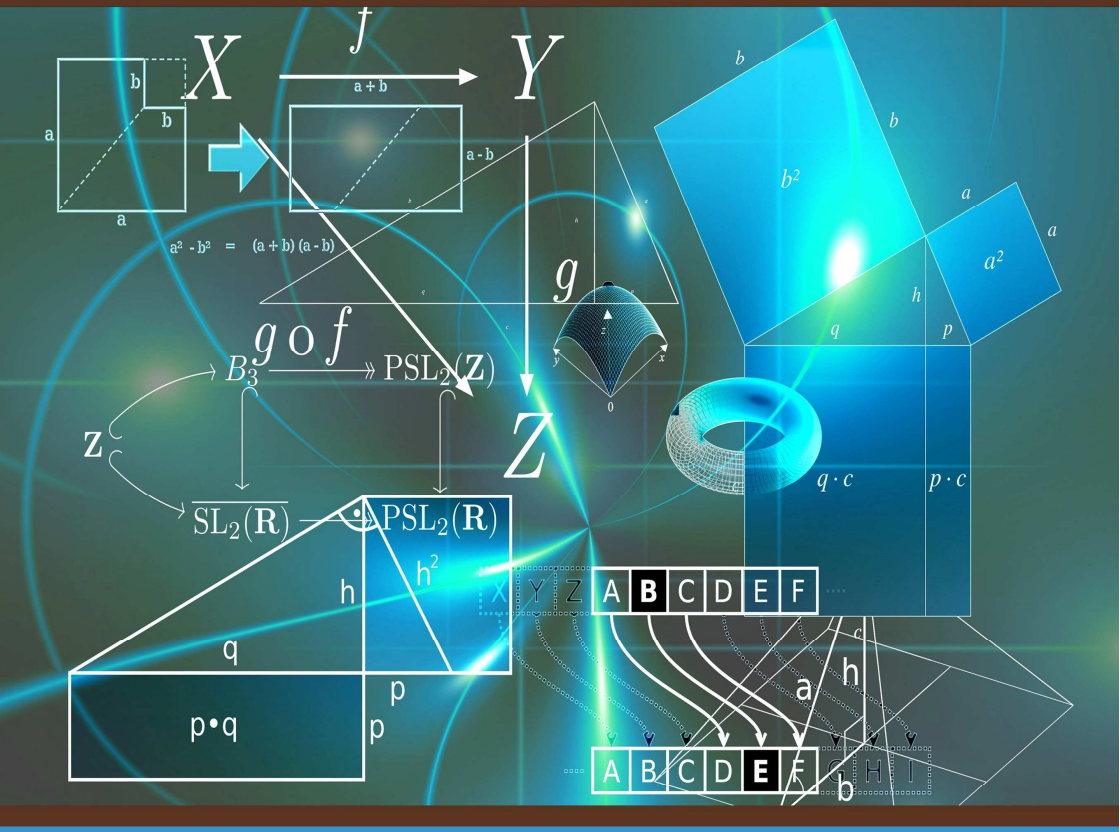
А.Н. Сапрыкин

МОДЕЛИ И МЕТОДЫ АНАЛИЗА ПРОЕКТНЫХ РЕШЕНИЙ



УЧЕБНОЕ ПОСОБИЕ

А.Н. САПРЫКИН

МОДЕЛИ И МЕТОДЫ АНАЛИЗА ПРОЕКТНЫХ РЕШЕНИЙ

Учебное пособие

Рекомендовано Научно-методическим советом ФГБОУ ВО «Рязанский государственный радиотехнический университет им. В.Ф. Уткина» в качестве учебного пособия для студентов высших учебных заведений, обучающихся по направлению подготовки 09.03.01 «Информатика и вычислительная техника» (квалификация «бакалавр»).

Рязань
Book Jet
2021

УДК 519.8
ББК 22.171
С19

Рецензенты:
Фаддеев А.О. – д-р техн. наук, главный научный сотрудник (Московский государственный университет им. Н.Э. Баумана);
Таганов А.И. – д-р техн. наук, профессор (Рязанский государственный радиотехнический университет им. В.Ф. Уткина)

Сапрыкин А.Н.
С19 Модели и методы анализа проектных решений: учебное пособие – Рязань: ИП Коняхин А.В. (Book Jet), 2021. – 104 с.

ISBN 978-5-907400-77-1

В учебном пособии рассмотрены основные математические модели и методы, используемые при проектировании и моделировании сложных технических систем на разных иерархических уровнях. Описаны основные методы их получения и анализа. Приводятся сведения о системах массового обслуживания и применяемых в них математических моделях.

Предназначено для студентов высших учебных заведений, обучающихся по направлениям подготовки 09.03.01 «Информатика и вычислительная техника» (квалификация «бакалавр»).

УДК 519.8
ББК 22.171

ISBN 978-5-907400-77-1

СОДЕРЖАНИЕ

ГЛАВА 1. ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ. КЛАССИФИКАЦИЯ МОДЕЛЕЙ И ВИДОВ МОДЕЛИРОВАНИЯ

РАЗДЕЛ 1.1. Основные понятия и определения

Полное и всестороннее исследование сложных технических систем на всех этапах их разработки от этапа НИР до этапа внедрения и эксплуатации невозможно без различных видов моделирования. Моделирование позволяет при сравнительно небольших затратах в приемлемые сроки решить многие проблемы разработки сложных технических систем (СТС).

Практически любая СТС непременно ведет к созданию, выбору и исполнению ее моделей. Исследование СТС состоит в определении параметров, изучении свойств и характеристик, выявлении зависимости данных характеристик от параметров.

Для того чтобы указанное исследование было менее дорогостоящим, более простым и проводилось в приемлемые сроки, в ходе процесса замещения объекта-оригинала объектом-моделью и исследования свойств объекта оригинала используют различные виды моделирования. Так как любая СТС по определению состоит из большого числа элементов и имеет сложную структуру, а ее многочисленные свойства зависят от очень большого числа параметров, то в одной модели практически невозможно учесть все параметры и свойства объекта.

В каждой модели следует учитывать только те параметры и свойства, которые являются наиболее важными только на данном этапе проектирования, а от остальных свойств следует абстрагироваться. Свойства, от которых абстрагируются, учитываются в других моделях.

В зависимости от того, какие свойства объекта-оригинала являются наиболее важными для разработки на данном этапе проектирования, могут создаваться различные модели одного и того же объекта. Отметим, что даже для одного и того же объекта на одном и том же этапе проектирования может использоваться множество моделей. Например, в зависимости от сложности СТС, требуемой точности и этапа проектирования могут использоваться аналитические, имитационные или натуральные модели одного и того же объекта, в зависимости от исследуемых режимов – статические и динамические модели, в зависимости от уровня исследования – модели микро (ДУЧП – дифференциальные уравнения частных производных), макро (ОДУ – обыкновенные дифференциальные уравнения) и метауровня (сетевые модели) и т.д.

Во многом справедливо и обратное утверждение: одна и та же модель может отображать функционирование процессов в различных объектах. Например, ОДУ могут отображать колебательные процессы в электрических цепях и в механических системах, ДУЧП могут отображать теплофизические процессы в элементах конструкций, элементах электрических цепей, процессы полей механических напряжений и т.д.

Исходной информацией при построении и выборе моделей являются данные о назначении и условиях работы проектируемого объекта. Это объясняется тем, что указанные данные и рассматриваемые этапы проектирования определяют основную цель моделирования и позволяют сформировать требования к моделям. При этом уровень абстрагирования от тех или иных свойств и параметров в модели в основном зависит от цели моделирования, которая определяет круг тех проблем и задач, которые разработчик должен решить с помощью данной модели. Таким образом, сама модель определяет, какие свойства и параметры следует учитывать.

РАЗДЕЛ 1.2. Классификация моделей и видов моделирования

Модели можно классифицировать по большому числу признаков. На рисунке 1.2.1 представлена классификация моделей по характеру изучаемых процессов.

Отметим, что в САПР под задачами статики понимают исследование и анализ, под задачами динамики – задачи оптимизации.

По форме представления объекта модели подразделяются на мысленные и реальные. Мысленные модели применяются, когда невозможно физически реализовать процесс либо в заданном интервале времени, либо по физическим условиям (например, процессы микромира). Реальные исследования проводятся либо на самом объекте, либо на специальных физических установках. При реальном моделировании могут задаваться как нормальные режимы, так и любые другие. Входные параметры могут меняться в широком диапазоне, вплоть до границ устойчивости. Реальное моделирование дает более достоверные результаты, однако возможность провести его существует не всегда.

При наглядном моделировании в зависимости от имеющихся сведений об объекте строятся наглядные модели. В основу гипотетической модели закладывается та или иная гипотеза, отображающая закономерности, протекающие в объекте. Она основана на причинно-следственных связях между входом и выходом объекта.

Гипотетическое моделирование используется тогда, когда явно недостаточно сведений для построения формализованных моделей.

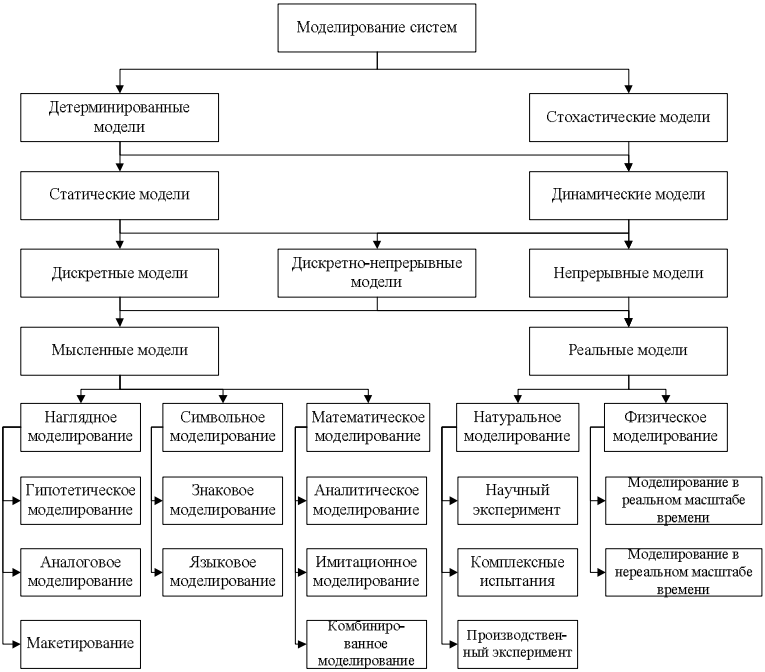


Рис. 1.2.1. Классификация моделей по характеру изучаемых процессов.

Аналоговое моделирование использует аналогии различных уровней. Наивысшим уровнем является полная аналогия. Учитываются все основные свойства объекта, однако полная аналогия может быть исследована только для простых объектов. Если учитывается только часть свойств объекта, то аналогию называют неполной. Если только одно качество – то приближенной.

Существенное место при мысленном моделировании занимают наглядные макеты. Наглядное макетирование либо предшествует другим видам моделирования, либо используется тогда, когда воспроизвести физические свойства не представляется возможным. Наглядное макетирование также основано на аналогиях, но они основываются на процессах, протекающих в объекте (например, планетарная модель атома).

Символьное моделирование – искусственный процесс создания логического объекта, который с помощью системы знаков и символов отображает некоторые отношения процессов, протекающих в объекте.

Например, если какие-то понятия обозначить знаками и между ними ввести операции, то с помощью них можно подставлять различные выражения (например, знаки теории множеств). При использовании языковых моделей применяются технические языки (тезаурус). Каждое слово в этом языке обязательно однозначно.

При исследовании объектов математическими методами сначала производится формализация процессов, протекающих в объекте, и строится математическая модель. Математическая модель представляет собой формализованное описание объекта с помощью абстрактного языка математических соотношений. В качестве описания может быть использованы все средства математики (алгебры, дифференциального, интегрального счисления, конечно-разностные соотношения, теория алгоритмов и т.д.). По существу вся математика создана для исследования различных объектов и процессов.

При натуральном моделировании исследования проводятся либо на самом объекте, либо на его части, с последующей обработкой результатов. В настоящее время при изучении процессов в сложных объектах при научных и производственных экспериментах широко используются средства автоматики и обработки информации. Эксперимент отличается от произвольного процесса тем, что при эксперименте параметры изменяются в более широких пределах и исследуются различные критические свойства (например, устойчивость).

Комплексные испытания представляют собой все возможные испытания новых образцов.

Физическое моделирование отличается от натурального тем, что оно выполняется не на самих объектах, а на специально созданных установках. В них сохраняется физическое подобие процессов в объекте.

Аналитическими моделями являются любые аналитические выражения, которые отображают функционирование исследуемого объекта. Эти модели основаны на базе некоторой математической теории (например, теории цепей Маркова, теории массового обслуживания и т.д.). Достоинством аналитических моделей является то, что они менее трудоемки, менее дорогостоящи и в большинстве случаев позволяют мгновенно получать результаты моделирования. Решение может быть получено тремя основными способами:

1) численный;
2) аналитический;
3) качественный.

Если зависимость основных характеристик объекта от его параметров можно выразить аналитически в явном виде, значит, для рассматриваемой задачи имеется аналитическое решение. Преимущества аналитического решения состоит в том, что с его помощью можно исследовать характеристики практически мгновенно в широком диапазоне изменения входных параметров.

Если аналитическую модель (зависимость) в явном виде разрешить не удается, то используются численные методы. Качественный способ используется тогда, когда значения самих характеристик получать не обязательно, а достаточно ограничить границы устойчивости.

Имитационные модели систем описывают функционирование объекта в виде последовательности операций или групп операций, которые происходят при его функционировании. Они состоят из описаний двух типов:

1) описание элементов, составляющих систему;

2) описание структуры системы.

При имитационном моделировании разрабатывается алгоритм, отображающий функционирование объекта. Таким образом, имитационная модель сводится к разработке программ и их реализации на множестве входных данных. В зависимости от типа входных данных выделяют:

1) трассоориентированные имитационные модели (входные параметры задаются в виде трассы);

2) статические имитационные модели.

Под трассой понимается поток событий, которые происходят при функционировании объекта. События учитываются в хронологическом порядке, учитывая при необходимости моменты их возникновения. При этом трасса должна по возможности наиболее полно отображать типовые решения.

В статических имитационных моделях входные данные задают искусственно с помощью датчика случайных чисел. Можно задавать входные данные, распределенные по любому закону, а также изменять их в широких пределах.

Таким образом, имитационные модели приближенно воспроизводят сам процесс-оригинал, его функционирование во времени. Влияние случайных факторов при этом учитывается с помощью случайных чисел. Средствами формализованного описания могут служить либо универсальные, либо специальные языки программирования.

Имитационные модели могут быть созданы для гораздо более широкого круга систем, нежели аналитические. Преимуществом по сравнению с аналитическими моделями является то, что они принудительно могут быть использованы для анализа объекта любой сложности: может быть достигнута любая точность, в пределе которой возможно получить принципиально полное совпадение исследуемого объекта и модели. В трассоориентированных имитационных моделях события в объекте учитываются с большей точностью, в то время как в статических проводится больше экспериментов.

Однако практика показывает, что в имитационных моделях зачастую затруднительно добиться высокой точности, так как с увеличением требований к точности резко возрастает трудоемкость данных моделей и вероятность влияния непредвиденных факторов.

Если часть устройств исследуемого объекта исследуется аналитическими моделями, а часть – имитационными, то такие модели называются комбинированными.

РАЗДЕЛ 1.3. Блочно-иерархический подход к проектированию

Параграф 1.3.1. Сущность блочно-иерархического подхода

Проектирование является сложным и трудоемким процессом, так как оно может быть распределено как по времени, так и в пространстве (например, между различными проектными организациями). Распределение процесса проектирования по времени осуществляется на основе этапов проектирования.

Распределение процесса проектирования между подразделениями осуществляется за счет блочно-иерархического подхода (БИП). При БИП процесс проектирования разбивается на ряд уровней и аспектов (рисунок 1.3.1.1). Описание всей системы разбивается на ряд иерархических уровней. Иерархические уровни отличаются друг от друга степенью детализации, т.е. степенью подробности описания системы и ее частей. В каждый иерархический уровень включаются задания и вопросы, имеющие общую физическую основу (например, логическое проектирование и т.д.).

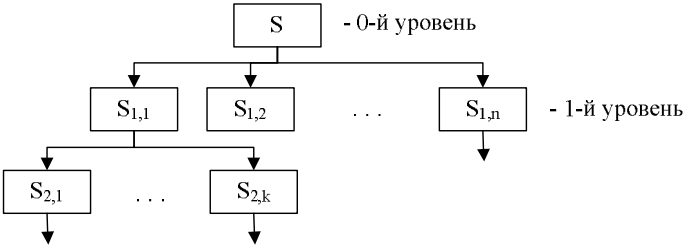


Рис. 1.3.1.1. Иерархические уровни БИП.

Каждому иерархическому уровню присущи свои формы документации, методы получения модели, методы реализации алгоритмов и т.д. По сути иерархический уровень представляет собой совокупность языков, моделей, постановок задач, способов получения описаний (в том числе первичного и конечного).

На самом высшем иерархическом уровне используются описания систем, учитывающие только самые общие свойства. На последних уровнях степень детализации описаний возрастает, и система рассматривается уже не как единое целое, а в виде нескольких блоков (рисунок 1.3.1.2). На дальнейших уровнях в свою очередь происходит разделение на последующие блоки данного уровня. Разбивка осуществляется таким образом, чтобы каждый блок был независимым. Это необходимо для того, чтобы разработкой каждого блока могли заниматься отдельные подразделения.

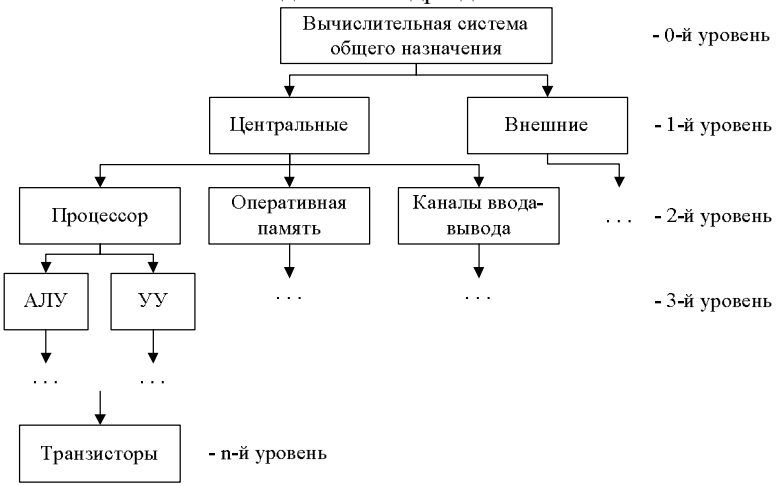


Рис. 1.3.1.2. Пример БИП.

На каждом уровне иерархии ставятся задачи приемлемой сложности. Иначе говоря, при БИП сложная задача проектирования разбивается на множество задач приемлемой сложности. На каждом иерархическом уровне существуют свои представления о системе и элементах.

То, что на i-ом уровне считалось элементом, на следующем более низком уровне считается системой. Элементы самого низшего уровня обычно называются базовыми элементами. По отношению к проектируемой СТС базовые элементы – это те элементы, которые не подлежат дальнейшему разбиению при данном проектировании.

На i-ом уровне блок разбивается на ряд элементов, между ними устанавливаются связи. На i+1-ом уровне каждый элемент i-го уровня становится системой. Каждый из них разбивается на элементы и между ними опять устанавливаются связи. Подход, при котором сложная задача разбивается на ряд более простых, называется декомпозиционным.

Проектирование СТС может вестись в различных направлениях. Если сначала проектируются верхние уровни, а затем нижние, такое проктирование назвется нисходящим; если сначала нижние, а затем вышестоящие – то восходящим.

При нисходящем проектировании составляется техническое задание (ТЗ) на всю систему. На основе этого ТЗ составляются ТЗ на отдельные устройства, узлы, постепенно переходя на все более низкие иерархические уровни. ТЗ на всю систему не подвергается формализации и составляется группой опытных разработчиков. ТЗ на все последующие уровни формализуется и поэтому составляется проще и быстрее.

При восходящем проектировании исходной информацией является ТЗ на разработку самого нижнего уровня, и только потом ТЗ для более высоких уровней иерархии. При восходящем проектировании составление ТЗ для каждого уровня не подвергается формализации. На практике часто используется смешанное проектирование.

Параграф 1.3.2. Уровни и аспекты проектирования

Ранее расписанные иерархические уровни проектирования, отличающиеся друг от друга степенью проработанности описания, называются горизонтальными уровнями или уровнями абстрагирования. Можно выделить уровни по характеру отображения свойств, например, выделить задачи технологического

проектирования. Такие уровни называются вертикальными уровнями или аспектами.

При проектировании вычислительных систем (ВС) выделяют три аспекта:

1) функциональный аспект;
2) конструкторский аспект;
3) технологический аспект.

Функциональный аспект отражает физические и (или) информационные процессы, протекающие в объекте при его функционировании. Конструкторский аспект характеризует структуру, расположение в пространстве и форму составных частей объекта. Технологический аспект – технологичность, возможности и способы изготовления объекта в заданных условиях.

При проектировании ВС также выделяют алгоритмический или программный аспект проектирования. Каждый из вертикальных уровней разбивается еще и на горизонтальные. Рассмотрим связь между ними.

	Аспекты (вертикальные уровни)			
Уровни абстрагирования (горизонтальные уровни)	***Функциональ-ный***	***Алгоритмический***	***Конструкторский***	***Технологический***
	Системный	Программирование систем	Шкаф, стойка	Принципиальная схема технологического процесса
	Логический	Программирование модулей	Вставные блоки, панели, ТЭЗы	Маршрутная технология
	Схемотехнический			
	Компонентный	Проектирование микропрограмм	Интегральные схемы	Технологические операции

Отметим, что можно выделить и другие аспекты проектирования в зависимости от природы. На системном уровне определяется общая структурная схема ВС, состав укрупненных устройств и связи между ними, типы устройств, общие принципы организации вычислительного процесса.

На логическом уровне разрабатываются логические схемы узлов и блоков, вся логика. На схемотехническом уровне – все схемы блоков и устройств. На компонентном уровне – микросхемы, схемы базовых элементов и т.д.

РАЗДЕЛ 1.4. Математические модели в САПР

Параграф 1.4.1. Классификация моделей в САПР

Классификация моделей может выполняться по очень большому числу признаков:

1) по характеру отображаемых свойств;
2) по способу получения;
3) по степени детализации;
4) по уровню иерархии.

По характеру отображаемых свойств модели разделяют на функциональные и структурные. Функциональные модели предназначены для отображения физических или информационных процессов, протекающих при функционировании или изготовлении объекта. Чаще всего эти модели имеют вид системного уровня. В зависимости от того, какие физические процессы они отображают, функциональные делятся на: электрические, тепловые, механические и т.д.

Структурные чаще всего относятся к конструкторскому аспекту проектирования. Среди них выделяют геометрические и топологические модели. Топологические отображают структуру связей, соединения, размещение элементов и т.д. Они имеют форму графов, матриц, списков и т.д.

Геометрические модели помимо размещения определяют форму и размеры конструктивных элементов. Они имеют вид алгебраических уравнений, уравнений линий, поверхностей и т.д. Иногда помимо функциональных и структурных моделей также выделяют надежностные, стоимостные и т.д.

Функциональные модели сложнее структурных и зачастую они учитывают и структурные особенности модели.

По способу получения выделяют теоретические и экспериментальные. Теоретические модели получаются путем изучения закономерностей процессов, протекающих в объекте. С помощью ряда допущений получают аналитические модели. Экспериментальные модели получают в результате информации, полученной на самих объектах, их частях или физических установках. При этом может использоваться как активный, так и пассивный эксперимент.

По степени детализации – полные и макромодели. На каждом уровне иерархии можно выделить модели системы и модели элементов.

Модель, полученная в результате объединения всех моделей, называется **полной**. С увеличением числа элементов при переходе к моделям более высоких уровней полные модели становятся трудоемкими, поэтому при переходе к моделям верхних уровней практически всегда используются не полные модели, а их аппроксимированный вариант, укрупненные модели. Учитываются не все свойства и связи элементов, а только те, которые необходимы на рассматриваемом уровне.

Модель, полученная в результате аппроксимации нижнего уровня, называется макромоделью. Понятие макромодели имеет большое значение при блочно-иерархическом подходе (БИП), так как позволяет перейти от полных неразрешаемых моделей к модели с приемлемой сложностью. На практике часто встречаются задачи, в которых одновременно учитываются задачи нескольких уровней. Тогда при анализе одних узлов используются полныемодели, а других – макромодели. Такие модели называют многоуровневыми или смешанными, если для различных уровней в них используются различные типы уравнений.

Заметим, что параметры, характеризующие систему, элементы системы и воздействие внешней среды, называют соответственно выходными, внутренними и внешними параметрами. В моделях каждого уровня иерархии используются независимые переменные (время, частота и пространственные координаты). Зависимые переменные называют фазовыми. Выходные параметры в моделях в явном виде в большинстве случаях не участвуют. Они легко определяются в результате решения модели – определения зависимых переменных.

Фазовые переменные характеризуют состояние объекта. Их также можно назвать переменными состояния.

По уровню иерархии. При проектировании СТС можно выделить то или иное количество иерархических уровней. Каждому иерархическому уровню соответствуют свои модели. Сами модели также распределяются по иерархическим уровням.

Чем больше иерархических уровней, тем проще модель на каждом из этих уровней. Однако при этом существенно возрастает степень согласования результатов моделирования между уровнями. В САПР выделяют 3 крупных уровня математических моделей:

1) микроуровень;
2) макроуровень;
3) метауровень.

Если при проектировании возникает больше трех уровней, то каждый из них либо совпадает с одним из названных, либо входит в какой-нибудь уровень или подуровень.

