}$ – среднее суммарное время пребывания сообщения в системе, складывающееся из среднего времени ожидания в очереди и среднего времени передачи:

$$T_{сист} = \frac{\rho}{\lambda \cdot (1 - \rho)} = \frac{0.56}{0.04 \cdot (1 - 0.56)} = 31.1.$$

Относительная пропускная способность Q межузловой ветви с бесконечным накопителем очереди численно равна доле сообщений в поступающем потоке, принимаемых для передачи (не получаемых отказа). При отсутствии перегрузки канала связи ($\rho < 1$) она будет равна единице.

Абсолютная пропускная способность межузловой ветви с бесконечным накопителем очереди, при условии отсутствия перегрузки канала связи ($\rho < 1$), будет равна интенсивности потока λ , поступающих для передачи сообщений.

3.5. Многоканальная модель с пуассоновским входным потоком и экспоненциальным распределением длительности обслуживания

На практике, в большинстве своем, системы массового обслуживания являются многоканальными, т. к. при этом есть возможность одновременного обслуживания некоторого количества заявок, а, следовательно, можно повысить скорость обслуживания. При этом режим функционирования того или иного обслуживающего канала не влияет на режим функционирования других обслуживающих каналов системы. Пусть процесс массового обслуживания, описываемый данной моделью, характеризуется интенсивностью входного потока λ , при этом параллельно может обслуживаться не более n клиентов (заявок). Средняя продолжительность обслуживания одной

заявки равняется $1/\mu$. Входной и выходной потоки являются пуассоновскими. Длительность процедуры обслуживания каждым из каналов является случайной величиной, подчиненной экспоненциальному закону распределения.

Вначале рассмотрим более простой вариант, когда n -канальная система работает с отказами, т. е. *СМО с отказами*.

Будем нумеровать состояния СМО по числу занятых каналов.

S_0 – все каналы свободны;

S_1 – занят один канал, остальные свободны;

.....

S_k – заняты ровно k каналов, остальные свободны;

.....

S_n – заняты все n каналов.

Граф состояний показан на рис. 7.

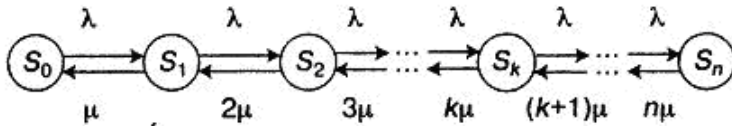


Рис. 7. Граф состояний многоканальной СМО с отказами

Обозначим $\rho = \lambda / \mu$ – приведенная интенсивность потока, иначе среднее число заявок за среднее время обслуживания.

Уравнения Колмогорова для вероятностей состояний системы $P_0, \dots, P_k, \dots, P_n$ будут иметь следующий вид:

$$\begin{cases} \frac{dP_0}{dt} = -\lambda \cdot P_0 + \mu \cdot P_1 \\ \dots \dots \dots \\ \frac{dP_k}{dt} = \lambda \cdot P_{k-1} - (\lambda + k\mu) \cdot P_k + \mu \cdot (k+1) \cdot P_{k+1}, & 1 \leq k \leq n-1. \\ \dots \dots \dots \\ \frac{dP_n}{dt} = \lambda \cdot P_{n-1} - \mu \cdot n \cdot P_n \end{cases}$$

Начальные условия решения системы таковы:

$$P_0(0) = 1, \quad P_1(0) = P_2(0) = \dots = P_k(0) = \dots = P_1(0) = 0.$$

Стационарное решение системы имеет вид:

$$P_k = \frac{\frac{\rho^k}{k!}}{\sum_{k=0}^n \frac{\rho^k}{k!}} = \frac{\rho^k}{k!} \cdot P_0, \quad (3.5.1)$$

где

$$P_0 = \frac{1}{\sum_{k=0}^n \frac{\rho^k}{k!}} = \left[1 + \left(\frac{\lambda}{\mu} \right) + \frac{\left(\frac{\lambda}{\mu} \right)^2}{2!} + \frac{\left(\frac{\lambda}{\mu} \right)^3}{3!} + \dots + \frac{\left(\frac{\lambda}{\mu} \right)^n}{n!} \right]^{-1}. \quad (3.5.2)$$

Формулы для вычисления вероятностей P_k (3.5.1) называются формулами Эрланга.

Зная предельные вероятности $P_0, P_1, \dots, P_n$, найдем характеристики СМО, вероятность отказа, относительную пропускную способность, абсолютную пропускную способность:

1. Если все каналы заняты, заявка получает отказ и покидает систему.

Вероятность отказа:

$$P_{отк} = P_n = \frac{\rho^n}{n!} \cdot P_0.$$

Величина $P_{отк}$ характеризует полноту обслуживания входящего потока.

2. Вероятность того, что заявка будет принята к обслуживанию (она же – относительная пропускная способность системы q), т. е. противоположное событие:

$$q = 1 - P_{отк} = 1 - \frac{\rho^n}{n!} \cdot P_0.$$

3. Абсолютная пропускная способность

$$A = \lambda \cdot q = \lambda \cdot (1 - P_{отк}).$$

4. Найдем среднее число занятых каналов. Воспользуемся средним значением или математическим ожиданием дискретной случайной величины:

$$\bar{k} = 0 \cdot P_0 + 1 \cdot P_1 + 2 \cdot P_2 + \dots + n \cdot P_n.$$

Поскольку A (абсолютная пропускная способность) есть среднее число заявок, обслуженных в единицу времени, то

$$\bar{k} = \frac{A}{\mu} = \frac{\lambda(1 - P_{отк})}{\mu} = \rho \cdot (1 - P_{отк}).$$

Перейдем к системам с ожиданием. СМО с ожиданием и с ограниченной длиной очереди.

Процесс массового обслуживания при этом характеризуется следующим: входной и выходной потоки являются пуассоновскими с интенсивностями λ и μ соответственно; параллельно обслуживаться могут не более n клиентов, т. е. система имеет n каналов обслуживания. Средняя продолжительность обслуживания одного клиента равна $1/\mu$ для каждого канала. Число мест в очереди равно m . Эта система также относится к процессу гибели и размножения.

Состояния системы:

S_0 – все каналы свободны;

S_1 – занят один канал, остальные свободны;

.....

S_k – заняты ровно k каналов, остальные свободны;

.....

S_n – заняты все n каналов.

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

.....

S_{n+r} – заняты все n каналов, r заявок в очереди;

.....

S_{n+m} – заняты все n каналов, m заявок в очереди.

Предельные вероятности состояний, полагая $\rho = \lambda / \mu$:

$$P_1 = \frac{\rho}{1!} \cdot P_0; \quad P_2 = \frac{\rho^2}{2!} \cdot P_0; \quad \dots \quad P_n = \frac{\rho^n}{n!} \cdot P_0;$$

$$P_{n+1} = \frac{\rho^{n+1}}{n \cdot n!} \cdot P_0; \quad P_{n+m} = \frac{\rho^{n+m}}{n^m \cdot n!} \cdot P_0.$$

P_0 – вероятность того, что все каналы свободны:

$$P_0 = \left[1 + \frac{\rho}{1!} + \frac{\rho^2}{2!} + \frac{\rho^3}{3!} + \dots + \frac{\rho^n}{n!} + \frac{\frac{\rho}{n} - \left(\frac{\rho}{n}\right)^{m+1}}{1 - \frac{\rho}{n}} \right]^{-1}.$$

Найдем некоторые характеристики эффективности обслуживания.

1. Заявка получает отказ, если заняты все n каналов и все m мест в очереди:

$$P_{\text{отк}} = P_{n+m} = \frac{\rho^{n+m}}{n^m \cdot n!} \cdot P_0.$$

2. Относительная пропускная способность:

$$q = 1 - P_{\text{отк}}.$$

3. Абсолютная пропускная способность:

$$A = q \cdot \lambda.$$

4. Обозначим через Z среднее число занятых каналов. Каждый канал обслуживает μ заявок в единицу времени. СМО обслуживает A заявок в среднем в единицу времени. Отсюда число занятых каналов:

$$Z = A / \mu.$$

5. Среднее число заявок в очереди вычисляем как математическое ожидание случайной величины:

$$L_{оч} = 1 \cdot P_{n+1} + 2 \cdot P_{n+2} + \dots + m \cdot P_{n+m},$$

$$L_{оч} = \frac{\rho^{n+1}}{n \cdot n!} \cdot P_0 \cdot \frac{1 - \left(\frac{\rho}{n}\right)^m (1 + m - m \frac{\rho}{n})}{\left(1 - \frac{\rho}{n}\right)^2}.$$

6. Среднее число заявок в очереди $L_{оч}$ и среднее число занятых каналов дает среднее число заявок в системе:

$$L_s = L_{оч} + Z.$$

7. Среднее время ожидания в очереди:

$$T_{оч} = \frac{L_{оч}}{\lambda}.$$

8. Среднее время пребывания заявки в очереди (системе):

$$T_s = T_{оч} + \frac{q}{\mu},$$

где q/μ – среднее время обслуживания.

Теперь рассмотрим СМО с ожиданием и с неограниченной длиной очереди, т. е. очередь $m \rightarrow \infty$, $\rho/n < 1$.

Выражения для предельных состояний.

$$P_1 = \frac{\rho}{1!} \cdot P_0; \quad P_2 = \frac{\rho^2}{2!} \cdot P_0; \quad \dots; \quad P_n = \frac{\rho^n}{n!} \cdot P_0;$$

$$P_{n+1} = \frac{\rho^{n+1}}{n \cdot n!} P_0; \quad \dots; \quad P_{n+r} = \frac{\rho^{n+r}}{n^r n!} P_0; \quad \dots;$$

где r – число заявок в очереди.

$$P_0 = \left[1 + \frac{\rho}{1!} + \frac{\rho^2}{2!} + \frac{\rho^3}{3!} + \dots + \frac{\rho^n}{n!} + \frac{\rho^{n+1}}{n! (n - \rho)} \right]^{-1}.$$

Так как каждая заявка рано или поздно будет обслужена, то характеристики пропускной способности будут:

$$P_{отк} = 0; \quad q=1; \quad A=\lambda; \quad Z = A / \mu = \rho;$$

$$L_{оч} = \frac{\rho^{n+1} \cdot P_0}{n \cdot n! \cdot \left(1 - \frac{\rho}{n}\right)^2};$$

$$T_s = T_{оч} + \frac{q}{\mu} = T_{оч} + \frac{1}{\mu}.$$

Среднее число заявок под обслуживанием:

$$L_s = L_{оч} + Z.$$

Пример 3.3

Межузловая ветвь вторичной сети связи имеет $n = 4$ каналов. Поток сообщений, поступающих для передачи по каналам ветви связи, имеет интенсивность $\lambda = 8$ сообщений в секунду. Среднее время $t = 0,1$ передачи одного сообщения равно $t/n = 0,025$ с. В накопителе очереди ожидающих передачи сообщений может находиться до $m = 4$ сообщений. Сообщение прибывшее в момент, когда все m мест в очереди заняты, получает отказ передачи по ветви связи. Найти характеристики СМО:

$P_{отк.}$ – вероятность отказа передачи сообщений;

Q – относительную пропускную способность межузловой ветви;

A – абсолютную пропускную способность межузловой ветви;

Z – среднее число занятых каналов;

$L_{оч.}$ – среднее число сообщений в очереди;

$T_{ож.}$ – среднее время ожидания;

$T_{\text{сист.}}$ – среднее суммарное время пребывания сообщения в очереди и его передачи по ветви связи.

