ВОЕННО-КОСМИЧЕСКАЯ АКАДЕМИЯ
имени А.Ф. Можайского

Ю.И. Рыжиков

АЛГОРИТМИЧЕСКИЙ ПОДХОД К ЗАДАЧАМ МАССОВОГО ОБСЛУЖИВАНИЯ

Монография

Санкт-Петербург
2013

Рецензенты:
член-корреспондент РАН **Р.М. Юсупов**;
доктор физико-математических наук, профессор **А.Я. Перельман**

Рыжиков Ю.И.
Алгоритмический подход к задачам массового обслуживания: монография / Ю.И. Рыжиков. — СПб.: ВКА им. А.Ф. Можайского, 2013. — 496 с.

В книге описаны аналитические и численные методы расчета систем с очередями, в том числе многоканальных, немарковских и приоритетных, а также расчет немарковских сетей на основе их потокоэквивалентной декомпозиции. Значительная часть этих методов разработана автором и реализована в составе пакета МОСТ (**М**ассовое **О**бслуживание — **СТ**ационарные задачи). Численные результаты иллюстрируются таблицами и графиками.

Набор, верстка и корректура выполнены автором

Подписано к печати 20.11.2013 Формат печатн. листа 445x300/8
Гарнитура Arial Авт. печ. л. 30,75
Уч.-печ. л. 62,00 Заказ 2648 Бесплатно

Типография ВКА им. А.Ф. Можайского

Оглавление

Введение

0.1. Понятие о моделировании

По М. Минскому, объект А является моделью объекта В для наблюдателя С, если наблюдатель С с помощью А получает ответ на интересующий его вопрос относительно В.

Модель *концентрирует* в себе написанную на определенном языке (естественном, математическом, алгоритмическом) совокупность наших знаний, представлений и гипотез о соответствующем объекте или явлении. Поскольку эти знания никогда не бывают абсолютными, а гипотезы могут вынужденно или не учитывать некоторые эффекты, модель лишь приближенно описывает поведение реальной системы и является ее абстракцией. Прогресс науки и техники нашел свое наиболее точное выражение в развитии способности человека создавать модели естественных явлений, понятий и объектов. По авторитетному мнению акад. А. А. Самарского, «методология математического моделирования может и должна быть ядром информационных технологий, всего процесса информатизации общества». О многообразии применений математического моделирования можно судить хотя бы по перечню моделей, разработанных в ВЦ АН СССР:

- эффект распределения власти в иерархии;
- безопасность ядерного реактора;
- экологические последствия сжигания углеводородных топлив;
- тунгусский метеорит;
- климатический эффект «100-мегатонного» конфликта;
- магнитогидродинамическая теория солнечной активности;

- процессы в переходной экономике.

Постановка над моделью экспериментов с последующей интерпретацией их результатов применительно к моделируемой системе проводится в целях:

- прогнозирования будущего поведения системы;
- осмысления действительности;
- обучения и тренажа специалистов.

Модели играют важную и всевозрастающую роль в развитии общества. «Часто новые идеи (например, в управлении производством и экономикой) длительное время не находят применения на практике по той причине, что система, в которую они должны быть внедрены, обладает большой внутренней сложностью и последствия предлагаемых преобразований трудно предсказать. С помощью имитации можно организовать проверку и демонстрацию новых идей и обосновать, таким образом, их принятие или отклонение». В частности, в ВЦ АН СССР под руководством акад. А. А. Петрова в свое время были смоделированы последствия «шоковой терапии» российской экономики. Действительность полностью подтвердила мрачные прогнозы, которые были проигнорированы власть имущими.

С другой стороны, в США важность компьютерного моделирования осознана (хоть и недавно — июнь 2005 г.) на самом высшем уровне. Приведем данные из доклада *лично* Президенту США Комитета советников по информационным технологиям.

Главный вывод комитета состоит в том, что *использование продвинутых компьютерных возможностей для понимания и решения сложных проблем стало критическим для научного лидерства, экономической конкурентоспособности и национальной безопасности*. Комитет признал, что наука о вычислениях — единственное средство исследовать проблему, к которой иначе невозможно подступиться: от биохимических процессов в человеческом мозгу и фундаментальных физических сил, формирующих Вселенную, до анализа распространения инфекционного заболевания или распыленных террористами ядов, развития экономики, разработки профиля самолетного крыла (вместо длительных и дорогостоящих экспериментов в аэродинамической трубе). На математических моделях основаны и ключевые технологии XXI

века: новые материалы (включая полупроводники и сверхпроводники), альтернативные источники энергии, биотехнологии, суперкомпьютеры, нанотехнологии, микроэлектромеханические системы, оптоэлектроника, беспроводная связь, по-атомное создание материалов с фантастическими свойствами, миниатюризация приборов вплоть до квантового уровня. *Гонка будет выиграна теми, кто лучше знает современные компьютерные системы и их научные приложения*. Однако сейчас основные усилия сосредоточены на создании ЭВМ рекордной производительности, тогда как реальные проблемы компьютерного моделирования — прикладная математика и программное обеспечение.

Далее в упомянутом докладе для основных областей знания приведены примеры в высшей степени успешного моделирования.

Физика. Экспериментальные высокотемпературные сверхпроводники могут передавать электрический ток без существенного сопротивления при необычно высоких температурах. Совершенствование и внедрение новых материалов могли бы дать колоссальный экономический эффект, позволив, например, по немногим сверхпроводящим кабелям обеспечить электричеством целые города или породив новое поколение мощных и легких двигателей. Недавние алгоритмические разработки и существенное увеличение вычислительных возможностей позволили организовать широкомасштабные параллельные вычисления, открыв путь к решению существенной для обсуждаемой задачи квантовой проблемы многих тел.

Лазерная технология позволяет создать компактные ускорители высокой энергии для исследования субатомного мира — с целью изучения новых материалов и технологий и для медицинских применений. В теории частицы, ускоренные электрическими полями созданных лазером волн плазмы, могли достигнуть меньше чем на 100-метровом расстоянии энергий, получаемых на машинах длиной в мили при ускорении на радиочастотах. Эти экспериментальные результаты были проверены на «плазменной» программе моделирования. Модель позволила видеть детали развития процесса, что создало предпосылки для его оптимизации.

Полупроводники и другие неорганические кристаллы служат базисом для электроники и других технологий. Однако обычные методы производства необходимых мягких структур типа тонких пленок или процессов самосборки страдают от влияния подложки и других молекулярных взаимодействий. Комбинируя экспериментальные наблю-

дения и разработки с обширными работами по вычислительной химии, исследователи добились однородной перпендикулярной ориентации к подложке с увеличенными сроками люминесценции и фотостабильностью. Был открыт путь к молекулярной фотонике, технологии дисплеев и биотехнологии, созданию оптических наноструктур.

Астрономия. Спустя четыреста лет после наблюдения Галилеем массивной взорвавшейся звезды механизм взрыва суперновых все еще остается неизвестным. Развитие многомерных моделей сверхновых звезд позволило исследовать роль, которую могли бы сыграть в их возникновении конвекция, вращение и магнитные поля. Появилась надежда понять механизмы, ответственные за взрыв суперновой, и все явления, с этим связанные (вклад сверхновой звезды в синтез химических элементов; эмиссию нейтрино; рождение гравитационных волн).

Науки о Земле. Серьезные штормы в Соединенных Штатах порождают приблизительно 800 торнадо в год. Потери собственности и экономические потери измеряются миллиардами долларов — в дополнение к ежегодному среднему числу 1 500 искалеченных и 80 смертных случаев. Предсказатели погоды могли идентифицировать штормы с вихревым потенциалом, но три из четырех предупреждений о торнадо оказывались ложными. Разработанная в университете штата Оклахома модель охватывала область 50 километров по каждой стороне и 16 километров по высоте. Было использовано 24 часа вычислительного времени с 2048 процессорами, моделируемый шторм сформировал 20 терабайтов данных, хорошо совпавших с реальными. Ожидается, что с использованием новой модели количество ложных торнадо-тревог уменьшится втрое.

Южная область Калифорнии не испытывала серьезных землетрясений с 1690 г., и накопленное напряжение может привести к катастрофе магнитудой 7.7 балла. Очень важно обеспечить в случае такого бедствия сохранность построек и спасение жизней везде, где оно могло бы произойти. Потенциальные угрозы оценивались в вычислительном эксперименте. Моделировался объем 600 км длиной, 300 км шириной и 80 км глубиной, охватывающий все главные города региона.

Экология. Экологи и специалисты по распространению пожаров разработали модель лесных пожаров. Чтобы предсказать поведение пожара в реальном времени в течение критического периода пожарной опасности и развивать планы борьбы с огнем, ученые объединили данные о топографии, растительности и погодных условиях.

Макроэкономика. Федеральное резервное управление потребителей использовало макроэкономическое моделирование в течение более чем трех десятилетий, чтобы анализировать национальные и международные экономические процессы и оценить возможные воздействия изменений в валютной политике.

Моделирование сложных систем реального времени в человеческом обществе. Способность моделировать, например, распространение эпидемии болезни или ежедневный трафик столичного метрополитена обеспечивает должностные лица здравоохранения и службы быстрого реагирования мощным инструментом планирования, который дает визуальное представление взаимодействия сложных данных. Наблюдение укрупненной картины того, что могло бы произойти в ходе кризиса, помогает предвидеть и принимать решения заранее (например, в каких больницах и сколько больничных коек будут необходимы в течение распространения эпидемии). Появление болезней типа болезни Лайма, ВИЧ/СПИДа, hanta-вируса, Западного Нильского вируса и птичьего гриппа вынудило признать эпидемиологическое моделирование жизненно важным инструментом выработки тактики здравоохранения. Изучение распространения болезней в пространстве и во времени обеспечивает лучшее понимание механизмов передачи и факторов, которые сильнее всего влияют на их распространение; позволяет делать предсказания; помогает выбирать и оценивать стратегии управления и минимизировать ущерб. Эпидемиологи разработали агентно-ориентированную модель распространения инфекционной болезни среди населения. Такие программы помогут, например, оценить перед лицом пандемии эффективность прививки и карантина.

Индустрия. В крупной промышленности специализированное программное обеспечение, работающее в сетевых компьютерных системах, используется для управления потоками информации, материалов, финансов и логистическими операциями. При этом улучшается соотношение цена/эффективность и обеспечивается конкурентное преимущество. В финансовом секторе компьютерные модели стали главным инструментом макро- и микроанализа и прогнозов.

Турбины, применяемые в двигателях и для генерации энергии, используются при температурах, где топливо сгорает полностью. Газы, переходящие из камеры сгорания в первую ступень турбины, имеют температуры, на сотни градусов превышающие точку плавления компонентов турбины. Увеличение поверхностной температуры на несколько

десятков градусов может сократить жизнь лопатки наполовину, так что для сохранения долговечности и безопасности турбины ее охлаждение является критическим. Лопасти турбины и лезвия охлаждаются компрессором, прогоняющим воздух через внутренние проходы в лезвиях. Чтобы увеличить внутреннюю передачу тепла, в этих проходах усиливают турбулентность с помощью ребер, игл и т. п. Но турбулентный поток трудно точно предсказать стандартными методами. Новые модели вычисления успешно моделировали турбулентный поток и теплопередачу для этих сложных систем, обеспечив надежное предсказание характеристик проекта.