На микроуровне проектируемый объект рассматривается как непрерывный объект в пространстве-времени, т.е. как сплошная среда. На данном уровне анализируются электрические поля, тепловые поля, диффузионные (в слоях полупроводников), поля механических напряжений и т.п.

В качестве математических моделей на микроуровне используется аппарат математической физики, т.е. дифференциальные уравнения в частных производных (ДУЧП).

Например, уравнение Лапласа:

$$\frac{\partial^2 t}{\partial x^2}+\frac{\partial^2 t}{\partial y^2}+\frac{\partial^2 t}{\partial z^2}=0\ .$$

Математические модели в виде ДУЧП применимы для анализа только очень малых объектов: для анализа полей в пределах одного кристалла, 1 детали конструкции и т.п. Если объем исследуемого объекта увеличивается, например с одной детали до узла (принципиальной электрической схемы), то количество ДУЧП становится столь большим, что из-за трудоемкости ее решения она становится практически неприемлемой.

Выходные параметры представляют электрические и тепловые сопротивления, фокусное расстояние, жесткость конструкции и т.п. Если объектом проектирования является многокомпонентная среда, то модель является не распределенной, как на предыдущем уровне, а сосредоточенной.

Для анализа названных объектов прибегают к аппроксимации полных моделей, полученных на микроуровне. Внутренними параметрами в **моделях макроуровня** являются резисторы, индуктивности, емкости, а выходными являются коэффициент усиления по напряжению, коэффициент усиления по мощности и т.п. На этом уровне объектом исследования являются отдельные узлы, блоки, математические модели имеют вид ОДУ или САУ (система алгебраических уравнений) (при установившемся режиме). При дальнейшем укрупнении объекта проектирования размерность модели на этом уровне становится очень большой и практически нереализуемой.

Путем дальнейшего абстрагирования от тех или иных параметров переходят к **моделям метауровня**, иными словами, математические модели метауровня представляют собой

аппроксимированную полную модель, полученную из модели макроуровня.

Выходными параметрами модели метауровня являются время ответа системы, относительная и абсолютная пропускные способности. На данном уровне в основном анализируются информационные процессы.

Математическими моделями на метауровне могут быть системы логических уравнений, теория нечетких множеств, теория автоматического управления, теория массового обслуживания, сетевые модели и т.п.

Параграф 1.4.2. Математические модели на различных уровнях и в различных аспектах проектирования

Функциональный аспект.

1) Системный. Объект проектирования: вычислительные системы. Математические модели: теория массового обслуживания, сети Петри.

2) Логический. Объект проектирования: отдельные узлы, блоки, устройства. Математические модели: система логических уравнений, теория конечных автоматов.

3) Схемотехнический. Объект проектирования: схему узлов, блоков, устройств и т.д. Математические модели: СОДУ, САУ.

4) Компонентный. Объект проектирования: микросхемы, интегральные схемы. Математические модели: СДУЧП.

С точки зрения **конструкторского аспекта** проектирования, если объектом проектирования являются конструкции отдельных узлов и механические соединения между ними, то в качестве математической модели чаще всего используют ОДУ, САУ, уравнения линий и поверхностей. Если объектом проектирования является коммутационно-монтажное соединение, то математической моделью являются матрицы, графы, списки и т.п.

С точки зрения **технологического аспекта** проектирования объектом проектирования является технологический процесс изготовления системы, а математической моделью –регрессионные уравнения, матрицы, списки, САУ.

РАЗДЕЛ 1.5. Требования к математическим моделям и методам анализа проектных решений

Параграф 1.5.1. Точность и адекватность математических моделей

Под точностью понимают степень совпадения выходных параметров модели и объекта.

y – выходной параметр объекта,

$y_м$ – выходной параметр модели,

$$\varepsilon = \frac{y - y_м}{y}.$$

Существует 3 основные трудности корректной оценки погрешностей модели.

Трудность 1. С помощью модели оценивается несколько характеристик. Пусть их m, тогда

$$\varepsilon_j = \frac{y_j - y_{jм}}{y_j}, \quad j = 1,\ldots,m$$
$$\varepsilon = (\varepsilon_1, \varepsilon_2 \ldots \varepsilon_m) \qquad (1.5.1.1)$$

Трудность 2. Точность характеристик модели зависит от условий функционирования объекта. Определяются внешними параметрами.

Трудность 3. Любая модель, как правило, используется для различных режимов работы, учитывая при этом различные параметры. Точность при этом оказывается разной.

В выражении (1.5.1.1) значение выходного параметра y_j на практике может определяться либо экспериментально, либо с помощью более точных моделей.

$$\varepsilon = \max_{j=1,m} |\varepsilon_j|$$

$$\varepsilon = \sqrt{\sum_{j=1}^{m} \varepsilon_j^2}$$

Для устранения II и III трудностей при оценке погрешностей используют наиболее часто встречающиеся условия, т.е. тестовые ситуации.

Под адекватностью понимают способность модели отображать свойства с погрешностью, не выше заданной. Как уже указывалось, погрешность зависит от значений внешних параметров. Пусть задана предельно допустимая погрешность модели:

$\left|\varepsilon_{м}\right| \le \varepsilon_{доп}$.

Если выполнить это требование в пространстве внешних параметров, то получим область адекватности. Пусть у нас m выходных параметров. Если задать предельно допустимую погрешность по каждому из выходных параметров, то получим m неравенств:

$\left|\varepsilon_{jм}\right| \le \varepsilon_{j\,доп}$.

Покажем графически область адекватности при нескольких внешних параметрах (рисунок 1.5.1.1).

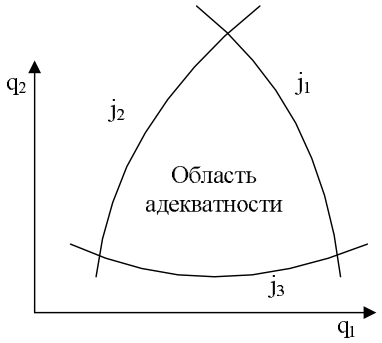


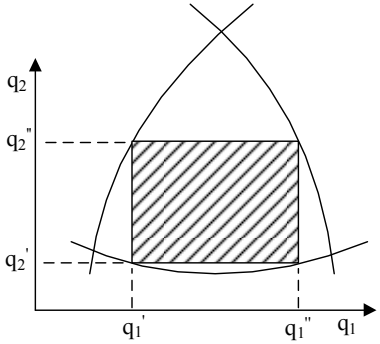
Рис. 1.5.1.1. Область адекватности.

Каждая j-ая линия определяет предельно допустимую погрешность по j-му выходному параметру в зависимости от внешних параметров.

Область адекватности дает проектировщику гораздо больше информации, чем точность, так как точность определяется только для точечных, случайных значений внешних параметров, а область адекватности предоставляет значения выходных параметров во всем допустимом диапазоне изменений внешних параметров.

Однако расчет области адекватности – трудоемкая задача, поэтому она никогда не определяется для вновь проектируемых объектов элементов. Если же элементы являются унифицированными и многократно используются при проектировании различных устройств, то определение области адекватности имеет смысл, так как эти элементы используются многократно, а область адекватности рассчитывается только один раз. Значение области адекватности позволяет более корректно выбирать модели в каждом конкретном случае из множества моделей из библиотеки.

Область адекватности может иметь произвольную форму, как в нашем примере. В этом случае она хуже воспринимается и является более громоздкой, поэтому часто вместо произвольной формы используют аппроксимированную. Эта аппроксимация чаще всего задается в виде гиперпараллелепипеда. В этом случае область адекватности задается в виде неравенств, в их пределах модель является адекватной (рисунок 1.5.1.2).



$$Q(q_1, q_2)$$
$$q_1' \le q_1 \le q_1'' \,,\; q_2' \le q_2 \le q_2''$$

Рис. 1.5.1.2. Аппроксимированная область адекватности.

Нередко используют другую аппроксимацию, например линеализацию зависимостей y_j. Область адекватности характеризуется размерами и расположением ее в координатном пространстве.

Параграф 1.5.2. Универсальность и экономичность математических моделей

Если модель окажется пригодной только для одной модели и одного режима работы, то для проектировщика станет считаться малопригодной, так как для другого случая будет требоваться другая модель. В каждой модели учитываются по возможности наибольшее число параметров.

Степень универсальности определяется числом параметров, которые учитываются в модели. Чем она универсальнее, тем больше область ее применения. Под экономичностью понимают трудоемкость решения. Очевидно, что одновременные требования высокой степени точности, большой области адекватности и экономичности противоречивы, поэтому в каждом случае принимается компромиссное решение. При анализе одного объекта параметры, пригодные для одних, оказываются малопригодными для других. Для

анализа одного и того же объекта используется множество математических моделей, в том числе многоуровневых и смешанных.

Трудоемкость модели также зависит от алгоритма ее решения.

Проектные решения – это получение промежуточного или конечного описания объекта. Это описание позволяет либо выбрать дальнейший ход проектирования, либо его окончание. Проектное решение получается при выполнении проектного процесса – формализованной совокупности действий. Трудоемкость проектного процесса и его погрешность также напрямую зависят от модели и алгоритма ее решения.

РАЗДЕЛ 1.6. Вопросы для самопроверки

1. Что является исходной информацией при построении и выборе моделей?
2. Опишите основные отличия мысленных моделей от реальных.
3. Что представляет собой аналоговое моделирование?
4. Перечислите основные способы получения результатов при аналитическом моделировании.
5. Охарактеризуйте основные виды имитационных моделей.
6. Поясните сущность блочно-иерархического подхода.
7. Перечислите три аспекта проектирования вычислительных систем.
8. На какие типы можно разделить модели по характеру отображаемых свойств?
9. На какие типы можно разделить модели по уровню иерархии?
10. Дайте определение полной модели.
11. Приведите пример модели объекта на микроуровне.
12. Что является объектом исследования в моделях макроуровня?
13. Какие существуют основные трудности корректной оценки погрешностей модели?
14. Что понимается под точностью и адекватностью математических моделей?
15. Охарактеризуйте область адекватности.

ГЛАВА 2. МАТЕМАТИЧЕСКИЕ МОДЕЛИ НА МИКРОУРОВНЕ

РАЗДЕЛ 2.1. Основные понятия и определения

Математические модели на микроуровне используются для анализа физического состояния и процессов в сплошных средах. Структура объектов не учитывается. Объектами проектирования являются участки резистивной пленки, объем участка кристалла или микросхемы и т.п. Независимыми переменными являются время и пространственные координаты. От них зависят фазовые переменные, которые в данных моделях представлены плотностью тепловых, электрических токов, напряженностью механических полей и прочими. Внутренними параметрами являются коэффициент электропроводности, геометрические размеры элементарных участков и т.п. Выходными параметрами могут быть тепловые, электрические, гидравлические сопротивления, жесткость пружин и т.д., то есть на этом уровне используются распределенные модели, чаще всего имеющие вид ДУЧП. Каждый процесс конкретной физической природы (например, тепловой, гидравлический и т.д.) описывается своим ДУЧП.

Результатом решения данных уравнений являются значения температур, электрических потенциалов, механических напряжений деформации в каждой точке пространства и в каждый момент времени. Для решения ДУЧП необходимо знать краевые условия. Рассмотрим получение и решение математических моделей микроуровня на примере ввода и решения уравнения теплопроводности.

Любая вычислительная система с точки зрения теплового режима описывается как:

$$E = E_1 + E_2 + E_3,$$

где E – подводимая энергия, E_1 – энергия полезного сигнала, E_2 – рассеиваемая энергия (энергия, рассеиваемая только с поверхности ЭС), E_3 – энергия на нагревание аппаратуры.

$$КПД = \frac{E_1}{E} \approx 10 - 12\%$$

Доля энергии, расходуемая на нагрев, растет во многих случаях. Температурные режимы являются единственным фактором, который ограничивает минимальные габариты ЭС.

Если разные части ЭС имеют разную температуру, то происходит процесс передачи энергии. Этот теплообмен осуществляется за счет 3 явлений:

1) Явление теплопроводности (кондукции).

2) Явление конвекции.

3) Явление излучения.

Кондукционный теплообмен – это теплообмен на уровне микрочастиц, молекул и атомов. Данный вид теплообмена в основном наблюдается в твердых телах. Конвекционный теплообмен происходит за счет передачи тепла вместе с массами среды и обычно сопровождается явлением теплопроводности. Явление излучения основано на способности тел излучать, отражать, поглощать и пропускать тепловую энергию в виде электромагнитных волн различной длины.

Следует иметь в виду, что при получении теоретических моделей основные законы следует учитывать в их наиболее фундаментальном виде. В этом случае удается получить достаточно универсальную математическую модель. В ином случае в модели могут оказаться неучтенными некоторые условия, которые проектировщик может не заметить. К наиболее фундаментальным законам относятся, например, законы сохранения энергии и массы.

РАЗДЕЛ 2.2. Элементы теории теплопроводности

Параграф 2.2.1. Основные понятия и определения

Температурное поле – это совокупность значений температур во всех точках рассматриваемого тела или части пространства в данный момент времени. Температурное поле в общем виде математически можно записать как

$$t = f(x, y, z, \tau),$$

где x, y, z – координаты тела в пространстве, τ – время.

Такое поле отвечает неустановившемуся тепловому режиму теплопроводности и называется нестационарным температурным полем, поскольку температура изменяется во времени.

Если тепловой режим является установившимся, то температура в каждой точке поля с течением времени остается неизменной, и такое температурное поле является функцией координат и называется стационарным:

$\frac{\partial t}{\partial \tau} = 0 \rightarrow t = f_1(x, y, z)$ – трехмерное стационарное температурное поле;

$\frac{\partial t}{\partial \tau} = \frac{\partial t}{\partial z} = 0 \rightarrow t = f_2(x, y)$ – двумерное стационарное температурное поле;

$\frac{\partial t}{\partial \tau} = \frac{\partial t}{\partial z} = \frac{\partial t}{\partial y} = 0 \rightarrow t = f_3(x)$ – одномерное стационарное температурное поле.

Изотермическая поверхность – геометрическое место точек, имеющих в данный момент времени одинаковую температуру. Изотермические поверхности могут заканчиваться на границах тела или замыкаться на самих себя, но никогда не пересекаются (рисунок 2.2.1.1).

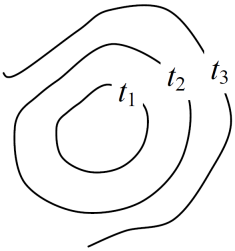


Рис. 2.2.1.1. Изотермические поверхности.

Температурный градиент – это вектор, направленный по нормали к изотермической поверхности в сторону возрастания температуры (рисунок 2.2.1.2) и численно равный производной от температуры по этому направлению.

$$\overline{grad}\ t = \lim_{\Delta n \to 0} \frac{\Delta t}{\Delta n} = \vec{n}_0 \frac{\partial t}{\partial n},$$

где n_0 – это единичный вектор нормали. |*grad t*|=[°С/м]

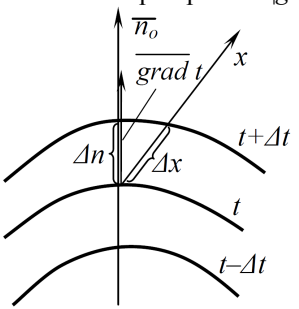


Рис. 2.2.1.2. Температурный градиент.

Количество теплоты – это тепловая энергия, передаваемая от одного тела к другому в течение какого-то времени: Q_τ = [Дж, кДж, ккал].

Тепловой поток – это количество теплоты, передаваемое в единицу времени:

Q = [Дж/с] = [Вт] = [ккал/ч].

Плотность теплового потока (удельный тепловой поток) – это количество теплоты, проходящее в единицу времени через единицу поверхности: q = [Дж/(с·м2)] = [Вт/м2].

Температурный напор – это разность температур между поверхностями тел, или телом и теплоносителем, или между теплоносителями: $\Delta t = t_1 - t_2$[°C].

Параграф 2.2.2. Закон Фурье. Коэффициент теплопроводности

Для распространения теплоты в любом теле (или в пространстве) необходимо наличие разности температур в различных точках тела, т. е. при передаче теплоты теплопроводностью $grad\ t \neq 0$.

Согласно гипотезе Фурье количество теплоты dQ_τ, проходящее через элемент изотермической поверхности dF за промежуток времени $d\tau$, пропорционально температурному градиенту:

$$dQ_\tau = -\lambda \frac{\partial t}{\partial n} dF d\tau, \qquad (2.2.2.1)$$

$$dQ_\tau = q dF d\tau, \qquad (2.2.2.2)$$

$$q = -\lambda \frac{\partial t}{\partial n} = -\lambda \cdot \overline{grad\ t}. \qquad (2.2.2.3)$$

Уравнение (2.2.2.3) называется ***законом Фурье***, знак «–» показывает, что направление удельного теплового потока противоположно направлению температурного градиента.

В уравнениях (2.2.2.1) и (2.2.2.3) λ – ***коэффициент теплопроводности*** – это тепловой поток, проходящий через единицу поверхности при единичном температурном градиенте. В этом состоит физический смысл коэффициента теплопроводности.

$$\lambda = \left[\frac{Вт}{м^2} / \frac{°C}{м} \right] = \left[\frac{Вт}{м \cdot °C} \right]$$

Чем больше значение λ, тем большей способностью проводить теплоту обладает тело. Коэффициент теплопроводности для данного тела не является постоянной величиной и зависит от физических свойств вещества, от температуры, давления, влажности.

Как показывают опыты, для многих материалов зависимость λ от температуры может быть принята линейной:

$$\lambda = \lambda_0(1 + bt),$$

где λ_0 – коэффициент теплопроводности при 0 °С, Вт/(м·°С); t – текущая температура, °С; b – постоянная, зависящая от свойств материала, 1/°С.

Однако в технических расчетах значения λ обычно принимаются постоянными, равными среднеарифметическим в данных пределах изменения температуры. Для большинства материалов λ определяется опытным путем, а для технических расчетов берется из справочных таблиц.

Для металлов λ = 2,3 – 420 Вт/(м·°С). С увеличением температуры коэффициент теплопроводности убывает. Это говорит о том, что холодный металл лучше проводит теплоту, чем нагретый.

Параграф 2.2.3. Дифференциальное уравнение теплопроводности. Условия однозначности

Изучение любого физического процесса связано с установлением зависимости между величинами, характеризующими данный процесс. Для сложных процессов, к которым относится передача теплоты теплопроводностью, при установлении зависимости между величинами удобно воспользоваться методами математической физики, которая рассматривает протекание процесса в элементарном объеме вещества в течение бесконечно малого отрезка времени.

При выводе дифференциального уравнения теплопроводности пренебрегают изменением некоторых величин и принимают следующие допущения:

- коэффициент теплопроводности $\lambda=const$, удельная теплоемкость тела $c=const$, плотность тела $\rho=const$;
- внутренние источники теплоты отсутствуют;
- тело однородно и изотропно;
- соблюдается закон сохранения энергии: разность между количеством теплоты, вошедшим в элементарный объем за время $d\tau$, и вышедшим из него за это же время, расходуется на изменение внутренней энергии рассматриваемого объема.

Выделим в теле элементарный параллелепипед с ребрами dx, dy, dz (рисунок 2.2.3.1). Температуры его граней различны, поэтому через них будет проходить теплота в направлении осей X, Y, Z.

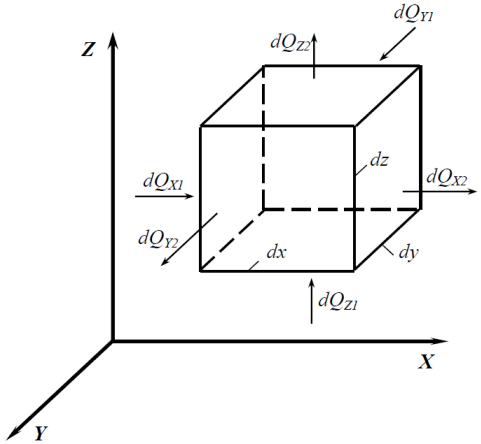


Рис. 2.2.3.1. Элементарный параллелепипед с разной температурой граней.

Через площадку $dxdy$ за время $d\tau$, согласно закону Фурье, проходит следующее количество теплоты:

$$dQ_{Z1} = -\lambda dxdyd\tau \frac{\partial t}{\partial z}. \qquad (2.2.3.1)$$

Через противоположную грань на расстоянии dz отводится количество теплоты, определяемое из выражения:

$$dQ_{Z2} = -\lambda dxdyd\tau \frac{\partial t}{\partial z}\left(t + \frac{\partial t}{\partial z}dz\right), \qquad (2.2.3.2)$$

где $\left(t + \frac{\partial t}{\partial z}dz\right)$ – температура второй грани; $\frac{\partial t}{\partial z}dz$ – изменение температуры в направлении оси Z.

Уравнение (2.2.3.2) можно записать как

$$dQ_{Z2} = -\lambda dxdyd\tau \frac{\partial t}{\partial z} - \lambda dxdydzd\tau \frac{\partial^2 t}{\partial z^2}. \qquad (2.2.3.3)$$

Приращение внутренней энергии в параллелепипеде в направлении оси Z будет равно:

$$dQ_Z = dQ_{Z1} - dQ_{Z2} = \lambda dxdydzd\tau \frac{\partial^2 t}{\partial z^2}. \qquad (2.2.3.4)$$

Для осей X и Y приращение внутренней энергии будет записываться аналогично:

$$dQ_X = dQ_{X1} - dQ_{X2} = \lambda dxdydzd\tau \frac{\partial^2 t}{\partial x^2}, \qquad (2.2.3.5)$$

$$dQ_Y = dQ_{Y1} - dQ_{Y2} = \lambda dxdydzd\tau \frac{\partial^2 t}{\partial y^2}. \qquad (2.2.3.6)$$

Полное приращение энергии в выделенном объеме равно

$$dQ = dQ_X + dQ_Y + dQ_Z = \lambda dxdydzd\tau \left(\frac{\partial^2 t}{\partial x^2} + \frac{\partial^2 t}{\partial y^2} + \frac{\partial^2 t}{\partial z^2} \right). \qquad (2.2.3.7)$$

С другой стороны, согласно закону сохранения энергии, можно записать:

$$dQ = dxdydzd \cdot p \cdot c \frac{\partial t}{\partial \tau} d\tau, \qquad (2.2.3.8)$$

где $dxdydz$ – объем параллелепипеда; ρ – плотность тела; c – удельная теплоемкость; $(\partial t/\partial \tau)\ d\tau$ – изменение температуры во времени.

Левые и правые части уравнений 2.2.3.7 и 2.2.3.8 равны, поэтому

$$\lambda dxdydzd\tau \frac{\partial t}{\partial z} \left(\frac{\partial^2 t}{\partial x^2} + \frac{\partial^2 t}{\partial y^2} + \frac{\partial^2 t}{\partial z^2} \right) = dxdydzd \cdot p \cdot c \frac{\partial t}{\partial \tau} d\tau, \qquad (2.2.3.9)$$

$$\frac{\partial t}{\partial \tau} = \frac{\lambda}{cp} \left(\frac{\partial^2 t}{\partial x^2} + \frac{\partial^2 t}{\partial y^2} + \frac{\partial^2 t}{\partial z^2} \right). \qquad (2.2.3.10)$$

Уравнение (2.2.3.10) называется **дифференциальным уравнением теплопроводности**, или **дифференциальным уравнением Фурье** для трехмерного температурного поля при отсутствии внутренних источников теплоты. Это уравнение устанавливает связь между пространственными и временными изменениями температуры в любой точке поля и является основным при изучении теплопроводности.

Для упрощения записи уравнения 2.2.3.10 вводят следующие обозначения:

$\frac{\lambda}{cp} = a$ – коэффициент температуропроводности, м^2/с;

$\left(\frac{\partial^2 t}{\partial x^2} + \frac{\partial^2 t}{\partial y^2} + \frac{\partial^2 t}{\partial z^2} \right) = \nabla^2 t$ – оператор Лапласа.

Коэффициент температуропроводности α, м^2/с, характеризует скорость изменения температуры и является мерой теплоинерционных свойств тела. Таким образом, уравнение 2.2.3.10 можно записать в более компактном виде:

$$\frac{\partial t}{\partial \tau} = a\nabla^2 t. \qquad (2.2.3.11)$$

Дифференциальное уравнение теплопроводности с источниками теплоты внутри тела будет иметь вид:

$$\frac{\partial t}{\partial \tau} = a\nabla^2 t. \qquad (2.2.3.11)$$

$$\frac{\partial t}{\partial \tau} = a\left(\frac{\partial^2 t}{\partial x^2} + \frac{\partial^2 t}{\partial y^2} + \frac{\partial^2 t}{\partial z^2}\right) + \frac{q_V}{cp}, \qquad (2.2.3.12)$$

где q_v – количество теплоты, выделяемой в единице объема вещества в единицу времени, Вт/м^3.

Дифференциальное уравнение описывает явление теплопроводности в самом общем виде, то есть описывает целый класс явлений. Для того чтобы из этого класса выделить конкретный процесс и дать его полное математическое описание, к дифференциальному уравнению необходимо присоединить математическое описание частных особенностей процесса. Эти частные особенности называются **условиями однозначности** или краевыми условиями.

Условия однозначности включают:

1. *Геометрические условия* – задают форму и линейные размеры тела, в котором протекает процесс.

2. *Физические условия* – задают физические параметры тела, также может быть задан закон распределения внутренних источников теплоты.

3. *Начальные условия* (для нестационарных процессов) – задают закон распределения температуры внутри тела в начальный момент времени:

$$t_{\tau=\tau_0} = f(x, y, z). \qquad (2.2.3.13)$$

При равномерном распределении температуры $\tau=0$, поэтому начальные условия упрощаются $t=t_0=const$.

4. *Граничные условия* – задают распределение физических параметров на поверхности тела для каждого момента времени.

Граничные условия бывают I, II и III рода. Граничные условия I рода задают распределение температуры на поверхности тела для каждого момента времени:

$$t_n = f(x, y, z, \tau), \qquad (2.2.3.14)$$

где t_n – температура поверхности тела.