Решение:

$$\mu = \frac{1}{t} = \frac{1}{0.1} = 10 \text{ с}^{-1},$$

$$\rho = \frac{\lambda}{\mu} = \frac{8}{10} = 0.8,$$

$$\frac{\rho}{n} = \frac{0.8}{4} = 0.2.$$

Найдём вначале вероятность нулевого состояния СМО:

$$P_0 = \left[1 + \frac{\rho}{1!} + \frac{\rho^2}{2!} + \frac{\rho^3}{3!} + \dots + \frac{\rho^n}{n!} + \frac{\frac{\rho}{n} - \left(\frac{\rho}{n}\right)^{m+1}}{1 - \frac{\rho}{n}} \right]^{-1},$$

$$P_0 = \left(1 + \frac{0.8}{1!} + \frac{0.8^2}{2!} + \frac{0.8^3}{3!} + \frac{0.8^4}{4!} + \frac{0.2 - 0.2^5}{1 - 0.2} \right)^{-1} = 0.413.$$

Вероятность отказа передачи по ветви связи будет равна:

$$P_{\text{отк.}} = \frac{0.8^{4+4}}{4^{4+4}!} \cdot 0.413 = 1.12777E-05.$$

Относительная пропускная способность:

$$Q = 1 - P_{\text{отк.}} = 1 - 1.128E-05 = 0.99998.$$

Абсолютная пропускная способность:

$$A = Q\lambda = 0.99998 \cdot 8 = 7.99984 \text{ сообщ./с.}$$

Среднее число занятых каналов связи:

$$Z = \frac{A}{\mu} = \frac{7.99984}{10} = 0.799984.$$

Среднее число сообщений в накопителе очереди определим по формуле:

$$L_{оч.} = \frac{\rho^{n+1}}{n \cdot n!} \cdot P_0 \cdot \frac{1 - \left(\frac{\rho}{n}\right)^m (1 + m - m \frac{\rho}{n})}{\left(1 - \frac{\rho}{n}\right)^2},$$

$$L_{оч.} = \frac{0,8^5}{4 \cdot 4!} \cdot 0,413 \cdot \frac{1 - (0,2)^4 (1 + 4 - 4 \cdot 0,2)}{(1 - 0,2)^2} = 0,0021.$$

Среднее время ожидания в очереди:

$$T_{оч.} = \frac{L_{оч.}}{\lambda} = \frac{0.0021}{8} = 0.00026 \text{ с.}$$

Среднее суммарное время пребывания сообщения в очереди и его передачи по ветви связи:

$$T_s = T_{оч.} + \frac{Q}{\mu} = 0.00018 + \frac{0.9998}{10} = 0.1 \text{ с.}$$

3.6. Замкнутая система массового обслуживания

Рассмотрим модель обслуживания машинного парка, которая представляет собой модель замкнутой системы массового обслуживания. До сих пор мы рассматривали только такие системы массового обслуживания, для которых интенсивность λ входящего потока заявок не зависит от состояния системы. В этом случае источник заявок является внешним по отношению к СМО и генерирует неограниченный поток требований. Рассмотрим системы массового обслуживания, для

которых λ зависит от состояния системы, при чем источник требований является внутренним и генерирует ограниченный поток заявок. Например, обслуживается машинный парк, состоящий из N машин, бригадой R механиков ($N > R$), причем каждая машина может обслуживаться только одним механиком. Здесь машины являются источниками требований (заявок на обслуживание), а механики – обслуживающими каналами. Неисправная машина после обслуживания используется по своему прямому назначению и становится потенциальным источником возникновения требований на обслуживание. Очевидно, что интенсивность λ зависит от того, сколько машин в данный момент находится в эксплуатации ($N - k$) и сколько машин обслуживается или стоит в очереди, ожидая обслуживания (k).

В рассматриваемой модели емкость источника требований следует считать ограниченной. Входящий поток требований исходит из ограниченного числа эксплуатируемых машин ($N - k$), которые в случайные моменты времени выходят из строя и требуют обслуживания. При этом каждая машина из ($N - k$) находится в эксплуатации. Генерирует пуассоновский поток требований с интенсивностью X независимо от других объектов, общий (суммарный) входящий поток имеет интенсивность $(N - k) \cdot \lambda$. Требование, поступившее в систему в момент, когда свободен хотя бы один канал, немедленно идет на обслуживание. Если требование застаёт все каналы занятыми обслуживанием других требований, то оно не покидает систему, а становится в очередь и ждет, пока один из каналов не станет свободным. Таким образом, в замкнутой системе массового обслуживания входящий поток требований формируется из выходящего. Состояние S_k системы характеризуется общим числом требований, находящихся на обслуживании и в очереди, равным k . Для рассматриваемой замкнутой системы, очевидно, $k = 0, 1, 2, \dots, N$. При этом если система находится в состоянии S_k , то число объектов, находящихся в эксплуатации, равно $(N - k)$. Если λ – интенсивность потока требований в расчете на одну машину, то:

$$\lambda_k = \begin{cases} (N - k) \cdot \lambda & 0 \leq k \leq N \\ 0, & k > N \end{cases},$$

$$\mu_k = \begin{cases} k \cdot \mu, & 0 \leq k \leq R \\ R \cdot \mu & R \leq k \leq N. \\ 0, & k > N \end{cases}$$

Система алгебраических уравнений, описывающих работу замкнутой СМО в стационарном режиме, выглядит следующим образом:

$$\begin{cases} 0 = -\rho \cdot N \cdot P_0 + P_0 \\ 0 = (N - k + 1) \cdot \rho \cdot P_{k-1} - [(N - k) \cdot \rho + k] \cdot P_k + (k + 1) \cdot P_{k+1}, \\ 0 = (N - k + 1) \cdot \rho \cdot P_{k-1} - [(N - k) \cdot \rho + R] \cdot P_k + R \cdot P_{k+1}, \\ 0 = \rho \cdot P_{N-1} - R \cdot P_N, \end{cases}$$

где $0 < k < R, R \leq k \leq N$.

Решая данную систему, находим вероятность k -го состояния:

$$P_k = \begin{cases} \frac{N! \cdot \rho^k}{k! \cdot (N - k)!} \cdot P_0, & 1 \leq k < R \\ \frac{N! \cdot \rho^k}{R! \cdot R^{k-R} \cdot (N - k)!} \cdot P_0, & R \leq k \leq N \end{cases}.$$

Величина P_0 определяется из условия нормирования $\sum P_k$ полученных результатов по формулам для $P_k, k = 0, 1, 2, \dots, N$.
Определим следующие вероятностные характеристики системы:

– среднее число требований в очереди на обслуживание

$$L_{oc} = \sum_{k=R}^N (k-R) \cdot P_k ;$$

– среднее число требований, находящихся в системе (на обслуживании и в очереди)

$$L_S = \sum_{k=1}^N k \cdot P_k ;$$

– среднее число механиков (каналов), «простаивающих» из-за отсутствия работы

$$\overline{R_n} = \sum_{k=0}^{R-1} (R-k) \cdot P_k ;$$

– коэффициент простоя обслуживаемого объекта (машины) в очереди

$$\alpha_1 = \frac{L_{oc}}{N} ;$$

– коэффициент использования объектов (машин)

$$\alpha_2 = 1 - \left(\frac{L_S}{N} \right) ;$$

– коэффициент простоя обслуживающих каналов (механиков)

$$\alpha_3 = \frac{\overline{R_n}}{R} ;$$

– среднее время ожидания обслуживания (время ожидания обслуживания в очереди)

$$T_{oc} = \frac{1}{\lambda} \cdot \left(\frac{1 - \alpha_2}{\alpha_2} \right) - \frac{1}{\mu} .$$

3.7. Марковские сети массового обслуживания

Быстрое развитие во второй половине XX века и особенно в конце его систем сбора, обработки, хранения, передачи и распределения информации, а также частичная или полная интеграция таких систем стали возможными благодаря взаимопроникновению техники связи и вычислительной техники, появлению и развитию инфокоммуникационных технологий (Интернет, IP телефония, электронная почта, глобальная телефонная сеть, включающая сети спутниковой связи, сети сотовой подвижной связи).

Аппаратом для аналитического моделирования таких сетей служит теория сетей массового обслуживания (СеМО). Теория СеМО основана на теории систем массового обслуживания (СМО) и является ее развитием, т. к. изучает взаимодействие нескольких СМО и циркулирующих между ними заявок. Цель моделирования СеМО – аналитического и имитационного – является прогноз производительности сети или ее фрагментов путем оценки вероятностно-временных характеристик использования ресурсов, длин очередей и задержек в них.

Поскольку сеть состоит из нескольких узлов, то предположим, что обслуживание заявки на ее маршруте в сети в различных узлах происходит последовательно. Это означает, что заявка не может одновременно обслуживаться в двух или более узлах. В накопителе любого из узлов ей выделяется одинаковая для всех заявок емкость, которую мы примем за единицу емкости. Так как все заявки одного типа, то для обслуживания в любом узле им предоставляется только один прибор (канал), т. е. любой узел представляет собой одноканальную СМО. Такие сети еще называются *многофазными системами*. Каждый узел такой сети может содержать СМО определенного типа, а заявки поступают в сеть в различных точках и, получив обслуживание в одном узле, переходят на обслуживание в другой для дальнейшей обработки.

Существуют открытые, замкнутые и смешанные сети. *Открытой* называется сеть массового обслуживания, состоящая из M узлов, причем хотя бы в один из узлов сети поступает извне

входящий поток заявок, и имеется сток заявок из сети. Кроме того в открытой сети интенсивность поступления заявок не зависит от состояния сети, т. е. числа заявок, уже поступивших в сеть. Открытые сети используются для моделирования ЛВС, работающих в неоперативном режиме. *Замкнутой* называется сеть массового обслуживания с множеством узлов M без источника и стока, в которой циркулирует постоянное число заявок. Замкнутые СеМО используются для моделирования таких ЛВС, источниками информации для которых служат абонентские терминалы, работающие в диалоговом режиме. Смешанной называется сеть массового обслуживания, в которой циркулирует несколько типов заявок, причем относительно одних типов заявок сеть работает как замкнутая, а относительно других – как открытая. Например, часть абонентов работает в диалоговом режиме (алгоритм работы замкнутой сети), а часть в неоперативном (открытая сеть).

Если все заявки одного типа, то сеть называется однородной, в противном случае – неоднородной (более общий случай). В дальнейшем будем рассматривать лишь наиболее простой случай однородных экспоненциальных СеМО.

При исследовании сети надо задать ее топологическую структуру, требуется описать маршруты отдельных заявок и вероятностные модели потоков заявок между узлами сети.

Рассмотрим простейшую последовательную систему с двумя узлами, рис. 8 [3].

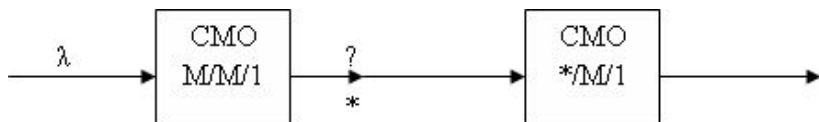


Рис. 8. Сеть из двух СМО

Это топологическая структура сети, а не диаграмма состояний. На вход сети поступает пуассоновский поток с

интенсивностью λ . На первом узле находится СМО типа М/М/1 с показательным распределением времени обслуживания со средним значением $1/\mu$. Будем считать, что второй узел СМО2 имеет также показательное распределение времени обслуживания с интенсивностью μ . Основная задача заключается в вычислении распределения промежутков времени между последовательными заявками, поступающими в узел 2, т. е. уходящими из узла 1.

Пусть $d(t)$ – плотность распределения вероятностей промежутков между заявками на выходе узла 1. Выразим эту функцию распределения для двух случаев:

1) когда при уходе заявки из первого узла СМО1 не пуста, т. е. сразу поступила следующая заявка на обслуживание в СМО1;

2) когда при выходе из СМО1 в буфере не было другой заявки, т. е. СМО1 пуста.