Исследователи давно знали, что микропузырьки примерно 50-500 микронов в диаметре могут в некоторых случаях сократить сопротивление движению судов на 80%, уменьшая расход топлива и увеличивая дальность плавания. В течение 30 лет системы микропузырьков изучались экспериментально: поршни проталкивали воздух через пористые пластины, представлявшие корпус судна в резервуарах с движущейся водой. Исследователи меняли местоположение пластин, увеличивали или уменьшали число и размер пузырьков. Однако микропузырьки мешают традиционным методам измерения деталей потока в экспериментальном резервуаре, поскольку оптические системы не могут видеть сквозь созданную пузырями турбулентность. Это не позволяло оптимально выбирать размеры и расположение пузырьков. Чтобы обойти эту проблему, были созданы новые вычислительные модели действия микропузырей, которые позволили отслеживать примерно 20000 пузырей вместо прежних 500.

Моделирование разрушения автомобилей ныне является фундаментальной компонентой в проектировании автомобиля всеми ведущими мировыми производителями. Корпорация Боинг в 1980-х гг. создала инструменты моделирования, которые позволили компании избавить процесс проектирования самолетов от дорогого физического испытания отдельных структур: крыла, двигателей, кабины пилотов и перейти к интегрированному компьютерному моделированию. При этом радикально уменьшились затраты и ускорился запуск производства.

В фармацевтической промышленности компьютерные возможности преобразовали поиск возможных лекарств и методов лечения, резко увеличив продуктивность и соревновательность в этой ключевой области.

Есть надежда, что важность компьютерного моделирования будет осознана и российскими властями всех уровней. Заместитель начальника вооружения ВС РФ в предисловии к [11] пишет:

> «Чтобы создать модель, требуются годы кропотливого труда высококвалифицированных системщиков, математиков и программистов. И когда модель удается создать, она начинает активно использоваться как для обоснования и проверки выполнимости системных требований, так и в процессе разработки для оценки и оптимизации технических решений и выявления «узких мест», а в процессе эксплуатации — для настройки технологических параметров системы.
>
> Отечественный и зарубежный опыт показывает, что использование моделей позволяет разработчику избежать нерациональных затрат, соизмеримых с затратами на эскизно-техническое проектирование системы...Если характеристики неудовлетворительные, то ущерб может оказаться несоизмеримым со стоимостью разработки всей системы».

Во втором предисловии Первого вице-президента Российской Академии ракетных и артиллерийских наук утверждается, что «ни один уважающий себя Главный конструктор не предлагает технических решений, не базирующихся на результатах моделирования, экспериментов или натурных испытаний».

0.2. Моделирование как прикладная математика

Важнейшая особенность модели состоит в возможности неограниченного накопления специализированных знаний без потери целостного взгляда на объект исследования. Моделирование процессов в природе, обществе и технических системах — это основная компонента системного подхода к познанию этих процессов и управлению ими. Математическое моделирование является важнейшим элементом *прикладного* исследования. От исследований в области теоретической математики прикладную математику отличают:

1) Социальный заказ на каждую конкретную разработку.

2) Жесткие сроки, требования экономичности и эффективности.

3) Сложность и комплексный характер как поставленных проблем, так и методов их решения.

4) Принципиальная необходимость в многообразном использовании ЭЦВМ.

5) Возможность некоторого смягчения традиционно принятых в математических исследованиях стандартов строгости:

   - широкое использование аналогий, правдоподобных рассуждений и эвристических решений;
   - поиск решений, единственность и даже существование которых пока не доказаны;
   - использование итерационных процедур, сходимость которых теоретически не доказана, и т. д.

Перечисленные выше «послабления» являются вынужденными. Прикладная математика «делает то, что нужно — так, как может» в отличие от чистой, которая «делает, как нужно — то, что может». Классик теории матриц и динамического программирования Р. Беллман справедливо полагал, что получение численного решения требует гораздо больше изобретательности и знаний, чем доказательство существования и единственности решения [233]. Признано (к сожалению, далеко не всеми), что хотя бы приближенный обсчет реалистичной модели имеет намного бóльшую ценность, чем строгое исследование оторванной от практики упрощенной абстракции (пример последней — обсуждаемые в главе 11 модели, подпадающие под условия теоремы ВСМР). Разумеется, результаты приближенного анализа должны быть тем или иным способом *верифицированы* — теоретическим обоснованием приемлемости частных допущений и/или сопоставлением конечных результатов с полученными на имитационной модели.

Решение прикладной проблемы с применением математики неизбежно требует математического моделирования. Общая схема этапов и компонент такого моделирования показана на рис. 1.

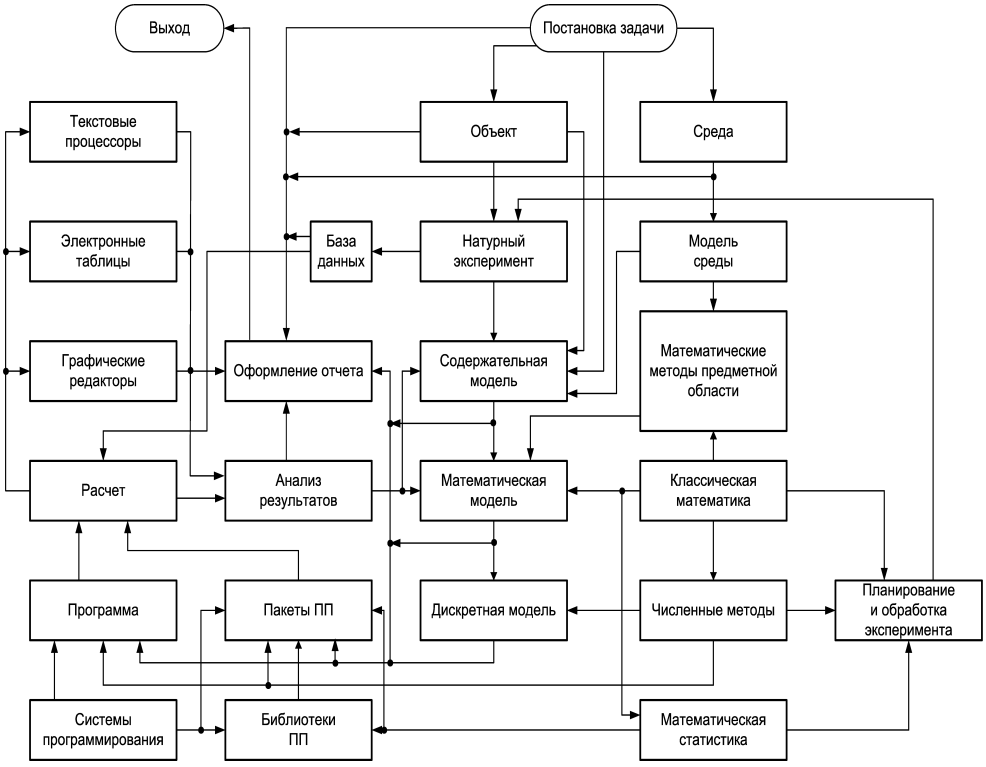

Рис. 1. Организация математического эксперимента

0.3. Свойства моделей и виды моделирования

Адекватность модели объекту всегда ограничена и зависит от цели моделирования. Рекомендуется выбирать модель минимальной сложности при заданной точности либо максимальной точности — при заданной сложности. Понятию сложности при этом придается весьма широкий смысл: количество внутрисистемных связей; количество операций, необходимых для получения псевдослучайных чисел; длина программы моделирования; ее сложность по Холстеду и т. п.

Принцип *баланса точности* требует соизмеримости погрешностей, вызываемых различными причинами: случайным характером результатов моделирования, неполным соответствием модели объекту, неточно-

стью задания исходных параметров модели.

При исследовании сложных систем может потребоваться разработка *набора моделей*, соответствующих различным иерархическим уровням рассмотрения и функциональным разрезам деятельности системы. Такое стратифицированное описание на каждом уровне использует свой набор концепций, понятий и терминов. Для предотвращения пропусков и бросовых затрат разработка иерархии моделей должна вестись «сверху вниз».

В сравнении с натурным экспериментом математическое моделирование имеет следующие преимущества:

1) Экономичность (в частности, сбережение ресурса реальной системы).

2) Возможность моделирования гипотетических (т. е. не реализованных в натуре) объектов и процессов.

3) Возможность реализации режимов опасных или трудновоспроизводимых (критический режим ядерного реактора, работа системы ПРО).

4) Возможность изменения масштаба времени.

5) Легкость многоаспектного анализа.

6) Большую прогностическую силу вследствие возможности выявления общих закономерностей.

7) Универсальность технического обеспечения проводимой работы (ЭЦВМ, системы программирования и пакеты прикладных программ широкого назначения).

Математическое моделирование на определенных этапах исследования может сочетаться с натурным. Классическим примером такого подхода является исследование динамики летательного аппарата (самолета, ракеты, спутника) на комплексе из математической модели самого аппарата, воспроизводимой на аналоговой или цифровой ЭВМ, и макета реальной аппаратуры управления.

Практическое использование модели возможно лишь после тщательного ее исследования и настройки, в процессе которых необходимо решить задачи проверки адекватности, идентификации параметров, оценки значимости параметров и структурного преобразования.

Адекватность модели устанавливается проверкой для нее основных законов предметной области типа законов сохранения и сопоставлением результатов моделирования частных вариантов с известными для этих вариантов аналитическими решениями или натурными наблюдениями.

Задачей *идентификации* является определение значений рабочих параметров модели по набору исходных данных, обычно получаемому в результате наблюдения над реальной системой. При этом тип модели предполагается известным. Эта задача обычно ставится в форме минимизации функционала отклонения траектории модели от траекторий исследуемой системы. На практике для ее решения традиционно применяются методы наименьших квадратов и наибольшего правдоподобия.

Анализ *чувствительности* выходных показателей по отношению к изменению входных параметров является важнейшим инструментом оптимизации. В частности, он позволяет уменьшить размерность пространства параметров. Планирование и обработка результатов таких исследований составляет предмет теории экспериментов.

Задачи *преобразования модели* обычно решаются после накопления некоторого опыта работы с моделью. К ним относятся упрощение, усложнение, структурное изменение. Изменение модели отражается как на потребных вычислительных ресурсах, так и на возможности получения тех или иных характеристик.

Математическому моделированию в любой его форме предшествует создание содержательной (*концептуальной*) модели, определяющей объект, цель и условия моделирования.

Разработка формальных моделей начинается с простейших подходов, при которых учитываются лишь наиболее существенные характеристики исследуемых процессов и важнейшие законы сохранения. Необходимая степень уточнения достигается последовательным усложнением модели с целью ее приближения к реальности. Каждое такое усложнение должно быть достаточно значимым (рекомендуется минимум удвоение сложности модели).

Завершающим этапом математического моделирования обычно является оптимизация, результаты которой затем переносятся на моделируемую систему. При поиске оптимальных значений параметров часто используются модели планирования эксперимента.

Математическое моделирование можно разделить на аналитическое, имитационное и комбинированное.

При *аналитическом* моделировании процессы функционирования элементов системы записываются в виде алгебраических, интегральных, дифференциальных, конечно-разностных и иных соотношений и логических условий. Аналитическое решение может быть:

- качественным (устанавливаются такие свойства решения, как существование, единственность, устойчивость в большом и малом, характер зависимости выходных параметров от входных и т. д.);
- аналитическим — в виде формульных зависимостей искомых характеристик от входных параметров.