Граничные условия II рода задают значение теплового потока для каждой точки поверхности тела и любого момента времени:

$$q_n = f(x, y, z, \tau), \qquad (2.2.3.15)$$

где q_n – плотность теплового потока на поверхности тела.

В простейшем случае плотность теплового потока по поверхности и во времени остается постоянной $q=q_0=const$, такой случай теплообмена имеет место при нагреве металлических изделий в высокотемпературных печах.

Граничные условия III рода задают температуру окружающей среды $t_{ж}$ и закон теплообмена между поверхностью тела и окружающей средой.

Процесс теплообмена между поверхностью тела и окружающей средой называется **теплоотдачей**. Теплоотдача является очень сложным процессом и зависит от большого количества параметров. Граничные условия III рода можно записать в виде:

$$\left(\frac{\partial t}{\partial n}\right)_c = -\frac{\alpha}{\lambda}(t_c - t_{ж}), \qquad (2.2.3.16)$$

где $(\partial t/\partial n)_c$ – температурный градиент на поверхности тела, м/℃; t_c – температура поверхности тела, ℃; α – коэффициент теплоотдачи, Вт/(м^2℃).

Таким образом, решение дифференциального уравнения теплопроводности при заданных условиях однозначности позволяет определить температурное поле во всем объеме тела для любого момента времени, т.е. найти функцию $t = f(x, y, z, \tau)$.

Параграф 2.2.4. Теплопроводность при стационарном режиме. Теплопроводность через однослойную плоскую стенку

Рассмотрим однослойную плоскую стенку, длина и ширина которой бесконечно велики по сравнению с толщиной δ, одинаковой по всей высоте (рисунок 2.2.4.1).

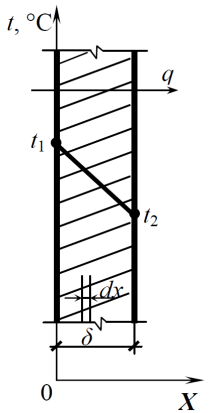


Рис. 2.2.4.1. Однослойная плоская стенка.

Температуры на поверхностях стенки t_1 и t_2 поддерживаются постоянными, т.е. поверхности являются изотермическими. Температура меняется только в направлении, перпендикулярном плоскости стенки, которые мы принимаем за ось X.

При стационарном тепловом режиме температура в любой точке тела неизменна и не зависит от времени, т.е. $\partial t/\partial\tau=0$. Тогда дифференциальное уравнение теплопроводности примет вид:

$$\left(\frac{\partial^2 t}{\partial x^2}+\frac{\partial^2 t}{\partial y^2}+\frac{\partial^2 t}{\partial z^2}\right)=0. \qquad (2.2.4.1)$$

Так температура изменится только в направлении оси X, тогда

$$\frac{\partial^2 t}{\partial y^2}=\frac{\partial^2 t}{\partial z^2}=0\,, \qquad (2.2.4.2)$$

$$\frac{\partial^2 t}{\partial x^2}=0. \qquad (2.2.4.3)$$

Проинтегрировав дважды уравнение 2.2.4.3 по x, получим:

$$\frac{\partial t}{\partial x}=const=A\,, \qquad (2.2.4.4)$$

$$t=Ax+B\,, \qquad (2.2.4.5)$$

где A, B – постоянные.

Зависимость 2.2.4.5 является уравнением прямой линии, т.е. при постоянном коэффициенте теплопроводности закон изменения температуры в однослойной плоской стенке будет линейным.

Добавим к уравнению 2.2.4.5 граничные условия:

1) при $x=0$ $t=t_1$, следовательно, подставив в уравнение 2.2.4.5, получим $B=t_1$;

2) при $x=\delta$ $t=t_2$, следовательно, подставив в уравнение 2.2.4.5, получим $A\delta+ t_1=t_2$.

Отсюда можно выразить постоянную А:

$$A = \frac{t_2 - t_1}{\delta} = \frac{\partial t}{\partial x}. \qquad (2.2.4.6)$$

Подставив значение градиента температуры в уравнение Фурье, найдем плотность теплового потока q, Вт/м^2.

$$q = -\lambda \frac{\partial t}{\partial x} = -.\lambda \frac{t_2 - t_1}{\delta}, \qquad (2.2.4.7)$$

$$q = \frac{.\lambda}{\delta}(t_1 - t_2). \qquad (2.2.4.8)$$

Уравнение 2.2.4.8 является уравнением теплопроводности для однослойной плоской стенки.

Зная удельный тепловой поток, можно вычислить общее количество теплоты Q_τ, Дж, которое передается через плоскую стенку с площадью поверхности F за время τ:

$$Q_\tau = \frac{\lambda}{\delta}(t_1 - t_2)F\tau. \qquad (2.2.4.8)$$

В уравнениях 2.2.4.8 и 2.2.4.9 отношение λ/δ называется **тепловой проводимостью стенки**, а обратная величина $\delta/\lambda=R$, (м^2℃)/Вт, называется тепловым или **термическим сопротивлением стенки**. Термическое сопротивление показывает величину падения температуры при прохождении через стенку удельного теплового потока, равного единице.

Параграф 2.2.5. Теплопроводность через многослойную плоскую стенку

На практике часто встречаются плоские стенки, состоящие из нескольких плоских слоев, выполненных из различных материалов. Для многослойной плоской стенки формулу теплопроводности можно вывести из уравнения теплопроводности для каждого отдельного слоя, считая, что тепловой поток, проходящий через эти слои, один и тот же.

Рассмотрим трехслойную плоскую стенку, толщины слоев которой равны δ_1, δ_2 δ_3, а коэффициенты теплопроводности слоев равны λ_1, λ_2, λ_3 (рисунок 2.2.5.1).

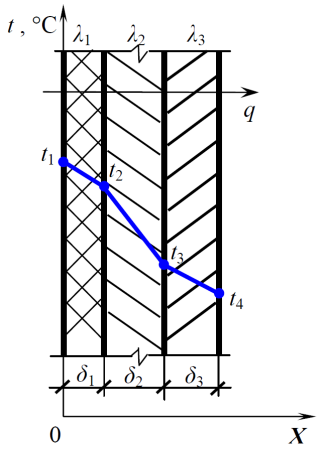


Рис. 2.2.5.1. Многослойная плоская стенка.

Для каждого слоя можно записать уравнение Фурье как для однослойной плоской стенки:

$$q = \frac{\lambda_1}{\delta_1}(t_1 - t_2),\ q = \frac{\lambda_2}{\delta_2}(t_2 - t_3),\ q = \frac{\lambda_3}{\delta_3}(t_3 - t_4). \qquad (2.2.5.1)$$

Можно решить эту систему относительно разности температур:

$$\frac{\delta_1}{\lambda_1} q = (t_1 - t_2);\ \frac{\delta_2}{\lambda_2} q = (t_2 - t_3),\ \frac{\delta_3}{\lambda_3} q = (t_3 - t_4),$$

$$q\left(\frac{\delta_1}{\lambda_1} + \frac{\delta_2}{\lambda_2} + \frac{\delta_3}{\lambda_3}\right) = t_1 - t_4\,. \qquad (2.2.5.2)$$

В итоге получим уравнение теплопроводности для трехслойной плоской стенки:

$$q = \frac{t_1 - t_4}{\frac{\delta_1}{\lambda_1} + \frac{\delta_2}{\lambda_2} + \frac{\delta_3}{\lambda_3}}\,. \qquad (2.2.5.3)$$

Величина, стоящая в знаменателе уравнения 2.2.5.3, представляет собой термическое сопротивление многослойной плоской стенки R_{λ}. Тогда уравнение можно переписать в виде

$$q = \frac{t_1 - t_4}{R_{\lambda}}. \qquad (2.2.5.4)$$

Неизвестные температуры t_2 и t_3 можно определить из условия постоянства теплового потока q=*const*:

$$q = \frac{t_1 - t_4}{R_\lambda} = \frac{t_1 - t_2}{R_1} \rightarrow t_2 = t_1 - \frac{R_1}{R_\lambda}(t_1 - t_4). \qquad (2.2.5.5)$$

Для плоской стенки, имеющей n слоев, уравнение примет вид:

$$q = \frac{t_1 - t_{n+1}}{\sum_{i=1}^{n} \frac{\delta_i}{\lambda_i}}, \qquad (2.2.5.6)$$

где δ_i и λ_i – толщина и коэффициент теплопроводности i-го слоя.

РАЗДЕЛ 2.3. Приближенные модели на микроуровне

Параграф 2.3.1. Основные понятия и определения

Аналитическое решение в явном виде возможно лишь для простейших частных процессов, например, для однослойной и многослойной цилиндрической и сферической поверхности. В общем случае точное решение краевых задач практически невозможно, поэтому при их решении практически всегда используют приближенные методы.

Чаще всего используют приближенные модели на основе интегральных уравнений или метод сеток. Модели на основе интегральных уравнений заключаются в том, что исходное ДУЧП заменяется эквивалентным интегральным уравнением. Затем оно преобразуется и приближенно решается одним из численных методов. Метод сеток гораздо чаще используется в САПР. Он основан на том, что искомая непрерывная функция аппроксимируется приближенными значениями.

Они определяются для неких конкретных значений аргумента, иными словами, непрерывный аргумент исходной функции заменяется его дискретными значениями. Их называют узлами, и они в свою очередь образуют сетку. Она является дискретной моделью области определения искомой функции. Наибольшее распространение получили 2 метода:

1) метод конечных элементов (МКЭ);

2) метод конечных разностей (МКР).

Оба метода состоят из трех этапов:

1. Дискретизация задачи: нанесение сетки на область определения.

2. Алгебраизация задачи: получение алгебраических уравнений относительно числовых значений функции.

3. Решение полученной системы уравнений (САУ).

МКР и МКЭ на первых двух этапах различны. Методы сеток позволяют свести задачи численного ДУЧП к решению системы алгебраических уравнений.

Рассмотрим МКЭ. Он имеет следующие достоинства:

1) Относительная доступность реализации.

2) Геометрическая гибкость и применимость метода к произвольным формам как в области применения, так и в области решения функций.

3) Единственность решения для любой области определения.

4) Метод позволяет строить эффективные алгоритмы на его основе.

К недостаткам следует отнести большую трудоемкость вычислений.

Параграф 2.3.2. Метод конечных элементов.

Основные этапы метода конечных элементов.

В МКЭ искомая функция аппроксимируется кусочно-непрерывными функциями. При этом область определения разбивается на части, каждая из которых называется конечным элементом. Искомая функция приближенно аппроксимируется другой функцией внутри каждого конечного элемента. Аппроксимирующая функция может иметь произвольный вид, но чаще всего для целей аппроксимации используют полиномы. Их подбирают таким образом, чтобы они сохраняли непрерывность в узлах сетки.

МКЭ состоит из следующих этапов:

1) выделение конечных элементов, то есть разбиение области определения на конечные элементы;

2) определение аппроксимирующей функции для конечных элементов;

3) объединение конечных элементов в ансамбль. Уравнения для каждого конечного элемента объединяются в общую систему для всей области определения;

4) определение вектора узловых значений.

Рассмотрим одномерный поток тепла в металлическом стержне с теплоизолированной боковой поверхностью (рисунок 2.3.2.1).

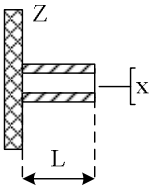


Рис. 2.3.2.1. Металлический стержень с теплоизолированной боковой поверхностью.

$$\frac{\partial t}{\partial y} = \frac{\partial t}{\partial z} = 0$$

Один конец стержня прикреплен к металлической плите. К стержню подводится тепловой поток заданной интенсивности $q = -\lambda \frac{\partial t}{\partial x}$. На свободном конце осуществляется конвекционный теплообмен $q = \alpha\left(t - t_{ср}\right)$.

Математическая постановка задачи имеет вид:

$$q + \lambda \frac{\partial t}{\partial x} = 0, \textit{ при } \mathrm{x} = 0,$$

$$\lambda \frac{\partial t}{\partial x} + \alpha(t - t_{ср}) = 0, \textit{ при } \mathrm{x} = \mathrm{L}.$$

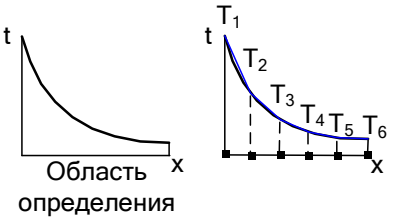


Рис. 2.3.2.2. Искомая и аппроксимированная функции.

Таким образом, искомая функция аппроксимируется кусочно-непрерывной функцией (рисунок 2.3.2.2).

Узловые значения T_i искомой функции не известны, но построить приближенную модель проще, если предположить, что они известны. Через них определяется аппроксимирующая функция, а сами значения T_i определяются на последнем этапе МКЭ.

Если область определения является двумерной, то в качестве конечных элементов выбираются плоские фигуры, чаще всего треугольники или четырехугольники, а аппроксимирующей функцией

будет кусочно-непрерывная поверхность. Если кусочная функция является трехмерной, то в качестве конечного элемента выбирается тетраэдр или параллелепипед.

Первый этап МКЭ. Выделение конечных элементов.

На этом этапе область определения искомой функции разбивается на множество конечных элементов. В двумерных областях чаще всего выделяются конечные элементы в виде треугольников. Границы этих областей могут быть и нелинейными. От формы конечных элементов существенно зависит точность решения. Обычно лучшую точность дают конечные элементы, близкие к равностороннему треугольнику.

Конечные элементы могут иметь различные размеры (например, если предполагается резкое изменение искомой функции). При произвольной области выбор узлов начинается с границы области (рисунок 2.3.2.3). Это позволяет наиболее точно аппроксимировать границы.

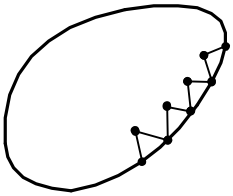


Рис. 2.3.2.3. Выбор узлов с границы области.

Очень часто разбиение производят в 2 этапа. На первом разбивают на крупные участки, подобласти. Как правило, разделение области определения на подобласти связано с резкими изменениями ее геометрической формы, либо с изменениями физических свойств материалов этой области (рисунок 2.3.2.4).

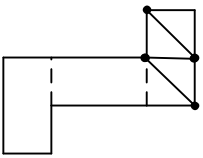


Рис. 2.3.2.4. Двухэтапное разбиение области.

На границе следует избегать резкого изменения размеров. Большое значение на этом этапе имеет правильная нумерация узлов области определения. Это связано с тем, что матрица коэффициентов

САУ в МКЭ, которая получается на III этапе, является сильно разряженной, потому что каждая ее строка (каждое уравнение в САУ) учитывает только те коэффициенты (узлы), которые относятся только к одному конечному элементу. А так как число конечных элементов очень велико, и в каждой строке оказывается только несколько ненулевых элементов, то эта матрица преобразуется в матрицу ленточной структуры.

$$\begin{bmatrix} a & a & a & 0 & 0 & 0 & 0 \\ a & a & a & a & 0 & 0 & 0 \\ a & a & a & a & a & 0 & 0 \\ 0 & a & a & a & a & a & 0 \\ 0 & 0 & a & a & a & a & a \\ 0 & 0 & 0 & a & a & a & a \\ 0 & 0 & 0 & 0 & a & a & a \end{bmatrix}$$

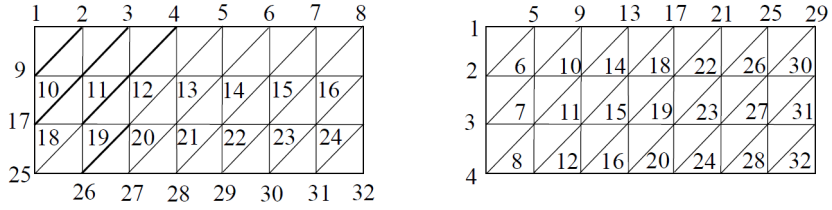
Для ленточных матриц требуется меньшая трудоемкость вычислений. В матрице для заданной системы все ненулевые коэффициенты заключены между линиями, параллельными главной диагонали матрицы.

Расстояние между главной диагональю и одной из этих линий называется шириной полосы матрицы и является показателем эффективности вычислений: чем ширина полосы уже, тем меньше размер требуемой машинной памяти и время вычислений. Ширина полосы существенно зависит от порядка нумерации узлов.

$A=(B+1)*C,$

где A – ширина полосы, C – число степеней свободы (число функций, которые определяются в одном узле), B – максимальная разность номеров узлов в пределах одного конечного элемента.

Таким образом, при нумерации узлов необходимо стремиться к тому, чтобы разница между номерами узлов в элементах области была как можно меньше. Для простой прямоугольной области проиллюстрируем это на рисунке 2.3.2.5.



а) *б)*

Рис. 2.3.2.5. Нумерация узлов.

Вариант *а)* менее эффективен, так как наибольшая разница между номерами узлов в элементе равна 8. Вариант *б)* более эффективен, соответствующая разница равна 4.

Очевидно, что информация этого этапа является исходной для последующих. Объем информации очень велик, так как велико число конечных элементов. В каждом узле должна храниться информация о его номере, координатах, принадлежности к конкретным конечным элементам, о типе каждого конечного элемента. Должны быть данные о тех функциях, которые определены для этого узла. Вручную перебрать такой объем информации без ошибок невозможно.

Однако часто с первого раза разбиение на элементы оказывается неудовлетворяющим требуемой точности, поэтому после реализации всех последующих этапов часто возвращаются к коррекции первого.

Второй этап МКЭ. Выбор аппроксимирующей функции конечного элемента.

Основные понятия.

В МКЭ очень важным является тот факт, что и форму аппроксимирующего конечного элемента, и его аппроксимирующую функцию часто выбирают только один раз, независимо от размера конечного элемента и его положения в области определения. В дальнейшем этот тип конечного элемента и выбранная ранее конечная функция могут многократно использоваться как при решении краевой задачи, так и многих других.

Как отмечалось, в качестве аппроксимирующих функция часто выбирают полиномы. Их порядок зависит от количества данных, используемых в узле. В зависимости от степени полинома выделяют:

- симплексэлементы;
- комплексэлементы;
- мультиплексэлементы.

Полиномы **симплексэлементов содержат** константы и линейные члены. Количество коэффициентов в них на один больше координатного пространства. Пример полинома: $\varphi=a_1+a_2x$. $\varphi=a_1+a_2x+a_3y$ (рисунок 2.3.2.6).

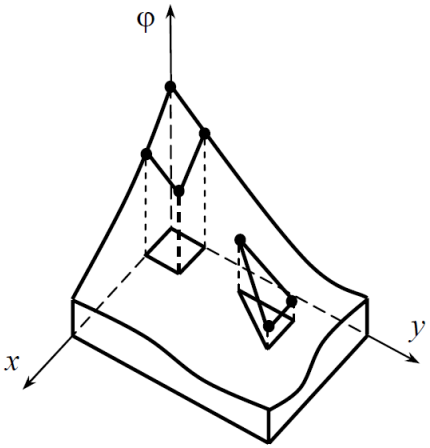


Рис. 2.3.2.6. Пример мультиплекс элемента.

Заметим, что число коэффициентов в полиноме, как правило, равно числу узлов. **Комплексэлементы содержат** константы, линейные члены, а также нелинейности. Количество коэффициентов в полиномах комплексэлементов больше, чем в полиномах симплексэлементов.

Конечные элементы для комплексэлементов могут иметь такую же форму, что и для симплексэлементов. Для учета нелинейности вводят дополнительные узлы:

$$\varphi = a_1 + a_2x + a_3y + a_4x^2 + a_5y^2 + a_6xy ,$$

$\varphi = a_1 + a_2x + a_3x^2$ (дуга).

Мультиплексэлементы помимо констант и линейных членов содержат и нелинейности, но на них накладываются дополнительные условия: их границы параллельны координатным осям.

Следует отметить, что чем резче меняется искомая функция, тем меньше размеры конечных элементов для получения одной и той же точности.

Одномерный симплекс элемент.

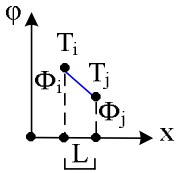


Рис. 2.3.2.7. Одномерный симплекс элемент.

$\varphi=a_1+a_2x$.

Будем считать, что узловые значения Φ_i и Φ_j известны. Из условия непрерывности можно заключить, что

$\varphi = \Phi_i$, *при* $x = X_i$; $\Phi_i = a_1 + a_2 X_i$

$\varphi = \Phi_j$, *при* $x = X_j$; $\Phi_j = a_1 + a_2 X_j$

$$a_1 = \frac{\Delta_1}{\Delta} = \frac{\begin{vmatrix} \Phi_i & x_i \\ \Phi_j & x_j \end{vmatrix}}{\begin{vmatrix} 1 & x_i \\ 1 & x_j \end{vmatrix}} = \frac{\Phi_i x_j - \Phi_j x_i}{x_j - x_i} = \frac{\Phi_i x_j - \Phi_j x_i}{L}$$

$$a_2 = \frac{\begin{vmatrix} 1 & \Phi_i \\ 1 & \Phi_j \end{vmatrix}}{\begin{vmatrix} 1 & x_i \\ 1 & x_j \end{vmatrix}} = \frac{\Phi_j - \Phi_i}{L}$$

$$\varphi = \frac{\Phi_i x_j - \Phi_j x_i}{L} + \frac{\Phi_j - \Phi_i}{L} x \qquad (2.3.2.1)$$

$$\varphi = \Phi_i \frac{x_j - x}{L} + \Phi_j \frac{x - x_i}{L} \qquad (2.3.2.2)$$

$$N_i = \frac{x_j - x}{L}; \; N_j = \frac{x - x_i}{L}. \qquad (2.3.2.3)$$

Выражения 2.3.2.3 называют функциями формы одномерного симплекс элемента. Подставим 2.3.2.3 в 2.3.2.2, тогда

$$\begin{aligned} \varphi &= N_i \Phi_i + N_j \Phi_j \\ \varphi &= N\Phi \end{aligned}, \qquad (2.3.2.4)$$

где N – матрица-строка, а Φ – матрица-столбец.

Функции формы обладают тем свойством, что каждая из них равна 1 только в своем узле и 0 во всех остальных узлах.

Двумерный симплекс элемент.

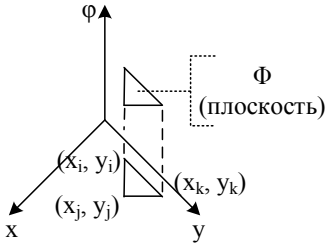


Рис. 2.3.2.8. Двумерный симплекс элемент.

Это плоский треугольник с прямоугольными сторонами и тремя узлами. Аппроксимирующий полином имеет вид:

$$\varphi = a_1 + a_2 x + a_3 y. \qquad (2.3.2.5)$$

Полагаем, что узловые значения функций Φ_i, Φ_j, Φ_k известны. Тогда условия непрерывности будут записаны как

$\varphi = \Phi_i$, *при* $x = x_i, y = y_i$

$\varphi = \Phi_j$, *при* $x = x_j, y = y_j$

$\varphi = \Phi_k$, *при* $x = x_k, y = y_k$

Подставим эти значения в выражение 2.3.2.5.

$$\Phi_i = a_1 + a_2 x_i + a_3 y_i$$

$$\Phi_j = a_1 + a_2 x_j + a_3 y_j$$

$$\Phi_k = a_1 + a_2 x_k + a_3 y_k$$

$$a_1 = \frac{\Delta_1}{\Delta} = \frac{\begin{vmatrix} \Phi_i & x_i & y_i \\ \Phi_j & x_j & y_j \\ \Phi_k & x_k & y_k \end{vmatrix}}{\begin{vmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{vmatrix}} =$$

$$= \frac{1}{\Delta}\left\{\left(x_j y_k - x_k y_j\right)\Phi_i + \left(x_k y_i - x_i y_k\right)\Phi_j + \left(x_i y_j - x_j y_i\right)\Phi_k\right\}$$

$$a_2 = \frac{\Delta_2}{\Delta} = \frac{\begin{vmatrix} 1 & \Phi_i & y_i \\ 1 & \Phi_j & y_j \\ 1 & \Phi_k & y_k \end{vmatrix}}{\begin{vmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{vmatrix}} = \frac{1}{\Delta}\left\{\left(y_j - y_k\right)\Phi_i + \left(y_k - y_i\right)\Phi_j + \left(y_i - y_j\right)\Phi_k\right\}$$

$$a_3 = \frac{\Delta_3}{\Delta} = \frac{\begin{vmatrix} 1 & x_i & \Phi_i \\ 1 & x_j & \Phi_j \\ 1 & x_k & \Phi_k \end{vmatrix}}{\begin{vmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{vmatrix}} = \frac{1}{\Delta}\left\{\left(x_k - x_j\right)\Phi_i + \left(x_i - x_k\right)\Phi_j + \left(x_j - x_i\right)\Phi_k\right\}$$

$c_i = \left(x_k - x_j\right); c_j = \left(x_i - x_k\right); c_k = \left(x_j - x_i\right).$

Подставляя значения коэффициентов в исходный полином и приводя подобные члены при узловых значениях функций, получим

$$\varphi = N_i\Phi_i + N_j\Phi_j + N_k\Phi_k \qquad (2.3.2.6)$$

$$N_i = \frac{1}{2s}\left(a_i + b_i x + c_i y\right), \Delta = 2s$$

$$N_j = \frac{1}{2s}\left(a_j + b_j x + c_j y\right)$$

$$N_k = \frac{1}{2s}\left(a_k + b_k x + c_k y\right)$$

$$a_i = x_j y_k - x_k y_j$$

$$b_i = y_i - y_k$$

$$c_i = x_j - x_k$$

Аналогично получаем выражения для a_j и т.д.

Выражение 2.3.2.6 можно записать в матричной форме:

$\varphi = N\Phi$

$$N = \left[N_i, N_j, N_k\right], \quad A = \begin{bmatrix} \Phi_i \\ \Phi_j \\ \Phi_k \end{bmatrix}$$

Элементы матрицы-строки – функции формы второго симплекс элемента. Эти функции формы должны быть равны 1 только в своем узле, а во всех других узлах равны 0.