В первом случае следующая заявка появится на выходе сразу через интервал времени обслуживания. Это событие будет происходить с вероятностью $(1 - P_0)$ (вероятность, что СМО не пуста). Плотность вероятности между выходами заявки из узла 1 будет совпадать с плотностью вероятности времени обслуживания в сервере 1 $b(x)$. Для второго случая следующая заявка покинет СМО1 через интервал, равный сумме интервала между входными событиями и времени обслуживания. Это событие будет происходить с вероятностью, равной вероятности простоя системы, т. е. $P_0 = (1 - \rho)$. Поскольку эти два интервала распределены независимо, то плотность распределения вероятностей их суммы будет равна свертке плотностей распределения суммируемых случайных величин. Для вычисления используется преобразование Лапласа (интегральное преобразование, связывающее функцию $F(s)$ комплексного переменного (модель) с функцией $f(x)$ – действительного переменного (оригинал)). Преобразование Лапласа функции действительной переменной $f(x)$, называется функция $F(s)$ комплексной переменной $s = \sigma + i\omega$, такая что

$$F(s) = \int_0^{\infty} e^{-st} f(t) dt.$$

Обозначим преобразование Лапласа для плотности вероятности суммы двух промежутков времени – время до появления следующего требования и время обслуживания этого требования – как $D_0(s)$. Соответственно преобразование Лапласа плотности распределения суммы равно произведению преобразований исходных плотностей распределения. Если обозначить преобразование Лапласа для плотностей вероятности интервалов между событиями входного потока $A(s)$, а для времени обслуживания в сервере – $B(s)$, то преобразование Лапласа плотности вероятности промежутка времени между выходными заявками для случая пустого узла I будет

$$D_0(s) = A(s) \cdot B(s).$$

Поскольку в системе М/М/1 распределения интервалов экспоненциальные как для входных событий, так и для времени обслуживания, то из преобразования Лапласа имеем:

$$A(s) = \int_0^{\infty} \lambda e^{-\lambda t} e^{-st} dt = \frac{\lambda}{s + \lambda},$$

$$B(s) = \int_0^{\infty} \mu e^{-\mu t} e^{-st} dt = \frac{\mu}{s + \mu}.$$

Общее распределение интервала между выходными заявками будет:

$$D(s) = (1 - P_0)B(s) + P_0 D_0(s) = \rho \frac{\mu}{s + \mu} + (1 - \rho) \frac{\lambda}{s + \lambda} \cdot \frac{\mu}{s + \mu} = \frac{\lambda}{s + \lambda}.$$

Следовательно, плотность вероятности распределения времени между заявками, покидающими узел I , является также случайной величиной с показательным законом распределения с тем же самым параметром. Это значит, что СМО типа М/М/1 превращает пуассоновский поток на входе в пуассоновский поток на выходе с тем же самым параметром. Данный результат

называется теоремой Беркё. На основе этой теоремы можно исследовать многофазные схемы.

Для расчетов сетей СМО обычно используется численный приближенный метод, при этом система должна работать в установившемся режиме и соблюдаться следующие условия [4]:

– все предположения относительно операционных переменных можно проверить измерениями на реальной системе или на её модели;

– в системе должен существовать баланс потоков: количество заявок, поступивших в систему за определенный промежуток времени, должно равняться количеству покинувших систему заявок за это же время;

– переходы требований от одного узла к другому не должны зависеть от длин очередей в узлах.

Для расчета используются основные операционные переменные:

q_{0j} ($j = 1, \dots, K$) – вероятность поступления требований в сеть извне к любому узлу. K – общее количество узлов, 0 – внешняя среда;

q_{kj} ($j = 1, \dots, K$) – вероятность перехода требований из узла k к узлу j , $k = 1, \dots, K$;

A_{0j} – количество требований, поступивших в узел j извне;

A_{jk} ($j = 1, \dots, K$) ($k = 1, \dots, K$) – количество требований, поступивших в узел k из узла j ;

C_{k0} – количество требований, покинувших узел k ;

C_{kj} – количество требований, покинувших узел k и поступивших в узел j ;

B_k – общее время обслуживания требований узлом k ;

T – общее время моделирования системы.

При этом

$$C_k = \sum_{j=1}^K C_{kj}, \quad A_0 = \sum_{j=1}^K A_{0j}, \quad C_0 = \sum_{j=1}^K C_{0j}. \quad (3.7.1)$$

Очевидно, если система замкнута, то $A_0 = C_0$.

Выводимые операционные переменные:

U_k – коэффициент использования узла

$$U_k = \frac{B_k}{T}; \quad (3.7.2)$$

S_k – среднее время обслуживания в узле k

$$S_k = \frac{B_k}{C_k}; \quad (3.7.3)$$

X_k – интенсивность выходящего потока требований из узла k

$$X_k = \frac{C_k}{T}; \quad (3.7.4)$$

$$q_{kj} = \frac{C_{kj}}{C_k}, \quad (k = 1, 2, \dots, K); \quad q_{kj} = \frac{A_{0j}}{A_k}, \quad (k = 0); \quad q_{kj} = \sum_{k=1}^K q_{kj} = 1.$$

Исходя из этих выражений, получим

$$U_k = X_k S_k. \quad (3.7.5)$$

Используя выводимые операционные переменные, можно найти среднее время пребывания требований в сети и количество требований в сети.

Для этого используется гипотеза баланса потоков в сети: количество требований, которые поступили в некоторый узел на протяжении продолжительного периода T , равняется количеству требований, которые покинули этот узел.

Возьмем некоторый узел j , т. к.

$$A_j = C_j = \sum_{k=0}^K C_{jk} = A_j, \quad j = 0, 1, \dots, K; \quad q_{kj} = \frac{C_{kj}}{C_k},$$

получим

$$C_j = \sum_{k=0}^K C_k q_{kj},$$

поделив обе части выражения на время T , придем к уравнению баланса потоков:

$$X_j = \sum_{k=0}^K X_k q_{kj}, \quad j = 0, 1, 2 \dots K. \quad (3.7.6)$$

Для разомкнутой сети уравнение (3.7.6) будет линейно зависимым.

Используя все вышеперечисленные формулы можно определить:

– производительность узла (интенсивность выходящего потока узла k)

$$X_k = \frac{II_k}{S_k}; \quad (3.7.7)$$

– коэффициент посещаемости узла k

$$V_k = \frac{X_k}{X_0}. \quad (3.7.8)$$

Поделив левую и правую части выражения (3.7.6) на X_0 , получим

$$V_0 = 1; \quad V_j = q_{0j} + \sum_{k=1}^K V_k q_{kj}; \quad j = 1, 2, \dots, K. \quad (3.7.9)$$

Пусть W_k – суммарное время ожидания и время обслуживания требования узлом k ;

R_k – среднее время пребывания требования в узле k ;

R – среднее время пребывания требования в сети.

Среднее время пребывания требования в узле:

$$R_k = \frac{W_k}{C_k}. \quad (3.7.10)$$

Среднее время пребывания требования в сети:

$$R = \sum_{k=1}^K V_k R_k . \quad (3.7.11)$$

Среднее количество требований в узле:

$$n_k = \frac{W_k}{T} . \quad (3.7.12)$$

Среднее количество требований в сети:

$$N = \sum_{k=1}^K n_k . \quad (3.7.13)$$

Среднее время пребывания требования в узле можно записать и как

$$R_k = \frac{n_k T}{C_k} = \frac{n_k}{\frac{C_k}{T}} = \frac{n_k}{X_k} - \text{закон Литтла.} \quad n_k = R_k X_k . \quad (3.7.14)$$

Среднее время пребывания в сети по закону Литтла

$$R = \sum_{k=1}^K \frac{n_k}{X_k} \cdot \frac{X_k}{X_0} = \frac{1}{X_0} \sum_{k=1}^K n_k = \frac{N}{X_0} ; \quad N = R X_0 . \quad (3.7.15)$$

Пример 3.4

Имеется фрагмент сети (рис. 9), где $M = 30$ устройств со средним временем обслуживания каждым $Z = 20$ с. Для узлов **а** и **в** сети частоты перехода к узлу **с** равняются $q_{ac} = 0,5$, $q_{vc} = 0,8$, а коэффициенты посещаемости этих узлов равняются $V_a = 15$, $V_v = 20$. Узел **с** используется на 60 % (U_c), среднее время обслуживания в нем требований $S_c = 20$ мс (0,020 с).

Необходимо найти среднее время пребывания и среднее количество требований в сети.

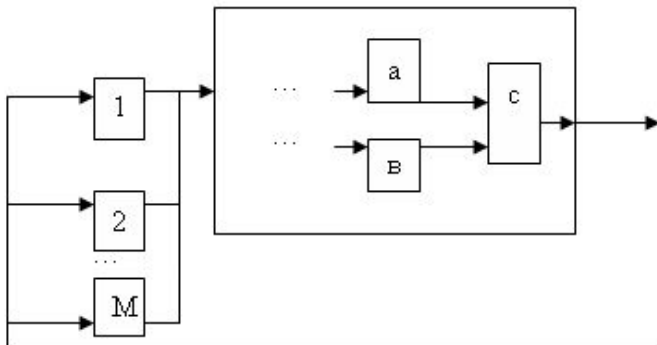


Рис. 9. Фрагмент сети

Для данной замкнутой схемы, где есть M устройств со средним временем обслуживания Z , среднее время пребывания требования в сети определяется с учетом того, что на обслуживание заявки уходит время обслуживания в одном из M устройств и время обслуживания во внешней сети, т. е. $Z + R$. Тогда по формуле Литтла, среднее количество обслуживающих устройств будет

$$M = (Z + R)X_0,$$

$$R = \frac{M}{X_0} - Z.$$

Определим коэффициент посещаемости узла c

$$V_c = V_a q_{ac} + V_b q_{bc} = 15 \cdot 0.5 + 20 \cdot 0.8 = 23.5$$

Интенсивность поступления требований в сеть, треб/с

$$X_0 = \frac{X_c}{V_c} = \frac{U_c}{V_c S_c} = \frac{0.6}{23.5 \cdot 0.020} = 1.3.$$

Тогда

$$R = 30/1.3 = 20 = 3.08 \text{ с}$$

По формуле Литтла среднее количество требований в сети

$$N = R X_0 = 3.08 \cdot 1.3 = 4.$$

3.8. Одноканальная многофазная система с отказами.

Метод приближенной оценки пропускных способностей многоканальной многофазной системы

При рассмотрении одноканальной многофазной системы массового обслуживания с отказами будем считать, что в каждой фазе имеется по одному прибору разной производительности. Пусть имеем три фазы. Конечные зависимости для установившегося процесса следующие [5]:

1. Вероятность того, что все фазы свободны:

$$P_{000} = \frac{\mu_1 \mu_2 \mu_3}{(\lambda + \mu_1)(\lambda + \mu_2)(\lambda + \mu_3)},$$

где μ_1, μ_2, μ_3 – производительность приборов соответственно в первой, второй и третьей фазах.