Качественные методы полезны и необходимы, но не дают конкретного решения практической задачи. Примером качественных методов анализа является общая теория дифференциальных уравнений.

Аналитическое решение всегда предпочтительно, но его обычно удается получить лишь после ряда упрощающих предположений. Возможность его достижения очень критична к изменениям модели, для нахождения решения требуются высокая квалификация и значительные творческие усилия разработчика.

Гораздо чаще удается применить *численные* методы. Здесь творческий элемент при наличии хорошей библиотеки подпрограмм может быть сведен к разумному минимуму. Численные методы дают частные результаты, по которым трудно делать обобщающие выводы. При оптимизации модели необходим многовариантный счет.

Имитационное моделирование усугубляет недостатки численных методов, но практически свободно от ограничений на класс решаемых задач. Здесь аналоги составляющих процесс элементарных явлений многократно воспроизводятся с сохранением его логической и временно́й структуры, после чего выполняется обработка накопленной статистики.

0.4. Задачи теории очередей

За последние десятилетия в разных областях техники, в организации производства и в военном деле возникла необходимость решения своеобразных вероятностных задач, связанных с работой *систем*

массового обслуживания разного вида требований. Термин «массовое обслуживание» предполагает многократную повторяемость ситуаций в том или ином смысле (много прибывших в систему и обслуженных заявок, большое число находящихся в эксплуатации аналогичных систем) и статистическую устойчивость картины. Выводы и рекомендации, получаемые методами теории массового обслуживания (ТМО), применимы лишь при наличии одного или обоих перечисленных факторов повторяемости.

По справедливому замечанию виднейшего советского специалиста в этой области Б. В. Гнеденко, легче указать ситуации, где не может быть использована ТМО (она же — теория очередей), чем перечислить все сферы ее потенциального применения. К таким сферам, в частности, относятся:

- в технике связи — проектирование телефонных станций и сетей связи, анализ протоколов обмена информацией;
- на транспорте — анализ процессов дорожного движения, прохождения туннелей, очередей у светофоров, технического осмотра автомашин, формирования железнодорожных составов на сортировочных станциях, регистрации пассажиров, взлета и посадки самолетов, диспетчерской службы в аэропортах, разгрузки и погрузки судов в речных и морских портах, резервирования билетов;
- в промышленности — при планировании сборочных операций (расчет числа линий и емкости бункеров), гибких автоматизированных производств, организации ремонта оборудования, расчете объема восстанавливаемого ЗИП, работе пунктов обслуживания (например, инструментальных складов), расчете зон обслуживания (в прядильном и ткацком производстве), определении числа резервных забоев в шахтах;
- в автоматизированных системах — для оценки своевременности обслуживания заявок на вычислительные работы и подготовку данных, обработки результатов экспериментов, управления технологическими процессами, при исследовании производительности вычислительных систем с асинхронно работающими устройствами;
- в торговле — для определения количества магазинов, продавцов, фасовщиц, узлов расчета, торговых и кассовых автоматов;

- в здравоохранении — при определении необходимого количества аптек, больничных коек, станций и бригад скорой помощи, врачей и младшего медицинского персонала, потребностей в диагностической и лечебной аппаратуре, пропускной способности специализированных санаториев;

- в службе быта — при расчете сети парикмахерских, мастерских разного назначения, количества техники и работников в них, транспортных средств обеспечения услуг населению, пунктов приема коммунальных платежей, количества ремонтников в жилищно-эксплуатационных конторах и трестах, сотрудников бюро по трудоустройству, по обмену жилой площади и т. д.;

- в органах юстиции и внутренних дел и МЧС — при расчете количества судебных органов и их штатной численности, емкости исправительно-трудовых учреждений, численности сотрудников следственного аппарата и оперативников, планировании работы контрольно-пропускных пунктов и пунктов таможенного досмотра (очереди грузовиков на пограничных пропускных пунктах Ваалимаа и Нуйямаа достигают 20 км — «СПб. Ведомости», 31.03.2008), диспетчерских пунктов противопожарной службы — см. [139];

- в сфере науки и образования — при исследовании некоторых видов природных процессов (регистрация элементарных частиц, фильтрация, диффузия, действие катализаторов химических реакций), запуске ИСЗ и обработке спутниковой информации, расчете количества лабораторных установок, экзаменаторов (в том числе автоматических), возможностей учебно-опытных производств, бытового и медицинского обслуживания студентов; при проектировании и анализе работы крупных библиотек;

- в издательском деле — при расчете численности сотрудников, участвующих в подготовке рукописей к изданию;

- в военном деле — при проектировании систем ПВО (каждая цель может рассматриваться как «заявка» на обслуживание, т. е. обстрел), организации охраны границ, расчете патрульных нарядов и во многих других случаях типа рассмотренных выше применительно к боевой и повседневной деятельности войск.

Сразу же подчеркнем, что задачи указанных типов приходится решать не только при проектировании вновь создаваемых систем и сетей обслуживания (СМО, СеМО), но и *в процессе эксплуатации имеющихся* — при увеличении нагрузки; изменении трудоемкости обработки заявок; выходе из строя, деградации или модернизации техники; снижении квалификации персонала; пересмотре требований к оперативности обработки заявок и т. п.

Характерной чертой массового обслуживания является случайный поток заявок и случайная продолжительность времени обслуживания. Поэтому *прогноз относительно единичного события может быть только вероятностным*, а математический аппарат ТМО строится на основе теории вероятностей.

Применение теории очередей особенно эффективно при большой роли случайных факторов, что подтверждается несколько экстремистским заявлением Г. Поттгофа: «...В этих опытах теория и наблюдения хорошо совпадали, так как благодаря большой нерегулярности или отсутствию должного порядка, что имелось в годы войны, были созданы особенно благоприятные предпосылки для применения теории вероятностей» [85, с.103]. Добавим, что столь же «благоприятно» в указанном смысле и разрегулированное состояние экономики. Впрочем, и в налаженной экономике проблема очередей актуальности не теряет. «Резонно спросить, какая польза от изучения такого неприятного явления. Ответ, конечно, состоит в том, что через понимание происходит сочувствие, а именно оно и необходимо, так как с развитием цивилизации люди все больше и больше времени проводят в очередях, и необходимо найти пути преодоления этих неприятностей» [49, т. 1, с. 17].

0.5. История и состояние теории очередей

Теория массового обслуживания берет свое начало в выполненных около 100 лет назад работах датского ученого А. К. Эрланга (1878–1929) — пионера применения математики в исследовании целенаправленной человеческой деятельности. В его честь интенсивность нагрузки (суммарное время обслуживания вызовов в единицу времени) измеряется в эрлангах.

Основными вехами в истории ТМО считаются [164, с. 4]: концепция статистического равновесия, которую неявно использовал еще Эрланг;

подход к анализу сетей обслуживания (Джексон и Келли); применение матричных методов (М. Ньютс). В становлении и развитии этой теории видная роль принадлежит советским ученым — Г. П. Башарину, А. А. Боровкову, Б. В. Гнеденко, Э. А. Даниеляну, Г. П. Климову, И. Н. Коваленко, А. Н. Колмогорову, А. А. Маркову, Ю. В. Прохорову, Б. А. Севастьянову, А. Я. Хинчину и сотням других. Библиография работ по ТМО насчитывает многие тысячи наименований. Современное состояние классической ТМО типично для хорошо развитой науки: простых задач с почти очевидными решениями в ней не осталось; некоторые направления (диффузионная аппроксимация, метод матрично-геометрической прогрессии, BCMP-сети) оказались практически тупиковыми (см. соответствующие разделы). Серьезную конкуренцию ТМО составляют имитационное моделирование — мощный, но тупой инструмент, позволяющий решить практически любую *частную* задачу, и сети Петри.

«Точками роста» теории являются матричные методы и их численная реализация; сети обслуживания при более реалистических допущениях о потоках и распределениях обслуживания; сети с «отрицательными» заявками, моделирующими негативные внешние воздействия; системы с «прогулками» (временным отключением обслуживающих устройств — как правило, на профилактику или для выполнения другой работы); обслуживание с повторными попытками — наличие дополнительной популяции с источниками заявок. Для современной теории связи важен расчет работы систем обслуживания в дискретном времени. С прикладной точки зрения весьма актуальны компьютерные применения (в частности, организация распределенных вычислений), производственные системы (конвейеры, гибкие автоматизированные производства), разнообразные службы быстрого реагирования.

Беспроводная передача данных открыла путь к мобильным системам связи. Появилась потребность в сетевых моделях, отражающих пространственное распределение пользователей; зависящее от пространственных факторов распределение обслуживания; распределения и перемещений последних в пространстве. В [167] дан краткий обзор результатов «пространственной» ТМО с 1995 г. Практически, однако, речь идет преимущественно об анализе потоков. Поскольку в мобильных сетях нет очередей (если линия занята, заявка уходит), рассматриваются только узлы с бесконечным числом каналов или с отказами.

С сожалением приходится констатировать, что использование результатов ТМО существенно отстает от уровня развития теории,

причем не только в России. По Р. Нельсону, «удивительно, что эти простые результаты не используются практиками» [229]. Учебная литература [2, 23, 34, 35, 39, 139], исключая университетские курсы [51, 68] и немногие книги предперестроечных лет, освещает в основном марковские СМО и в этом смысле недалеко ушла от первых работ Эрланга, да и в этих случаях содержит элементарные ошибки. Например, приведенное в [139] определение потока без последействия (с. 10) на самом деле относится к потоку с ограниченным последействием, регулярное обслуживание объявляется частным случаем гиперэкспоненциального (с. 12) и т. д. Монографии же перечисленных выше авторов написаны на недоступном инженеру уровне. Извлечение из них для практического использования конечных результатов не решает проблемы: в таких случаях эти результаты выглядят менее убедительными, условия их применимости — недостаточно ясными, а проведение каких-либо полезных аналогий становится невозможным. Публикуемые методы расчета немарковских систем [43, 48, 49, 50, 51, 55, 68, 86, 132] редко позволяют довести результаты до числа, а для многоканальных систем разбросаны по журнальным статьям. Достойно изумления, что в фундаментальной монографии [86] и в университетских учебниках [51, 68] нет ни таблиц, ни графиков результатов счета. Разумеется, в век высокоразвитой вычислительной техники последние утратили ценность как расчетный инструмент, но сохраняют полезность при сопоставлении различных подходов, оценке влияния входных параметров и т. п. С другой стороны, информация о пакетах программ для расчета СМО весьма скудна, имеет рекламный характер и не раскрывает реализованных в них алгоритмов.

Актуальность проблем ТМО в последние годы существенно возросла в связи с массовым созданием и использованием вычислительных сетей разного масштаба — от локальных до Интернет: «В условиях стремительного развития информационного взаимодействия между различными абонентами часто возникает необходимость в применении какого-либо метода, позволяющего предсказывать последствия при определенных изменениях в загрузке сети, ее топологии, а также позволяющего проводить априорную оценку некоторых параметров вновь развертываемой сети... Может быть сформировано представление о том, насколько сервер будет загружен... и насколько большим должен быть буфер сервера» [57].

«Крупнейшими пользователями информационных технологий являются финансовые и коммерческие структуры. Широкое использование кредитных карт и безналичных расчетов привело к тому, что в развитых странах системы передачи данных стали, образно говоря, кровеносной системой хозяйства. Конкурирующие фирмы предлагают пользователям сетей определенный набор услуг (по скорости передачи информации, надежности, конфиденциальности, стоимости и др.). Оптимальный или хотя бы сознательный выбор такого набора невозможен без знакомства с теорией массового обслуживания.