Заметим, что выражения 2.3.2.4 и 2.3.2.6 получены без конкретных размеров конечных элементов и их расположения в области определения. На этапе выбора аппроксимирующей функции аппроксимирующий полином преобразуется к виду 2.3.2.4 или 2.3.2.6. При этих преобразованиях коэффициенты полинома a_i выражаются через узловые значения искомой функции и координаты узлов конечного элемента. Функции формы легко определяются для любой точки внутри каждого элемента через координаты точек и узлов.

Пример: Пусть в результате решения задачи установили, что T_i=100 °С, T_j=60 °С. Координаты узлов конечного элемента равны: X_i=2 см, X_j=6 см, X=4 см, L=4 см. Аппроксимирующий полином имеет вид:

$$\varphi = t = N_i T_i + N_j T_j = T_i \frac{x_j - x}{L} + T_j \frac{x - x_i}{L}$$

$$\varphi = \frac{6-4}{4} 100\% + \frac{4-2}{4} 60\% = 80\% .$$

Третий этап МКЭ. Объединение конечных элементов в ансамбль.

Уравнения, полученные на втором этапе, имеют произвольные номера *i*, *j*, *k*. Для объединения данных различных конечных элементов необходимо заменить эти номера и получить САУ, решение которой позволяет получить значения функций для любых точек области определения.

Необходимо установить соответствие между произвольными и глобальными номерами. Рассмотрим следующий пример: для I конечного элемента i=1, j=2; для II конечного элемента i=2, j=3;…; для V конечного элемента i=5, j=6.

Тогда, подставляя глобальные номера в формулу 2.3.2.4, получим:

$$\begin{aligned}
\varphi^{(1)} &= N_1^{(1)} \Phi_1 + N_2^{(1)} \Phi_2 \\
\varphi^{(2)} &= N_2^{(2)} \Phi_2 + N_3^{(2)} \Phi_3 \\
\varphi^{(3)} &= N_3^{(3)} \Phi_3 + N_4^{(3)} \Phi_4 \\
\varphi^{(4)} &= N_4^{(4)} \Phi_4 + N_5^{(4)} \Phi_5 \\
\varphi^{(5)} &= N_5^{(5)} \Phi_5 + N_6^{(5)} \Phi_6
\end{aligned} \qquad (2.3.2.7)$$

Полученные системы уравнений и являются приближенной моделью. Число уравнений равно числу конечных элементов.

В уравнениях 2.3.2.7 фигурируют функции формы для каждого узла и каждого элемента. Выражения, определяющие значения функции формы тоже следует видоизменить, заменив произвольные номера на глобальные.

$$N_3^{(2)} = \frac{x - x_2}{l^{(2)}}; \quad N_3^{(3)} = \frac{x_4 - x}{l^{(3)}}.$$

Функции формы, относящиеся к одному узлу, но к различным элементам, не одинаковы даже при одинаковых размерах конечных элементов.

Пусть область определения двумерная (рисунок 2.3.2.9):

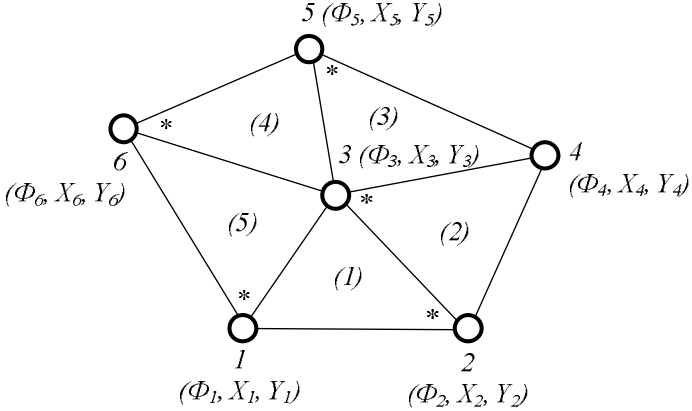


Рис. 2.3.2.9. Двумерная область определения.

$$\varphi^{(1)} = N_2^{(1)}\Phi_2 + N_3^{(1)}\Phi_3 + N_1^{(1)}\Phi_1$$
$$\cdots \qquad (2.3.2.8)$$
$$\varphi^{(5)} = N_1^{(5)}\Phi_1 + N_3^{(5)}\Phi_3 + N_6^{(5)}\Phi_6 .$$

В уравнениях 2.3.2.7 и 2.3.2.8 указаны глобальные номера, но каждое из этих уравнений относится только к одному конечному элементу, а не ко всей области определения. Такую форму уравнений называют сокращенной. Часто используется форма уравнений, в которой каждое уравнение относится ко всей области определения. Такую форму называют расширенной.

$$\varphi^{(1)} = N_1^{(1)}\Phi_1 + N_2^{(1)}\Phi_2 + N_3^{(1)}\Phi_3 + 0\cdot\Phi_4 + 0\cdot\Phi_5 + 0\cdot\Phi_6$$

…

$$\varphi^{(5)} = N_1^{(5)}\Phi_1 + 0\cdot\Phi_2 + N_3^{(5)}\Phi_3 + 0\cdot\Phi_4 + 0\cdot\Phi_5 + N_6^{(5)}\Phi_6 .$$

Расширенная форма уравнений показывает, что глобальная нумерация узлов в области определения влияет на ширину полосы. Заметим, что минимальная ширина полосы таких систем уравнений не может быть меньше максимального числа узлов в одном конечном элементе. Сокращенная форма уравнений используется непосредственно при расчетах, а расширенная часто применяется при минимизации, например, при преобразовании матрицы разряженной структуры в ленточную.

Четвертый этап МКЭ. Выбор вектора узловых значений.

Основные понятия.

Вариационное исчисление – это раздел математики, изучающий наибольшие или наименьшие значения величин (функционалов). Значения функционалов зависят от выбора одной или нескольких функций. Функционал устанавливает соответствие между множеством чисел и множеством функций (*J*). Вещественное число $t \epsilon T$ есть функционал от функции *f(x)* из множества функций *F(x)*.

Иначе говоря, функционал отображает класс функций в классе чисел.

$$J : F(x) \to T;\ J : \left[f(x), t\right],\ t = J\left[f(x)\right]$$

Понятие функционала часто используется в теории оптимального уравнения и сводится к выполнению следующего условия:

$$\min J\left[f(x)\right],\ f(x) \in F(x) .$$

В графическом виде это можно выразить через рисунок 2.3.2.10.

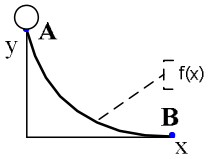


Рис. 2.3.2.10. Графическое пояснение функционала.

Необходимо подобрать такую функцию *f(x),* чтобы время, через которое шарик скатится из точки А в точку В, было минимальным.

Этап определения вектора узловых значений в свою очередь также делится на несколько этапов.

1 этап. Выбор функционала является сложной процедурой. Он требует не только знаний в вариационном исчислении, но и знаний физических процессов, протекающих в объекте. Для стационарных задач функционал, как правило, зависит от функции φ и ее производных по пространственным координатам. Чаще всего функционал выбирается в виде определенного интеграла.

$$F = \int_v f\left(\varphi, \frac{\partial\varphi}{\partial x}, \frac{\partial\varphi}{\partial y}, \frac{\partial\varphi}{\partial z}\right) dv$$

Так как область определения разбита на множество конечных элементов, то функционал для всей области определения представляет собой сумму функционалов по всем конечным элементам.

$$F = \sum_{L=1}^{n} \int_{v^{(L)}} f^{(L)}\left(\varphi^{(L)}, \frac{\partial\varphi^{(L)}}{\partial x}, \frac{\partial\varphi^{(L)}}{\partial y}, \frac{\partial\varphi^{(L)}}{\partial z}\right) dv^{(L)}$$

2 этап. Подстановка значений в функционал для функций φ и ее производных.

$$\varphi^{(L)} = N^{(L)}\Phi$$

$$\frac{\partial\varphi^{(L)}}{\partial x} = \frac{\partial N^{(L)}}{\partial x}\Phi$$

3 этап. Минимизация функционала.

$$\frac{\partial F}{\partial \varphi} = 0$$

В результате таких производных получается система математических уравнений.

$\Phi = \Lambda B$,

где Λ – это матрица теплопроводности, B – матрица тепловой нагрузки.

4 этап. Решение полученной системы уравнений.

Необходимо определить полученные узловые значения Φ. Их подставляют в систему уравнений, полученную на третьем этапе, т.е. в сокращенную или расширенную форму. Только после этого такую систему можно решать.

Определение узловых значений на примере нахождения температуры поля в стержне.

$$F = \int_V \frac{\lambda_x}{2}\left(\frac{\partial t}{\partial x}\right)^2 dV + \int_S \left[qt + \frac{\alpha}{2}\left(t - t_{cp}\right)^2\right] dS \qquad (2.3.2.9)$$

В вариационном исчислении доказано, что представленный функционал принимает минимальное значение, если удовлетворяется

следующее условие (дифференциальное уравнение вместе с граничными условиями):

$$\lambda_x \frac{\partial^2 t}{\partial x^2} = 0,\ q + \lambda_x \frac{\partial t}{\partial x} + \alpha\left(t - t_{cp}\right) = 0\ . \qquad (2.3.2.10)$$

Иными словами, подбирается такой функционал, что его минимальная функция совпадает с полученным дифференциальным уравнением и начальными условиями. Уравнения 2.3.2.10 идентичны исходным уравнениям, поэтому любое распределение температуры, при котором функционал *F*, определяемый формулой 2.3.2.9, становится минимальным, также удовлетворяет начальным дифференциальным уравнениям и является решением исходной задачи.

Уравнение 2.3.2.9 служит отправной точкой для определения температуры в каждом узле.

РАЗДЕЛ 2.4. Вопросы для самопроверки

1. Для решения каких задач используются математические модели на микроуровне?
2. Что такое кондукционный теплообмен?
3. Дайте определения следующим понятиям: температурное поле, изотермическая поверхность, температурный градиент.
4. В чем состоит чуть закона Фурье?
5. Выведите дифференциальное уравнение теплопроводности.
6. Перечислите условия однозначности.
7. Охарактеризуйте граничные условия I, II и III рода.
8. Запишите уравнение теплопроводности для однослойной плоской стенки.
9. Запишите уравнение теплопроводности для многослойной плоской стенки.
10. Опишите различия между методами конечных элементов и конечных разностей.
11. Перечислите основные этапы метода конечных элементов.
12. Как происходит выбор формы конечных элементов на первом этапе МКЭ?
13. Опишите процесс выбора аппроксимирующей функции конечного элемента.
14. Чем расширенная форма уравнения конечных элементов отличается от сокращенной?
15. На какие этапы делится процесс определения вектора узловых значений.

ГЛАВА 3. МАТЕМАТИЧЕСКИЕ МОДЕЛИ НА МАКРОУРОВНЕ

РАЗДЕЛ 3.1. Общая характеристика моделей макроуровня

В моделях микроуровня анализируются процессы в сплошных средах. В этих моделях не учитывается структуры. Такие модели пригодны для анализа элементов небольших объемов. В моделях более высоких уровней и в самих объектах исследований выделяют системы, элементы и условия внешней среды, в которых функционируют эти объекты. Модели данного уровня подразумевают анализ объектов с дискретной структурой.

Модели макроуровня относятся к схемотехническому уровню проектирования. Эти модели получаются, как правило, в результате объединения комплексных и топологических уравнений. Обычно имеют вид ОДУ. Независимыми переменными в них являются время и частота. Фазовыми переменными в них являются переменные вроде потока и напряжения. При моделировании на этом уровне задача состоит в том, чтобы при получении комплексных уравнений построить модель и определить ее фазовые и выходные переменные.

Комплексные уравнения отражают законы функционирования процессов, протекающих в элементах. Эти уравнения связывают разнородные переменные, относящиеся к одному и тому же элементу.

Топологические уравнения указывают на связь между отдельными элементами. При автоматизированном проектировании математическое обеспечение и программные средства должны предполагать применение элементов различной физической природы, при этом комплексные и топологические уравнения будет иметь универсальный вид (они должны быть пригодны для описания систем различной физической природы). Эта универсальность достигается за счет принципа аналогии.

РАЗДЕЛ 3.2. Сведения о начальных моментах случайных величин

В теории вероятностей наибольшую роль играет *математическое ожидание*, которое иногда называют просто **средним значением** случайной величины (СВ).

Подойдем к понятию математического ожидания исходя из механической интерпретации распределения дискретной СВ. Пусть единичная масса распределена между точками оси абсцисс

$x_1, x_2, \ldots x_n$, причем материальная точка x_i имеет массу p_i $(i = 1, 2, \ldots, n)$.

Нам требуется выбрать на оси абсцисс точку, характеризующую положение всей системы материальных точек с учетом их масс. Естественно в качестве такой точки взять центр массы системы материальных точек. Обозначим абсциссу центра массы $\mathbf{M}[X]$. Имеем:

$$\mathbf{M}[X] = \frac{x_1 p_1 + x_2 p_2 + \ldots + x_n p_n}{p_1 + p_2 + \ldots + p_n} = \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i},$$

или, учитывая, что $\sum_{i=1}^{n} p_i = 1$,

$$\mathbf{M}[X] = \sum_{i=1}^{n} x_i p_i .$$

Это есть **среднее взвешенное** значение СВ X, в которое абсцисса каждой точки x_i входит с «весом», равным соответствующей вероятности.

Полученное таким образом среднее значение случайной величины X называется ее **математическим ожиданием**. Это – одно из важнейших понятий теории вероятностей. Дадим ему словесную формулировку.

Математическим ожиданием дискретной случайной величины называется сумма произведений всех возможных ее значений на вероятности этих значений.

Теперь рассмотрим случай, когда число возможных значений дискретной СВ X не конечно, а бесконечно (образует счетное множество). Формула для математического ожидания остается такой же, только в верхнем пределе суммы n заменяется на бесконечность:

$$\mathbf{M}[X] = \sum_{i=1}^{\infty} x_i p_i .$$

Некоторая сложность заключается в том, что бесконечная сумма может и расходиться, т.е. соответствующая СВ X может и не иметь математического ожидания. Например, для СВ X с рядом распределения из таблицы 3.2.1 сумма расходится (равна ∞), и, значит, у такой СВ матического ожидания не существует.

Таблица 3.2.1.

	2	2^2	2^3	…	2^1	…
X:	1/2	$1/2^2$	$1/2^3$	…	$1/2^1$	…

Перейдем от дискретной СВ *X* к непрерывной с плотностью $f(x)$. Механическая интерпретация математического ожидания сохранит тот же смысл: центр массы для единичной массы, распределенной непрерывно на оси абсцисс с плотностью $f(x)$. Заменяя в предыдущей формуле «скачущий» аргумент x_i непрерывно меняющимся *x*, а вероятность p_i – элементом вероятности $f(x)\,dx$, получим:

$$\mathbf{M}[X] = \int_{-\infty}^{\infty} x f(x)\,dx .$$

Разумеется, те значения *x*, для которых $f(x) = 0$, можно выбросить из области интегрирования. Так же, как и сумма, интеграл может расходиться, а математическое ожидание – не существовать, но на практике обычно область значений СВ, для которых $f(x) \neq 0$, ограничена и математическое ожидание существует.

Математическое ожидание СВ связано тесной зависимостью со среднеарифметическим ее наблюденных значений при большом числе опытов. Действительно, пусть имеется дискретная СВ *X* с рядом распределения и *k* значениями (таблица 3.2.2), где $p_i = \mathrm{P}\{X = x_i\}$.

Таблица 3.2.2.

	x_1	x_2	…	x_i	…	x_k
X:	p_1	p_2	…	p_i	…	p_k

Пусть производится *n* независимых опытов, в каждом из которых СВ *X* принимает определенное значение из множества $\{x_1, x_2, \ldots, x_n\}$. Предположим, что значение x_1 появилось n_1 раз, значение x_2 – n_2 раз и т. д.; $\sum_{i=1}^{k} n_i = n$. Среднее арифметическое наблюденных значений СВ *X* обозначим $\mathbf{M}^*[X]$. Имеем:

$$\mathbf{M}^*[X] = \left(x_1 n_1 + x_2 n_2 + \ldots x_k n_k\right) / n = \sum_{i=1}^{k} \left(x_i n_i\right) / n .$$

Но n_i/n есть не что иное, как частота (или статистическая вероятность) события $\{X = x_i\}$; обозначим ее p_i^*:

$$\mathbf{M}^*[X] = \sum_{i=1}^{k} x_i p_i^*,$$

т.е. среднее арифметическое наблюденных значений СВ равно сумме произведений ее возможных значений на соответствующие им частоты.

Мы знаем, что при увеличении числа опытов *n* частота события p_i^* будет приближаться (сходиться по вероятности) к вероятности p_i этого события. Значит, и среднее арифметическое $\mathbf{M}^*[X]$ будет приближаться (сходиться по вероятности) к математическому ожиданию $\mathbf{M}[X]$ случайной величины *X*. Это значит, что при достаточно большом числе опытов среднее арифметическое наблюденных значений СВ *X* можно принимать приближенно равным ее математическому ожиданию.

Начальным моментом *s*-го порядка случайной величины *X* называется математическое ожидание *s*-той степени этой величины:

$$\mu_s[X] = M[X^s].$$

Для дискретной случайной величины *X* начальный момент *s*-го порядка выражается суммой:

$$\mu_s[X] = \sum_{i=1}^{n} x_i^s p_i,$$

где x_i – значения СВ *X*, p_i – соответствующие вероятности; для непрерывной – интегралом:

$$\mu_s[X] = \int_{-\infty}^{\infty} x^s f(x) dx,$$

где $f(x)$ – плотность распределения.

Ранее введенная характеристика положения – математическое ожидание СВ – есть не что иное, как ее **первый начальный момент**:

$$m_x = \mathbf{M}[X] = \mu_1[X].$$

Дисперсия выражается через **второй начальный момент**:

$$D_x = \mu_2 - m_x^2 = \mu_2 - \mu_1^2.$$

Или в других обозначениях:

$$D_x = \mathbf{M}[X^2] - (\mathbf{M}[X])^2,$$

т.е. *дисперсия случайной величины равна математическому ожиданию ее квадрата минус квадрат математического ожидания.*

Дисперсия случайной величины есть характеристика рассеивания, разбросанности СВ около ее математического ожидания. Само слово «дисперсия» означает рассеивание.

Дисперсия имеет размерность квадрата случайной величины, что не всегда удобно. Для наглядности в качестве характеристики рассеивания удобнее пользоваться числом, размерность которого совпадает с размерностью СВ. Для этого из дисперсии извлекают квадратный корень. Полученная величина называется **средним квадратическим отклонением** (иначе «стандартом» или «стандартным отклонением») случайной величины. Будем обозначать его $\sigma[X]$ (или σ_x):

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

В качестве корня допускаются только положительные значения.

Для упрощения записей мы часто будем пользоваться сокращением σ_x для среднеквадратического отклонения (или просто σ, если ясно, о какой СВ идет речь).

Для неотрицательной случайной величины X в качестве характеристики «степени ее случайности» иногда применяется **коэффициент вариации**, равный отношению среднеквадратического отклонения (СКО) к математическому ожиданию (МО):

$$v = \sigma / m\ .$$

Зная МО и СКО случайной величины X, можно составить приближенное представление о диапазоне ее возможных значений: значения случайной величины X лишь иногда выходят за пределы интервала

$$m \pm 3\sigma\ ,$$

и в большинстве случаев можно считать, что они укладываются в этот интервал. Это правило носит название «**правила трех сигм**». Согласно этому правилу для того, чтобы приближенно представить себе размах случайных отклонений СВ X от ее МО, достаточно отложить от точки m вправо и влево по отрезку, равному 3σ.

Неравенство Чебышева. Пусть случайная величина X имеет конечные математическое ожидание μ и дисперсию σ^2. Тогда

$\mathrm{P}\left(\left| X-\mu \right| \geq a\right) \leq \frac{\sigma^2}{a^2}$, где $a>0$. Если $a=k\sigma$, где σ – стандартное отклонение и $k>0$, то получаем $\mathrm{P}\left(\left| X-\mu \right| \geq k\sigma\right) \leq \frac{1}{k^2}$.

В частности, случайная величина с конечной дисперсией отклоняется от среднего больше чем на 2 стандартных отклонения с вероятностью меньше 25 %. Она отклоняется от среднего на 3 стандартных отклонения с вероятностью меньше 11,2 %.

Неравенство Высочанского-Петунина. Единственным ограничением на функцию плотности распределения вероятности является то, что она должна быть одномодальной и иметь конечную дисперсию. Это неравенство справедливо, в том числе, и для резко асимметричных распределений, тем самым устанавливая границы для множества значений случайной величины, попадающих в определенный интервал.

Пусть X случайная величина с одномодальным распределением, средним значением μ и конечной ненулевой дисперсией σ^2. Тогда для любого $k>\sqrt{8/3}=1{,}63299\ldots$ будет справедливо следующее неравенство:

$$\mathrm{P}\left(\left| X-\mu \right| \geq k\sigma\right) \leq \frac{4}{9k^2}.$$

В приложениях математической статистики очень часто используется эвристическое правило, при котором $k=3$, что соответствует верхней границе вероятности 4/81= 0,04938…, и таким образом строится граница, которая включает 95,06% значения случайной величины. В случае нормального распределения оценка улучшается до 99,73%.

РАЗДЕЛ 3.3. Основные понятия цепей Маркова

Рассмотрим независимые испытания, которые можно описать следующим образом. Задано множество возможных исходов $E_1, E_2, \ldots$ (в конечном или бесконечном числе), и каждому из них соотнесена некоторая вероятность p_k; вероятности последовательностей исходов определяются по правилу умножения: $\mathbf{P}\left\{E_{j_0}, E_{j_1}, \ldots, E_{j_n}\right\} = p_{j_0} p_{j_1} \ldots p_{j_n}$. В теории цепей Маркова мы рассматриваем простейшее обобщение этой схемы, которое состоит в

том, что для любого испытания допускается *зависимость его от непосредственно предшествующего ему испытания (и только от него)*. С исходом E_k не связана более фиксированная вероятность p_k, но зато каждой паре E_j, E_k теперь соответствует условная вероятность p_{jk}; при условии, что E_j появился в некотором испытании, вероятность появления E_k в следующем испытании равна p_{jk}. Помимо p_{jk} должны быть заданы вероятности a_k исходов в начальном испытании. Чтобы p_{jk} имели приписанный им смысл, вероятности последовательностей исходов, соответствующих двум, трем или четырем испытаниям, должны быть определены равенствами

$$\mathbf{P}\{E_j, E_k\} = a_j p_{jk}, \quad \mathbf{P}\{E_j, E_k, E_r\} = a_j p_{jk} p_{kr},$$

$$\mathbf{P}\{E_j, E_k, E_r, E_s\} = a_j p_{jk} p_{kr} p_{rs},$$

$$\mathbf{P}\{E_{j_0}, E_{j_1}, \ldots, E_{j_n}\} = a_{j_0} p_{j_0 j_1} p_{j_1 j_2} \ldots p_{j_{n-2}} p_{j_{n-1}} p_{j_{n-1}} p_{j_n}. \qquad (3.3.1)$$

Здесь начальному испытанию присвоен нулевой номер, так, что испытание номер один является вторым.

Последовательность испытаний с возможными исходами $E_1, E_2, \ldots$ называется **цепью Маркова**, если вероятности последовательностей исходов определяются формулой (3.3.1) через распределение вероятностей $\{a_k\}$ для E_k в начальном (или нулевом) испытании и через фиксированные условные вероятности p_{jk} появления E_k при условии, что в предыдущем испытании появился E_j.

Для цепей Маркова удобнее несколько видоизмененная технология. Возможные исходы E_k обычно называются возможными **состояниями системы**; вместо того, чтобы говорить, что *n*-е испытание окончилось появлением E_k, говорят, что *n*-й шаг приводит к состоянию E_k или что система попадает в E_k на *n*-м шаге. Наконец, p_{jk} называется вероятностью перехода из E_j в E_k. Как обычно, мы считаем, что испытания происходят через равные интервалы времени, так что номер шага служит временным параметром.

Вероятности перехода p_{jk} будут расположены в матрицу переходных вероятностей

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ p_{31} & p_{32} & p_{33} & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix}.$$

где первый индекс означает номер строки, а второй – номер столбца. Ясно, что P – квадратная матрица с неотрицательными элементами и единичными суммами по строкам. Такая матрица (конечная или бесконечная) называется **стохастической матрицей**. Любая стохастическая матрица может служить матрицей переходных вероятностей; вместе с начальным распределением $\{a_k\}$ она полностью определяет цепь Маркова с состояниями $E_1, E_2, \ldots$.

В некоторых частных случаях бывает удобно нумеровать состояния, начиная с 0, а не с 1. Тогда к матрице P следует добавить нулевые строку и столбец.

Рассмотрим несколько пояснительных примеров.

Пример 1. Когда у цепи есть только два возможных состояния E_1 и E_2, матрица переходных вероятностей с необходимостью имеет вид

$$P = \begin{bmatrix} 1-p & p \\ \alpha & 1-\alpha \end{bmatrix}.$$

Подобная цепь могла бы быть реализована в следующем мысленном эксперименте. Частица движется вдоль оси x таким образом, что абсолютная величина ее скорости остается постоянной, но направление движения может меняться на противоположное. Говорят, что система находится в состоянии E_1, если частица движется направо, и в состоянии E_2, если она движется налево. Тогда p – вероятность поворота, когда частица движется направо, а α – вероятность поворота при движении налево.