2. Вероятность того, что занята обслуживанием только последняя фаза:

$$P_{001} = \frac{\lambda \mu_1 \mu_2}{(\lambda + \mu_1)(\lambda + \mu_2)(\lambda + \mu_3)}.$$

3. Вероятность того, что первая фаза свободна, а вторая и третья заняты:

$$P_{011} = \frac{\mu_1^2 \mu_2 \lambda^2 (\mu_1 + \mu_2 + \mu_3)}{(\lambda + \mu_1)(\lambda + \mu_2)(\lambda + \mu_3)(\mu_1 + \mu_2)(\mu_1 + \mu_3)(\mu_2 + \mu_3)}.$$

4. Вероятность того, что свободен прибор второй фазы, а первой и третьей заняты обслуживанием:

$$P_{101} = \frac{\mu_1 \mu_2 \lambda^2 (\lambda + \mu_1 + \mu_2 + \mu_3)}{(\lambda + \mu_1)(\lambda + \mu_2)(\lambda + \mu_3)(\mu_1 + \mu_2)(\mu_1 + \mu_3)(\mu_2 + \mu_3)}.$$

5. Вероятность того, что вторая фаза занята обслуживанием, а первая и третья свободны:

$$P_{010} = \frac{\lambda \mu_1 ((\lambda + \mu_3)(\mu_1 + \mu_2)(\mu_1 + \mu_3)(\mu_2 + \mu_3) + \lambda \mu_1 \mu_2 (\mu_1 + \mu_2 + \mu_3))}{(\lambda + \mu_1)(\lambda + \mu_2)(\lambda + \mu_3)(\mu_1 + \mu_2)(\mu_1 + \mu_3)(\mu_2 + \mu_3)}.$$

6. Вероятность того, что первая фаза занята обслуживанием, а вторая и третья свободны:

$$P_{100} = \frac{\lambda \mu_2 \mu_3 [\lambda (\lambda + \mu_1 + \mu_2 + \mu_3) + (\mu_1 + \mu_2)(\mu_1 + \mu_3)]}{(\lambda + \mu_1)(\lambda + \mu_2)(\lambda + \mu_3)(\mu_1 + \mu_2)(\mu_1 + \mu_3)(\mu_2 + \mu_3)}.$$

7. Вероятность того, что заняты обслуживанием первая и вторая фазы, а третья свободна:

$$P_{110} = \frac{\lambda^2 \mu_1 [(\lambda + \mu_3)(\mu_1 + \mu_2)(\mu_2 + \mu_3) - \lambda \mu_1 \mu_2]}{(\lambda + \mu_1)(\lambda + \mu_2)(\lambda + \mu_3)(\mu_1 + \mu_2)(\mu_1 + \mu_3)(\mu_2 + \mu_3)}.$$

8. Вероятность того, что все фазы заняты обслуживанием:

$$P_{111} = \frac{\lambda \mu_2 \mu_1^2}{(\lambda + \mu_1)(\lambda + \mu_2)(\lambda + \mu_3)(\mu_1 + \mu_2)(\mu_1 + \mu_3)(\mu_2 + \mu_3)}.$$

Как видим из выражений, вероятность состояния не зависит от распределения приборов разной производительности по фазам, а зависит от производительности этих приборов.

Если обозначить через r количество фаз системы, то вероятности состояний можно записать следующим образом:

Вероятность, когда все приборы свободны:

$$P_{0(r)} = \prod_{i=1}^r \frac{\mu_i}{(\lambda + \mu_i)}.$$

Если все приборы заняты обслуживанием:

$$P_{1(r)} = \lambda^r \frac{\prod_{i=1}^{r-1} \mu_i^{r-1}}{\prod_{j=1}^r (\lambda + \mu_j) \prod_{\substack{i=1 \\ k \neq i}}^r (\mu_k + \mu_i)}.$$

Вероятность, когда заявка обслуживается в последней фазе, а остальные свободны:

$$P_{01(r)} = \frac{\lambda}{\mu_r} \prod_{i=1}^r \frac{\mu_i}{(\lambda + \mu_i)}.$$

Вероятность обслуживания каждой заявки всеми фазами:

$$P_{оо\phi r} = \frac{\prod_{i=1}^r \mu_i \left[\prod_{\substack{j=1 \\ i \neq j}}^r (\mu_i + \mu_j) + (\lambda + \sum_{i=1}^r \mu_i)^{r-2} \lambda^{r-2} (\sum_{i=1}^r \mu_i)^{r-2} \right]}{\prod_{i=1}^r (\lambda + \mu_i) \prod_{\substack{j=1 \\ i \neq j}}^r (\mu_i + \mu_j)}.$$

Если производительность одной из фаз очень мала по сравнению с другими, то эта фаза и будет определять пропускную способность системы.

Для многоканальной многофазной системы используется метод замены группы приборов одним прибором с эквивалент-

ной пропускной способностью. Пусть в первой фазе имеется n_1 приборов, а в следующих фазах по одному, тогда вероятность обслуживания заявки во всех фазах будет:

$$P_{обсл} = \frac{\frac{\lambda(1-P_{отк})}{P_{отк}} \prod_{i=2}^n \mu_i \left\{ \prod_{j=1}^n \left[\frac{\lambda(1-P_{отк})}{P_{отк}} + \mu_j \right] \cdot \prod_{j=1}^n (\mu_i + \mu_j) \right\} + \frac{\lambda}{P_{отк}} \prod_{i=2}^n (\lambda + \mu_i) \cdot \prod_{j=1}^n \left[\frac{\lambda(1-P_{отк})}{P_{отк}} + \mu_j \right] \times}{\left(\frac{\lambda}{P_{отк}} + \sum_{i=2}^n \mu_i \right)^{n-2} \lambda^{n-2} \left[\frac{\lambda(1-P_{отк})}{P_{отк}} + \sum_{i=2}^n \mu_i \right]^{n-2} \times \prod_{j=1}^n (\mu_i + \mu_j)}$$

где $P_{отк}$ определяется по формуле Эрланга.

3.9. Модели систем массового обслуживания с потерей. Модель коммутационного узла

На практике при использовании способа коммутации каналов (КК) сеть должна предоставить физический канал (электрическую цепь) от источника к получателю на время сеанса связи. Эта физическая цепь «из конца в конец» может состоять из нескольких звеньев передачи (каналов), которые соединяются друг с другом в устройство коммутации (УК) с помощью коммутационных полей. Звенья передачи могут быть представлены каналами одного из двух типов – каналами тональной частоты (КТЧ) аналоговых систем передач (СП) или каналами цифровых систем передач (ЦСП) с временным разделением каналов. Большинство пользователей сетей с коммутацией каналов обслуживаются с блокировками вызовов. Под блокировкой вызова понимают отказ в предоставлении канала из конца в конец. Доля блокировок определяет качество сети с КК. Объем ресурсов сети и эффективность их использования зависят от допустимой доли блокировок. Характеристики

обслуживания вызовов описываются с помощью вероятности блокировки, времени установления и разъединения соединения. Для сети с КК установлены протоколы соединения и разъединения. Под протоколом соединения (разъединения) понимают:

а) состав (перечень) сигналов, которыми обмениваются абонентская установка с сетью и станции и узлы сети друг с другом;

б) логику обмена сигналами;

в) способ сигнализации (от звена к звену или из конца в конец);

г) параметры сигналов (длительности, уровни и др.). Известно, что при предоставлении обычных услуг телефонной связи требуется передавать десять видов сигналов. Для уяснения принципов коммутации в сетях с КК рассмотрим общую модель УК (рис. 10).



Рис. 10. Общая модель узла коммутации

Приведенная на рисунке модель описывает большое разнообразие систем коммутации (СК). Под СК понимают совокупность средств коммутации и управления, обеспечивающих установление физических соединений входящих линий (каналов) с исходящими. Так, например, М-входами могут быть абонентские линии, а N-выходами — исходящие каналы оконечной станции к одной из станций сети; на узле или на транзитной станции М-входами могут быть входящие каналы (линии) от одной из станции сети, а N-выходами — исходящие каналы к другой станции сети.

Рассмотрим приведенную выше модель. Любой из М-входов может быть либо свободен в течение интервала времени, распределенного по экспоненциальному (показательному) закону со средним значением $1/\lambda$, либо генерировать вызов. Этот вызов может быть обслужен в течение случайного интервала времени, который распределен по экспоненциальному закону со средним значением $1/\mu$. Вызов, поступивший на любой вход, занимает любой свободный выход (такая полная доступность всех выходов пучка характерна для узлов и станций с программным управлением). Если все выходы направления связи заняты, то вызов блокируется, отказывает ему в обслуживании, и уходит из системы массового обслуживания (СМО). Любая СК является СМО, т. к. предоставляет общие ресурсы (обычно ограниченные) большой массе пользователей. Если в СМО установлено n соединений, то она перейдет в стационарный, установившийся режим. Вероятностные характеристики этого режима не будут зависеть от времени. Именно этот режим работы СК интересует нас, поскольку расчеты требуемого количества каналов М выполняются для часа наибольшей нагрузки (ЧНН), когда уже установлено большое количество соединений. В этом режиме на входы СМО поступают вызовы с интенсивностью λ_n и уходят из системы с интенсивностью μ_n . Система, находящаяся в состоянии n , описывается двумя переменными:

$$\lambda_n = (M - n)\lambda = \begin{cases} 0 \leq n \leq N \\ N < M \end{cases},$$

$$\mu_n = n\mu, \quad 1 \leq n \leq N.$$

Стационарный режим СМО описывается уравнением равновесия. В нем устанавливается вероятностная зависимость перехода в состояние $(n + 1)$ от интенсивности поступления вызовов λ_n в состоянии n и от вероятности наличия в модели n установленных соединений для любого $n \geq 0$:

$$\mu_{n+1}p_{n+1} = \lambda_n p_n, \quad n \geq 0,$$

где p_n , p_{n+1} – вероятность существования в СМО n или $(n + 1)$ установленных соединений соответственно.

Вероятности состояний СМО описываются закономерностями, параметры которых существенно зависят от соотношения между M и N . Так, например, для часто встречающегося в практике случая, когда $M \gg N$ (N конечно) и M очень велико, поступление вызовов на входы описывается распределением (законом) Эрланга. Этот закон описывает поведение некоторой случайной величины X (для рассматриваемой СМО – это появление вызовов на входах). Положения этого закона таковы:

- если вызовы расположить на оси времени t (точки на рис. 11), то вероятность попадания того или иного числа вызовов на отрезок L зависит только от длины этого отрезка, а не от положения отрезка на оси времени. Последнее указывает на то, что вызовы распределены во времени с одинаковой средней плотностью (λ), которая характеризует среднее количество вызовов в единицу времени;

- вызовы распределяются во времени независимо друг от друга. Это значит, что вероятность попадания заданного числа вызовов на выбранный отрезок времени не зависит от того, сколько вызовов попало на любой другой отрезок, не перекрывающийся с ним;

- вероятность попадания двух или более вызовов на малый участок Δt пренебрежимо мала по сравнению с вероятностью попадания одного вызова (это эквивалентно невозможности одновременного появления двух вызовов).

Для модели СМО с такими свойствами потока вызовов вероятность блокировки (отказа в обслуживании вызова из-за занятости всех V -выходов) описывается *распределением Эрланга*:

$$E_N(\rho) = \frac{\frac{\rho^N}{N!}}{\sum_{n=0}^N \frac{\rho^n}{n!}},$$

где $\rho = \lambda/\mu$, $E_N(\rho)$ – вероятность занятости (блокировки) всех N -выходов при нагрузке от любого из M -источников. Это выражение верно при $M = \infty$. Использование его при инженерных

расчетах схем с большим количеством входов дает небольшую погрешность.

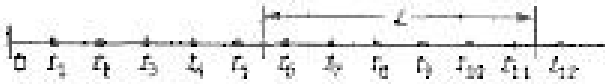


Рис. 11. Моменты поступления потока вызовов Эрланга

Блокировку еще называют потерей вызова (вызов уходит из СМО, теряется), долей потерянных вызовов на практике оценивают качество обслуживания систем с блокировками.

Эта формула, называемая В-формулой Эрланга, играет важную роль в телефонии, ее значения табулированы, и существуют таблицы обратного расчета. С ее помощью можно вычислить вероятность потери запросов (блокировки каналов) при фиксированном числе каналов, интенсивности входящего потока и интенсивности обслуживания запросов. Можно решать также двойственные задачи, например, расчет необходимого числа каналов или допустимого потока запросов, исходя из заданной максимально допустимой вероятности потери запроса. Также для расчетов можно использовать специальный программный калькулятор, который предоставляется на сайте компании Westbay по адресу <http://www.erlang.com/calculator/erlb/>.

С ростом величины N число ρ^N и $N!$ будут быстро расти, может произойти переполнение или исчезновение величины $\rho^N/N!$, поэтому вычисление $E_N(\rho)$ производится с помощью следующей рекуррентной процедуры

$$E_N(\rho) = \frac{\rho E_{N-1}(\rho)}{N + \rho E_{N-1}(\rho)},$$

где $N = 1, 2, \dots$

В первых автоматических системах коммутации использовался исключительно алгоритм обслуживания заявок с потерями. В следующих поколениях коммутационной техники

стали применяться различные виды алгоритмов с ожиданием. Тем не менее, для пучков СЛ чаще всего используется алгоритм обслуживания с потерями. Если свободной СЛ нет, то вызов теряется. Такой способ обслуживания иногда называется алгоритмом с явными потерями.