В борьбу за клиента в современной экономике вкладываются огромные средства. По оценкам западных экономистов, завоевание фирмой нового клиента обходится ей в 6 раз дороже, чем удержание существующих покупателей. А если клиент ушел неудовлетворенным, то на его возвращение приходится потратить в 25 раз больше средств. Во многих случаях неудовлетворенность клиента вызвана неудачной организацией обслуживания (слишком долгое ожидание в очереди, отказ в обслуживании и др.). Использование теории массового обслуживания позволяет избежать подобных неприятностей» [67, с. 5].

К сожалению, молодые специалисты по компьютерным сетям убеждены в том, что информационные технологии вполне заменяют знание основ ТМО. Автор цитированной выше статьи М. Кульгин полагает, что «аналитик может провести анализ очередей в заданной сетевой структуре, используя уже готовые таблицы очередей или простые компьютерные программы, которые занимают несколько строк кода», и далее демонстрирует математическую безграмотность и полное непонимание проблемы.

Для длительности обслуживания им предлагается «закон интервалов, или экспоненциальный закон». Здесь родовое понятие отождествляется с его частным случаем. Формула $\Pr[\Theta \geq \vartheta] = 100e^{-\mu\vartheta}$ из-за невесть откуда появившейся сотни может давать значения вероятностей до 100 включительно. Далее утверждается, что «μ — уровень обслуживания, в данном случае равный коэффициенту загрузки ρ». Поскольку ρ — величина безразмерная, показатель степени приобретает размерность времени — совершенно вопиющий абсурд. На самом деле μ есть $интенсивность\ обслуживания$ (величина, обратная его средней длительности).

Формулы для $m_{T_q}(r)$ и $m_{T_\omega}(r)$ в табл. 3 из [57] неверны уже потому, что эти величины в табл. 1 [57] определены как вероятности, а имеют

размерность времени.

Утверждается, далее, что «если существует среда, в которой есть разделяемые каналы связи, то производительность такой системы обычно изменяется по экспоненциальному закону при увеличении нагрузки ... При дальнейшей загрузке системы ее производительность будет резко снижаться». На самом деле производительность системы измеряется не временем ответа, а числом заявок, обслуженных в единицу времени. Она равна интенсивности входящего потока — следовательно, с ростом коэффициента загрузки ρ будет *возрастать*, пока последний не достигнет единицы. После этого производительность системы будет постоянна и равна ее максимальной производительности. Экспоненциальному же закону при докритическом режиме подчиняется не производительность системы, а распределение *времени пребывания* заявки в ней. На самом деле среднее время ответа при простейших допущениях меняется по *гиперболическому* закону с вертикальной асимптотой в точке $\rho = 1$.

В тезисе «Наихудшую производительность показывает система с экспоненциальным распределением, наилучшую — с постоянным» опять смешиваются производительность и реактивность. Кроме того, распределения с коэффициентом вариации, бóльшим единицы (гамма-, Вейбулла и гиперэкспонента при частных значениях их параметров), дают реактивность худшую, чем экспонента. Данное М. Кульгиным определение коэффициента загрузки не работает для многоканального сервера. Обещанное рассмотрение приоритетного обслуживания так и не состоялось. Приведенные им условия пуассоновости процесса недостаточны (см. учебники по теории вероятностей). Наконец, в статье фактически рассматривается расчет *не сетей, а систем обслуживания*. М. Кульгин путает допуски с допущениями, которые приходится делать; отождествляет программное моделирование (к которому относится и счет по формулам) с имитационным.

Все перечисленные недостатки перенесены М. Кульгиным в «Энциклопедию» того же автора [58] под названием «Технологии корпоративных сетей».

Теория очередей является идейной и математической основой *имитационных моделей*, с помощью которых можно исследовать не поддающиеся аналитическим и численным методам ситуации. Естественно потребовать от «имитаторов» уверенного владения этой теорией. Анализ докладов на четырех Всероссийских конференциях по имита-

ционному моделированию [69, 120] показал, что дело, к сожалению, обстоит иначе. К примеру, принималось экспоненциальное распределение длительностей обслуживания, «поскольку разброс времени обслуживания относительно среднего невелик». Но ведь это распределение имеет единичный, т. е. весьма значительный, коэффициент вариации, а из упомянутого факта вытекает гипотеза о $регулярном$ обслуживании! Утверждалось, что «распределение $A(t)$ задает время поступления заявки» (вместо интервала между поступлениями), и что «второй момент распределения можно выразить в форме коэффициента вариации». В «военно-медицинском» докладе о помощи жертвам нападения на аэродром было заявлено, что «очередь образуется при интенсивности потока 13 раненых в час и более». Но ведь процесс-то случайный, и надо говорить об $ожидаемой$ длине очереди, отличной от нуля при сколь угодно малой интенсивности потока.

В ряде докладов игнорировался присущий анализируемой ситуации $групповой$ характер заявок, радикально меняющий результаты расчета. С помощью аппарата для стационарных ситуаций исследовались нестационарные (та же военно-медицинская проблема; обращения граждан в учреждения — как правило перед началом обслуживания уже имеется очередь, которая слабо пополняется со временем). К тому же, в обоих обсуждаемых случаях надо было моделировать не систему, а $сеть$ обслуживания.

Непонимание базовых идей особенно ощущалось при обсуждении $производительности$ СМО, где повторялись ошибки М. Кульгина.

Отчетливо просматривалось незнание возможностей современных численных методов теории очередей. Один из докладчиков назвал аналитическое решение задачи Эрланга (марковская система с отказами) непомерно трудным и требующим «программного обеспечения, обрабатывающего большие числа». Неужели 10^{4932} (расширенный формат для чисел при работе ПК с двойной точностью) будет недостаточно? К тому же, известны простые асимптотические зависимости. Почему-то считались проблемой расчет приоритетных режимов для одноканальных систем; моделирование с распределениями Парето и Вейбулла; оценка влияния высших моментов на показатели работы системы GI/M/1. Посредством имитации тщательно исследовалась погрешность приближенных формул для расчета системы с известным $точным$ решением. Демонстрировалось влияние высших моментов на среднюю длину очереди в системе GI/M/1, вытекающее из давно известного теоретического

решения. Утверждалось, что аналитическое моделирование требует лишь знания Excel (было даже заявлено — профессором! — что «заниматься имитационным моделированием можно, не зная ни высшей математики, ни теории очередей»).

Опыт таких «занятий» имеется, но прискорбный. Надо понимать, что имитация практических задач хотя и дает дополнительные расчетные возможности, но *подчинена тем же общевероятностным и специальным законам*. Их незнание «имитаторами» есть явный непрофессионализм. Такие сведения остро необходимы нынешней компьютерной молодежи, натасканной на технологии, но слабо оснащенной математически и не осознавшей базовых идей. В связи с этим нельзя не привести цитату из доклада на одной из конференций ИММОД: «Ни один производственник или управленец не будет работать с моделью, требующей знания языка программирования (моделирования) и работы с математическими формулами». Хочется спросить: кого, как и зачем готовят инженерно-экономические вузы («университеты»)?

В заключение этого обзора подчеркнем, что численно-аналитические подходы к исследованию систем и сетей с очередями не могут быть полностью заменены имитационными моделями. Прежде всего отметим, что вера в возможность добиться сколь угодно высокой точности соответствующим увеличением числа испытаний основана на гипотезе об идеальной работе датчиков псевдослучайных чисел (ДСЧ). О ее несостоятельности убедительно свидетельствует табл. 1 погрешностей вычисления π методом статистических испытаний — через долю точек, попавших во вписанный в единичный квадрат круг. К этому добавим рост погрешностей статистической оценки среднего времени ожидания в системах обслуживания обратно пропорционально $(1-\rho)^2$, где ρ — коэффициент загрузки.

Таблица 1. Погрешности статистического определения числа π

N	δ	N	δ	N	δ
1 тыс.	9.8e-2	50 тыс.	2.8e-3	2 млн	-1.4e-4
2 тыс.	8.6e-2	100 тыс.	8.3e-3	5 млн	-1.0e-5
5 тыс.	5.6e-3	200 тыс.	4.3e-3	10 млн	2.9e-4
10 тыс.	1.8e-2	500 тыс.	4.3e-3	20 млн	4.2e-5
20 тыс.	2.2e-3	1 млн	2.0e-3	50 млн	1.2e-4

Отметим, наконец, что наличие статистических погрешностей ухудшает оценку градиента целевой функции и затрудняет (если не исключает вообще) оптимизацию исследуемых процессов по данным имитационного моделирования. Все это вынуждает стремиться к максимальному использованию *численных* методов расчета систем и сетей обслуживания.

0.6. Об этой книге

В свете вышеизложенного актуальны *машинные алгоритмы расчета* стационарных режимов в СМО и СеМО, которые, вопреки мнению М. Кульгина и его единомышленников, не удается «свести к нескольким строкам программного кода». В их основу положены:

- аппроксимации вероятностных распределений по методу моментов [92, 188] — глава 1;
- систематическое использование и развитие «законов сохранения» [92, 170], описанных в главе 3;
- модификация и обобщения итерационной [265], рекуррентной [7] и матрично-геометрической [193] схем расчета многофазных систем (глава 7);
- работы [51, 55, 68, 86], излагающие теорию приоритетных систем;
- обзоры [9, 36, 41, 42, 165] теории сетей обслуживания.

Этот материал не претендует на давно ставший невозможным исчерпывающий охват проблемы и содержит преимущественно методы и зависимости, реализованные в разработанном автором пакете программ. Первые версии этого пакета были описаны в [71, 91]. Информация о последних — на ПЛ/1 для ЕС ЭВМ и Фортране 77 для IBM-совместимых ПЭВМ — приводится в главах 12 и 13.

Книга дает преподавателю возможность варьировать изучаемый материал по объему, глубине и уровню изложения (вопросы исследования сходимости и сопоставления методов будут уместны лишь в дисциплинах повышенного типа). Подчеркнем, что рассмотрение детально хотя бы одного доказательства помогает демистифицировать теорию, развить

некоторое чутье и использовать аналогичные приемы при решении других задач [196, с. 36]. Частные модели могут быть предложены как темы для курсовых проектов, дипломных работ и магистерских диссертаций.