Пример 2. Случайное блуждание с поглощающими экранами. Пусть возможными состояниями будут $E_0, E_1, \ldots, E_\rho$; рассмотрим матрицу переходных вероятностей

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ q & 0 & p & 0 & \dots & 0 & 0 & 0 \\ 0 & q & 0 & p & \dots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \dots & q & 0 & p \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{bmatrix}.$$

Из каждого "внутреннего" состояния $E_1, \dots, E_{\rho-1}$ возможны переходы в правое и левое соседние состояния (с вероятностями $p_{i,i+1} = p$ и $p_{i,i-1} = q$). Однако ни из E_0, ни из E_ρ невозможны переходы в какое-либо иное состояние; система будет переходить из одного состояния в другое, но как только будет достигнуто E_0 или E_ρ, система останется неизменной навсегда.

Пример 3. Отражающие экраны. Интересный вариант предыдущего примера представляет собой цепь с возможными состояниями $E_1, \dots, E_\rho$ и переходными вероятностями

$$P = \begin{bmatrix} q & p & 0 & 0 & \dots & 0 & 0 & 0 \\ q & 0 & p & 0 & \dots & 0 & 0 & 0 \\ 0 & q & 0 & p & \dots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \dots & q & 0 & p \\ 0 & 0 & 0 & 0 & \dots & 0 & q & p \end{bmatrix}.$$

Эту цепь можно интерпретировать на языке азартных игр, рассматривая двух игроков, ведущих игру с единичными ставками и с соглашением, что каждый раз, когда один из игроков проигрывает свой последний доллар, тот немедленно возвращается ему его противником, так, что игра может продолжаться бесконечно. Мы предполагаем, что игроки вместе имеют $\rho + 1$ долларов, и говорим, что система находится в состоянии E_k, если их капиталы равны k и $\rho - k + 1$ соответственно. Тогда переходные вероятности даются матрицей *P*.

РАЗДЕЛ 3.4. Вероятность перехода за несколько шагов в цепях Маркова

Обозначим через $p_{jk}^{(n)}$ **вероятность перехода** из E_j в E_k ровно за *n* шагов. Иными словами, $p_{jk}^{(n)}$ есть условная вероятность попадания в E_k на *n*-м шаге при условии, что начальным состоянием было E_j; она равна сумме вероятностей всех путей $E_j E_{j_1} \dots E_{j_{n-1}} E_k$ длины n, начинающихся в E_j и оканчивающихся в E_k. В частности, $p_{jk}^{(1)} = p_{jk}$ и

$$p_{jk}^{(2)} = \sum_{\nu} p_{j\nu} p_{\nu k} . \qquad (3.4.1)$$

По индукции мы получаем общую рекуррентную формулу:

$$p_{jk}^{(n+1)} = \sum_{\nu} p_{j\nu} p_{\nu k}^{(n)} . \qquad (3.4.2)$$

Дальнейшая индукция по *m* приводит к основному тождеству:

$$p_{jk}^{(m+n)} = \sum_{\nu} p_{j\nu}^{(m)} p_{\nu k}^{(n)} , \qquad (3.4.3)$$

которое является частным случаем уравнения Колмогорова-Чепмена. Оно отражает тот простой факт, что первые m шагов приводят из E_j в некоторое промежуточное состояние E_ν и что вероятность последующего перехода из E_ν в E_k не зависит от того, каким образом было достигнуто E_ν.

Так же как и в случае p_{jk}, образовавших матрицу P, мы расположим $p_{jk}^{(n)}$ в матрицу, которую обозначим P^n. Тогда (3.4.2) утверждает, что для того, чтобы получить элемент $p_{jk}^{(n+1)}$ матрицы P^{n+1}, мы должны умножить элементы *j*-й строки P на соответствующие элементы *k*-го столбца P^n и сложить полученные произведения. Эта операция называется умножением матриц P и P^n и выражается символически равенством $P^{n+1} = PP^n$. Данное определение позволяет назвать P^n *n*-й степенью P; уравнение (3.4.3) выражает известный закон $P^{m+n} = P^m P^n$.

Для того чтобы (3.4.3) было справедливо для всех $n \geq 0$, определим $p_{jk}^{(0)}$, положив $p_{ij}^{(0)} = 1$ и $p_{jk}^{(0)} = 0$ при $j \neq k$.

Пример. Независимые испытания. Обычно бывает трудно получить явные выражения для вероятностей перехода за несколько шагов, но как правило они не представляют особого интереса. Как важное, хотя и тривиальное исключение, отметим частный случай независимых испытаний. Он имеет место тогда, когда все строки P тождественно совпадают с данным распределением вероятностей, отсюда следует равенство $P^n = P$ при всех n.

Рассмотрим *безусловные вероятности***.** Пусть снова a_j означает вероятность состояния E_j в начальном (нулевом) испытании. Тогда (безусловная) вероятность попадания в E_k на *n*-м шаге равна

$$a_k^{(n)} = \sum_j a_j p_{jk}^{(n)}. \qquad (3.4.4)$$

Обычно мы считаем, что процесс начинается из фиксированного состояния E_i, т.е. полагаем $a_i = 1$. В этом случае $a_k^{(n)} = p_{ik}^{(n)}$. Влияние начального состояния должно постепенно ослабевать, так как при больших *n* распределение (3.4.4) должно быть почти независимым от начального распределения $\{a_j\}$. Так оно и будет, если (как в последнем примере) $p_{ik}^{(n)}$ сходится к независящему от *j* пределу, т.е. если P^n сходится к матрице с одинаковыми строками. Мы видим, что обычно это действительно так, хотя и придется еще принимать в расчет некоторые исключения, обусловленные периодичностью.

Рассмотрим пример вероятности перехода за несколько шагов (рисунок 3.4.1). Вероятности перехода за несколько шагов проиллюстрируем сначала путем возведения матрицы в степень, оперируя стохастической матрицей.

Для определения всевозможных путей достижения нужного состояния (нужной вершины в графе) проделаем подобное возведение матрицы в степень с элементами, являющимися мнемоническими обозначениями путей.

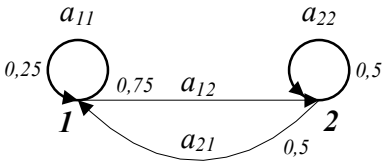


Рис. 3.4.1. Вероятности переходов.

$$\mathbf{P} = \begin{bmatrix} 0{,}25 & 0{,}75 \\ 0{,}5 & 05 \end{bmatrix};\ (\textit{за один шаг})$$

$$\mathbf{PP} = \mathbf{P}^2 = \begin{bmatrix} 0{,}25 & 0{,}75 \\ 0{,}5 & 05 \end{bmatrix} \cdot \begin{bmatrix} 0{,}25 & 0{,}75 \\ 0{,}5 & 05 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}0625 + 0{,}375 & 0{,}1875 + 0{,}375 \\ 0{,}125 + 0{,}25 & 0{,}375 + 0{,}25 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}4375 & 0{,}5625 \\ 0{,}375 & 0{,}625 \end{bmatrix}.\ (\textit{за 2 шага})$$

$$\mathbf{P} \cdot \mathbf{P}^2 = \begin{bmatrix} 0{,}25 & 0{,}75 \\ 0{,}5 & 05 \end{bmatrix} \cdot \begin{bmatrix} 0{,}4375 & 0{,}5625 \\ 0{,}375 & 0{,}625 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}25 \cdot 0{,}4375 + 0{,}75 \cdot 0{,}375 & 0{,}25 \cdot 0{,}5625 + 0{,}75 \cdot 0{,}625 \\ 0{,}5 \cdot 0{,}4375 + 0{,}5 \cdot 0{,}375 & 0{,}5 \cdot 0{,}5625 + 0{,}5 \cdot 0{,}625 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}390625 & 0{,}609375 \\ 0{,}40625 & 0{,}59375 \end{bmatrix}.\ (\textit{за 3 шага})$$

$$\mathbf{P} \cdot \mathbf{P}^3 = \begin{bmatrix} 0{,}25 & 0{,}75 \\ 0{,}5 & 05 \end{bmatrix} \cdot \begin{bmatrix} 0{,}390625 & 0{,}609375 \\ 0{,}40625 & 0{,}59375 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}25 & 0{,}75 \\ 0{,}5 & 0{,}5 \end{bmatrix} \cdot \begin{bmatrix} 0{,}390625 & 0{,}609375 \\ 0{,}40625 & 0{,}59375 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}25 \cdot 0{,}39 + 0{,}75 \cdot 0{,}406 & 0{,}25 \cdot 0{,}61 + 0{,}75 \cdot 0{,}59 \\ 0{,}5 \cdot 0{,}39 + 0{,}5 \cdot 0{,}406 & 0{,}5 \cdot 0{,}61 + 0{,}5 \cdot 0{,}59 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}40234375 & 0{,}59765625 \\ 0{,}3984375 & 0{,}6015625 \end{bmatrix} \approx \begin{bmatrix} 0{,}4 & 0{,}6 \\ 0{,}4 & 0{,}6 \end{bmatrix}.\ (\textit{за 4 шага})$$

С символьными обозначениями для отслеживания путей вероятности переходов будут определяться следующим образом:

$$\mathbf{P} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},\ (\textit{за один шаг})$$

$$\mathbf{PP} = \mathbf{P}^2 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = .$$

$$= \begin{bmatrix} a_{11}^{\,2} + a_{12} \cdot a_{21} & a_{11} \cdot a_{12} + a_{12} \cdot a_{22} \\ a_{21} \cdot a_{11} + a_{22} a_{21} & a_{21} \cdot a_{12} + a_{22}^{\,2} \end{bmatrix}$$. (*за 2 шага*) и т.д.

РАЗДЕЛ 3.5. Простейшие стохастические процессы с непрерывным временем

Цепи Маркова могут быть описаны как стохастические процессы, в которых будущее зависит лишь от настоящего состояния, но не от прошлого, или того способа, которым было достигнуто настоящее состояние. Эти процессы имеют только счетное множество значений (состояний) $E_1, E_2, \ldots$ и зависят от дискретного временного параметра: изменения могут происходить лишь в фиксированные моменты времени $t = 0, 1, \ldots$ Примерами могут выступить такие явления как телефонные вызовы, радиоактивный распад и расщепление хромосом, в которых изменения могут происходить в любой момент времени. С математической точки зрения мы будем иметь дело со стохастическими процессами со счетным множеством состояний, но зависящими уже от непрерывного временного параметра. В рамках дискретных вероятностей описание таких процессов невозможно, и мы на самом деле не в состоянии формально определить интересующий нас класс марковских процессов.

Выражение “будущее развитие не зависит от прошлой истории” имеет очевидное интуитивное значение (по крайней мере, по аналогии с дискретными цепями Маркова). Переходной вероятности $p_{jk}^{(n)}$ для цепей Маркова теперь соответствует переходная вероятность $P_{jk}(t)$, а именно условная вероятность состояния E_k в момент $t+s$ при условии, что в момент $s < t+s$ система находилась в состоянии E_j. Как показывает обозначение, предполагается, что эта вероятность зависит только от продолжительности t временного интервала, но не от его положения на оси времени. Такие переходные вероятности называются **стационарными** или однородными по времени. Основным соотношением является **уравнение Колмогорова-Чепмена**

$$P_{ik}(\tau + t) = \sum_j P_{ij}(\tau) P_{jk}(t), \qquad (3.5.1)$$

которое основано на следующем рассуждении. Предположим, что в момент времени 0 система находится в состоянии E_i. Тогда j-й член в правой части представляет вероятность сложного события,

состоящего в том, что система в момент времени τ находится в состоянии E_j, а в более поздний момент $\tau+t$ – в состоянии E_k. Но переход из состояния E_i в момент времени 0 в состояние E_k в момент $\tau+t$ с необходимостью происходит через некоторое промежуточное состояние E_j в момент времени τ, и, суммируя по всем возможным состояниям E_j, мы видим, что (3.5.1) должно выполняться для произвольных (фиксированных) $\tau>0$ и $t>0$.

Рассмотрим решения основного уравнения (3.5.1). Простые постулаты, приспособленные к конкретным ситуациям, приводят к системам дифференциальных уравнений для $P_{ik}(t)$, и что из этих уравнений, даже не решая их, можно получить интересные результаты. И эти результаты имеют смысл, потому что наши решения действительно являются переходными вероятностями марковского процесса, который однозначно определяется этими вероятностями и начальным положением в момент времени 0. Этот факт мы примем без доказательства.

Для фиксированных j и t переходные вероятности $P_{jk}(t)$ определяют обычное дискретное распределение вероятностей. Оно зависит от непрерывного параметра *t*. Технически рассуждения последующих параграфов остаются в рамках дискретных вероятностей, но это искусственное ограничение является для многих целей слишком строгим. Этот момент может проиллюстрировать распределение Пуассона $\left\{ e^{-\lambda t}(\lambda t)^n/n! \right\}$. Нулевой член $e^{-\lambda t}$ этого распределения можно интерпретировать как вероятность того, что за интервал времени фиксированной длины *t* не поступило ни одного телефонного вызова. Но тогда $e^{-\lambda t}$ будет также вероятностью того, что время ожидания первого вызова превышает *t*, и поэтому мы косвенно имеем дело с непрерывным распределением вероятностей на оси времени.

Пуассоновский процесс можно рассматривать с различных точек зрения, и здесь мы рассмотрим его в качестве прототипа всех процессов. Последующий вывод распределения Пуассона наилучшим образом подходит для наших обобщений, однако он никоим образом не является лучшим и в других контекстах.

В качестве эмпирических предпосылок возьмем такие случайные события, как распад частиц, поступающие телефонные вызовы, расщепление хромосом под воздействием вредной радиации.

Предполагается, что все наблюдаемые события однотипны, и мы интересуемся полным числом $Z(t)$ событий, происшедших в течение произвольного интервала времени длины t. Каждое событие представляется точкой на оси времени, и поэтому мы в действительности рассматриваем некоторые случайные размещения точек на прямой. Лежащие в основе нашей математической модели физические предположения состоят в том, что силы и воздействия, управляемые процессом, остаются постоянными так, что **вероятность любого отдельного события** *одна и та же для всех интервалов времени продолжительности t и не зависит от прошлого развития процесса.* Математически это означает, что наш процесс является однородным по времени марковским процессом. Выведем основные вероятности:

$$P_n(t) = P\{Z(t) = n\}.$$

Чтобы ввести понятия, подходящие и для других процессов, выберем начало отсчета времени и будем говорить, что в момент времени $t > 0$ система находится в состоянии E_n, если между 0 и t произошло ровно n скачков функции $Z(t)$. Тогда $P_n(t)$ равняется вероятности состояния E_n в момент t, однако $P_n(t)$ может быть также описана как вероятность перехода из произвольного состояния E_j в произвольный момент времени s в состояние E_{j+n} к моменту $s + t$. Теперь преобразуем данное нестрогое описание процесса в свойства вероятностей $P_n(t)$.

Разобьем временной интервал единичной длины на N подинтервалов длины $h = N^{-1}$. Вероятность скачка внутри любого из этих подинтервалов равна $1 - P_0(h)$, и поэтому математическое ожидание числа интервалов, содержащих скачки, равно $h^{-1}\left[1 - P_0(h)\right]$. При $h \to 0$ это число должно стремиться к математическому ожиданию **числа скачков** внутри произвольного интервала времени единичной длины, и поэтому естественно предположить, что существует число $\lambda > 0$ такое, что

$$h^{-1}\left[1 - P_0(h)\right] \to \lambda. \qquad (3.5.2)$$

Физическая картина процесса требует также, чтобы скачок обязательно приводил из состояния E_j в соседнее состояние E_{j+1}, тогда математическое ожидание числа подынтервалов (длины h),

содержащих более чем один скачок, должно стремиться к 0. Можно предположить, что при $h \to 0$

$$h^{-1}\left[1 - P_0(h) - P_1(h)\right] \to 0. \qquad (3.5.3)$$

Чтобы окончательно сформулировать постулаты, запишем (3.5.2) в виде $P_0(h) = 1 - \lambda h + o(h)$, где $o(h)$ обозначает величину, по порядку меньшую чем h. (Точнее говоря, $o(h)$ означает такую величину, что $h^{-1}o(h) \to 0$ при $h \to 0$). С учетом этого (3.5.3) эквивалентно соотношению $P_1(h) = \lambda h + o(h)$. Сформулируем теперь следующие постулаты.

Постулаты пуассоновского процесса. Процесс начинается в момент времени 0 в состоянии E_0(i). Непосредственный переход из состояния E_j возможен только в состояние $E_{j+1}(ii)$. Каково бы ни было состояние E_j процесса в момент времени t, (условная) вероятность скачка внутри последующего короткого интервала времени между t и $t+h$ равна $\lambda h + o(h)$, тогда как (условная) вероятность наличия в нем более чем одного скачка есть $o(h)$.

Данные постулаты носят исключительно аналитический характер, и их достаточно, чтобы показать следующее:

$$P_n(t) = \left[(\lambda t)^n / n!\right] e^{-\lambda t}. \qquad (3.5.4)$$

Для доказательства этого возьмем $n \geq 1$ и рассмотрим событие, состоящее в том, что в момент времени $t+h$ система находится в состоянии E_n. Вероятность этого события равна $P_n(t+h)$, и осуществиться оно может тремя взаимоисключающими способами.

Во-первых, в момент времени t система может находиться в состоянии E_n, и между t и $t+h$ не произойдет ни одного скачка. Вероятность этой возможности равна

$$P_n(t)P_0(h) = P_n(t)\left[1 - \lambda h\right] + o(h).$$

Вторая возможность состоит в том, что в момент времени t система находится в состоянии E_{n-1} и между t и $t+h$ происходит в точности один скачок. Вероятность этого равна

$$P_{n-1}(t) \cdot \lambda h + o(h).$$

Любое другое состояние в момент t потребует более одного скачка в интервале между t и $t+h$, и вероятность подобного события есть $o(h)$.

Следовательно, мы должны иметь:

$$P_n(t+h)=P_n(t)(1-\lambda h)+P_{n-1}(t)\lambda h+o(h), \qquad (3.5.5)$$

Это соотношение можно переписать в виде:

$$\left[P_n(t+h)-P_n(t)\right]/h=-\lambda P_n(t)+\lambda P_{n-1}(t)+o(h)/h. \qquad (3.5.6)$$

При $h\to 0$ последний член стремится к нулю; следовательно, предел левой части существует и равен

$$P_n'(t)=-\lambda P_n(t)+\lambda P_{n-1}(t), \quad n\ge 1. \qquad (3.5.7)$$

При $n=0$ вторая и третья из упомянутых выше возможностей не возникают, и поэтому (3.5.5) следует заменить на

$$P_0(t+h)=P_0(t)(1-\lambda h)+o(h), \qquad (3.5.8)$$

Это приводит к

$$P_0'(t)=-\lambda P_0(t). \qquad (3.5.9)$$

Отсюда и из $P_0(0)=1$ получаем $P_0(t)=e^{-\lambda t}$. Подставляя это значение $P_0(0)$ в (3.5.7) при $n=1$, мы получим обыкновенное дифференциальное уравнение для $P_1(t)$.

Легко проверить путем подстановки в формулу (3.5.7) и проверки тождества левой и правой частей, что $P_n(t)=\dfrac{\lambda t}{n}\cdot P_{n-1}(t)$, $n=1,2,\ldots$. Поскольку $P_1(0)=P_2(0)=P_3(0)=\ldots=0$, мы легко находим, что

$$P_1(t)=\frac{\lambda t}{1!}P_0(t)=\frac{\lambda t}{1!}e^{-\lambda t},$$

$$P_2(t)=\frac{\lambda t}{2}P_1(t)=\frac{\lambda t}{2}\cdot\lambda t e^{-\lambda t}=\frac{(\lambda t)^2}{2!}e^{-\lambda t},$$

$$P_3(t)=\frac{\lambda t}{3}P_2(t)=\frac{\lambda t}{3}\cdot\frac{(\lambda t)^2}{2!}e^{-\lambda t}=\frac{(\lambda t)^3}{3!}e^{-\lambda t},$$

… ,

а это полностью согласуется с (3.5.4).

РАЗДЕЛ 3.6. Марковские процессы с дискретными состояниями и непрерывным временем

Нам будет удобно считать, что переходы системы S из состояния в состояние происходят под воздействие каких-то потоков событий, например, «поток отказов», «поток восстановлений» и т. д..

Как только произошло первое после момента t_0 событие, осуществляется переход из состояния в состояние (последующие события потока не учитываются).

Предполагаем, что переходы из состояния в состояние происходят под воздействием пуассоновских потоков событий (не обязательно стационарных).

Отсутствие последействия в пуассоновском потоке позволит при фиксированном настоящем (состояние s_i системы в момент t) не заботиться о том, когда и как система оказалась в этом состоянии.

Пусть на графе состояний системы S существует стрелка, ведущая из состояния s_i в одно из соседних состояний (рисунок 3.6.1).

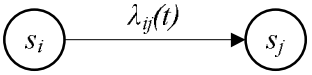


Рис. 3.6.1. Граф состояний.

Будем считать, что переход из состояния s_i в состояние s_j осуществляется под воздействием пуассоновского потока с интенсивностью $\lambda_{ij}(t)$. Переход из s_i в s_j происходит в момент, когда наступает первое событие потока.

Рассмотрим на оси $0-t$ элементарный участок времени Δt, примыкающий к t (рисунок 3.6.2), и найдем вероятность того, что за время Δt система S перейдет из состояния s_i в состояние s_j (в предположении, что в момент времени t система S находилась в состоянии s_i).

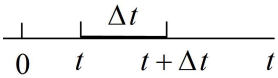


Рис. 3.6.2. Временная ось с участком времени Δt.

Эта вероятность с точностью до бесконечно малых величин высших порядков равна $\lambda_{ij}(t)\Delta t$. Случайная величина $X(t,\Delta t)$, равная числу событий потока попадающих на элементарный участок Δt, имеет математическое ожидание $\lambda_{ij}(t)\Delta t$, и с точностью до бесконечно малых высших порядков равна вероятности p_i попадания на элементарный участок одного (а, значит, хотя бы одного) события

(вероятностью попадания на участок $(t,\ t+\Delta t)$ более одного события пренебрегаем).

Итак, **вероятность перехода** системы S из состояния s_i, в котором она находилась в момент времени t, в состояние s_j за элементарный промежуток времени Δt, непосредственно примыкающий к t, приближенно равна $\lambda_{ij}(t)\Delta t$, где $\lambda_{ij}(t)$ – интенсивность пуассоновского потока событий переводящего систему из s_i в s_j.

Можно доказать, что если известны все потоки событий, переводящие систему из состояния в состояние – пуассоновские и независимые – то процесс протекающий в системе S, будет марковским.

Если известны все интенсивности пуассоновских потоков событий, переводящих систему из состояния в состояние, то можно составить дифференциальные уравнения для вероятностей состояний.

Рассмотрим систему S, имеющую n возможных состояний: $s_1, s_2, \ldots, s_i, \ldots, s_j, \ldots, s_n$. Пусть для любой пары состояний s_i, s_j известна интенсивность $\lambda_{ij}(t)$ пуассоновского потока событий переводящего систему S из любого состояния s_i в любое другое состояние $s_j\ (i \neq j)$; будем полагать эту интенсивность равной нулю, если непосредственный переход из состояния s_i в состояние s_j невозможен. Обозначим $p_i(t)$ – вероятность того, что в момент t система находится в состоянии $s_i\ (i = 1, 2, \ldots, n)$. Теперь придадим t приращение Δt и найдем вероятность $p_i(t+\Delta t)$ того, что в момент $t+\Delta t$ система будет находиться в состоянии s_i. Обозначим это событие A: $A = \{S(t+\Delta t)\}$.

Зададим вопрос, как это событие может произойти? Двумя способами: либо произойдет событие B, состоящее в том, что в момент t система уже была в состоянии s_i и за время Δt не вышла из этого состояния; либо произойдет событие C, состоящее в том, что в момент t система была в одном из соседних состояний s_j, из которых возможен переход в $s_i\ (\lambda_{ji}(t) \neq 0)$, и за время Δt перешла из состояния s_j в s_i.

Очевидно $A = B + C$. Найдем вероятности событий B и C. Согласно правилу умножения вероятностей вероятность события B равна вероятности $p_i(t)$ того, что система в момент t была в состоянии S_i, умноженной на условную вероятность того, что за время Δt она не выйдет из этого состояния, т.е. в суммарном потоке событий, выводящих систему из состояния s_i, не появится ни одного события. Так как суммарный поток событий, выводящий систему из состояния s_i, как и все его слагаемые – пуассоновский с интенсивностью, равной сумме интенсивностей слагаемых потоков:

$\sum_{j=1}^{n} \lambda_{ij}(t)$ $(i \neq j)$, то условная вероятность того, что на участке времени Δt появится хотя бы одно событие, равна (приближенно)

$\sum_{j=1}^{n} \lambda_{ij}(t)\,\Delta t$ $(i \neq j)$ (для простоты приближенные равенства, становящиеся точными при $\Delta t \to 0$, будем записывать просто как равенства, не оговаривая их приближенность), а условная вероятность противоположного события равна $1 - \sum_{j=1}^{n} \lambda_{ij}(t)\,\Delta t$. Таким образом,

$$\mathbf{P}(B) = p_i(t)\left[1 - \sum_{j=1}^{n} \lambda_{ij}(t)\,\Delta t\right]. \qquad (3.6.1)$$

Найдем теперь вероятность события C. Представим его в виде суммы несовместных вариантов:

$$C = \sum_j C_j\,, \qquad (3.6.2)$$

где суммирование распространяется на все состояния s_i, из которых возможен непосредственный переход в s_i (т. е. для которых $\lambda_{ji}(t) \neq 0$). События C_j, в силу ординарности потоков, можно считать несовместимыми. По правилу сложения вероятностей

$$\mathbf{P}(C) = \sum_j \mathbf{P}(C_j). \qquad (3.6.3)$$

По правилу умножения вероятностей
$\mathbf{P}(C_j) = p_j(t)\,\lambda_{ji}(t)\Delta t$, откуда

$$\mathbf{P}(C)=\sum_{j=1}^{n} p_j(t)\,\lambda_{ji}(t)\Delta t \quad (i\neq j), \tag{3.6.4}$$

следовательно

$$\mathbf{P}(A)=\mathbf{P}(B)+\mathbf{P}(C)=p_i(t)\left[1-\sum_{j=1}^{n}\lambda_{ij}(t)\Delta t\right]+\sum_{j} p_j(t)\,\lambda_{ji}(t)\Delta t\,.$$

Таким образом,

$$p_i(t+\Delta t)=p_i(t)\left[1-\sum_{j=1}^{n}\lambda_{ij}(t)\Delta t\right]+\sum_{j} p_j(t)\,\lambda_{ji}(t)\Delta t\,. \tag{3.6.5}$$

Вычитая из (3.6.5) $p_i(t)$, получим приращение функции на участке $(t, t+\Delta t)$:

$$p_i(t+\Delta t)-p_i(t)=\sum_{j=1}^{n} p_j(t)\lambda_{ji}(t)\Delta t-\sum_{j=1}^{n}\lambda_{ij}(t)\Delta t\,p_i(t).$$

Деля приращение функции на приращение аргумента Δt и устремляя Δt к нулю, получим в пределе производную функции $p_i(t)$:

$$\frac{dp_i(t)}{dt}=\sum_{j=1}^{n} p_j(t)\lambda_{ji}(t)-p_i(t)\sum_{j=1}^{n}\lambda_{ij}(t) \quad (i=1, 2, \ldots, n). \tag{3.6.6}$$

Первая сумма в правой части формулы (3.6.6) распространяется на те значения j, для которых возможен непосредственный переход из состояния s_j в s_i (т. е. для которых $\lambda_{ji}(t)\neq 0$), а вторая – на те значения j, для которых возможен непосредственный переход из s_i в s_j (т. е. $\lambda_{ij}(t)\neq 0$).