Если коммутационное устройство работает в режиме с явными потерями, то необходимо различать потери по вызовам, по нагрузке и по времени.

Потери по вызовам $P_{\text{в}}(t_1, t_2)$ за промежуток времени $[t_1, t_2]$ представляют собой отношение числа потерянных вызовов к числу поступающих вызовов:

$$P_{\text{в}}(t_1, t_2) = \frac{c_{\text{т}}(t_1, t_2)}{c(t_1, t_2)}.$$

Для стационарных потоков потери по вызовам

$$P_{\text{в}} = \frac{\mu_{\text{п}}}{\mu},$$

где $\mu_{\text{п}}$ – математическое ожидание числа потерянных вызовов в единицу времени;

μ – математическое ожидание числа поступающих вызовов в единицу времени.

Потери по нагрузкам $P_{\text{н}}(t_1, t_2)$ есть отношение величины потерянной к величине поступающей телефонной нагрузки:

$$P_{\text{н}} = \frac{Y_{\text{п}}(t_1, t_2)}{Y(t_1, t_2)}.$$

Потери по времени $P_{\text{т}}(t_1, t_2)$ представляют собой долю того промежутка, в течение которого все соединительные пути, доступные источнику вызовов, заняты.

Для стационарных потоков

$$P_{\text{т}} = P_{\text{н}},$$

где $P_{\text{в}}$ – вероятность занятости всех соединительных путей.

$$P_{\text{н}} \leq P_{\text{в}} \leq P_{\text{т}}.$$

Указанное неравенство справедливо для любых коммутационных устройств обслуживания (полнодоступные, неполнодоступные, звеньевые включения), на которые поступают простейшие потоки. Потери по вызовам могут быть измерены в условиях станционного эксперимента. Остальные виды потерь используются при расчетах и представляют теоретический интерес.

Пример 3.5

К АТС подключено 10^4 абонентов, каждый из которых создает нагрузку в 0,04 Эрл, причем 10 % вызовов являются междугородними. Требуется найти количество каналов N для междугородних вызовов при условии, что вероятность их блокировки не превысит 1 %.

Решение:

Полагая $M = 10^4$, получим для интенсивностей ρ_1 общей и ρ_2 междугородней нагрузки на АТС следующие значения:

$$\rho_1 = 10^4 \cdot 0,04 = 400 \text{ Эрл}, \quad \rho_2 = 0,1 \cdot \rho_1 = 40 \text{ Эрл},$$

согласно условию задачи

$$E_N(\rho) - \frac{\rho_2 E_{N-1}(\rho_2)}{N + \rho E_{N-1}(\rho_2)} \leq 0,01,$$

отсюда следует

$$N \geq 99 \rho_2 E_{N-1}(\rho_2).$$

Используя таблицы для $E_N(\rho)$ или рекуррентную процедуру, получим при $\rho_2 = 40$ Эрл условие задачи, которое будет выполняться, если $N \geq 53$ канала.

4. МОДЕЛИРОВАНИЕ ТЕЛЕТРАФИКА

4.1. Вычисление трафика

Теория телетрафика – раздел теории массового обслуживания. Основы теории телетрафика заложил датский ученый А.К. Эрланг. Его работы были опубликованы в 1909 – 1928 гг. Дадим важные определения, используемые в теории телетрафика (ТТ). Термин «трафик» (англ. traffic) соответствует термину «телефонная нагрузка». Подразумевается нагрузка, создаваемая потоком вызовов, требований, сообщений, поступающих на входы СМО. Объемом трафика называется пропущенная тем или иным ресурсом величина суммарного, интегрального интервала времени, в течение которого данный ресурс был занят за анализируемый период времени. Единицей работы можно считать секундо-занятие ресурса. Иногда можно прочесть о часо-занятии, а порой и просто – секундах или часах. Однако рекомендации ИТУ дают размерность объема трафика в эрланго-часах. Чтобы понять смысл такой единицы измерения, надо рассмотреть еще один параметр трафика – интенсивность трафика. При этом чаще говорят о средней интенсивности трафика (нагрузки) на некотором заданном пуле (наборе) ресурсов. Если в каждый момент времени t из заданного интервала (t_1, t_2) число занятых обслуживанием трафика ресурсов из данного набора равно $A(t)$, то средняя интенсивность трафика будет

$$A(t_1, t_2) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} A(t) dt .$$

Величина интенсивности трафика характеризуется как среднее число ресурсов, занятых обслуживанием трафика на заданном интервале времени. Единицей измерения интенсивности нагрузки является один Эрланг (1 Эрл , $1 E$), т. е. 1 Эрл – это такая интенсивность трафика, которая требует полной занятости одного ресурса, или, иначе говоря, при которой

ресурсом выполняется работа величиной в одно секундо-занятие за время в одну секунду. Например, если мини-АТС связана с ТФоП по 10 линиям, а в данный момент заняты только 5 из них, то нагрузка на внешние линии будет 5 Эрл. Точно также нагрузку создают и внутренние линии: снятие трубки на внутреннем телефоне создает дополнительную нагрузку на АТС в 1 Эрл. В американской литературе иногда можно встретить другую единицу измерения, называемую CCS – Centrum (or hundred) Calls Second (гектосекундозанятие). Число CCS отражает время занятия серверов в 100 секундных интервалов за 1 час. Интенсивность, измеренную в CCS, можно пересчитать в Эрланги по формуле $36\text{CCS} = 1 \text{ Эрл}$.

Трафик, создаваемый одним источником и выраженный в часо-занятиях, равен произведению числа попыток вызовов c за определенный интервал времени T на среднюю длительность одной попытки t : $y = c \cdot t$ (ч-з). Трафик можно вычислить тремя разными способами:

1) пусть число вызовов c в течение часа равно 1800, а средняя длительность занятия $t = 3 \text{ мин} = 0,05 \text{ ч}$, тогда $Y = 1800 \cdot 0,05 = 90 \text{ Эрл}$;

2) пусть в течение времени T фиксируются длительности t_i всех n занятий выходов некоторого пучка, тогда трафик определяется так:

$$Y = \frac{1}{T} \cdot \sum_{i=1}^n t_i ;$$

3) пусть в течение времени T выполняется наблюдение через равные промежутки времени за количеством одновременно занятых выходов некоторого пучка, по результатам наблюдений строят (рис. 12) ступенчатую функцию времени $x(t)$.

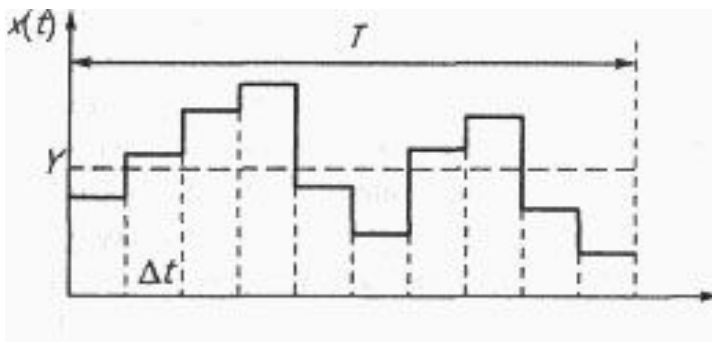


Рис. 12. Отсчеты одновременно занятых выходов пучка

Трафик в течение времени T может быть оценен как среднее значение $x(t)$ за это время:

$$Y = \frac{1}{T} \cdot \sum_{i=1}^n x_i(t) \Delta t_i,$$

где n – число отсчетов одновременно занятых выходов. Величина Y есть среднее количество одновременно занятых выходов пучка в течение времени T .

Телефонная нагрузка меняется в течение дня (колебания можно проследить по дням недели, месяцам, годам). Расчет оборудования, необходимого для обслуживания телефонного трафика, осуществляется для часа наибольшей нагрузки (ЧНН). В некоторых публикациях предлагается – в дополнение к ЧНН – выделять период наибольшей нагрузки. Его длительность (T_{MAX}) может быть различной в зависимости от цели, для которой необходимо знать период максимальной нагрузки. Она может быть меньше или больше одного часа. В данном случае T_{MAX} составляет три часа (рис. 13).

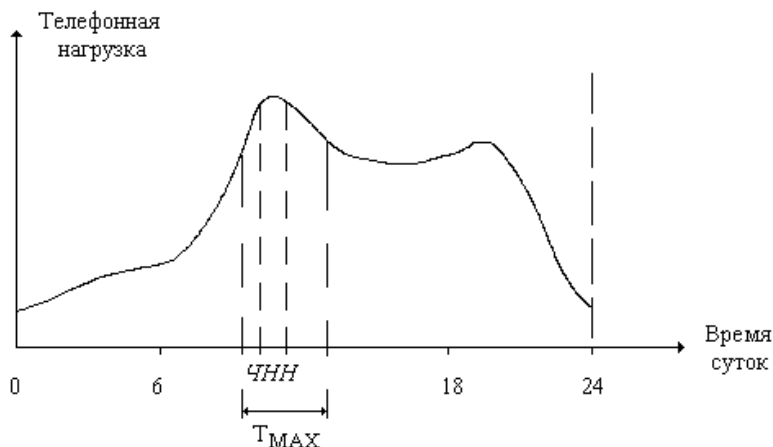


Рис. 13. Колебания трафика в течение суток

Выделение ЧНН или периода максимальной нагрузки преследует конкретную цель. Установленные Администрацией связи показатели качества обслуживания телефонной нагрузки должны соблюдаться и в периоды высокого трафика. Естественно, что в другое время (в частности, ночью) ресурсы телефонной сети будут использоваться не так эффективно как в ЧНН. Правда, Оператор с помощью технических и маркетинговых операций может обеспечить экономичное функционирование инфокоммуникационной системы.

Правила измерения трафика разработаны МСЭ-Т и используются администрациями национальных сетей для того, чтобы удовлетворить требованиям качества предоставляемых услуг, как для абонентов своей сети, так и для абонентов других сетей, связанных с ней. Теорию телетрафика можно использовать для практических расчетов потерь или объема оборудования станции (узла) только в том случае, если трафик стационарный (статистически установившийся). Этому условию приближенно удовлетворяет трафик в ЧНН. Величина нагрузки, поступающая за сутки на АТС, влияет на профилактику и ре-

монт оборудования. Неравномерность поступления нагрузки на станцию в течение суток определяется коэффициентом концентрации:

$$k_k = \frac{Y_{\text{чнн}}}{Y_{\text{сут}}},$$

где $Y_{\text{чнн}}$ – нагрузка в ЧНН,

$Y_{\text{сут}}$ – нагрузка за сутки.

Более строго определение ЧНН производится следующим образом. Рекомендация ITU E.500 предписывает проанализировать данные об интенсивности за 12 месяцев, выбрать из них 30 наиболее загруженных дней, найти в эти дни наиболее загруженные часы и усреднить результаты измерения интенсивности на этих интервалах. Такой расчет интенсивности трафика (нагрузки) называется нормальной оценкой интенсивности трафика в ЧНН или уровнем А. более жесткой оценки можно проводить усреднение за 5 самых загруженных дней выбранного 30-дневного периода. Такая оценка носит название повышенной или оценкой по уровню В.

Процесс создания трафика. Как известно каждому пользователю телефонной сети, не все попытки установления соединения с вызываемым абонентом заканчиваются успешно. Иногда приходится делать несколько неудачных попыток, прежде чем будет установлено желаемое соединение.

Рассмотрим возможные события при моделировании установления соединения между абонентами А и Б (рис. 14). Статистические данные о вызовах в телефонных сетях таковы: доля состоявшихся разговоров составляет 70 – 50 %, доля несостоявшихся – 30 – 50 %. Любая попытка абонента занимает вход СМО. При удачных попытках (когда разговор состоялся) время занятия коммутационных приборов, устанавливающих соединения входов с выходами, больше, чем при неудачных попытках. Абонент может в любой момент времени прервать попытки установления соединения. Повторные попытки могут быть вызваны следующими причинами:

- номер набран неправильно;
- предположение об ошибке в работе сети;
- степень срочности разговора;
- неудачные предыдущие попытки;
- знание привычек абонента Б;
- сомнение в правильности набора номера.