Теория очередей, как и любая отрасль прикладной математики, требует *доведения результатов до числа*. Это обстоятельство должно заставить исследователя заботиться о верификации предлагаемых им методов, которые, бывает, при всей своей внешней заманчивости не выдерживают испытания расчетом: идея оказывается неправильной или неэффективной (как в случае расчета знаменателя матрично-геометрической прогрессии по Ньютону). Первоначальная неудача ставит новую и конкретную проблему и стимулирует дальнейший прогресс; успех вдохновляет, укрепляет доверие к полученным результатам, повышает шансы на публикацию и внедрение. К сожалению, в последние десятилетия бездумное увлечение «информационными технологиями» и демагогия о «непрограммирующих пользователях» заметно ухудшили программистскую подготовку будущих математиков и инженеров в их основной массе. *Прикладную математику нельзя изучать в отрыве от программной реализации ее методов*, с помощью которой гораздо лучше уясняются смысл теории, логика алгоритмов и ограничения на их применение [118]. Работа в области прикладной математики без уверенного программирования не только бессмысленна, но и опасна как для общества, так и для самого горе-специалиста. Практическое программирование является идеальным средством развития логического мышления, поскольку

> «Именно программисты непосредственно упираются в пределы человеческого познания в виде алгоритмически неразрешимых проблем и глубоких тайн работы головного мозга.
>
> Собственный стек (магазин) программиста должен быть глубины не в 5–6 позиций, как это обнаружили психологи у среднего человека, а той же, что и стек в его очередной задаче, подлежащей программированию, плюс еще две–три позиции.
>
> Программист должен обладать способностью первоклассного математика к абстракции и логическому мышлению в сочетании с эдисоновским талантом сооружать все, что угодно, из нуля и единицы. Он должен сочетать аккуратность бух-

> галтера с проницательностью разведчика, фантазию автора детективных романов с трезвой практичностью экономиста.
>
> Необходимость постоянного преодоления этих трудностей делает труд программиста весьма нелегким, но в то же время придает ему захватывающий интерес. Программирование обладает богатой, глубокой и своеобразной эстетикой, которая является основой внутреннего отношения программиста к своей профессии и служит источником интеллектуальной силы, ярких переживаний и глубокого удовлетворения.
>
> В своей творческой природе программирование идет намного дальше большинства других профессий, приближаясь к математике и писательскому делу.
>
> Очень важным эстетическим принципом программирования является его высочайшая требовательность к законченности продукта. Эта стопроцентность программирования — источник его трудности и в то же время величайшей удовлетворенности сделанной работой. В минуты, когда снабженная новой программой машина ведет себя разумно, программист видит материальное воплощение своих интеллектуальных усилий, становящихся отныне общим достоянием. Это торжество интеллекта — наверное, самая сильная и специфическая сторона программирования».

(А. П. Ершов. О человеческом и эстетическом факторах в программировании. — Кибернетика, 1972, № 5).

В связи с этим автор настоятельно рекомендует читателю запрограммировать и отладить заинтересовавшие его расчетные методики на современном Фортране. Выбор именно этого языка определяется «категорическим императивом» наличия в нем

- встроенных средств работы с комплексными переменными,
- многомерных динамических массивов,
- внутренних процедур.

Потребность во внутренних процедурах порождается большим объемом промежуточных результатов счета, которые удобнее передавать через глобальные переменные.

Обратим особое внимание читателей на рациональную *технику программирования*. Из общих требований к последней отметим:

- примат надежности программы перед всеми прочими требованиями к ней;
- сохранение резервной копии исходного текста до завершения отладки улучшенного;
- эффективность программы (чистка циклов, рекуррентный счет, запроцедуривание как средство минимизации объема самой программы; экономия памяти — например, размещение обратных матриц на месте исходных);
- наглядность (достаточность, своевременность и смысловая насыщенность комментариев, структурированная запись, сохранение в программе исходных обозначений алгоритма);
- потребность в разнообразных средствах вычислительной математики;
- зависимость программы не только от алгоритма, но и от структуры данных, которая в ряде случаев является вынужденной и весьма нетривиальной;
- необходимость защиты как от некорректных исходных данных (перегрузка системы, отрицательные дисперсии распределений, особые случаи, ненормированность строк матриц вероятностей переходов), так и от недопустимых промежуточных результатов;
- постоянную озабоченность проблемой тестирования программ и разнообразие способов тестирования;
- системный подход к созданию программных комплексов и гибкую технологию их использования посредством сборочного программирования.

Поскольку работа над книгой продолжалась около 40 лет, данные по трудоемкости счета окрашены «исторической динамикой»: ссылками на стационарные машины от М-220 и ЕС-1030 до ЕС-1066 и позднее — на ПЭВМ от «четверки» до Pentium IV. Это обстоятельство не влияет на результаты сравнения альтернативных алгоритмов, всякий раз выполняемого при *общей* базе.

Кратко изложим содержание глав. В $первой$ главе излагаются элементы теории и аппроксимации вероятностных распределений (раметим, что вероятностные модели сложных явлений — это всегда аппроксимации [164]). Здесь обсуждаются показательное распределение, его марковское свойство и проистекающие из этого расчетные преимущества. Далее обосновывается необходимость аппроксимации одних распределений другими, аргументируется принцип аппроксимации (метод моментов), вводятся коэффициенты немарковости для указания степени отличия заданного распределения от показательного с тем же средним. Рассмотрена связь преобразования Лапласа—Стилтьеса (ПЛС) с моментами распределения. Обсуждается технология свертки непрерывных распределений в моментах и свертки дискретных распределений. Остальная часть главы посвящена выводу расчетных соотношений для аппроксимаций распределениями фазового типа, в том числе с помощью матричных представлений, а также гамма-плотностью и распределением Вейбулла с поправочными многочленами.

Перечисленные распределения закладывают «технологическую основу» работы: фазовые обеспечивают марковизацию переходов в СМО; гамма-плотность служит для быстрого вычисления факториальных моментов — коэффициентов вида

$$q_j = \int_0^\infty \frac{(\lambda t)^j}{j\,!} e^{-\lambda t}\, dB(t), \qquad j = 0, 1, \ldots$$

и, в частности, преобразований Лапласа; распределение Вейбулла позволяет легко табулировать дополнительную функцию распределения.

Во $второй$ главе описана математическая модель ТМО. Здесь определены основные понятия, связанные с потоком заявок, устанавливаются свойства простейшего потока и его особая роль в ТМО. Далее рассматривается проблема расчета распределения $\{q_j\}$ числа заявок простейшего потока за случайный интервал времени и ее обобщение на случай потока $пачек\ заявок$ случайного объема. Обсуждается необходимая для расчета сетей обслуживания технология случайного прореживания и суммирования рекуррентных потоков (последняя — в двух вариантах). На числовых примерах демонстрируются известные предельные тенденции сходимости результирующего потока к простейшему. В то же время показано, что такая сходимость может быть весьма медленной и традиционное представление о простейшем результирующем

потоке — несостоятельным. В этой главе обсуждаются также остальные элементы математической модели ТМО: каналы обслуживания (в частности, суммирование потоков заявок, обслуженных многоканальной системой заявок), организация и продвижение очереди, показатели эффективности работы СМО.

В *третьей* главе формулируются законы сохранения для стационарных режимов в СМО и получены важные следствия из них. В частности, из закона сохранения требований выводятся формула для вероятности свободного состояния однолинейной СМО и обобщающее ее соотношение между начальными вероятностями, средней интенсивностью потока и средней длительностью обслуживания в многоканальной системе. На основании закона сохранения стационарной очереди установлена общая связь между распределением числа заявок в системе перед прибытием очередной заявки и распределением времени ожидания (пребывания) в системе, которая затем конкретизируется для показательного распределения интервалов между заявками. Из закона сохранения объема работы непосредственно выведена формула Полячека—Хинчина для среднего времени ожидания заявки в системе $M/G/1$.

В *четвертой* главе описана традиционная расчетная схема Эрланга для марковской системы с параметрами входящего потока и процесса обслуживания, зависящими от числа заявок в системе, и приводятся формулы для стационарного распределения числа заявок в системах $M/M/n$ и $M/M/1$. Далее дается непосредственный вывод тех же результатов с применением законов сохранения, демонстрирующий корректность и эффективность этого подхода и создающий психологическую основу для его широкого использования в дальнейшем — при рассмотрении более сложных случаев. Решается задача о выборе экономически оптимального быстродействия, исследуется чувствительность к коэффициенту загрузки. Обсуждается неожиданный масштабный эффект одновременного увеличения интенсивностей потока и обслуживания. Глава заканчивается обзором (на уровне идей) основных методов марковизации систем: линейчатых марковских процессов (введение дополнительной переменной), вложенных цепей Маркова и фиктивных фаз.

Пятая глава посвящена методу вложенных цепей Маркова. Здесь указаны этапы метода, на основе законов сохранения предлагается общий принцип установления связи между финальными вероятностями состояний вложенной цепи Маркова и стационарным распределением числа

заявок в системе. Далее выводятся все необходимые соотношения для расчета систем $M/G/1$ (разомкнутой, с порогом включения и с разогревом, с неординарным входящим потоком, замкнутой), $GI/M/1$, $GI/M/n$, $GI/E_k/1$ и $E_k/G/1$ и отмечается роль предложенной в главе 1 аппроксимации распределений гамма-плотностью с поправочным многочленом в численном решении этих задач. На примерах иллюстрируется влияние вида распределения на конечные результаты.

В *шестой* главе описаны методы расчета многоканальных систем с простейшим и эрланговским входящим потоком и детерминированным обслуживанием, основанные на известном подходе Кроммелина, и обсуждается техника их численной реализации.

В *седьмой* главе излагается расчет многоканальных систем фазового типа. Глава начинается с обсуждения иллюстрирующих идею метода фиктивных фаз диаграмм перехода и записанных на их основе векторно-матричных уравнений глобального баланса переходов. Затем разбираются два метода их решения: итерационный для векторов условных вероятностей и матрично-геометрической прогрессии; даны их сравнительная оценка и рекомендации по применению.

В этой же главе рассмотрена специфическая система $M/\overrightarrow{M_k}/\overrightarrow{n_k}$ (с каналами нескольких типов, различающихся интенсивностью обслуживания). Описан расчет необходимого для анализа сетей распределения интервалов между обслуженными заявками в терминах преобразований Лапласа и непосредственно в моментах. На числовых примерах иллюстрируются справедливость теоремы Берке о выходящем потоке марковской системы, зависимость результатов от количества и загрузки каналов, возможность заметного отличия выходящих потоков от простейших, линейная зависимость коэффициентов немарковости выходящих потоков от аналогичных коэффициентов входящих. Рассмотрен расчет систем обслуживания с прибытием заявок пачками случайного объема.

В *восьмой* главе описаны методы расчета временны́х характеристик СМО по распределению числа заявок в ней. Здесь рассмотрен известный метод накопления сверток для показательно распределенной длительности обслуживания и его матричное обобщение на многофазные распределения. Разработаны варианты расчетной схемы «интегрального» метода, основанного на законе сохранения стационарной очереди. Описан новый метод пересчета, в основу которого положена гипотеза о зависимости типа распределения условного распределения времени ожидания в системе $GI/G/n$ только от типа распределения времени обслужива-

ния, причем характер зависимости определяется формулой Полячека—Хинчина в терминах коэффициентов немарковости. Проведено сравнение методов и указаны области их предпочтительного применения. Результаты применения сопоставлены с известными из литературы [149] приближенными оценками среднего времени ожидания и показано крайне низкое качество последних. Предложена и проверена экспериментально новая формула для системы $M/G/n$, приемлемая для прикидочных расчетов. Обсуждается подход к анализу временны́х характеристик бесприоритетного обслуживания неоднородного потока заявок, основанный на разделении общего ожидания и дифференцированных по типам заявок длительностей чистого обслуживания. Рассмотрен метод расчета временных характеристик СМО со $случайным$ выбором на обслуживание заявок из очереди.

Глава заканчивается описанием способов решения двух наиболее актуальных нестационарных задач ТМО: процесса накопления очереди при перегрузке системы и ее последующего рассасывания.

$Девятая$ глава содержит элементарный вывод основных расчетных соотношений для среднего времени ожидания и пребывания в одноканальной системе с относительным, абсолютным (с дообслуживанием прерванной заявки) и смешанным (в схеме классов) приоритетами. Предложен итерационный алгоритм решения этой задачи для ограниченных популяций заявок. Обсуждается задача с динамическими (линейно возрастающими по времени ожидания) приоритетами.