Таким образом, для вероятностей $p_i(t)$ мы получили систему обыкновенных дифференциальных уравнений (3.6.6) с переменными (в общем случае) коэффициентами. Эти уравнения называются уравнениями **Колмогорова**.

Систему дифференциальных уравнений (3.6.6) решают при начальных условиях, задающих вероятности состояний в начальный момент времени при $t=0$:

$$p_1(0), p_2(0), \ldots, p_n(0), \tag{3.6.7}$$

причем для любого момента времени t выполняется нормировочное условие

$$\sum_{i=1}^{n} p_i(t) = 1 \quad (t \geq 0). \tag{3.6.8}$$

Это следует из того, что в любой момент времени *t* события

$\{S(t) = s_1\}, \{S(t) = s_2\}, \ \ldots \ , \{S(t) = s_n\}$

образуют полную группу несовместных событий. Нормировочное условие (3.6.8) можно использовать вместо одного (любого) из дифференциальных уравнений (3.6.6).

При составлении системы дифференциальных уравнений (3.6.6) удобно пользоваться **размеченным графом состояний** (рисунок 3.6.3) системы, где возле каждой стрелки, ведущей из состояния s_i в состояние s_j, стоит интенсивность $\lambda_{ij}(t)$ пуассоновского потока событий, переводящего систему из состояния s_i в s_j. Если $\lambda_{ij}(t) \equiv 0$, ни стрелка, ни соответствующая интенсивность на размеченном графе не ставятся.

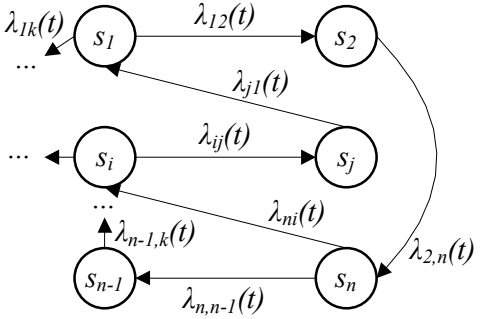


Рис. 3.6.3. Размеченный граф состояний.

При составлении уравнений Колмогорова по графу состояний удобно ввести понятие «потока вероятности». **Потоком вероятности**, переводящим систему из состояния s_i в состояние s_j, будем называть произведение вероятности $p_i(t)$ состояния s_i, из которого исходит стрелка, на интенсивность $\lambda_{ij}(t)$ потока событий переводящего систему по этой стрелке.

Уравнения Колмогорова (3.6.6) составляются по следующему правилу: производная вероятности любого состояния равна сумме потоков вероятности, переводящих систему в это состояние, минус сумма всех потоков вероятности, выводящих систему из этого состояния.

Все интенсивности $\lambda_{ij}(t)$ в уравнении (3.6.6) можно записать в виде квадратной матрицы:

$$\left\| \lambda_S(t) \right\| = \begin{Vmatrix} 0 & \lambda_{12}(t) & \dots & \lambda_{1j}(t) & \dots & \lambda_{1n}(t) \\ \lambda_{21}(t) & 0 & \dots & \lambda_{2j}(t) & \dots & \lambda_{2n}(t) \\ \dots & & & & & \\ \lambda_{i1}(t) & \lambda_{i2}(t) & \dots & \lambda_{ij}(t) & \dots & \lambda_{in}(t) \\ \dots & & & & & \\ \lambda_{n1}(t) & \lambda_{n2}(t) & \dots & \lambda_{nj}(t) & \dots & 0 \end{Vmatrix}, \qquad (3.6.9)$$

где $\lambda_{ii}(t) \equiv 0 \ \ (i = 1, 2, \dots, n)$. По главной диагонали этой матрицы размерности $n \times n$ стоят нули, а на пересечении *i*-й строки и *j*-го столбца стоит функция $\lambda_{ij}(t)$ – интенсивность пуассоновского потока событий, переводящих в систему *S* из состояния s_i в состояние s_j.

Матрицу интенсивностей (3.6.9) удобно иллюстрировать с помощью размеченного графа состояний системы *S*, на котором указываются только те ребра между состояниями s_i и s_j для которых соответствующие интенсивности не равны нулю, а около каждого ребра проставляется соответствующая интенсивность потока событий. Между матрицей интенсивностей (3.6.9) и размеченным графом состояний системы $G(\mathbf{S})$ существует однозначное соответствие.

Зная размеченный граф состояний системы $G(\mathbf{S})$ (или матрицу интенсивностей $\left\| \lambda_S(t) \right\|$), можно, воспользовавшись мнемоническим правилом, записать систему дифференциальных уравнений для вероятностей состояний системы (3.6.6).

Если все интенсивности потоков $\lambda_{ij}(t)$ не зависят от аргумента t $\left(\lambda_{ij}(t) = \lambda_{ij}\right)$, то марковский процесс называется **однородным**. Если хотя бы одна из интенсивностей в матрице (3.6.9) зависит от времени, то такой марковский процесс называется **неоднородным**. У однородного марковского процесса коэффициенты в системе дифференциальных уравнений (3.6.6) являются постоянными.

Таким образом, для исследования марковского случайного процесса нужно знать: 1) матрицу интенсивностей $\left\| \lambda_S(t) \right\|$ (или размеченный граф состояний системы $G(\mathbf{S})$) и 2) начальные условия:

$$p_1(0),\ p_2(0),\ \dots\ ,\ p_n(0), \tag{3.6.10}$$

$$\sum_{i=1}^{n} p_i(0) = 1, \quad p_i(0) \ge 0 \quad (i = 1, 2, \dots, n).. \tag{3.6.11}$$

Рассмотрим следующий пример. Размеченный граф состояний системы имеет вид, показанный на рисунке 3.6.4.

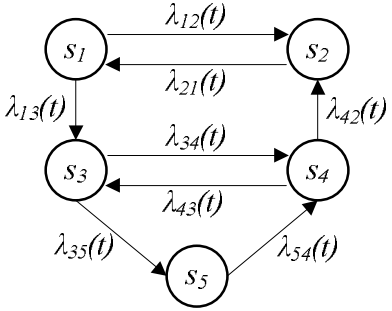


Рис. 3.6.4. Размеченный граф состояний.

Необходимо составить уравнения Колмогорова для вероятностей состояний и указать, при каких начальных условиях их нужно решать, если в начальный момент времени система S с вероятностью ½ находится в состоянии s_1 и с вероятностью ½ – в состоянии s_2.

Решение. Уравнения Колмогорова имеют вид

$$\begin{aligned}
&dp_1(t)/dt = p_2(t)\lambda_{21} - p_1(t)(\lambda_{12} + \lambda_{13}),\\
&dp_2(t)/dt = p_1(t)\lambda_{12} + p_4(t)\lambda_{42} - p_2(t)\lambda_{21},\\
&dp_3(t)/dt = p_1(t)\lambda_{13} + p_4(t)\lambda_{43} - p_3(t)(\lambda_{34} + \lambda_{35}),\\
&dp_4(t)/dt = p_3(t)\lambda_{34} + p_5(t)\lambda_{54} - p_4(t)(\lambda_{43} + \lambda_{42}),\\
&dp_5(t)/dt = p_3(t)\lambda_{35} - p_5(t)\lambda_{54}.
\end{aligned} \tag{3.6.12}$$

Любое из этих уравнений может быть отброшено, а соответствующая ему вероятность $p_i(t)$ $(i = 1, 2, 3, 4, 5)$ выражена через остальные с помощью нормировочного условия:

$$p_1(t) + p_2(t) + p_3(t) + p_4(t) + p_5(t) = 1. \tag{3.6.13}$$

Начальные условия, при которых надо будет решать систему дифференциальных уравнений, будут следующими:

$$p_1(0) = p_2(0) = 0{,}5; \quad p_3(0) = p_4(0) = p_5(0) = 0. \tag{3.6.14}$$

Уравнения (3.6.12) как при постоянных, так и переменных интенсивностях λ_{ij} (совместно с нормировочным условием (3.6.13)), можно решать на ЭВМ при начальных условиях (3.6.14) любым из перечисленных методов.

РАЗДЕЛ 3.7. Модели очередей в вычислительных системах и сетях

С целью повышения загрузки (уменьшения простоев) программных и аппаратных ресурсов вычислительных систем (ВС) современная организация вычислительного процесса предусматривает возможность создания к ним очередей. Примером могу служить очередь заданий, ожидающих распределения памяти, очереди заданий к центральному процессору и на ввод-вывод. Ожидающие того или иного вида обслуживания задания (в других случаях это могут быть запросы, сообщения, задачи процессы или программы) будем называть **заявками** (запросами или требованиями), а устройство, предназначенное для их обслуживания (например, память, центральный процессор (ЦП), устройство ввода-вывода), – **обслуживающим устройством**.

В ВС возможны очереди, в которых заявки не являются заданиями в обычном смысле этого слова. Так, например, в мультипроцессорных ВС, как правило, работать данным модулем памяти (производить считывание-запись) в каждый момент времени может только какой-нибудь один ЦП. Таким образом, если в процессе работы одного из ЦП с некоторым модулем памяти к тому возникает запрос от другого ЦП, то он должен подождать освобождения этого модуля памяти. Понятно, что в приведенном примере заявками являются запросы от ЦП, а обслуживающими устройствами – блоки памяти.

При количественном анализе очередей в ВС требуется дать ответ по крайней мере на два вопроса: насколько загружено рассматриваемое обслуживающее устройство и каково время ожидания заявок в очереди? Оба крайних случая, когда обслуживающее устройство занято мало, т.е. подолгу простаивает, и когда загрузка чрезмерно велика, вследствие чего заявки длительное время ожидают обслуживания, требуют принятия корректирующих решений в управлении вычислительным процессом. Поскольку в ВС многие ресурсы взаимосвязаны, излишняя загрузка одного из них и недостаточная загрузка другого могут привести к уменьшению пропускной способности ВС в целом.

Методы решения задач количественного анализа очередей составляют предмет одного из разделов теории вероятностей, известного под названием **теория очередей** или **теория массового обслуживания**.

Структура системы массового обслуживания.

Хотя ВС представляет собой взаимосвязанную совокупность вычислительных ресурсов, в ряде случаев основной интерес представляет задача оценки загруженности одного из этих ресурсов, например, центрального процессора, накопителя на магнитных дисках или оператора вычислительной установки (например, сетевого оператора). Эту задачу можно решать в рамках моделей систем массового обслуживания с одним обслуживающим устройством, методы исследований которых составляют наиболее развитый и завершенный раздел теории.

Основные элементы системы массового обслуживания (СМО) показаны на рисунке 3.7.1.

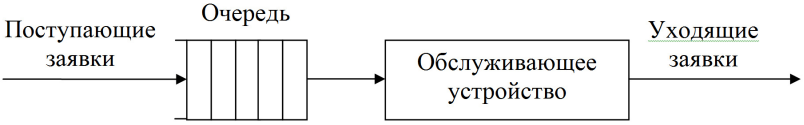


Рис. 3.7.1. Основные элементы СМО.

Обслуживаемой единицей в СМО является заявка. Заявки поступают на обслуживающее устройство. Если поступающие в СМО заявки не могут быть удовлетворены немедленно, то возникает очередь. Очередь присуща не всякой СМО. Существуют такие СМО, в которых очередь не допускается, и заявка, заставшая обслуживающее устройство занятым, теряется. В других СМО если в момент поступления заявки обслуживающее устройство занято, то заявка занимает очередь к нему, где ожидает начала обслуживания.

Выбор заявки на обслуживание в какой-то момент времени производится в соответствии с некоторым правилом, которое называется **дисциплиной обслуживания**. Далее выполняется обслуживание заявки, и после завершения обслуживания заявка покидает систему. Выходящий поток обслуженных заявок может оказаться весьма важным в тех случаях, когда он является входящим для другой СМО. Так, например, программы могут попеременно требовать обслуживания центрального процессора и процессора ввода-вывода.

О таких элементах СМО, как входящий поток заявок, механизм обслуживания и дисциплина обслуживания, можно сделать различные предположения. Остановимся на некоторых из них.

1. Входящий поток заявок.

Пусть теперь $\tau_1, \tau_2, \ldots$ – моменты поступления заявок пуассоновского потока. Тогда для любого $k \geq 1$ функция распределения принимает следующий вид:

$$F_k(t) = \mathrm{P}\{\tau_k - \tau_{k-1} \leq t\} \text{ или } F_k(t) = 1 - e^{-\lambda t}, t \geq 0 .$$

Таким образом, для пуассоновского входящего потока промежутки времени между моментами поступления заявок статистически независимы и имеют одинаковое экспоненциальное распределение.

Для пуассоновского входящего потока имеет место важное свойство **отсутствия последействия**: время ожидания поступления новой заявки не зависит от того, когда появилась предыдущая заявка. Поскольку интервалы между моментами поступления заявок имеют экспоненциальное распределение, точная формулировка этого свойства является следующей. Пусть случайная величина Z распределена по экспоненциальному закону, т.е.

$$\mathrm{P}\{Z \leq t\} = \begin{cases} 1 - e^{-\lambda t}, t \geq 0, \\ 0, t < 0. \end{cases}$$

Тогда для любого числа $a \geq 0$

$$\mathrm{P}\{Z - a \leq t \mid Z > a\} = \mathrm{P}\{Z \leq t\}. \qquad (3.7.1)$$

В общем случае входящий поток заявок определяется посредством задания для каждого $n \geq 0$ совместного распределения случайных величин $\zeta_1, \zeta_2, \ldots, \zeta_n$, где $\zeta_k = \tau_k - \tau_{k-1}$ $(k > 1, \tau_0 = 0)$, а $\tau_1, \tau_2, \ldots$ – моменты поступления заявок $(0 \leq \tau_1 \leq \tau_2 \leq \ldots)$. В том случае, когда случайные величины $\zeta_1, \zeta_2, \ldots, \zeta_n$ независимы в совокупности, для определения входящего потока достаточно задать набор одномерных функций распределения $F_k(t) = P\{\zeta_k \leq t\}, k \geq 1$. Такой входящий поток называется **потоком с ограниченным последействием**. Естественным обобщением пуассоновского потока является поток, для которого $F_k(t) = F(t), k \geq 1$. Такой поток называется **рекуррентным**.

Рассмотрим на рисунке 3.7.2 ординарный поток однородных событий.

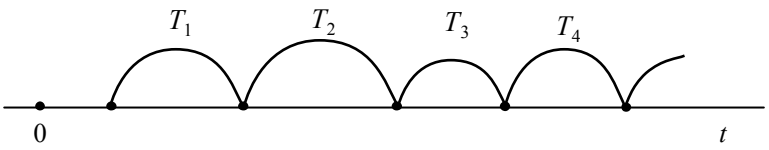


Рис. 3.7.2. Ординарный поток однородных событий.

Данный поток называется **потоком с ограниченным последействием** (или **потоком Пальма**), если промежутки времени между последовательными событиями T_1, T_2, … представляют собой независимые случайные величины.

Очевидно, простейший поток является частным случаем потока Пальма: в нем расстояния T_1, T_2, … представляют собой независимые случайные величины, распределенные по показательному закону.

Рассмотрим примеры потоков Пальма.

1. Некоторая деталь технического устройства (например, радиолампа) работает непрерывно до своего отказа (выхода из строя), после чего она мгновенно заменяется новой. Срок безотказной работы детали случаен; отдельные экземпляры выходят из строя независимо друг от друга. При этих условиях поток отказов (или поток «восстановлений») представляет собой поток Пальма. Если, к тому же, срок работы детали распределен по показательному закону, то поток Пальма превращается в простейший.

2. Колонна машин едет по автомагистрали в виде единой «колонны» с одинаковой для всех машин скоростью V.

Каждый автомобиль, кроме ведущего, обязан выдерживать заданный строй, т. е. держаться на заданном расстоянии L от впереди идущей машины. Это расстояние, вследствие погрешностей измерений, выдерживается с ошибками. Моменты пересечения авмобилями заданного рубежа образуют поток Пальма, так как случайные величины $T_1 = \frac{L_1}{V}$; $T_2 = \frac{L_2}{V}$; … независимы. Заметим, что тот же поток не будет потоком Пальма, если каждый из автомобилей стремится выдержать заданное расстояние не от соседа, а от ведущего.

Интересным примером потоков с ограниченным последействием являются так называемые **потоки Эрланга**. Они образуются «просеиванием» простейшего потока.

Рассмотрим простейший поток и выбросим из него каждую вторую точку (на рисунке 3.7.3 выброшенные точки отмечены крестами).

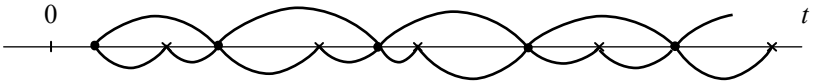


Рис. 3.7.3. Поток Эрланга первого порядка.

Оставшиеся точки образуют поток; этот поток называется потоком Эрланга первого порядка $Э_1$. Очевидно, этот поток есть поток Пальма: поскольку независимы промежутки между событиями в простейшем потоке, то, независимы и величины $T_1, T_2, \ldots$, получающиеся суммированием таких промежутков по два.

Поток Эрланга второго порядка получится, если сохранить в простейшем потоке каждую третью точку, а две промежуточные выбросить (рисунок 3.7.4).

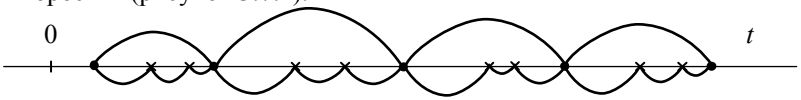


Рис. 3.7.4. Поток Эрланга второго порядка.

Таким образом, потоком Эрланга k-го порядка $Э_k$ называется поток, получаемый из простейшего, если сохранить каждую $k+1$-ю точку, а остальные выбросить. Очевидно, простейший поток можно рассматривать как поток Эрланга нулевого порядка.

2. Механизм обслуживания СМО.

Второй компонентой СМО является количественная характеристика обслуживания, требуемого отдельной заявкой. Назовем эту характеристику **длиной заявки**. Единица измерения длины заявки меняется в зависимости от природы обслуживающего устройства и заявок. Если обслуживающее устройство – центральный процессор, а заявки – программы, то длина может измеряться в командах. Если обслуживающее устройство – линия передачи данных, а заявки – передаваемые сообщения или данные, то длина может измеряться в битах или байтах. Если совокупность заявок однородна, то предполагается, что длины различных заявок являются независимыми в совокупности и одинаково распределенными случайными величинами. В более сложных ситуациях заявки можно разделить на несколько различных типов, каждый из которых составит однородную совокупность заявок.

Чтобы задать механизм обслуживания полностью, помимо распределения длин заявок необходимо также задать **быстродействие** обслуживающего устройства. Обозначим величину быстродействия через C. Единица измерения быстродействия зависит от типа обслуживания. Если обслуживающее устройство – центральный

процессор, то быстродействие измеряется в операциях в секунду. Если обслуживающее устройство – канал или линия передачи данных, то быстродействие, т.е. скорость передачи данных, измеряется в битах в секунду.

Если длина заявки равна S [единиц обслуживания] и она обслуживается устройством с быстродействием C [единиц обслуживания в секунду], то отношение S/C [секунд] называется **длительностью обслуживания заявки**. Его среднее значение $\overline{S}/C$ [секунд] называется **средней длительностью обслуживания**, а обратная к ней величина $\mu = C/\overline{S}$ называется **интенсивностью обслуживания**.

Если C постоянно, то можно не делать различия между длиной заявки и длительностью ее обслуживания, и в этом случае будем полагать, что C=1. Тем самым длина заявки измеряется в единицах времени. Это соглашение принимается всюду далее, если не оговорено противное.

Пусть Y_k – длительность обслуживания k-й заявки. Если случайные величины Y_k, $k \geq 1$ независимы в совокупности, одинаково распределены и не зависят от входящего потока, то такое обслуживание называется **рекуррентным**. В дальнейшем, как правило, будут рассматриваться СМО с рекуррентным обслуживанием.

В некоторых случаях быстродействие меняется в зависимости от загрузки обслуживающего устройства. В качестве примера рассмотрим СМО с l обслуживающими устройствами и общей очередью. Поступившая заявка обслуживается любым свободным обслуживающим устройством. Для простоты предположим, что все обслуживающие устройства имеют одинаковое быстродействие, скажем, C. Определим состояние СМО как число находящихся в ней заявок n (как на обслуживании, так и в очереди). Тогда общее быстродействие станции обслуживания, состоящей из l обслуживающих устройств, зависит от состояния n и определяется формулой $C(n) = C\min\{n, l\}$.

3. Дисциплина обслуживания.

Наиболее простой и хорошо известной является дисциплина обслуживания «**первый пришел – первый обслужен**», при которой заявки обслуживаются полностью без прерываний в порядке их поступления, причем заявка, поступившая в момент простоя обслуживающего устройства, сразу же начинает обслуживаться. Легко

представить себе ситуацию, когда эта дисциплина нежелательна. Например, часто бывает, что одни заявки важнее других и заслуживают предпочтительного обслуживания. Разделение заявок на группы по степени их важности осуществляется с помощью приоритетных дисциплин обслуживания, и соответствующая система массового обслуживания называется **системой с приоритетами**. Правило назначения приоритетов определяет порядок, в котором будут обслуживаться ожидающие заявки. Приоритетные дисциплины обслуживания бывают двух типов: с **абсолютными приоритетами** и с **относительными приоритетами**. Если обслуживание текущей заявки прерывается при появлении заявки с более высоким приоритетом и последняя немедленно начинает обслуживаться, то говорят, что имеет место дисциплина обслуживания с абсолютными приоритетами. Если прерывание обслуживания не допускается, то имеет место дисциплина с относительными приоритетами.

Далее, если не оговорено противное, рассматриваются СМО, в которых обслуживание заявок осуществляется в порядке их поступления.

Краткие обозначения. Для определения типа системы массового обслуживания часто используются обозначения вида *A/B/l*, где символы *A* и *B* обозначают входящий поток и распределение длительности обслуживания соответственно, а *l* – число параллельных устройств обслуживания в СМО. Чтобы отличить СМО, в которой нет ограничений на допустимое число заявок, от СМО, в которой не может находиться более *m* заявок, для последней используются обозначения вида *A/B/l/m*. Приведем некоторые из общепринятых обозначений для часто используемых распределений:

M – экспоненциальное распределение, которое приводит к “марковскому” свойству СМО;

D – обозначает вырожденное распределение (*deterministic*), при котором интервалы между моментами поступления или моментами начала и завершения обслуживания заявок являются постоянными;

E_k – распределение Эрланга (*Erlang*) *k*-го порядка;

H_k – гиперэкспоненциальное (*hyperexponetial*) распределение *k*-го порядка;

G – произвольное (*general*) распределение;

GI – рекуррентный входящий поток (*general independent*).

Таким образом, под системой *M/M/1* понимается СМО с одним обслуживающим прибором, пуассоновским входящим потоком и экспоненциально распределенной длительностью обслуживания. Аналогично, под системой $GI/H_2/1$ понимается СМО с одним

обслуживающим устройством, рекуррентным входящим потоком и гиперэкспоненциальным распределением второго порядка длительности обслуживания.

РАЗДЕЛ 3.8 Формула Литтла

Выведем формулу, связывающую (для предельного стационарного режима) среднее число заявок $\overline{N}$, находящихся в СМО (т.е. обслуживаемых или стоящих в очереди), и среднее время пребывания заявки в системе $\overline{T}$. Рассмотрим любую СМО (одноканальную, многоканальную, марковскую, немарковскую, с неограниченной или ограниченной очередью) и связанные с ней два потока событий: поток заявок, прибывающих в СМО, и поток заявок покидающих СМО. Если в системе установился предельный, стационарный режим, то среднее число заявок, прибывающих в СМО за единицу времени, равно среднему числу заявок, покидающих ее: оба потока имеют одну и ту же интенсивность λ.

Обозначим: *X(t)* – число заявок, прибывших в СМО до момента *t*, *Y(t)* – число заявок, покинувших СМО до момента *t*. И та, и другая функция являются случайными и меняются скачком (увеличиваются на единицу) в моменты прихода заявок *X(t)* и уходов заявок *Y(t)*. Вид функций *X(t)* и *Y(t)* показан на рисунке 3.8.1.