Рис. 14. Диаграмма событий при установлении соединения между абонентами

Повторная попытка может быть предпринята в зависимости от следующих обстоятельств:

- степени срочности;
- оценки причины неуспеха;
- оценки целесообразности повторения попыток;
- оценки приемлемого интервала между попытками.

Отказ от повторной попытки может быть связан с низкой степенью срочности. Различают несколько видов трафика, создаваемого вызовами: *поступающий (предложенный)* Y_n и *пропущенный* Y_{np} . Трафик Y_n включает все успешные и неуспешные попытки, трафик Y_{np} , являющийся частью Y_n , включает успешные и часть неуспешных попыток:

$$Y_{np} = Y_p + Y_{nt} ,$$

где Y_p – разговорный (полезный) трафик, а Y_{np} – трафик, созданный неудачными попытками. Равенство $Y_n = Y_p$ возможно лишь в том идеальном случае, если нет потерь, ошибок вызывающих абонентов и нет ответов вызываемых абонентов.

Разность между поступающей и пропущенной нагрузками за определенный промежуток времени будет являться потерянной нагрузкой.

Прогнозирование трафика. Ограниченность ресурсов приводит к необходимости поэтапного расширения станции и сети. Администрация сети делает прогноз увеличения трафика в течение этапа развития, учитывая, что:

- доход определяется частью пропущенного трафика Y_p , затраты определяются качеством обслуживания при наибольшем трафике;

- большая доля потерь (низкое качество) бывает в редких случаях и характерна для конца периода развития;

- наибольший объем пропущенного трафика приходится на периоды, когда потери практически отсутствуют. Если потери меньше 10 %, то абоненты на них не реагируют. При планировании развития станций и сети проектировщик должен ответить на вопрос, каковы требования к качеству предоставления услуг (к потерям). Для этого нужно проводить измерения трафика потерь по принятым в стране правилам.

Пример измерения трафика. Сначала рассмотрим, как можно отображать работу СМО, имеющую несколько ресурсов, которые одновременно обслуживают некоторый трафик. Будем далее говорить о таких ресурсах, как о серверах, которые обслуживают поток заявок или требований. Одним из наиболее наглядных и часто употребляемых способов изображения процесса обслуживания заявок пулом серверов является диаграмма Ганта (Gantt). Эта диаграмма представляет собой прямоугольную систему координат, ось абсцисс которой изображает время, а на оси ординат помечаются дискретные точки, соответствующие серверам пула. На рис. 15 изображена диаграмма Ганта для системы с тремя серверами.

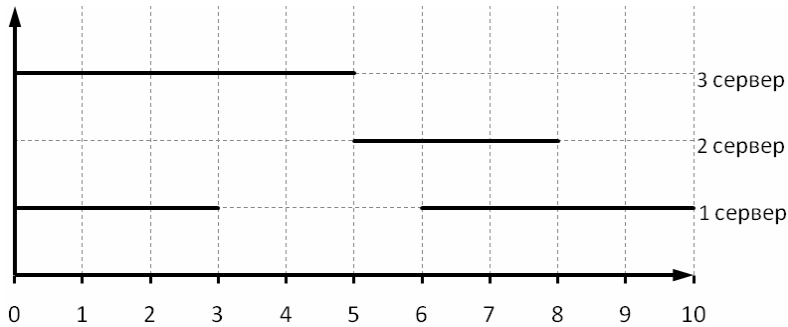


Рис. 15. Диаграмма Ганта для системы с тремя серверами

Построенная диаграмма позволяет произвести расчеты объема трафика и его интенсивности. Диаграмма отражает только обслуженный или пропущенный трафик, поскольку ничего не говорит о том, поступали ли в систему заявки, которые не смогли быть обслужены серверами.

В первые три интервала времени (считаем их секундой) заняты первый и третий серверы, следующие две секунды – только третий, затем одну секунду работает второй, потом, две секунды второй и первый, и последние две секунды – только первый. Пусть поток заявок и обслуживания простейшие.

Объем пропущенного трафика вычисляется как суммарная длина всех отрезков диаграммы Ганта. Объем за время 10 с:

$$Y(0,10) = 3 \cdot 2 + 2 \cdot 1 + 1 \cdot 1 + 2 \cdot 2 + 2 \cdot 1 = 15 \text{ (с} \cdot \text{з)}$$

Свяжем с каждым временным интервалом, отложенным по оси абсцисс, целое число, равное количеству серверов, занятых на этом единичном интервале. Эта величина $A(t)$ – мгновенная интенсивность. Для нашего примера

$$A(t) = (2, 2, 2, 1, 1, 1, 2, 2, 1, 1).$$

Найдем теперь среднюю за период 10 с интенсивность трафика

$$A = \frac{1}{10} \sum_{t=1}^{10} A(t) = \frac{2+2+2+1+1+1+2+2+1+1}{10} = 1.5 E.$$

Таким образом, средняя интенсивность трафика, пропущенного рассматриваемой системой из трех серверов, равна 1,5 Эрл.

4.2. Основные параметры нагрузки

Телефонной связью пользуются различные категории абонентов, которые характеризуются:

- числом источников нагрузки – N ;
- средним числом вызовов от одного источника за определенное время (ЧНН обычно) – c ;
- средней длительностью одного занятия коммутационной системы при обслуживании одного вызова – t .

Величина интенсивности нагрузки будет

$$Y = Nct.$$

Определим различные источники вызовов. Например,

$c_{\text{учр.}}$ – среднее число вызовов в ЧНН от одного учрежденческого аппарата;

$c_{\text{инд.}}$ – среднее число вызовов от одного квартирного индивидуального аппарата;

$c_{\text{колл.}}$ – то же от аппарата коллективного пользования;

$c_{\text{м.а}}$ – то же от одного монетного аппарата;

$c_{\text{с.л.}}$ – то же от одной соединительной линии.

Тогда среднее число вызовов от одного источника:

$$c = \frac{c_{\text{учр.}} N_{\text{учр.}} + c_{\text{инд.}} N_{\text{инд.}} + c_{\text{колл.}} N_{\text{колл.}} + c_{\text{м.а}} N_{\text{м.а}} + c_{\text{с.л.}} N_{\text{с.л.}}}{N}. \quad (4.2.1)$$

Существуют приблизительные данные для среднего числа вызовов от одного источника, соответствующей категории:

$$c_{\text{учр.}} = 3,5 - 5; \quad c_{\text{инд.}} = 0,5 - 1; \quad c_{\text{кол.}} = 1,5 - 2; \\ c_{\text{м.а}} = 15 - 30; \quad c_{\text{сл}} = 10 - 30.$$

Различают следующие виды соединений, которые в зависимости от исхода соединения создают на станции различную по величине телефонную нагрузку:

k_p – коэффициент, показывающий долю соединений, окончившихся разговором;

k_3 – соединения, не окончившиеся разговором из-за занятости вызываемого абонента;

$k_{\text{но}}$ – коэффициент, выражающий долю соединений, не окончившихся разговором по причине неответа вызываемого абонента;

$k_{\text{ош}}$ – соединения, не окончившиеся разговором из-за ошибок вызывающего абонента;

$k_{\text{тех.}}$ – вызовы, не окончившиеся разговором по техническим причинам.

При нормальной работе сети значения этих коэффициентов равны:

$$k_p = 0,60 - 0,75; \quad k_3 = 0,12 - 0,15; \quad k_{\text{но}} = 0,08 - 0,12; \\ k_{\text{ош}} = 0,02 - 0,05; \quad k_{\text{тех.}} = 0,005 - 0,01.$$

Средняя длительность занятия зависит от видов соединений. Например, если соединение закончилось разговором, средняя длительность занятия приборов $t_{\text{сост}}$ будет равна

$$t_{\text{сост}} = t_{\text{уст.соед.}} + t_{\text{в}} + t_{\text{р}}, \quad (4.2.2)$$

где $t_{\text{уст.соед.}}$ – длительность установления соединения;

$t_{\text{сост.}}$ – состоявшийся разговор;

$t_{\text{в}}$ – длительность послышки вызова в телефонный аппарат вызываемого абонента;

$t_{\text{р}}$ – длительность разговора.

$$t_{\text{уст.соед.}} = t_{\text{co}} + 1,5n + t_{\text{c}}, \quad (4.2.3)$$

где t_{co} – сигнал ответа станции;

$1,5n$ – время набора номера вызываемого абонента (n – количество знаков в номере);

t_{c} – время, необходимое для установления соединения коммутационными механизмами и разъединения соединения после окончания разговора.

Примерные значения рассмотренных величин:

$$t_{\text{co}} = 3 \text{ с}; \quad t_{\text{c}} = 1 - 2,5 \text{ с}; \quad t_{\text{в}} = 8 - 10 \text{ с}; \quad t_{\text{p}} = 90 - 130 \text{ с}.$$

Вызовы, не окончившиеся разговором, тоже создают телефонную нагрузку.

Среднее время занятия приборов при занятости вызываемого абонента равно

$$l_{\text{з}} = l_{\text{уст.соед.}} + l_{\text{зз}}, \quad (4.2.4)$$

где $t_{\text{уст.соед.}}$ определяется по (4.2.3),

$t_{\text{зз}}$ – время слушания зуммера занятости, $t_{\text{зз}} = 6 \text{ с}$.

Средняя длительность занятия приборов при неответе вызываемого абонента равна

$$t_{\text{пз}} = t_{\text{уст.соед.}} + t_{\text{пв}}, \quad (4.2.5)$$

где $t_{\text{пв}}$ – время слушания сигнала контроля посылки вызова, $t_{\text{пв}} = 20 \text{ с}$.

Если разговора не было из-за ошибок абонента, то в среднем $t_{\text{ош}} = 30 \text{ с}$.

Длительность занятий, не окончившихся разговором по техническим причинам, не определена, т. к. процент таких занятий мал.

Из всего вышесказанного следует, что полная нагрузка, создаваемая группой источников за ЧНН, равна сумме нагрузок отдельных видов занятий.

$$Y = Y_{\text{сост}} + Y_{\text{зн}} + Y_{\text{но}} + Y_{\text{ош}} + Y_{\text{тех}} = \alpha k_p N c (t_{\text{уст.соед.}} + t_{\text{в}} + t_p), \quad (4.2.6)$$

где α – коэффициент, учитывающий слагаемые $Y_{\text{зн}}$, $Y_{\text{ош}}$, $Y_{\text{но}}$ и $Y_{\text{тех}}$ как доли $Y_{\text{сост}}$, $\alpha = 1,05 - 1,15$.