Далее приводится обоснование схем расчета преобразований Лапласа распределений и моментов распределения времени пребывания в системе для упомянутых и более сложных дисциплин обслуживания. Сопоставляются результаты их численной реализации. В заключительном разделе главы разработан приближенный метод расчета в средних $многоканальных$ приоритетных систем.

В $десятой$ главе рассмотрены идея квантованного обслуживания и простейшая циклическая схема при экспоненциально распределенной трудоемкости. Далее обсуждаются: схема усреднения вероятности возврата, позволяющая перейти к произвольно распределенной трудоемкости; режим разделения процессора; многоуровневая схема (дисциплина FB_N и ее варианты).

$Одиннадцатая$ глава посвящена сетям обслуживания. В ней дается обзор основных применений, понятий и определений, отмечается центральная роль проблемы декомпозиции сети и приведены условия

строгой декомпозиции (теорема BCMP), при которых распределение вероятностей состояний сети имеет мультипликативную форму. Кратко описаны метод средних и метод свертки по Бузену в их простейших вариантах.

Далее формулируется подход к немультипликативным сетям. Здесь перечисляются исключающие мультипликативность особенности задачи; общие требования к расчетным методам (достаточная точность, умеренная трудоемкость и робастность); рассмотрены способы их обеспечения. Затем обсуждаются реализующие их итеративные методы приближенной декомпозиции сети. Подробно описан алгоритм расчета разомкнутой (однородной и неоднородной) сети, в том числе с учетом преобразования потоков. Далее излагается подход к расчету замкнутых (однородных и неоднородных) и смешанных сетей, основанный на аппроксимации среднего времени ожидания в узлах предложенным в главе 8 эмпирическим обобщением формулы Полячека—Хинчина. Обсуждаются способы учета блокировок и одновременного использования нескольких ресурсов.

Проблема расчета распределения времени пребывания заявки в сети решается на двух уровнях. Для расчета среднего времени пребывания указаны варианты формулы Литтла применительно к разомкнутой и замкнутой сетям. Высшие моменты (в предположении о независимости распределений времени пребывания в узлах) предлагается определять численным дифференцированием соответствующего ПЛС, для которого получено компактное представление через матричные произведения. Обсуждаются неоднородный вариант расчетной схемы, в том числе для приоритетного обслуживания в узлах, и техника моделирования процессов с подзадачами, отражающая процессы ветвления и соединения заданий. Сообщаются начальные сведения о G-сетях с «отрицательными заявками» и *пространственных* процессах обслуживания. Глава заканчивается обсуждением задач оптимизации сети по минимуму среднего времени пребывания при ограничении на стоимость.

В *двенадцатой* главе из разнообразия, сложности и необходимости решения практических задач расчета очередей выводится необходимость создания соответствующих пакетов прикладных программ (ППП); приводится типовая структура подобных пакетов; дается краткий обзор истории создания пакетов для расчета систем с очередями по данным зарубежной и отечественной печати. Далее на уровне проблемного программиста описан ППП МОСТ, разработанный автором на языке ПЛ/1 для ЕС ЭВМ и переведенный на Фортран 90 для IBM-совместимых ПЭВМ

(180 процедур и 120 тестов). Работа с пакетом организуется по технологии сборочного программирования. Указаны возможности сокращенной версии «персонального» МОСТа для непрофессионалов, которая в процессе диалога с пользователем автоматически строит вызывающую нужные модули пакета главную Фортран-программу.

В *тринадцатой* главе описана идеология тестирования пакета и приведены результаты его — в частности, средняя длина очереди в различных СМО, полученная с помощью разных процедур при исходных данных из пересечения областей их применения. Приводится описание нескольких лабораторных работ по ТМО, проводившихся автором на базе пакета.

В *список литературы* включены 277 источников, в том числе доступных через Интернет полнотекстовых электронных книг, изданных до 2010 г. включительно.

Все главы книги содержат не публиковавшиеся в общедоступной монографической (а в ряде случаев — даже в периодической) литературе результаты, полученнные *лично автором*. К их числу относятся:

- введение и использование «коэффициентов немарковости» распределений, расчетная схема варианта гиперэрланговской аппроксимации, представление непрерывной плотности распределения гамма-плотностью с поправочным многочленом, способ расчета преобразования Лапласа и его обращения в моментах, теория аппроксимации ДФР посредством ДФР Вейбулла с поправочным многочленом; оценка влияния числа и порядка выравниваемых моментов на аппроксимацию ДФР; применение алгоритма Хэмминга для подбора H_k-аппроксимаций для $k \geq 2$ — гл. 1;

- эффективные алгоритмы расчета распределения $\{q_j\}$ числа событий простейшего и обобщенного простейшего потока за случайный интервал времени; алгоритм суммирования потоков; иллюстрация скорости сходимости результирующего потока к простейшему в процессе суммирования и прореживания исходных рекуррентных потоков; расчет интервалов между обслуживаниями в многоканальной системе для общего случая; пример комплексного подхода к выбору показателя эффективности СМО в задаче о восстанавливаемом ЗИП — гл. 2;

- вербальная формулировка законов сохранения; основное уравнение закона сохранения стационарной очереди — гл. 3;

- альтернативный расчет системы $M/M/1$ с применением законов сохранения, иллюстрация метода линейчатых процессов, дифференциальные версии «законов сохранения» — гл. 4;

- обоснование с помощью законов сохранения общего принципа перехода от финальных вероятностей вложенной цепи к стационарным вероятностям состояний; расчет системы с порогом включения и разогревом; новые методы расчета систем с потоком групповых заявок, $GI/E_k/1$ и $E_k/G/1$; выявление и разработка особого случая в алгоритме Такача для $GI/M/n$; иллюстрация влияния дисперсии и высших моментов распределений на результаты расчета — гл. 5;

- численный метод расчета $E_k/D/n$ — гл. 6;

- практически все расчетные схемы главы 7, их анализ и оптимизация — кроме общих идей итерационного метода (Такахаси—Таками) и метода матрично-геометрической прогрессии (Ивэнс), расчета ПЛС распределения интервалов между обслуженными заявками для простейшей модели $M/M/n$ (Хомоненко); оценка эффектов масштабирования и дробления производительности;

- «интегральный» метод и метод пересчета для вычисления временных характеристик обслуживания, сопоставление приближенных оценок среднего времени ожидания, приближенное обобщение формулы Полячека—Хинчина на многоканальный случай, общий алгоритм расчета временны́х характеристик СМО при случайном выборе заявок из очереди — гл. 8;

- расчет в средних многоканальных приоритетных систем, в том числе замкнутых и с динамическим приоритетом; исправленная формула для расчета среднего времени ожидания j-заявки в системе с динамическим приоритетом, формулы для высших моментов распределения периода непрерывной занятости (в том числе с разогревом), схема расчета приоритетных режимов обслуживания в моментах, упрощение расчета ПЛС активного времени для дисциплины RW, иллюстрация влияния типов приоритет-

ных дисциплин на показатели обслуживания, приближенные методы расчета многоканальных приоритетных систем — гл. 9;

- алгоритм усреднения, позволяющий применить циклическую схему к произвольным распределениям времени обслуживания; новая схема расчета многоуровневой системы с учетом системных потерь и ее верификация — гл. 10;

- обоснование требований, обеспечивающих приемлемую точность, простоту и робастность приближенных методов декомпозиции сетей обслуживания; потокоэквивалентные алгоритмы расчета разомкнутых сетей с преобразованием потоков, замкнутых и смешанных сетей; расчет высших моментов распределения времени пребывания заявки в сети; процессы с подзадачами; оптимизация расчетной матрицы — гл. 11;

- весь текст глав 12 и 13.

Книга написана на основе курса лекций, многократно читанного автором в Военно-космической академии им. А. Ф. Можайского. Все предлагаемые методы проверены на ЭВМ — см. главу 12.

Книга предназначена для специалистов в области исследования операций, системного анализа, связи, вычислительного дела, городского хозяйства и может быть использована для научно-исследовательских и проектных работ на стадии системного проектирования, для прогнозирования работы существующих систем с очередями при изменении условий их функционирования; при изучении ТМО в вузах с повышенной математической подготовкой и на курсах переподготовки инженеров.

Компьютерный набор книги выполнен автором в издательской системе MiK$\TeX$.

Глава 1

Элементы теории и аппроксимации распределений

В порядке комментария к названию данной главы (и к содержанию всей книги) отметим, что «вероятностные модели случайных явлений — всегда аппроксимации» [164, с. 175]. Выбор конкретного распределения рождается в сопоставлении преимуществ точного анализа и наших возможностей. Подчеркнем, что целью любых вероятностных расчетов является прогнозирование *не отдельных событий, но их вероятностей*.

Практические проблемы теории очередей обычно связываются с достижением высоких вероятностей своевременного решения задачи, мало чувствительных к изменению параметров. В таких случаях удобнее работать с *дополнительными* функциями распределения (ДФР)

$$\bar{F}_T(t) \stackrel{\text{def}}{=} P(T > t).$$

Индекс имени случайной величины там, где это не ведет к неоднозначности, мы будем опускать.

1.1. Равномерное и треугольное распределения

На первый взгляд простейшим является распределение, равномерное на отрезке $[a-l, a+l]$. Во всяком случае, именно оно обычно постулируется при недостатке реальных данных и фигурирует в большей части примеров из руководств к популярной системе имитационного моделирования GPSS. Плотность этого распределения равна $1/(2l)$ внутри упомянутого интервала и нулю — вне его. Начальные моменты равномерного распределения

$$f_k = \frac{(a+l)^{k+1} - (a-l)^{k+1}}{2l(k+1)}, \qquad k = 1, 2, \ldots$$

Его дисперсия $D = l^2/3$.

Можно показать, что сумма двух равномерно распределенных величин имеет треугольное распределение (Симпсона), границы которого равны суммам левых и правых границ слагаемых соответственно. В симметричном случае для отрезка $[a-l, a+l]$ максимальное значение плотности в точке a равно $1/l$. Начальные моменты этого распределения

$$f_k = \frac{(a+l)^{k+2} + (a-l)^{k+2} - 2a^{k+2}}{l^2(k+1)(k+2)}, \qquad k = 1, 2, \ldots,$$

а его дисперсия $D = l^2/6$.

Обсуждаемые в этом разделе распределения имеют *конечный* размах. Это позволяет применять их для описания рекуррентных потоков с технологическими ограничениями на интервалы между смежными заявками.

Остальные используемые в данной книге распределения определены на полуоси $[0, \infty)$. Это обстоятельство в дальнейшем дополнительно не оговаривается.