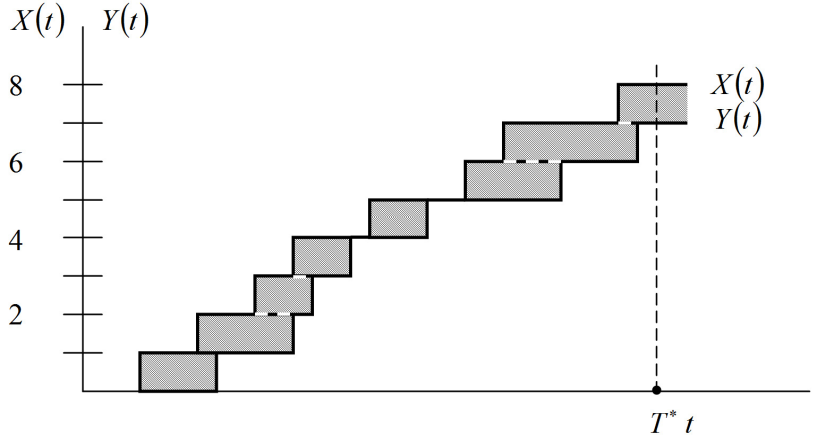


Рис. 3.8.1. Число заявок в СМО.

Обе линии – ступенчатые, верхняя – *X(t)*, нижняя – *Y(t)*. Очевидно, что для любого момента *t* их разность *Z(t) = X(t) - Y(t)* есть

не что иное, как число заявок, находящихся в СМО. Когда линии *X(t)* и *Y(t)* сливаются, в системе нет заявок.

Рассмотрим очень большой промежуток времени T^* (мысленно продолжив график далеко за пределы чертежа) и вычислим для него среднее число заявок, находящихся в СМО. Оно будет равно интегралу от функции *Z(t)* на этом промежутке, деленному на длину интервала T^*:

$$\overline{N} = \frac{1}{T^*}\int_0^{T^*} Z(t)dt \,. \qquad (3.8.1)$$

Однако этот интеграл представляет собой не что иное, как площадь фигуры, заштрихованной на рисунке. Фигура состоит из прямоугольников, каждый из которых имеет высоту, равную единице, и основание, равное времени пребывания в системе соответствующей заявки (первой, второй и т.д.). Обозначим эти времена t_1, t_2, … . Стоит отметить, что под конец промежутка T^* некоторые прямоугольники войдут в заштрихованную фигуру не полностью, а частично, но при достаточно большом T^* это не будет играть роли. Таким образом, можно считать, что

$$\int_0^{T^*} Z(t)dt = \sum_i t_i \,, \qquad (3.8.2)$$

где сумма распространяется на все заявки, пришедшие за время T^*.

Разделим правую и левую части (3.8.2) на длину интервала T^*. Получим с учетом (3.8.1)

$$\overline{N} = \frac{1}{T^*}\sum_i t_i \,. \qquad (3.8.3)$$

Разделим и умножим правую часть (3.8.3) на интенсивность λ: $\overline{N} = \frac{1}{T^*\lambda}\sum_i t_i\lambda$. Но величина $T^*\lambda$ есть ни что иное, как среднее число заявок, пришедших за время T^*. Если мы разделим сумму всех времен t_i на среднее число заявок, то получим **среднее время пребывания заявки в системе** $\overline{T}$. Итак

$\overline{N} = \lambda\overline{T}$, откуда

$$\overline{T} = \frac{1}{\lambda}\overline{N} \,. \qquad (3.8.4)$$

Это и есть **формула Литтла**: для любой СМО, при любом характере потока заявок, при любом распределении времени обслуживания, при любой дисциплине обслуживания среднее время

пребывания заявки в системе равно среднему числу заявок в системе, деленному на интенсивность потока заявок.

Точно таким же образом выводится вторая формула Литтла, связывающая **время пребывания заявки в очереди** W и **среднее число заявок в очереди** Q:

$$\overline{W} = \frac{1}{\lambda}\overline{Q}. \qquad (3.8.5)$$

Для вывода достаточно вместо нижней линии на рисунке взять функцию $U(t)$ – количество заявок, ушедших до момента t не из системы, а из очереди (если заявка, пришедшая в систему, не становится в очередь, а сразу идет под обслуживание, можно все же считать, что она становится в очередь, но находится в ней нулевое время).

РАЗДЕЛ 3.9 Модели, описываемые процессами рождения и гибели. Простейшая система M/M/1

Модели, описываемые процессами рождения и гибели.

Число заявок в системе $N(t) = A(t) - D(t)$.

Если в каждый данный момент рассматривать значение $N(t)$ как размер некоторой популяции, то $A(t)$ можно интерпретировать как общее число рождений до момента времени t, а $D(t)$ – как число погибнувших членов популяции. Отсюда процесс $N(t)$ можно назвать процессом рождения и гибели.

Ранее было получено:

$$P_n'(t) = -(\lambda_n + \mu_n)P_n(t) + \lambda_{n-1}P_{n-1}(t) + \mu_{n+1}P_{n+1}(t). \qquad (3.9.1a)$$

Эти уравнения выполняются при $n \geq 1$. При n=0 аналогичным образом выводится уравнение

$$P_0'(t) = -\lambda_0 P_0(t) + \mu_1 P_1(t). \qquad (3.9.1a)$$

Если в начальный момент времени $N(0)=i$, то должны выполняться начальные условия $P_i(0)=1$, $P_n(0)=0$, при $n \neq i$.

Мы будем искать установившееся решение системы (3.9.1), которого вполне достаточно для многих приложений. Установившееся (стационарное) решение определяется как не зависящее от t распределение вероятностей $P_0, P_1, \ldots, P_n$, удовлетворяющее системе (3.9.1). Если такое распределение существует, оно единственно и для каждого состояния n

$$\lim_{t\to\infty} P_n(t) = P_n.$$

Для нахождения P_n можно использовать систему линейных уравнений

$$\lambda_n P_n - \mu_{n+1} P_{n+1} = \lambda_{n-1} P_{n-1} - \mu_n P_n, \qquad (3.9.2)$$

которая получается из уравнений (3.9.1а), если в них $P_n^{'}(t) = 0$. Преобразуя уравнения системы (3.9.2), получим

$$\lambda_{n-1} P_{n-1} - \mu_n P_n = c, \qquad (3.9.3)$$

где c – постоянная. Из (3.9.1б) находим, что $\lambda_0 P_0 - \mu_1 P_1 = 0$.

Отсюда c=0, и из (3.9.3) получается следующая система рекуррентных уравнений:

$$\mu_n P_n = \lambda_{n-1} P_{n-1}. \qquad (3.9.4)$$

Уравнению (3.9.4) можно дать следующую интерпретацию. Его левая часть представляет собой интенсивность перехода из состояния n в состояние n-1, и эта величина балансируется правой частью, представляющей собой интенсивность перехода из состояния n-1 в состояние n. Граф переходов, отвечающий уравнениям баланса (3.9.4), изображен на рисунке 3.9.1.

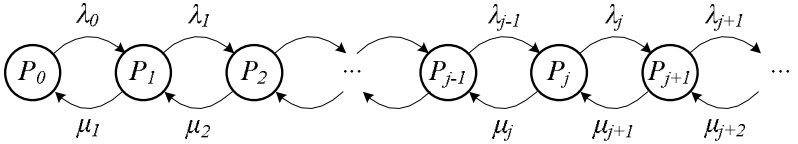


Рис. 3.9.1. Диаграмма уравнений баланса для процесса рождения и гибели.

Стационарные вероятности теперь вычисляются рекуррентно:

$$P_n = \frac{\lambda_{n-1}}{\mu_n} P_{n-1} = \frac{\Lambda(n)}{M(n)} P_0, \qquad (3.9.5)$$

где

$$\Lambda(n) = \prod_{i=1}^{n} \lambda_{i-1}, \; M(n) = \prod_{i=1}^{n} \mu_i. \qquad (3.9.6)$$

Вероятность P_0 определяется из того условия, что $\sum_{n=0}^{\infty} P_n = 1$. Таким образом, если ряд

$$1 + \sum_{n=1}^{\infty} \frac{\Lambda(n)}{M(n)} = 1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}. \qquad (3.9.7)$$

сходится, то, обозначая его сумму через θ, получим

$$P_0 = \frac{1}{\theta}. \qquad (3.9.8)$$

Система M/M/1.

Рассмотрим СМО с одним обслуживающим устройством, пуассоновским входящим потоком с параметром λ и экспоненциально распределенной с параметром μ длительностью обслуживания. Легко видеть, что число заявок *N(t)*, находящихся в системе *M/M/1* в момент времени *t*, описывается процессом рождения и гибели с $\lambda_n = \lambda$ и $\mu_n = \mu$. В этом случае рекуррентное соотношение (3.9.5) принимает вид

$$P_n = \rho P_{n-1} = \rho^n P_0,$$

где $\rho = \lambda/\mu$. Если $\rho<1$, то ряд сходится и

$$P_0 = \left(1 + \sum_{n=1}^{\infty} \rho^n\right) = 1 - \rho.$$

Таким образом, стационарная вероятность того, что в системе находится *n* заявок, равна

$$P_n = (1-\rho)\rho^n, n \geq 0. \qquad (3.9.9)$$

Стационарное распределение (3.9.9) является геометрическим распределением. Его среднее легко вычисляется:

$$\overline{N} = \sum_{n=0}^{\infty} nP_n = \frac{\rho}{(1-\rho)}. \qquad (3.9.10)$$

Среднее время ответа $\overline{T}$ можно легко вычислить из (3.9.10), используя первую из формул Литтла.

Среднее число заявок в системе $\overline{N} = \frac{\rho}{(1-\rho)}$.

Применяя первую формулу Литтла, найдем среднее время пребывания заявки в системе:

$$\overline{T} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu - \lambda}.$$

Найдем среднее число заявок в очереди $\overline{Q}$. Будем рассуждать так: число заявок в очереди равно числу заявок в системе минус число заявок, находящихся на обслуживании. Значит (по правилу сложения математических ожиданий), среднее число заявок в очереди $\overline{Q}$ равно среднему числу заявок в системе $\overline{N}$ минус среднее число заявок на обслуживании. Число заявок на обслуживании может быть либо нулем (если обслуживающее устройство свободно), либо единицей (если оно занято). Математическое ожидание такой случайной величины равно

вероятности того, что обслуживающее устройство занято (обозначим ее $P_{зан}$). Очевидно, $P_{зан}$ равно единице минус вероятность P_0 того, что обслуживающее устройство свободно:

$$P_{зан} = 1 - P_0 = \rho .$$

Следовательно, среднее число заявок, находящихся на обслуживании, равно $L_{об} = \rho$, отсюда

$$\overline{\mathrm{Q}} = \overline{N} - \rho = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho} .$$

По второй формуле Литтла найдем среднее время пребывания заявки в очереди:

$$\overline{\mathrm{W}} = \frac{\rho^2}{\lambda(1-\rho)} .$$

РАЗДЕЛ 3.10 Система M/M/m. m-канальная СМО с отказами

Имеется m каналов (линий связи), на которые поступает поток заявок с интенсивностью λ. Поток обслуживаний имеет интенсивность μ. Необходимо найти финальные вероятности состояний СМО, а также **характеристики ее эффективности**:

$A_{\text{абс}}$ – абсолютную пропускную способность, т.е. среднее число заявок, обслуживаемых в единицу времени;

$A_{\text{отн}}$ – относительную пропускную способность, т.е. среднюю долю пришедших заявок, обслуживаемых системой;

$P_{\text{отк}}$ – вероятность отказа, т.е. того, что заявка покинет СМО необслуженной;

$\overline{k}$ – среднее число занятых каналов.

Состояния системы E будем нумеровать по числу заявок, находящихся в системе (в данном случае оно совпадает с числом занятых каналов):

E_0 – в СМО нет ни одной заявки;

E_1 – в СМО находится одна заявка (один канал занят, остальные свободны);

E_k – в СМО находится k заявок (k –каналов заняты, остальные свободны);

E_m – в СМО находится m заявок (все m каналов заняты).

Граф состояний СМО соответствует схеме размножения и гибели и представлен на рисунке 3.10.1.

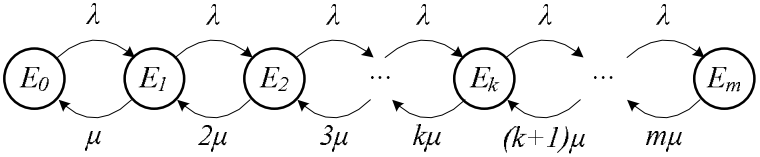


Рис. 3.10.1. Граф состояний СМО.

Из E_0 в E_1 систему переводит поток с интенсивностью λ (как только приходит заявка, система переходит из состояния E_0 в состояние E_1). Тот же поток заявок переводит систему из любого левого состояния в соседнее правое.

Поставим интенсивности у нижних стрелок. Пусть система находится в состоянии E_1 (работает один канал). Он производит μ обслуживаний в единицу времени. Проставляем у стрелки $E_1 \to E_0$ интенсивность μ. Теперь представим себе, что система находится в состоянии E_2 (работают два канала). Чтобы ей перейти в E_1, нужно, чтобы закончил обслуживание либо первый канал, либо второй; суммарная интенсивность их потоков обслуживаний равна 2μ; проставляем ее у соответствующей стрелки. Суммарный поток обслуживаний, даваемый тремя каналами, имеет интенсивность 3μ, k каналами – $k\mu$. Проставляем эти интенсивности у нижних стрелок на рисунке.

А теперь, зная все интенсивности, воспользуемся уже готовыми формулами для финальных вероятностей в схеме рождения и гибели:

$$P_0 = \left(1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2\mu^2} + \frac{\lambda^3}{2\cdot 3\mu^3} + \ldots + \frac{\lambda^k}{k!\mu^k} + \ldots + \frac{\lambda^m}{m!\mu^m}\right)^{-1}. \qquad (3.10.1)$$

Члены разложения $\frac{\lambda}{\mu}, \frac{\lambda^2}{2\mu^2}, \ldots, \frac{\lambda^m}{m!\mu^m}$ представляют собой коэффициенты при P_0 в выражениях для $P_1, P_2, \ldots, P_m$:

$$P_1 = \frac{\lambda}{\mu}P_0,\ P_2 = \frac{\lambda^2}{2\mu^2}P_0, \ldots, P_k = \frac{\lambda^k}{k!\mu^k}P_0, \ldots, P_m = \frac{\lambda^m}{m!\mu^m}P_0. \qquad (3.10.2)$$

Заметим, что в формулы (3.10.1) и (3.10.2) интенсивности λ и μ входят не по отдельности, а только в виде отношения λ/μ. Обозначим

$$\lambda/\mu = \rho \qquad (3.10.3)$$

и будем называть величину ρ "приведенной интенсивностью потока заявок". Ее смысл – среднее число заявок, приходящее за среднее время обслуживания одной заявки. Тогда

$$P_0 = \left(1 + \rho + \frac{\rho^2}{2!} + \ldots + \frac{\rho^k}{k!} + \ldots + \frac{\rho^n}{m!}\right)^{-1}, \qquad (3.10.4)$$

$$P_1 = \rho P_0,\ P_2 = \frac{\rho^2}{2} P_0, \ldots, P_k = \frac{\rho^k}{k!} P_0, \ldots, P_m = \frac{\rho^m}{m!} P_0. \qquad (3.10.5)$$

Формулы (3.10.4) и (3.10.5) для финальных вероятностей состояний называются формулами Эрланга.

Найдем $P_{\text{отк}}$ – вероятность того, что пришедшая заявка получит отказ (не будет обслужена). Для этого нужно, чтобы все m каналов были заняты, значит,

$$P_{\text{отк}} = P_m = \frac{\rho^m}{m!} P_0 = \frac{(\lambda/\mu)^m / m!}{\sum_{k=0}^{m} (\lambda/\mu)^k / k!}. \qquad (3.10.6)$$

Вероятность P_m описывает долю времени, когда все m приборов заняты. Последнее равенство называют **формулой потерь Эрланга** и обозначают $B(m, \lambda/\mu)$.

Находим **относительную пропускную способность** – вероятность того, что заявка будет обслужена:

$$A_{\text{отн}} = 1 - P_{\text{отк}} = 1 - \frac{\rho^m}{m!} P_0. \qquad (3.10.7)$$

Абсолютную пропускную способность мы получим, умножая интенсивность потока заявок λ на $A_{\text{отн}}$:

$$A_{\text{абс}} = \lambda A_{\text{отн}} = \lambda \left(1 - \frac{\rho^m}{m!} P_0\right). \qquad (3.10.8)$$

Осталось найти среднее число занятых каналов $\bar{k}$. Эту величину можно было бы найти как математическое ожидание

дискретной случайной величины с возможными случайными значениями $0,1,\ldots,m$ и вероятностями этих значений $P_0, P_1, \ldots, P_m$:

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

Подставляя сюда выражения (3.10.5) для $P_k \ (k = 0, 1, \ldots, m)$ и выполняя соответствующие преобразования, мы, в конце концов, получили бы верную формулу для $\bar{k}$. Но выведем ее гораздо проще. Нам известна абсолютная пропускная способность $A_{\text{абс}}$. Это – ни что иное, как интенсивность потока обслуженных системой заявок. Каждый занятый канал обслуживает в среднем μ заявок. Значит, среднее число занятых каналов равно

$$\bar{k} = A_{\text{абс}} \,/\, \mu , \qquad (3.10.9)$$

или учитывая (3.10.8),

$$\bar{k} = \rho \left(1 - \frac{\rho^m}{m!} P_0 \right).$$

РАЗДЕЛ 3.11. Система M/M/m с неограниченной очередью

Нумерация состояний – опять по числу заявок, находящихся в системе:

E_0 – в СМО заявок нет (все каналы свободны);

E_1 – занят один канал, остальные свободны;

E_2 – занято два канала, остальные свободны;

E_k – занято *k* каналов, остальные свободны;

E_m – заняты все *m* каналов (очереди нет);

E_{m+1} – заняты все *m* каналов, одна заявка стоит в очереди;

E_{m+r} – заняты все *m* каналов, *r* заявок стоит в очереди.

Граф состояний показан на рисунке 3.11.1.

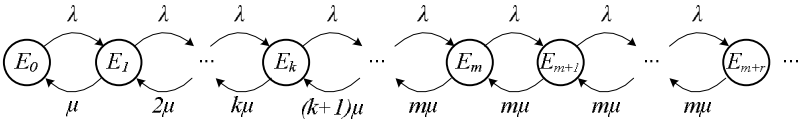


Рис. 3.11.1. Граф состояний СМО.

Это есть схема размножения и гибели, но с бесконечным числом состояний. Сообщим без доказательства естественное условие существования финальных вероятностей: $\rho/m<1$. Если $\rho/m\geq 1$, очередь растет до бесконечности.

Предположим, что условие $\rho/m<1$ выполнено, и финальные вероятности существуют. Применяя все те же формулы для схемы размножения и гибели, найдем эти финальные вероятности. В выражении для P_0 будет стоять ряд членов, содержащих факториалы, плюс сумма бесконечно убывающей геометрической прогрессии со знаменателем ρ/m. Суммируя ее найдем

$$\left.\begin{aligned} P_0 &= \left(1+\frac{\rho}{1!}+\frac{\rho^2}{2!}+\ldots+\frac{\rho^n}{m!}+\frac{\rho^{n+1}}{m!(m-\rho)}\right)^{-1} \\ P_1 &= \frac{\rho}{1!}P_0,\ \ldots,\ P_k=\frac{\rho^k}{k!}P_0,\ \ldots,\ P_n=\frac{\rho^m}{m!}P_0, \\ P_{m+1} &= \frac{\rho^{m+1}}{m\cdot m!}P_0,\ \ldots,\ P_{m+r}=\frac{\rho^{m+r}}{m^r\cdot m!}P_0,\ \ldots \end{aligned}\right\}.$$

Теперь найдем характеристики эффективности СМО. Из них легче всего находится среднее число занятых каналов $\overline{k}=\lambda/\mu=\rho$ (это справедливо для любой СМО с неограниченной очередью). Найдем среднее число заявок в системе $\overline{N}$ и среднее число заявок в очереди $\overline{Q}$. Из них легче вычислить второе, по формуле $\overline{Q}=\sum_{r=1}^{\infty} r\,P_{n+r}$; выполняя преобразования с дифференцированием ряда, получим:

$$\overline{Q}=\frac{\rho^{m+1}P_0}{m\cdot m!\,(1-\rho/m)^2}.$$

В самом деле

$$\overline{Q}=\sum_{r=1}^{\infty} r\,P_{m+r}=1\cdot P_{m+1}+2\cdot P_{m+2}+3\cdot P_{m+3}+\ldots=$$

$$=1\cdot\frac{\rho^{m+1}}{m\cdot m!}P_0+2\cdot\frac{\rho^{m+2}}{m^2\cdot m!}P_0+3\cdot\frac{\rho^{m+3}}{m^3\cdot m!}P_0+4\cdot\frac{\rho^{m+4}}{m^4\cdot m!}P_0+\ldots=$$

$$=\frac{\rho^{m+1}}{m\cdot m!}P_0\left(1+2\cdot\frac{\rho}{m}+3\cdot\frac{\rho^2}{m^2}+4\cdot\frac{\rho^3}{m^3}\ldots\right).$$

Введем обозначение $\rho/m=x<1$, тогда

$$\overline{Q} = \frac{\rho^{m+1}}{m \cdot m!} P_0 \left(1 + 2 \cdot x + 3 \cdot x^2 + 4 \cdot x^3 \ldots\right) =$$

$$= \frac{\rho^{m+1}}{m \cdot m!} P_0 \frac{d}{dx}\left(1 + x + x^2 + x^3 + x^4 \ldots\right) =$$

$$= \frac{\rho^{m+1}}{m \cdot m!} P_0 \frac{d}{dx}\left(\frac{1}{1-x}\right) = \frac{\rho^{m+1}}{m \cdot m!} P_0 \frac{1}{(1-x)^2} = \frac{\rho^{m+1}}{m \cdot m!} P_0 \frac{1}{(1-\rho/m)^2}.$$

Прибавляя к нему среднее число заявок под обслуживанием (оно же – среднее число занятых каналов) $\overline{k} = \rho$, получим:

$$\overline{N} = \overline{Q} + \rho = \frac{\rho^{m+1} P_0}{m \cdot m! (1-\rho/m)^2} + \rho .$$

Деля выражения для $\overline{Q}$ и $\overline{N}$ на λ, по формуле Литтла получим средние времена пребывания заявки в очереди и системе:

$$\overline{W} = \frac{1}{\lambda}\overline{Q}\ ,\ \overline{T} = \frac{1}{\lambda}\overline{N}\ .$$

РАЗДЕЛ 3.12. Система M/M/1/K: конечный накопитель

Рассмотрим теперь СМО, для которой фиксировано максимальное число ожидающих заявок (или требований); в частности, предположим, что в системе могут находиться самое большее K требований (включая и то требование, которое находится на обслуживании) и что любое поступившее сверх этого числа требование получает отказ и немедленно покидает систему без обслуживания. Поступление новых требований происходит по закону Пуассона, но в систему допускаются только те, которые застают в ней строго меньше, чем K требований. В телефонии требования, получившие отказ, считаются потерянными; систему с $K = 1$ (в которой места для ожидания отсутствуют совсем) называют системой с удалением заблокированных вызовов и одним обслуживающим прибором.

Будем перекрывать входящий пуассоновский поток на время, когда система заполняется, следующим образом:

$$\lambda_k = \begin{cases} \lambda, & k < K; \\ 0, & k \geq K; \end{cases}$$

$$\mu_k = \mu, \quad k = 1, 2, \ldots, K.$$

Диаграмма интенсивностей переходов для рассматриваемой конечной цепи Маркова показана на рисунке 3.12.1.

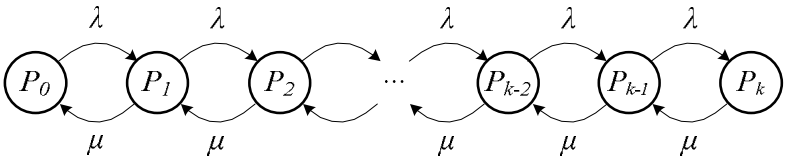


Рис. 3.12.1. Диаграмма интенсивностей переходов.

Как и ранее,

$$P_k = P_0\left(\frac{\lambda}{\mu}\right)^k = P_0\,\rho^k, \quad k \le K\ .$$

Кроме того, $P_k = 0, \quad k > K$.

Используя условие нормировки, находим значение P_0:

$$P_0 = \left[1+\sum_{k=1}^{K}\rho^k\right]^{-1} = \left[1+\frac{\rho\left(1-\rho^K\right)}{1-\rho}\right]^{-1},$$

и, следовательно, $P_0 = \dfrac{1-\rho}{1-\rho^{K+1}}$.

Таким образом, окончательно имеем

$$P_k = \begin{cases} \dfrac{1-\rho}{1-\rho^{K+1}}\rho^k, & 0 \le k \le K; \\ 0 \quad \text{в остальных случаях.} \end{cases}$$

Для системы с удалением заблокированных вызовов $\left(K=1\right)$ имеем

$$P_k = \begin{cases} \dfrac{1}{1+\rho}, & k = 0, \\ \dfrac{\rho}{1+\rho}, & k = 1 = K, \\ \\ 0 \quad \text{в остальных случаях.} \end{cases}$$

РАЗДЕЛ 3.13. Марковские сети массового обслуживания

До сих пор рассматривались марковские системы, в которых каждое требование проходило одну операцию обслуживания. Такие

системы можно называть **однофазными**. Далее рассматриваются системы с **многофазным** обслуживанием, в которых требование получает обслуживание более чем в одном приборе (узле). Таким образом, можно говорить о сети узлов, каждый из которых представляет собой СМО (некоторые узлы могут иметь несколько обслуживающих приборов) с накопителем для образования очереди. Требования поступают в систему в различных точках, ждут в очередях обслуживания и, покинув один узел, поступают в другой для дальнейшего обслуживания.