Пример 4.1

На телефонной сети с семизначной нумерацией запроектирована АТС, структурный состав абонентов которой следующий:

$$N_{\text{учр.}} = 4000, N_{\text{инд.}} = 1000, N_{\text{кол.}} = 2000, N_{\text{м.а}} = 400, N_{\text{с.л.}} = 400.$$

Определить нагрузку, поступающую на станцию Y , среднюю длительность занятия t , если известно, что

$$c_{\text{учр.}} = 4, c_{\text{инд.}} = 1, c_{\text{кол.}} = 2, c_{\text{м.а}} = 10, c_{\text{с.л.}} = 12, t_p = 120 \text{ с}, t_{\text{в}} = 10 \text{ с}, k_p = 0,6, t_c = 1 \text{ с}, \alpha = 1,1.$$

Решение:

Среднее число вызовов, поступающее от одного источника в ЧНН, равно

$$c = \frac{N_{\text{учр.}} \cdot c_{\text{учр.}} + N_{\text{инд.}} \cdot c_{\text{инд.}} + N_{\text{кол.}} \cdot c_{\text{кол.}} + N_{\text{м.а}} \cdot c_{\text{м.а}} + N_{\text{с.л.}} \cdot c_{\text{с.л.}}}{N_{\text{учр.}} + N_{\text{инд.}} + N_{\text{кол.}} + N_{\text{м.а}} + N_{\text{с.л.}}} =$$

$$= \frac{4000 \cdot 4 + 1000 \cdot 1 + 2000 \cdot 2 + 400 \cdot 10 + 400 \cdot 12}{4000 + 1000 + 2000 + 400 + 400} = 3,8.$$

По формулам (4.2.3) и (4.2.6) находим нагрузку

$$Y = \alpha k_p N c (t_{\text{соед.}} + 1,5n + t_c + t_{\text{в}} + t_p) = 2826767 \text{ с} \cdot \text{з} = 785,2 \text{ ч} \cdot \text{з}.$$

Средняя длительность занятия t из формулы $Y = Nct$

$$t = \frac{Y}{Nc} = \frac{2826767}{7800 \cdot 3,8} = 95,4 \text{ с}.$$

4.3. Методы оценки параметров и моделирования трафика мультисервисной сети

Мультисервисная сеть связи — это единая телекоммуникационная инфраструктура для переноса, коммутации трафика произвольного типа, порождаемого взаимодействием потребителей и поставщиков услуг связи с контролируемыми и гарантированными параметрами трафика. Это сеть, которая использует единый канал для передачи данных разных видов трафика. Мультисервисная сеть является результатом сближения компьютерных и телекоммуникационных сетей разных видов. Прежде всего, можно видеть сближение видов услуг, предоставляемых клиентам. Первой попыткой создания мультисервисной сети, способной оказывать различные услуги, в том числе услуги телефонии и передачи данных, была технология цифровых сетей с интегрированным обслуживанием (Integrated Services Digital Network, ISDN). Сегодня на роль глобальной мультисервисной сети нового поколения, часто называемой Next Generation Network (NGN), или New Public Network (NPN), претендует Интернет.

Технологическое сближение сетей происходит на основе цифровой передачи информации, метода коммутации пакетов и программирования услуг. Благодаря возможности представлять голос в цифровой форме, можно телефонный и компьютерный трафик передавать по одним и тем же цифровым каналам.

Существует три различные схемы коммутации абонентов в сетях: коммутация каналов, коммутация пакетов и коммутация сообщений. Коммутация пакетов — основной способ передачи данных, что обусловлено малыми задержками при передаче данных через системы передачи данных. Кроме того коммутация пакетов и сообщений позволяет передавать данные по линиям связи с любой пропускной способностью, дает возможность мультиплексирования потоков данных. Например, паузы в телефонном разговоре могут составлять до 40 % общего времени соединения, а пакетная коммутация позволяет «вырезать» паузы и использовать высвободившуюся

пропускную способность канала для передачи трафика других абонентов. Коммутация каналов требует, чтобы все соединительные линии, из которых формируется канал, имели одинаковую пропускную способность, что ужесточает требования к структуре системы передачи данных. Однако нельзя говорить, что методы коммутации каналов морально устарели, и у них нет будущего. Они будут находить свое применение, но уже в новых технологиях.

Выбор методов коммутации – сложная оптимизационная задача. Она решается исходя из требований к транспортной сети, которые в свою очередь определяются особенностями трафика, классом пользователей и показателями качества их обслуживания.

Мультисервисная сеть отличается надежностью, характерной для телефонных сетей (в противоположность негарантированному качеству связи через Интернет), и обеспечивает низкую стоимость передачи в расчете на единицу объема информации (приближающуюся к стоимости передачи данных по Интернету).

Мультисервисная сеть представляет собой многоуровневую конструкцию, определяемую требованиями трех категорий пользователей:

- корпоративный пользователь;
- малый бизнес;
- частный пользователь.

Основным фактором, который определяет уровень развития сети, является базовая сеть коммутации и переноса трафика. С ней тесно связана сеть абонентского доступа. Управление ресурсами сети (базовая сеть + доступ) включает в себя не только мониторинг и сбор статистики, но и принимает участие в процессе функционирования базовой сети и абонентского доступа: обеспечивает живучесть сети, управляет перегрузками, очередями, маршрутизацией трафика и ведение единой базы данных сети, ведение соглашений о сервисе и т.п. При распределении сервисных информационных потоков уровень

управления сервисами и уровень управления ресурсами сети должны выступать согласованно.

Регулирование трафика в мультисервисных сетях

Как известно, существует полезный трафик, несущий информационные сообщения, и служебный трафик, обеспечивающий целостное функционирование сети. Полезный трафик – поток битов телефонных сообщений, пакетов, видеовещания, электронная почта, HTTP-пакеты и т. д.

В мультисервисных сетях особое значение приобретает анализ основных параметров качества обслуживания (QoS): надежность доставки, задержка, флуктуации задержки и скорость передачи, по отношению к каждому потоку.

В зависимости от типа приложения требуются разные параметры QoS. В табл. 4.1 показаны некоторые виды приложений с оценкой требований к различным параметрам QoS для них [3].

Таблица 4.1
Характеристики приложений

Приложения	Надежность	Задержка	Флуктуации	Скорость
Электронная почта	высокая	низкая	низкие	низкая
Web-дступ	высокая	средняя	низкие	средняя
Аудио по заказу	низкая	низкая	высокие	средняя
Телефония	низкая	высокая	высокие	низкая
Видеоконферен-ция	низкая	высокая	высокие	высокая

Задержка (delay) – характеризует интервал между приемом и передачей пакетов; флуктуация задержки (jitter) – параметр, описывающий возможные отклонения от времени задержки при передаче пакетов; готовность предоставляемого сервиса определяет надежность соединений пользователей с сервис-провайдером; производительность, или пропускная способность

– скорость передачи пакетов в сети (выделяют среднюю и пиковую скорости).

Для надежности доставки обычно используется вероятность блокировки пакетов, для задержки – среднее время пребывания пакета в системе, анализ флуктуаций сводят к оцениванию дисперсии времени ожидания пакетов в буфере, анализ пропускной способности – определение максимального значения параметра входного потока, при котором обеспечивается допустимое значение других параметров QoS.

В зависимости от типа предоставляемых услуг выделяют две категории мультимедийного трафика [6]:

- трафик реального времени, предоставляющий услуги пользователям в реальном масштабе времени (IP-телефония, видеовещание, видеотелефония, видеоконференция и т. д.);
- трафик обычных данных, который образуется традиционными распределенными услугами современной телекоммуникационной сети (электронная почта, передача файлов, и т. д.).

Современные сети связи являются сложными динамическими системами, поэтому можно применить систему массового обслуживания, которая описывает поведение быстропротекающих процессов. Эта методика моделирования может применяться для прогнозирования, анализа и контроля качества работы мультисервисной сети.

На сегодняшний день не существует единой методики для расчета, прогнозирования и анализа трафика мультисервисных сетей.

Основные характеристики трафика

При решении задачи распределения сетевых ресурсов между различными службами, абонент каждой службы характеризуется, с одной стороны, традиционными параметрами трафика:

- интенсивностью входящего потока заявок на предоставление услуг k -й службы $\eta^{(k)}$, выз/ч;
- средней длительностью сеанса связи $T_c^{(k)}$, с;
- удельной интенсивностью нагрузки $\gamma^{(k)}$, Эрл.

С другой стороны, параметрами случайного процесса, характеризующими конкретного абонента k -й службы, ШЦСИО являются:

- пиковая (максимальная) битовая скорость передачи $B^{(k)}_{max}$, бит/с;
- средняя битовая скорость передачи B_{cp} , бит/с;
- пачечность $\kappa^{(k)}_{II}$, определяемая отношением $B^{(k)}_{max} / B_{cp}$, бит/с;
- среднее время пика $T_p^{(k)}$, с.

Реальный размер передаваемых по сети данных складывается из непосредственно данных и необходимого информационного обрамления, составляющего накладные расходы на передачу. Многие технологии устанавливают ограничения на минимальный и максимальный размеры пакета. Так, например, для технологии X.25 максимальный размер пакета составляет 4096 байт, а в технологии Frame Relay максимальный размер кадра – 8096 байт. Можно выделить четыре наиболее общие характеристики трафика:

- «взрывоопасность»;
- терпимость к задержкам;
- время ответа;
- емкость и пропускная способность.

Эти характеристики с учетом маршрутизации, приоритетов, соединений и т. д. как раз и определяют характер работы приложений в сети. Больше всего проблем возникает при попытке «собрать» множество однофункциональных сетей в одну гибкую мультисервисную сеть. Еще труднее получить такую сеть, которая бы могла разрешить абсолютно все проблемы, хотя бы в обозримом будущем.

Характеристики законов формирования трафика

Существует несколько законов формирования изменяющегося трафика: Uniform, Exponential, Normal, Log Normal, Gamma, Erlang, Weibull. Объем передаваемой информации может быть установлен в битах, байтах, килобитах, килобайтах, мегабитах, мегабайтах, гигабитах или гигабайтах. Были исследованы законы распределения объемов передаваемых сообщений: Constant, Exponential, Uniform,

Normal, Weibull в сочетании с законами распределения интенсивностей Constant, Exponential, Uniform, Normal, Lognormal, Gamma, Erlang, Weibull.

Анализ характеристик основных сетевых технологий

Методы коммутации и передачи данных. В информационно-вычислительных сетях (ИВС) коммутация пакетов – основной способ передачи данных. Представление данных пакетами создает наилучшие условия для мультиплексирования потоков данных. Благодаря этому эффективно используется линия связи, соединяющая узлы связи и компьютеры с СПД, и одна линия обеспечивает работу многих взаимодействующих абонентов. Экономичность коммутации пакетов несколько снижается из-за размножения заголовков, сопровождающих каждый пакет, но эти потери окупаются за счет эффекта мультиплексирования сильно пульсирующего потока данных, характерных для ИВС. Малая длина пакетов позволяет выделять для промежуточного хранения передаваемых данных меньшую емкость памяти, чем требуется для сообщений. Также использование пакетов упрощает задачу управления потоками данных, поскольку для приема потока пакетов в узлах связи требуется меньше резервной памяти. Пакеты значительно лучше согласуются с механизмом перезапросов, чем сообщения, и обеспечивают наилучшее использование пропускной способности линии связи, работающей в условиях помех. Выбор длины пакетов производится исходя из размера сообщения с учетом влияния длины пакетов на время доставки данных, пропускную способность ЛС, емкость памяти и загрузку компьютера. Наиболее широко используются пакеты длиной 1024 бит (128 байт). При такой длине все управляющие сообщения и большинство сообщений, генерируемых в режиме диалоговой обработки, «вкладываются» в один пакет.

Методы доступа и кадры для сетей Ethernet, Token Ring, ATM. Обработку кадров, передаваемых по сети, выполняют сетевой адаптер, устанавливаемый в слот расширения станции, и соответствующий ему драйвер. Сетевой адаптер (СА) и драйвер СА реализуют следующие функции:

- поддерживают метод доступа в сети;
- формируют и анализируют кадры, передаваемые по сети.

В зависимости от поддерживаемого метода доступа и типа кадра сетевые адаптеры можно разделить на несколько групп: Ethernet, Token Ring, ARCNet, FDDI и др. Сети, где устанавливаются перечисленные адаптеры, имеют те же названия: сети Ethernet, сети Token Ring и т. д.

При моделировании нагрузки мультисервисной сети необходимо учитывать такие особенности, как:

- метод коммутации;
- используемые протоколы на разных уровнях модели OSI;
- обработка пакета протоколами нижележащих уровней.

Каждое сетевое приложение формирует свой поток данных, который необходимо передать по сети. К основным сетевым приложениям относятся:

- Internet – приложения (FTP, HTTP, E-mail);
- CAD/CAM;
- Telnet;
- сетевые базы данных;
- файловый обмен.