1.2. Показательное распределение

Показательным (экспоненциальным) называется распределение с плотностью

$$f(t) = \mu e^{-\mu t}, \tag{1.2.1}$$

дополнительной функцией распределения

$$\bar{F}(t) = e^{-\mu t} \tag{1.2.2}$$

и начальными моментами

$$f_k = k\,!/\mu^k, \qquad k = 1, 2, \ldots \tag{1.2.3}$$

Примем для определенности, что речь идет о процессе обслуживания. Условная плотность распределения длительности остатка обслуживания

$$\tilde{f}(t|\tau) = \frac{f(t+\tau)}{\bar{F}(\tau)} = \frac{\mu e^{-\mu(t+\tau)}}{e^{-\mu\tau}} = \mu e^{-\mu t}$$

независимо от уже истекшей длительности обслуживания. Говорят, что показательное распределение обладает *марковским* свойством — отсутствием последействия (памяти). Параметр потока обслуживаний с функцией распределения $B(t)$

$$\mu(\tau) \stackrel{\text{def}}{=} \lim_{\Delta t \to 0} \frac{B(\tau+\Delta t) - B(\tau)}{\bar{B}(\tau)\Delta t} = -\frac{\bar{B}'(\tau)}{\bar{B}(\tau)} \tag{1.2.4}$$

в случае показательно распределенной длительности равен μ и *не зависит* от уже истекшего времени τ. Соответственно вероятность завершения обслуживания на малом интервале длины Δt

$$P(t, t+\Delta t) = \mu\Delta t + o(\Delta t)$$

не зависит от положения этого интервала на оси времени. Отмеченное уникальное свойство показательного распределения делает его исключительно удобным в аналитических выкладках, связанных с описанием процессов обслуживания — в отличие от всех прочих распределений.

Можно показать, что минимум из n показательно распределеннных величин также имеет показательное распределение с суммарным параметром, причем условная вероятность реализации конкретной величины равна доле ее параметра в суммарном. Далее мы будем систематически опираться на это свойство — например, при решении задач о выходящем потоке.

1.3. Гамма-распределение

Гамма-распределение имеет плотность

$$f(t) = \frac{\mu(\mu t)^{\alpha-1}}{\Gamma(\alpha)} e^{-\mu t}, \tag{1.3.1}$$

где $\Gamma(\alpha)$ — гамма-функция со свойствами

$$\begin{array}{lcl} \Gamma(\alpha) & \stackrel{\text{def}}{=} & \int\limits_0^\infty x^{\alpha-1} e^{-x}\, dx, \\ \Gamma(1) & = & 1, \\ \Gamma(\alpha+1) & = & \alpha\Gamma(\alpha). \end{array} \tag{1.3.2}$$

Показательное распределение является его частным случаем при $\alpha = 1$.

Моменты Г-распределения вычисляются по формулам

$$f_k = \alpha(\alpha+1)\ldots(\alpha+k-1)/\mu^k, \qquad k = 1, 2, \ldots \tag{1.3.3}$$

Его дисперсия $D = f_2 - f_1^2 = \alpha/\mu^2$.

1.4. Аппроксимация распределений

Выравнивание статистических распределений, т. е. подбор теоретических зависимостей, описывающих фактически наблюдавшиеся данные, обычно проводится при условии сохранения значений интегральных числовых характеристик распределений, аккумулирующих его основные свойства. В качестве таких характеристик чаще всего выступают начальные моменты

$$f_i \stackrel{\text{def}}{=} \int\limits_0^\infty t^i f(t)\, dt, \qquad i = 1, 2, \ldots$$

Как отмечается в [47, с.127], «можно ожидать, что если два распределения имеют некоторое число одинаковых моментов, то они в какой-то степени схожи и аппроксимируют друг друга. Обоснованием этого суждения является совпадение их аппроксимаций по методу наименьших квадратов при конечном размахе и разложений по ортогональным многочленам — при бесконечном ... Практически аппроксимация такого рода оказывается очень хорошей, даже если совпадают только первые три или четыре

момента». В частности, двухпараметрическое гамма-распределение позволяет обеспечить точное выравнивание $\textit{двух}$ моментов, если положить

$$\mu = f_1/D, \qquad \alpha = \mu f_1. \tag{1.4.1}$$

Наращивание числа учитываемых моментов не всегда сопровождается монотонным улучшением качества аппроксимации. Статистические оценки моментов реальных распределений состоятельны, но могут быть смещенными.

Существует обширная литература (см., например, [4]) по проблеме моментов: определению условий, при которых заданный набор чисел может служить моментами некоторого распределения (требуется гарантировать нормировку к единице и неотрицательность плотности $f(t)$ для всех t). Если определитель

$$\begin{vmatrix} 1 & \mu_1 & \dots & \mu_n \\ \mu_1 & \mu_2 & \dots & \mu_{n+1} \\ \dots & \dots & \dots & \dots \\ \mu_n & \mu_{n+1} & \dots & \mu_{2n} \end{vmatrix} \le 0$$

для всех целых n, то существует распределение с моментами $\{\mu_i\}$ (в случае строгого неравенства — без точек концентрации). Для единственности распределения на положительной полуоси установлено $\textit{достаточное}$ условие (Карлемана)

$$\sum_{i=0}^{\infty} \mu_i^{-1/2i} = \infty.$$

Если распределение имеет плотность, то необходимо и достаточно условие Крейна

$$\int_{-\infty}^{\infty} \frac{\ln(f(x))}{1+x^2}\, dx = -\infty.$$

Логнормальное распределение с моментами $\mu_r = \sigma^2 r^2/2$ не удовлетворяет обоим этим условиям. В [192, с. 693–694] приводятся семейство плотностей Хейде (Heyde) и плотность Кендалла—Стьюарта $f(x) = \frac{1}{4} e^{-\sqrt{|x|}}(1 + a\cos(\sqrt{|x|})$ — с параметром $a \in [-1, 1]$, а также еще две $\textit{различных}$ плотности с совпадающими моментами.

Отмеченные обстоятельства дают некоторые основания характеризовать метод моментов как «наивный» и «рискованный» — [232, 233]

(риск порождается работой с распределениями, имеющими ограниченное число конечных моментов — например, с распределением Парето).Но критики этого метода ничего не предлагают взамен. Известно, что оценки параметров распределений на основе фактической статистики по методу моментов не являются эффективными (т. е. имеющими минимально возможную дисперсию) и в этом смысле уступают методу максимального правдоподобия. Однако последний в $теоретических$ выкладках неприменим вообще. При всех своих недостатках (чаще потенциальных, чем реальных) метод моментов остается наиболее удобным в работе и обычно применяемым методом подбора распределений и замены одного распределения другим. Количество сохраняемых моментов рассматривается как порядок аппроксимации. Оно равно числу свободных параметров теоретической кривой. Очевидно, что на значения высших моментов наибольшее влияние оказывает поведение плотности распределения при больших аргументах. Соответственно при построении по моментам аппроксимации ДФР следует учитывать моменты возможно более высокого порядка (в разделе об H_2-аппроксимации мы покажем это на числовых расчетах).

Существуют методы, позволяющие автоматически выбрать тип выравнивающего распределения в заданном семействе (например, одну из кривых К. Пирсона — по четырем моментам [88]). Критерии согласия (χ^2 Р. Фишера, $n\omega^2$ Н. В. Смирнова, А. Н. Колмогорова) статистических и теоретических распределений с несколькими общими моментами мало чувствительны к виду распределений. Это позволяет в классе функций с заданными несколькими моментами выбирать дающие те или иные вычислительные преимущества — например, допускающие эффективное вычисление ДФР или облегчающие расчет функционалов специального вида. Именно такие распределения и будут обсуждаться в дальнейшем.

Поскольку аппроксимирующими распределениями служат показательное и его «ближайшие родственники», для выбора типа аппроксимации важно знать степень отличия исходного распределения от показательного при равенстве средних значений. В первом приближении мерой отличия может служить коэффициент вариации

$$v = \sqrt{D}/f_1, \tag{1.4.2}$$

где D — дисперсия. Для показательного закона $v = 1$. Однако с целью единообразного представления отличий в высших моментах удобно ввести набор $коэффициентов\ немарковости$

$$\xi_i = f_i/f_1^i - i\,!, \qquad i = 2, 3, \ldots \tag{1.4.3}$$

С помощью (1.2.3) легко убедиться, что для показательного распределения все они равны нулю. Второй коэфициент немарковости

$$\xi_2 = v^2 - 1.$$

Значения $\xi_2 > 0$ имеют «сверхслучайные» распределения с аномально большой долей малых значений (плотность имеет J-образную форму с вертикальной асимптотой в нуле). Соответственно чаще встречаются и аномально большие значения (плотность имеет «толстый хвост»). Примером «сверхслучайной» величины является интервал между машинами, следующими по узкому шоссе. Для «субслучайных» распределений $\xi_2 < 0$. В случае регулярных (детерминированных) процессов $\xi_2 = -1,\ \xi_3 = -5$.

Для вычисления высших моментов распределения по заданному первому и коэффициентам немарковости используется обращенный вариант формулы (1.4.3):

$$f_i = f_1^i(\xi_i + i\,!), \qquad i = 2, 3, \ldots \tag{1.4.4}$$

1.5. Распределения максимума и минимума

Упомянутые задачи приходится решать при суммировании потоков заявок и завершений обслуживания и во многих других ситуациях. Для минимума ДФР есть произведение ДФР, для максимума ФР равна произведению ФР компонент [165, с. 30–31]. Приведем окончательные формулы для симметричного случая:

$$F_{\max}(x) = [F(x)]^n; \qquad \bar{F}_{\min}(x) = [\bar{F}(x)]^n.$$

1.6. Преобразование Лапласа, свертки и моменты распределения

1.6.1. ПЛС и моменты

Задачи теории очередей часто ставятся и/или решаются в терминах преобразований Лапласа-Стилтьеса (ПЛС) от распределений:

$$\varphi(s) \stackrel{\text{def}}{=} \int_0^\infty e^{-st}\, dF(t). \tag{1.6.1}$$

Разлагая экспоненту в степенной ряд и интегрируя почленно, устанавливаем связь между ПЛС и моментами распределения:

$$\varphi(s) = \sum_{i=0}^{\infty} (-s)^i f_i / i\,!. \tag{1.6.2}$$

С другой стороны, нетрудно вывести обратное соотношение:

$$f_i = (-1)^i \varphi^{(i)}(s)\Big|_{s=0}. \tag{1.6.3}$$

Ценную информацию можно получить с помощью предельных формул:

$$\begin{aligned} \lim_{t\to\infty} f(t) &= \lim_{s\to 0} s\varphi(s), \\ \lim_{t\to 0} f(t) &= \lim_{s\to\infty} s\varphi(s). \end{aligned}$$

Многие расчеты (в особенности для приоритетных режимов) удается довести только до получаемых алгоритмически преобразований Лапласа, вследствие чего расчет моментов распределений приходится выполнять численно. Такой расчет основывается на построении интерполяционного многочлена для ПЛС и последующем численном дифференцировании этого многочлена в нуле. Программная реализация определяется видом используемой в этой технологии интерполяционной формулы.

1.6.2. Формула Ньютона

Интерполяционная формула $Ньютона$ имеет вид

$$y(x) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \frac{q(q-1)(q-2)}{3!}\Delta^3 y_0 + \ldots, \quad (1.6.4)$$

где

$$q = x/h. \quad (1.6.5)$$

Вычисление коэффициентов многочленов основано на представлении многочлена при k-й разности в форме

$$P_k(q) = a_{k,0} + a_{k,1}q + a_{k,2}q^2 + \cdots + a_{k,k}q^k,$$

начальном условии $P_1 = q$ и очевидном из (1.6.4) законе пересчета

$$P_k = P_{k-1}\frac{q-k+1}{k} = P_{k-1}\Big[\frac{q}{k} - (1-1/k)\Big], \qquad k = \overline{2,K}.$$

В цикле по k идет собственно вычисление производных. Здесь сначала почленно дифференцируются все еще не обратившиеся в константу многочлены, а затем суммируются произведения их свободных членов на соответствующую разность (все остальные слагаемые при $Q = 0$ обращаются в нуль). Поскольку производные отыскиваются по $x = qh$, для получения каждой очередной производной делитель B умножается на H.