При исследовании сетей возникает много новых аспектов. Например, важной становится топологическая структура сети, так как она определяет возможные переходы между узлами. Требуется также каким-нибудь способом описать пути отдельных требований. Большое значение имеет описание природы вероятностных потоков с помощью основных вероятностных процессов; например, в случае последовательной цепочки СМО, при которой требование, покидающее *i*-й узел, сразу поступает на обслуживание в $(i+1)$-й узел; очевидно, что промежутки времени между последовательными уходами требований из предыдущего узла равны промежуткам времени между последовательными поступлениями в следующий узел.

Рассмотрим простейшую последовательную систему с двумя узлами, показанную на рисунке 3.13.1.

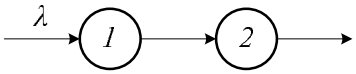


Рис. 3.13.1. Последовательная система с двумя узлами.

Каждый овал на этом рисунке изображает СМО, состоящую из очереди и обслуживающего прибора; внутри каждого овала указан номер узла. Предположим, что входящий поток является пуассоновским с интенсивностью λ, причем каждое требование поступает сначала в узел 1; предположим также, что этот узел содержит единственный обслуживающий прибор, время обслуживания которого распределено по показательному закону с интенсивностью μ. Таким образом, узел 1 представляет собой в точности СМО типа *M/M/1*. Предположим далее, что узел 2 также состоит из единственного обслуживающего прибора с показательным временем обслуживания с интенсивностью μ. Основная задача состоит в вычислении распределения промежутков времени между последовательными требованиями, поступающими в узел 2; это

эквивалентно задаче вычисления распределения промежутков времени между последовательными требованиями, уходящими из узла 1.

Пусть $d(t)$ означает плотность распределения вероятностей промежутков между последовательными требованиями на выходе узла 1, а $D^*(s)$ пусть означает ее преобразование Лапласа. Перейдем к вычислению $D^*(s)$ в момент, когда требование покидает узел 1. Возможно одно из двух событий: либо в очереди имеется второе требование, готовое немедленно поступить на обслуживание в узел 1, либо требования нет (накопитель пуст). В первом случае промежуток времени, через которое это следующее требование покинет узел 1, распределен точно так же, как и время обслуживания $b(t)$, и в этом случае получаем

$$D^*(s)\Big|_{\text{узел 1 не пуст}} = B^*(s) .$$

С другой стороны, если при уходе рассматриваемого первого требования узел оказывается пустым, то приходится ожидать в течение двух промежутков времени: первый промежуток – время до поступления следующего требования и второй – время обслуживания этого требования. Так как эти два промежутка распределены независимо, то плотность распределения вероятностей их суммы равна свертке плотностей распределения суммируемых величин. Соответственно, преобразование Лапласа плотности распределения суммы равно произведению преобразований исходных плотностей распределения, и следовательно,

$$D^*(s)\Big|_{\text{узел 1 пуст}} = \frac{\lambda}{\lambda + s} B^*(s),$$

где для преобразования Лапласа плотности распределения промежутков между поступающими требованиями уже известно явное выражение. Так как время обслуживания является показательно распределенной случайной величиной, то можно записать $B^*(s) = \mu/(s+\mu)$; кроме того, как было показано ранее, вероятность того, что требование покинет систему пустой равна $1-\rho$. Это позволяет записать следующее безусловное преобразование Лапласа для плотности распределения промежутков времени между уходящими требованиями:

$$D^*(s) = (1-\rho)D^*(s)\Big|_{\text{узел 1 пуст}} + \rho D^*(s)\Big|_{\text{узел 1 не пуст}} .$$

Используя проделанные ранее вычисления, получаем

$$D^*(s)=(1-\rho)\left(\frac{\lambda}{s+\lambda}\right)\left(\frac{\mu}{s+\mu}\right)+\rho\left(\frac{\mu}{s+\mu}\right).$$

Простые алгебраические преобразования дают

$$D^*(s)=\frac{\lambda}{s+\lambda}$$

и, следовательно, распределение промежутков времени между уходящими требованиями

$$D(t)=1-e^{-\lambda t},\ t\geq 0.$$

Итак, $D^*(s)=(1-\rho)\left(\frac{\lambda}{s+\lambda}\right)\left(\frac{\mu}{s+\mu}\right)+\rho\left(\frac{\mu}{s+\mu}\right)=$

$$=\frac{\mu}{\mu+s}\left[\frac{\lambda(1-\rho)}{s+\lambda}+\rho\right]=\frac{\mu}{\mu+s}\left(\frac{\lambda-\lambda\rho+\lambda\rho+\rho s}{s+\lambda}\right)=$$

$$=\frac{\mu}{\mu+s}\cdot\frac{\lambda+\rho s}{\lambda+s}.$$

Делая подстановку $\mu=\lambda/\rho$, получаем

$$D^*(s)=\frac{\mu}{\mu+s}\cdot\frac{\lambda+\rho s}{\lambda+s}=\frac{\lambda}{\rho\left(\frac{\lambda}{\rho}+s\right)}\cdot\frac{\lambda+\rho s}{\lambda+s}=$$

$$=\frac{\lambda}{\lambda+\rho s}\cdot\frac{\lambda+\rho s}{\lambda+s}=\frac{\lambda}{\lambda+s}.$$

Таким образом, приходим к замечательному **выводу** о том, что промежутки времени между уходящими требованиями, так же как и промежутки времени между поступающими требованиями, распределены по показательному закону с тем же самым параметром. Иначе говоря, в случае стационарной СМО входящий поток, протекая через обслуживающий прибор с показательным распределением времени обслуживания, порождает выходящий пуассоновский поток. Этот основополагающий результат обычно называют **теоремой Берке**.

Фактически теорема Берке значит больше: что исходящий поток стационарной СМО типа *M/M/m* с пуассоновским входящим потоком с параметром λ и показательным распределением времени обслуживания с параметром μ в каждом из m приборов является пуассоновским с тем же параметром λ.

Доказано также, что система *M/M/m* является единственной системой с обслуживанием в порядке поступления, обладающей таким свойством. Возвращаясь к рис. 1, видим, что в узел 2 поступает

независимый пуассоновский поток, и, следовательно, этот узел также представляет собой систему *M/M/1*, что позволяет рассматривать его независимо от узла 1. Тем самым теорема Берке говорит о том, можно соединить последовательно несколько узлов, состоящих из многолинейных СМО (с показательным распределением времени обслуживания для каждого прибора), и при этом будет сохраняться описанное свойство разложения на отдельные узлы.

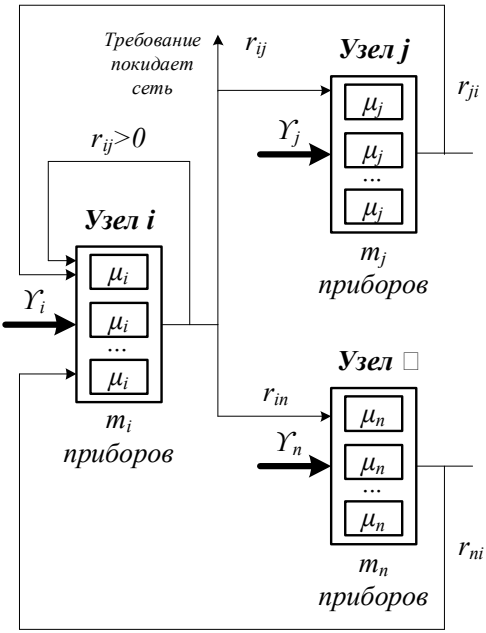


Рис. 3.13.2. Последовательная система с двумя узлами.

Этот вопрос для произвольной сети массового обслуживания исследовал **Джексон**. Он рассматривал сеть (рисунок 3.13.2), содержащую N узлов, причем каждый i-й узел состоит из m_i обслуживающих приборов с показательным временем обслуживания с параметром μ_i; в каждый i-й узел извне поступает пуассоновский поток требований с интенсивностью γ_i. Таким образом, при $N=1$ получаем обычную систему *M/M/m*. Покидая i-й узел, требование с вероятностью r_{ij} поступает в j-й узел, причем в формулировке задачи допускается, что $r_{ii} \geq 0$. С другой стороны, вероятность того, что

после обслуживания в i-м узле требование покинет сеть (и никогда не вернется обратно), равна $1-\sum_{j=1}^{N} r_{ij}$.

Необходимо вычислить полную интенсивность потока требований в заданный узел. Для этого нужно просуммировать (пуассоновские) потоки, поступающие извне, и потоки требований, поступающие от других узлов сети. Обозначая через λ_i полную интенсивность потока, входящего в i-й узел, легко показать, что это множество параметров должно удовлетворять следующей системе уравнений:

$$\lambda_i = \gamma_i + \sum_{j=1}^{N} \lambda_j r_{ji},\ i = 1, 2, \ldots, N .$$

Для того, чтобы все узлы сети описывались эргодическими цепями Маркова, для всех i должно выполняться требование $\lambda_i < m_i \mu_i$. Еще раз обратим внимание, что понятие узла, используемое в данном пособии, и понятие состояния системы, которое ранее изображалось узлом графа, должны быть дифференцированы. Джексон доказал, что каждый узел ведет себя в сети так, как если бы он был независимой системой *M/M/m* с входящим пуассоновским потоком с параметром λ_i. В общем случае полный входящий поток не является пуассоновским.

РАЗДЕЛ 3.14. Система M/G/1

Рассмотрим систему массового обслуживания с одной очередью, в которой требования поступают в соответствии с пуассоновским процессом с интенсивностью λ, но длительности обслуживания требований имеют произвольное распределение – не обязательно экспоненциальное, как в системе *M*/*M*/1. Предположим, что требования обслуживаются в порядке поступления, X_i – длительность обслуживания i-го требования, случайные величины $(X_1, X_2, \ldots)$ одинаково распределены, взаимно независимы и не зависят от интервалов поступления.

Пусть

$\overline{X} = E\{X\} = \frac{1}{\mu}$ – средняя длительность обслуживания,

$\overline{X^2} = E\{X^2\} = \dfrac{1}{\mu^2}$ – второй момент длительности обслуживания.

Наша цель состоит в том, чтобы получить и понять **формулу Поллачека-Хинчина**:

$$\overline{W} = \frac{\lambda \overline{X^2}}{2(1-\rho)}, \qquad (3.14.1)$$

где $\overline{W}$ – математическое ожидание времени пребывания требования в очереди, а $\rho = \lambda/\mu = \lambda \overline{X}$. Согласно (3.14.1), общее время пребывания в очереди и в обслуживающем приборе равно

$$\overline{T} = \overline{X} + \frac{\lambda \overline{X^2}}{2(1-\rho)}. \qquad (3.14.2)$$

Применяя формулу Литтла для $\overline{W}$ и $\overline{T}$, получим математическое ожидание числа требований в очереди $\overline{N}_Q$ и математическое ожидание числа требований в системе:

$$\overline{N}_Q = \frac{\lambda^2 \overline{X^2}}{2(1-\rho)}, \qquad (3.14.2)$$

$$\overline{N} = \rho + \frac{\lambda^2 \overline{X^2}}{2(1-\rho)}. \qquad (3.14.2)$$

Например, если длительности обслуживания распределены экспоненциально, как в системе $M/M/1$, то имеем $\overline{X^2} = 2/\mu^2$ и равенство (3.14.1) сводится к формуле

$$\overline{W} = \frac{\rho}{\mu(1-\rho)}.$$

Если длительности обслуживания одинаковы для всех требований (система $M/D/1$, где D означает детерминированность), имеем $\overline{X^2} = 1/\mu^2$ и

$$\overline{W} = \frac{\rho}{2\mu(1-\rho)}. \qquad (3.14.5)$$

Так как в случае $M/D/1$ при данном μ получается минимально возможное значение $\overline{X^2}$, из этого следует, что при одинаковых

значениях λ и μ величины $\overline{W}$, $\overline{T}$, $\overline{N}_Q$ и $\overline{N}$ для системы массового обслуживания $M/D/1$ являются нижними границами для соответствующих величин в системе $M/G/1$. Интересно заметить, что $\overline{W}$ и $\overline{N}_Q$ для системы $M/D/1$ равны половине их значений в системе $M/M/1$. Вместе с тем значения $\overline{T}$ и $\overline{N}$ при малых ρ для $M/D/1$ такие же, как в системе $M/M/1$, и приближаются к половине их значений в системе $M/M/1$ по мере того, как ρ приближается к 1. Это происходит от того, что математическое ожидание длительности обслуживания одно и то же в обоих случаях и при малых ρ бо́льшую часть времени пребывания в системе требования находятся в обслуживающем приборе, а при больших ρ бо́льшую часть времени требования стоят в очереди.

РАЗДЕЛ 3.15. Системы массового обслуживания с приоритетами

Рассмотрим систему $M/G/1$ с тем отличием, что поступающие требования делятся на n классов различного приоритета, причем 1-й класс имеет наивысший приоритет, 2-й класс – второй по величине приоритет и т. д. Скорость поступления и первые два момента времени обслуживания для класса k обозначаются соответственно через λ_k, $\overline{X}_k = 1/\mu_k$ и $\overline{X}_k^2$. Предполагается, что процессы поступления для всех классов независимые, пуассоновские и не зависят от длительностей обслуживания.

Приоритет без прерывания обслуживания.

Сначала рассмотрим правило приоритета без прерывания обслуживания, согласно которому требованию, находящемуся на обслуживании, позволяется завершить обслуживание без прерывания, даже если во время его обслуживания поступает требование более высокого приоритета. Отдельные очереди формируются для каждого приоритета. В момент, когда освобождается обслуживающий прибор, первое требование из непустой очереди наивысшего приоритета поступает на обслуживание. Это правило приоритета является одним из самых подходящих для описания систем пакетной передачи.

Мы получим формулы для средней задержки требований каждого приоритета, которые аналогичны формуле Поллачека-Хинчина и допускают аналогичное доказательство. Введем обозначения:

N_Q^k – среднее число требований в очереди k-го приоритета;

W_k – среднее время ожидания в очереди для требования k-го приоритета;

$\rho_k = \lambda_k / \mu_k$ – коэффициент использования системы для k-го приоритета;

R – среднее остаточное время обслуживания.

Предположим, что общий коэффициент использования системы меньше единицы, т. е. $\rho_1 + \rho_2 + \dots \rho_n < 1$.

Если это предположение не выполняется, появляется некоторый класс приоритета k, такой, что средняя задержка требований приоритета k и более низких будет бесконечной, тогда как средняя задержка требований с более высокими приоритетами будет конечной.

Как и в приведенном ранее выводе формулы Поллачека-Хинчина для наивысшего приоритета, имеем

$$W_1 = R + \frac{1}{\mu_1} N_Q^1 .$$

Исключая N_Q^1 с помощью теоремы Литтла, получаем

$N_Q^1 = \lambda_1 W_1$, а также $W_1 = R + \rho_1 W_1$ и, наконец,

$$W_1 = \frac{R}{1-\rho_1} . \qquad (3.15.1)$$

Для второго приоритета имеем аналогично выражение для времени ожидания в очереди W_2 за исключением того, что необходимо учесть дополнительную задержку, возникающую из-за требований более высокого приоритета, которые поступают в то время, когда требование ожидает в очереди. Это учитывает последний член в формуле

$$W_2 = R + \frac{1}{\mu_1} N_Q^1 + \frac{1}{\mu_2} N_Q^2 + \frac{1}{\mu_1} \lambda_1 W_2 .$$

Согласно теореме Литтла ($N_Q^k = \lambda_k W_k$), получаем

$W_2 = R + \rho_1 W_1 + \rho_2 W_2 + \rho_1 W_2$,

откуда следует, что

$$W_2 = \frac{R + \rho_1 W_1}{1 - \rho_1 - \rho_2} .$$

Используя ранее полученное выражение для $W_1 = R/(1-\rho_1)$, окончательно имеем

$$W_2 = \frac{R}{(1-\rho_1)(1-\rho_1-\rho_2)}.$$

Для всех приоритетов $k>1$ эта формула получается аналогично и среднее время ожидания в очереди равно

$$W_k = \frac{R}{(1-\rho_1-\ldots-\rho_{k-1})(1-\rho_1-\ldots-\rho_k)}. \qquad (3.15.2)$$

Выражение для средней задержки требования k-го приоритета имеет вид

$$T_k = \frac{1}{\mu_k} + W_k. \qquad (3.15.3)$$

Теперь нужно найти среднее остаточное время обслуживания R. Как и ранее, при выводе формулы Поллачека-Хинчина имеем

$$R = \frac{1}{2}\left(\sum_{i=1}^{n}\lambda_i\right)\overline{X^2}, \qquad (3.15.4)$$

где $\overline{X^2}$ обозначает второй момент времени обслуживания, усредненный по всем классам приоритета:

$$\overline{X^2} = \frac{\lambda_1}{\sum_{i=1}^{n}\lambda_i}\overline{X_1^2} + \ldots + \frac{\lambda_n}{\sum_{i=1}^{n}\lambda_i}\overline{X_n^2}.$$

Разные интенсивности λ_i – это числа событий (считаются события, когда мы фиксируем остаточное время) разных приоритетов пуассоновского процесса за время t (интенсивности потоков разных приоритетов неодинаковы). Величина $\sum_{i=1}^{n}\lambda_i$ есть общее число событий. Тогда $\lambda_i \Big/ \sum_{i=1}^{n}\lambda_i$ есть относительное число остаточных времен i-го приоритета.

Подставляя это в равенство (3.15.4), находим

$$R = \frac{1}{2}\sum_{i=1}^{n}\lambda_i\overline{X_i^2}. \qquad (3.15.5)$$

Среднее время ожидания в очереди и средняя задержка требования для каждого приоритета получается из равенств (3.15.2), (3.15.3) и (3.15.5)

$$W_k = \frac{\sum_{i=1}^{n} \lambda_i \overline{X_i^2}}{2\left(1-\rho_1-\ldots-\rho_{k-1}\right)\left(1-\rho_1-\ldots-\rho_k\right)}. \qquad (3.15.6)$$

Заметим, что имеется возможность оказывать влияние на среднюю задержку требования путем соответствующего выбора приоритета. Как правило, средняя задержка уменьшается, если требования с малым временем обслуживания получают более высокий приоритет. Примером из повседневной жизни является работа магазина самообслуживания, в котором имеется специальная касса для покупателей с небольшим числом покупок. Аналогичную ситуацию можно наблюдать в очередях на копировальную машину, где часто получают приоритет по отношению к другим те люди, которым необходимо сделать только небольшое число копий.

Аналитическое обоснование этого можно получить путем рассмотрения системы без прерывания с двумя классами требований A и B и соответствующими скоростями поступления и скоростями обслуживания λ_A, μ_A и λ_B, μ_B. Непосредственное вычисление по приведенным выше формулам показывает, что если $\mu_A > \mu_B$ (т. е. требование A меньшей длины), то средняя задержка требования (усредненная по обоим классам), равная

$$T = \frac{\lambda_A T_A + \lambda_B T_B}{\lambda_A + \lambda_B},$$

будет меньше в случае, когда A имеет приоритет над B (чем в случае, когда B имеет приоритет над A).

Приведенный анализ нельзя распространить на случай системы с несколькими обслуживающими приборами в первую очередь потому, что нет простой формулы для среднего остаточного времени.

Приоритеты с прерыванием и дообслуживанием.

Одной из характерных особенностей приоритета без прерывания является то, что средняя задержка данного класса приоритета зависит от скорости поступления в классах более низкого приоритета. Это ясно из равенства (3.15.6) и следует из того, что требования более высоких приоритетов должны ждать завершения обслуживания требований более низких приоритетов. Эта зависимость отсутствует в дисциплине приоритета с прерыванием и дообслуживанием, в которой обслуживание требования прерывается,

когда поступает требование более высокого приоритета, и возобновляется от момента прерывания сразу же, после того как обслужены все требования более высокого приоритета.

Следует помнить, что меньшие значения k соответствуют более высоким значениям приоритета.

При вычислении T_k – среднего времени пребывания в системе требований k-го приоритета, будем иметь в виду, что присутствие требований с приоритетами от $k+1$ до n не влияет на это вычисление. Следовательно, можно рассматривать каждый класс приоритета в системе как самый низкий.

Среднее время пребывания в системе T_k состоит из трех частей.

Первая часть – среднее время обслуживания требования, равное $1/\mu_k$.

Вторая часть – среднее время, необходимое для обслуживания требований с приоритетом от 1 до k, которые уже находятся в системе в момент поступления данного требования с приоритетом k с учетом средней продолжительности неоконченной работы, соответствующая приоритетам от 1 до k. Этот член равен среднему времени ожидания в соответствующей обычной системе $M/G/1$ (без приоритетов), в которой требования с приоритетами от $k+1$ до n не учитываются, т. е.

$$\frac{R_k}{1-\rho_1-\ldots-\rho_k},$$

где R_k – среднее остаточное время:

$$R_k = \frac{\sum_{i=1}^{k}\lambda_i X_i^2}{2}. \qquad (3.15.7)$$

Это справедливо, так как в любой момент времени неоконченная работа (сумма оставшихся длительностей обслуживания всех требований в системе) в системе $M/G/1$ не зависит от дисциплины приоритета.

Третья часть в выражении для T_k равна среднему времени ожидания для требований с приоритетами от 1 до $k-1$, которые поступили во время пребывания в системе данного требования с приоритетом k. Этот член равен

$$\sum_{i=1}^{k-1}\frac{1}{\mu_i}\lambda_i T_k = \sum_{i=1}^{k-1}\rho_i T_k$$

при $k>1$ и 0 при $k=1$ (здесь $\frac{1}{\mu_i}$ – средняя длительность обслуживания *i*-го требования; $N_i=\lambda_i T_i=\lambda_i T_k$ – среднее число требований более высокого приоритета *i*, которые поступили во время пребывания в системе данного требования с приоритетом k).

Используя эти части, получим равенство:

$$T_k=\frac{1}{\mu_k}+\frac{R_k}{1-\rho_1-\ldots-\rho_k}+\left(\sum_{i=1}^{k-1}\rho_i\right)T_k. \qquad (3.15.8)$$

Окончательный результат

$$T_1=\frac{\left(1/\mu_1\right)\left(1-\rho_1\right)+R_1}{1-\rho_1} \qquad (3.15.9)$$

при $k=1$ и

$$T_k=\frac{\left(1/\mu_k\right)\left(1-\rho_1-\ldots-\rho_k\right)+R_k}{\left(1-\rho_1-\ldots-\rho_{k-1}\right)\left(1-\rho_1-\ldots-\rho_k\right)} \qquad (3.15.10)$$

при $k>1$, где R_k дается формулой (3.15.7). Так же, как и для системы без прерываний обслуживания, эту формулу непросто обобщить на случай нескольких обслуживающих приборов за исключением случая, когда длительности обслуживания имеют одинаковое экспоненциальное распределение.

РАЗДЕЛ 3.16. Вопросы для самопроверки

1. Какие объекты анализируются в моделях макроуровня?
2. Дайте понятие математического ожидания дискретной случайной величины.
3. Поясните смысл правила трех сигм.
4. Запишите неравенство Чебышева.
5. Дайте определение цепям Маркова.
6. Как рассчитывается вероятность перехода в цепях Маркова?
7. Запишите и поясните уравнение Колмогорова-Чепмена.
8. Каковы основные особенности пуассоновского процесса?
9. В чем заключается отличие однородного марковского процесса от неоднородного?
10. Перечислите основные элементы систем массового обслуживания.
11. Какие обозначения используются для определения типа системы массового обслуживания?

12. Что показывает формула Литтла? Запишите ее.

13. Что представляет собой СМО M/M/1?

14. Выведите формула расчета среднего числа заявок в системе M/M/1.

15. Перечислите характеристики эффективности СМО M/M/m.

16. Запишите выражения для абсолютной и относительной пропускных способностей системы M/M/m.

17. Опишите особенности системы M/M/m с неограниченной очередью.

18. В чем состоит отличие однофазной и многофазной СМО?

19. Сформулируйте теорему Берке.

20. Объясните сущность систем массового обслуживания с приоритетами.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Вентцель Е.С. Исследование операция. – М.:Сов. Радио, 1972. – 552 с.

2. Герасимов А.И. Теория и практическое приложение стохастических цепей. – М.:Радио и связь, 1994.

3. Клейнрок Л. Вычислительные системы с очередями. – М.: Мир, 1979. – 600 с.

4. Корячко В.П., Курейчик В.М., Норенков И.П. Теоретические основы САПР. – М.:Энергоатомиздат, 1987. – 400 с.

5. Норенков И.П. Введение в автоматизированное проектирование технических устройств и систем. – М.: Высшая школа, 1986. – 304 с.

6. Норенков И.П. Основы автоматизированного проектирования: учеб. для вузов. 2-е изд., перераб. и доп. – М.: Изд-во МГТУ им. Н. Э. Баумана, 2002. – 336 с.: ил. - (Сер. Информатика в техническом университете).

7. САПР: в 9-ти книгах. Книга 4. Трудоношин В.А., Пивоваров Н.В. САПР: математические модели технических объектов. Ред. Норенков И.П. – М.: Высшая школа, 1986. – 158 с.

8. Советов Б.Я., Яковлев С.А. Моделирование систем: учебник для вузов – 3-е изд., перераб и доп. – М.: Высш. школа, 2001. – 343 с.

А.Н. САПРЫКИН

МОДЕЛИ И МЕТОДЫ АНАЛИЗА ПРОЕКТНЫХ РЕШЕНИЙ

Учебное пособие

Подписано в печать 04.10.21. Формат бумаги 60x84 1/16.
Бумага офсетная. Печать струйная. Усл. печ. л. 6,5.
Тираж 100 экз. Заказ № 4594.

Издательство ИП Коняхин А.В. (BookJet)

Отпечатано в типографии BookJet
390005, г. Рязань, ул. Пушкина, д. 18
Сайт: http://bookjet.ru
Почта: info@ bookjet.ru
тел.:+7 (4912) 466-151