При поступлении пакета от приложения протокол TCP/IP на транспортном уровне оценивает его размер и делит его на пакеты (если размер его слишком велик), которые передаются межсетевому уровню (т. е. протоколу IP). Последний формирует свои IP-пакеты. Затем происходит их «упаковка» в кадры (frame), приемлемые для данной физической среды передачи информации. При этом каждый пакет снабжается заголовком данного протокола. Размер заголовка протокола TCP насчитывает 12 байт. Размер заголовка протокола IP также насчитывает 12 байт. При прохождении пакета через протоколы IP и TCP их размер увеличивается на величину, равную размеру служебной информации. Также происходит и на канальном уровне, где пакеты еще снабжаются заголовками протоколов данного уровня. Величину, характеризующую данное явление обозначим как k_m , где m обозначает уровень семиуровневой модели OSI. Величины данных коэффициентов рассчитаны как

отношение общего размера пакета на уровне m (K_m) к размеру полезных данных приложения

$$k_m = K_m / k_{\text{прил}}.$$

Величины коэффициентов, рассчитанные для протокола TCP/IP и для наиболее распространенных и перспективных сетевых технологий Ethernet, Token Ring и ATM представлены в табл. 4.2.

Таблица 4.2

Величины коэффициентов k_m

$k_{\text{Eth.II}}$	$k_{\text{Eth.802.3}}$	$k_{\text{Eth.802.2}}$	$k_{\text{Eth.SNAP}}$	k_{TR}	k_{ATM}	k_{TCP}	k_{IP}
1,034	1,034	1,038	1,044	1,005	1,104	1,0156	1,0156

За основу разрабатываемой методики была взята методика Назарова для ATM. Если известно количество абонентов на каждом объекте и интенсивность потоков пакетов, генерируемые абонентами каждой службы, то ожидаемую нагрузку на объекте i k -ой службы можно определить по формуле:

$$B_{i\text{np}}^k = N_{\text{аб}i}^k \cdot T_c^k \cdot \gamma_i^k \cdot B_{\text{max}}^k,$$

где $N_{\text{аб}i}^k$ – число абонентов k -ой службы на объекте i ;

γ_i^k – число заявок, поступающих от абонента k -ой службы в единицу времени, равно: $\gamma_i^k = l/t$;

T_c^k – средняя длительность сеанса связи абонента в единицу времени;

B_{max}^k – максимальная скорость k -ой службы. Тогда ожидаемая нагрузка, генерируемая абонентами i -го узла связи: $B_{\text{np}} = B_{i\text{np}} \cdot k$, где k – количество служб. Следует напомнить, что найденная нагрузка предъявляет требования к пропускной способности, необходимой на прикладном уровне. На транспортном уровне данная нагрузка увеличится на величину k_{TCP} . Ожидаемая нагрузка на транспортном уровне будет равна:

$$B_{\Sigma\text{mp}} = k_{\text{TCP}} \cdot B_{\Sigma\text{np}}.$$

Аналогично рассчитывается нагрузка на сетевом и канальном уровнях

$$B_{\Sigma \text{сет}} = k_{\text{лр}} \cdot B_{\Sigma \text{лр}},$$

$$B_{\Sigma \text{кан}} = k_{\text{кан}} \cdot B_{\Sigma \text{сет}}.$$

Отметим, что предыдущие расчеты нагрузки относились к сетям с коммутацией пакетов. Если сеть использует коммутацию каналов, то расчет нагрузки рекомендуется проводить по следующей методике. Каждая индивидуальная АЛ i -ой категории характеризуется в ЧНН интенсивностями нагрузок при местной связи в пределах зоны $Y_{\text{вн},i}$. Выделяют 6 абонентских категорий, а т.к. статистические данные отсутствуют, то используются рекомендуемые значения для прогнозирования нагрузок абонентов этих категорий.

Для расчета интенсивности нагрузок внутризональной связи $Y_{\text{вн},j}$ используют формулу:

$$Y_{\text{вн},j} = \sum_i (N_{i,j} \cdot y_{\text{вн},i}),$$

где $N_{i,j}$ – число абонентов i -й категории на станции j .

Расчет интенсивностей нагрузок межзональной связи будем осуществлять, исходя из соотношения

$$Y_{\text{внеш},j} = Y_{\text{ис},j} + Y_{\text{вх},j} = 0,5 \cdot Y_{\text{вн},j},$$

а межрайонной

$$Y_{\text{внеш},p} = 0,2 \cdot Y_{\text{вн},p},$$

где $Y_{\text{внеш},p}$, $Y_{\text{вн},p}$ – соответственно внешняя и внутренняя интенсивность нагрузки всего проектируемого района.

ГЛОССАРИЙ

Анализ трафика – обзор информации, касающейся связи между пользователями (например, наличие/отсутствие, частота, направление, тип, объем и т. д.).

Аналоговый сигнал – непрерывный сигнал непрерывного времени.

Информация – сведения, являющиеся объектом передачи, распределения, преобразования, хранения или непосредственного использования.

Аппроксимация – математический метод, состоящий в замене одних математических объектов другими, в том или ином смысле близкими к исходным, но более простыми. Аппроксимация позволяет исследовать числовые характеристики и качественные свойства объекта, сводя задачу к изучению более простых и удобных объектов.

АТМ – Asynchronous Transfer Mode (асинхронный режим передачи).

Стандартизованная ITU технология коммутации пакетов фиксированной длины - ячеек (cell). Режим АТМ является асинхронным в том смысле, что ячейки от отдельных пользователей передаются аperiодически.

Дисперсия случайной величины – характеризует меру рассеяния случайной величины.

КК (коммутация каналов) – совокупность операций по соединению каналов для получения сквозного канала, связывающего через узлы коммутации один конечный пункт с другим. При КК сначала организуется через узлы коммутации сквозной канал передачи сообщений между взаимодействующими абонентами, а затем осуществляется передача сообщений.

Коммутатор – коммутационная схема (матрица) с n -входами и m -выходами, в которой соединение входа с выходом устанавливается с помощью одного коммутационного элемента.

Коммутационное поле – коммутационная схема станции (узла), обеспечивающая коммутацию входящих каналов (линий) с исходящими каналами (линиями).

КП (коммутация пакетов) – разновидность коммутации с накоплением, при которой длинные сообщения передаются не целиком, а разбиваются на относительно короткие части – пакеты.

Коммутация сообщений – разновидность коммутации с накоплением, при которой сообщения передаются целиком.

Корреляция – взаимозависимость, связь переменных, при которой одному значению одного признака соответствует несколько значений другого признака, отклоняющегося в ту или иную сторону от своего среднего значения.

ЛВС – локальная вычислительная сеть.

Математическое ожидание – среднее значение случайной величины.

Сеть связи – совокупность узлов и соединительных трактов, предусматривающих соединение между двумя или более точками обеспечения связи между ними.

Синхронизация – процесс установления и поддержания определенных временных соотношений между двумя или несколькими процессами.

Скорость передачи информации – число бит, передаваемых в секунду.

СЛ – соединительные линии.

СМО – система массового обслуживания.

Сообщение – форма представления информации.

Среднеквадратичное отклонение – (стандартное отклонение), корень второй степени из дисперсии.

Трафик телефонный – телефонная нагрузка, создаваемая вызовами пользователей.

ЧНН – час наибольшей нагрузки – интервал в течение суток, когда телефонная нагрузка на станции или в направлении связи является стационарной и наибольшей.

БИБЛИОГРАФИЧЕСКОЕ СПИСОК

1. Ушаков И.А. Вероятностные модели надежности информационных вычислительных систем. М.: Радио и связь, 1991.
2. Лившиц Б.С. и др. Теория телетрафика. М.: Связь, 1979.
3. Крылов В.В., Самохвалова С.С. Теория телетрафика и ее приложения. – СПб: БХВ-Петербург, 2005.
4. Туманбаева К.Х. Моделирование систем телекоммуникаций: Учеб. пособие. – Алматы: АИЭС, 2007.
5. Новиков О.А., Петухов С.И. Прикладные вопросы теории массового обслуживания. – М.: Наука, 1989.
6. Щитникова А.Н. Автореф. магистр. раб.: Разработка методов оценки параметров трафика мультисервисной сети. Донецк: ДонНТУ, 2006.
7. Лившиц Б.С., Мамонтова Н.П. Теория телефонных сообщений: Учеб. пособие – Л. 1970.
8. Вишневский В.М. Теоретические основы проектирования компьютерных сетей. – М.: Техносфера, 2003.
9. Величко В.В., Субботин Е.А., Шувалов В.В., Ярославцев А.Ф. Телекоммуникационные системы и сети. Т. 3. Мультисервисные сети. М: Горячая линия – Телеком, 2005.
10. Филимонов А.Ю. Построение мультисервисных сетей Ethernet. – СПб: БХВ – Петербург, 2007.
11. Башарин Г.П. Лекции по математической теории телетрафика. – М: Изд-во. РУДН, 2004.
12. Советов Б.Я., Яковлев С.А. Моделирование систем. Практикум: Учеб. пособие для вузов. М.: Высш. шк., 1999.
13. Шварц М. Сети связи. Протоколы, моделирование и анализ. Ч1и Ч2 /Пер. с англ. – М.: Мир, 1992.
14. Гургенидзе А.Т., Корше В.И. Мультисервисные сети и услуги широкополосного доступа. СПб., 2003.
15. Лифшиц А.Л., Мальц Э.А. Статистическое моделирование систем массового обслуживания. М.: Наука, 1978.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....	3
1. МОДЕЛИРОВАНИЕ. МОДЕЛИ ДАННЫХ ДЛЯ ОПИСАНИЯ ТЕЛЕТРАФИКА.....	5
1.1. Моделирование. Виды моделирования.....	5
1.2. Системы и сети электросвязи – объекты моделирования. Основные термины и определения.....	11
1.3. Модели потоков событий.....	17
1.4. Основные принципы моделирования потоков событий.....	24
2. МЕТОД СТАТИСТИЧЕСКОГО МОДЕЛИРОВАНИЯ...	28
2.1. Статистическое моделирование. Метод Монте-Карло.....	28
2.2. Моделирование случайных чисел. Псевдослучайные числа.....	35
3. ОСНОВЫ ТЕОРИИ СИСТЕМ МАССОВОГО ОБСЛУЖИВАНИЯ.....	42
3.1. Основы построения систем массового обслуживания (СМО). Понятие СМО. Классификация моделей систем массового обслуживания.....	42
3.2. Марковские системы массового обслуживания.....	48
3.3. Основные показатели эффективности и характеристики СМО.....	52
3.4. Одноканальная модель с пуассоновским входным потоком с экспоненциальным распределением длительности обслуживания.....	56
3.5. Многоканальная модель с пуассоновским входным потоком и экспоненциальным распределением длительности обслуживания.....	67
3.6. Замкнутая система массового обслуживания.....	75
3.7. Марковские сети массового обслуживания.....	79

3.8. Одноканальная многофазная система с отказами. Метод приближенной оценки пропускных способностей многоканальной многофазной системы..	88
3.9. Модели систем массового обслуживания с потерей. Модель коммутационного узла.....	91
4. МОДЕЛИРОВАНИЕ ТЕЛЕТРАФИКА.....	98
4.1. Вычисление трафика.....	98
4.2. Основные параметры нагрузки.....	106
4.3. Методы оценки параметров и моделирования трафика мультисервисной сети.....	110
ГЛОССАРИЙ.....	119
БИБЛИОГРАФИЧЕСКИЙ СПИСОК	121

Учебное издание

Ниеталина Жаннат Жаксылыковна
Байкенова Гульжан Маматовна

МОДЕЛИРОВАНИЕ СИСТЕМ И СЕТЕЙ ЭЛЕКТРОСВЯЗИ

Учебное пособие

Нач. РО НТИЦ
Редактор
Компьютерная верстка

З.А. Губайдулина
Ш.Т. Касымбекова
А.Н. Оразалиева

Подписано в печать ***** . 2011 г.

Тираж 500 экз. Формат 60х84 1/16. Бумага типографская №1.

Усл. п.л. 7,3. Уч. изд.л. 7,8. Заказ №***. Цена договорная

Издание Казахского национального технического университета
имени К.И. Сатпаева

Научно-технический издательский центр
г. Алматы, ул. Ладыгина, 32

ISBN 978-601-228-299-3



9