Альтернативным вариантом является безразностная формула конечных разностей. Из формул

$$\begin{aligned} \Delta_0 &= y_1 - y_0, \\ \Delta_1 &= y_2 - y_1, \end{aligned}$$

следует

$$\Delta_0^2 = \Delta_1 - \Delta_0 = y_2 - 2y_1 + y_0.$$

Соответственно

$$\Delta_1^2 = \Delta_2 - \Delta_1 = y_3 - 2y_2 + y_1.$$

Теперь ясно, что

$$\Delta_0^3 = \Delta_1^2 - \Delta_0^2 = y_3 - 3y_2 + 3y_1 - y_0.$$

Продолжая последовательные подстановки, приходим к символической формуле

$$\Delta_0^k = (y-1)^k,$$

в которой после развертывания бинома показатели степени переносятся в индексы.

Преимуществами этого подхода являются возможность обойтись одномерным массивом разностей и исключение вычитания близких чисел при малом шаге построения таблицы. Этап дифференцирования не меняется.

В основу процедуры DIFSTIR положена интерполяционная формула Стирлинга

$$\begin{aligned} y(q) &= y_0 + \frac{q}{1!}\tilde{\Delta}^1 + +\frac{q^2}{2!}\tilde{\Delta}^2 + \frac{q(q^2-1^2)}{3!}\tilde{\Delta}^3 + \frac{q(q^2-1^2)}{4!}\tilde{\Delta}^4 \\ &+ \frac{q(q^2-1^2)(q^2-2^2)}{5!}\tilde{\Delta}^5 + \frac{q(q^2-1^2)(q^2-2^2)}{6!}\tilde{\Delta}^6 + \dots, \end{aligned}$$

в которой $q = x/h$, а центральные разности $\{\tilde{\Delta}^i\}$ выражаются через обычные начальные формулами

$$\tilde{\Delta}^i = \begin{cases} \Delta^i_{-i/2}, & \text{если } i \text{ четное,} \\ [\Delta^i_{(1-i)/2} + \Delta^i_{-(1+i)/2}], & \text{если } i \text{ нечетное.} \end{cases} \tag{1.6.6}$$

Для полиномов при нечетных разностях

$$A_{k+2}(q) = A_k(q)\frac{q^2-(k-1)^2/4}{(k+1)(k+2)}.$$

Полагая

$$\begin{aligned} d_{k+2} &= 1/((k+1)(k+2)), \\ e_{k+2} &= (k-1)^2/(4(k+1)(k+2)), \end{aligned} \tag{1.6.7}$$

выводим рекуррентные зависимости

$$a_{k+2,j} = \begin{cases} d_{k+2}a_{k,k}, & j = k+2, \\ d_{k+2}a_{k,j-2} - e_{k+2}a_{k,j}, & j = 3,5,\dots, \\ -e_{k+2}a_{k,1}, & j = 1. \end{cases} \tag{1.6.8}$$

Здесь $k = 1,3,\dots$.

При любом k все коэффициенты при четных степенях q равны нулю. Необходимые для реализации формул (1.6.8) начальные значения

$$a_{1,1} = 1, \qquad a_{1,0} = 0.$$

Для полиномов при четных разностях имеет место пересчет по закону

$$A_{k+2}(q) = A_k(q)\frac{q^2 - (k/2-1)^2}{(k+1)(k+2)}.$$

Зависимость между коэффициентами смежных полиномов дается формулами

$$a_{k+2,j} = \begin{cases} d_{k+2}a_{k,k}, & j = k+2, \\ d_{k+2}a_{k,j-2} - e_{k+2}a_{k,j}, & j = 4, 6, \ldots k, \\ -e_{k+2}a_{k,2}, & j = 2. \end{cases} \tag{1.6.9}$$

Для d_{k+2} сохраняется формула (1.6.6), а

$$e_{k+2} = (k-2)^2/[4(k+1)(k+2)].$$

Начальные коэффициенты здесь

$$a_{2,2} = 1/2; \qquad a_{2,0} = a_{2,1} = 0.$$

В процессе отыскания производных все упомянутые полиномы дифференцируются почленно по q. Поскольку производные определяются при $q = 0$, суммируются свободные члены производных. Результат для получения m-й производной делится на h^m.

1.6.3. Тестирование процедур дифференцирования

Процедуры численного дифференцирования тестировались на задаче расчета моментов показательного распределения с ПЛС вида $\mu/(\mu + s)$, вычисленным в 9 точках. В ходе тестирования было определено рациональное значение шага построения таблицы ПЛС $h \approx 10^{-3}\mu$, при котором распределение погрешностей четырех младших моментов «наиболее гармонично» (табл. 1.1).

Рост относительной погрешности статистических моментов считается пропорциональным порядку момента. Если погрешность 5 % для первого момента признать допустимой, то *все* полученные результаты представляются вполне приемлемыми.

Таблица 1.1. Тестирование процедур численного дифференцирования

№	Точное значение производной	погрешности		
		`difnewt`	`difndif`	`difstir`
1	-.333333333333333e+01	-.760e-12	-.285e-11	-.225e-12
2	.222222222222222e+02	-.431e-08	-.184e-07	-.620e-09
3	-.222222222222222e+03	-.754e-05	-.418e-04	.114e-05
4	.296296296296296e+04	-.517e-02	-.442e-01	.229e-02
5	-.493827160493827e+05	-.122e+01	-.229e+02	-.130e+01
6	.987654320987655e+06	-.143e-02	-.472e+04	-.228e+04
7	-.230452674897119e+08	.286e-01	.193e+06	.409e+06
8	.614540466392319e+09	.460e+01	.145e+09	.668e+09

1.6.4. Свертки распределений в моментах

В задачах ТМО часто приходится выполнять *свертку распределений*, т. е. находить распределение суммы независимых случайных величин. Примером такой задачи служит построение функции распределения времени пребывания заявки в системе $V(t)$ по распределениям $W(t)$ времени ожидания и $B(t)$ чистой длительности обслуживания. В преобразованиях Лапласа

$$\nu(s) = \omega(s)\beta(s) \tag{1.6.10}$$

(ПЛС свертки равно произведению ПЛС составляющих). Свертка может быть выполнена непосредственно в моментах на основе символического разложения

$$v^k = (w + b)^k, \tag{1.6.11}$$

в котором после развертывания бинома показатели степени переводятся в индексы соответствующих моментов. Простота этих соотношений — серьезный аргумент в пользу применения метода моментов.

На основе этой же идеи построена процедура `FASTCONV`, которая (по аналогии с ускоренным возведением в целую степень) выполняет быструю свертку распределения с самим собой, управляемую двоичным разложением кратности.

Данная задача имеет эффектное решение и в виде *рекурсивной* процедуры: на каждом шаге она сводится к свертке двух половинных кратностей для четного аргумента или исходного распределения и кратности на единицу меньшей — для нечетного.

Процедуры тестировались на Г-распределении, для которого свертка порождает также Г-распределение с параметром формы α, умноженным на кратность свертки. Правильность работы второй процедуры автоматически гарантирует правильность первой. Поэтому можно ограничиться тестированием `fastconv`. Обязательно должны проверяться граничный случай $k = 1$, четная и нечетная кратности.

Таблица 1.2. Результаты сверток

Кратность свертки		
1	8	11
.12500e+1	.10000e+2	.13750e+2
.21875e+1	.10500e+3	.19594e+3
.49219e+1	.11550e+4	.28901e+4
.13535e+2	.13283e+5	.44074e+5
.43989e+2	.15939e+6	.69416e+6
.16496e+3	.19924e+7	.11280e+8

Результаты полностью совпали с эталонными значениями.

1.7. Остаточные распределения

В задачах ТМО важную роль играют распределения $остатка$ временного интервала с функцией распределения $B(t)$, отсчитываемого от произвольного момента времени. Очевидно, его плотность

$$f(t + \Delta t) = f(t)[1 - \mu(t)\Delta t] + o(\Delta t),$$

где $\mu(t)$ — мгновенное значение параметра потока, определяемое согласно (1.2.4). Легко получить дифференциальное уравнение

$$f'(t)/f(t) = -\mu(t)$$

с решением

$$f(t) = C \cdot \exp(-\int_0^t \mu(\tau)\, d\tau).$$

Но

$$-\int_0^t \mu(\tau)\,d\tau = -\int_0^t \frac{B'(\tau)\,d\tau}{1-B(\tau)} = \ln[1-B(\tau)],$$

так что $f(t) = C[1-B(t)]$. Интегрируя обе части последнего равенства в пределах от 0 до ∞, из условия $1 = Cb_1$ определяем постоянную C. Итак,

$$f(t) = [1-B(t)]/b_1. \tag{1.7.1}$$

Вычисляя ПЛС от левой и правой частей (1.7.1), получаем

$$\varphi(s) = [1-\beta(s)]/(sb_1). \tag{1.7.2}$$

Разложив ПЛС в этой формуле по моментам соответствующих распределений согласно (1.6.2) и приравнивая коэффициенты при одинаковых степенях s в равенстве $s\varphi(s) = [1-\beta(s)]/b_1$, получаем соотношение

$$f_k = b_{k+1}/[(k+1)b_1], \qquad k = 1, 2, \ldots \tag{1.7.3}$$

В частности,

$$f_1 = b_2/(2b_1). \tag{1.7.4}$$

Для показательного распределения остаточное распределение совпадает с исходным. Прочие распределения могут быть классифицированы на «стареющие» $(f_1 < b_1)$ и «молодеющие» $(f_1 > b_1)$. Иначе они называются распределениями с возрастающей и убывающей функциями интенсивности (ВФИ и УФИ соответственно). Стареющие распределения имеют коэффициент вариации $v < 1 \quad (\xi_2 < 0)$, молодеющие — наоборот.

1.8. Распределения фазового типа

К этим распределениям относятся порождаемые системой подлежащих прохождению фаз обслуживания с показательно распределенной длительностью пребывания в каждой из них. При фиксации номера фазы такие распределения приобретают марковское свойство, что и определяет целесообразность их использования для аппроксимации исходных распределений (в этом случае расщепление на фазы является фиктивным). Идея метода фиктивных фаз была выдвинута еще А. К. Эрлангом.

В [168, с. 177] даны сводка фазовых распределений и выражения для моментов; в [165, с. 21] — таблица распределений и их ПЛС.

Многофазный процесс обслуживания в принципе отличается от сетевого тем, что новая заявка может быть принята в данный канал лишь *после завершения обработки предыдущей*.

1.8.1. Распределение Эрланга

Распределение Эрланга r-го порядка с плотностью

$$f(t) = \frac{\mu(\mu t)^{r-1}}{(r-1)!}e^{-\mu t}, \tag{1.8.1}$$

дополнительной функцией распределения

$$\bar{F}(t) = \sum_{i=0}^{r-1} \frac{(\mu t)^i}{i!}e^{-\mu t} \tag{1.8.2}$$

и моментами

$$f_k = r(r+1)\cdots(r+k-1)/\mu^k, \qquad k = 1, 2, \ldots \tag{1.8.3}$$

предполагает последовательное прохождение r фаз (см. рис. 1.1) и соответственно является r-кратной сверткой показательного закона.

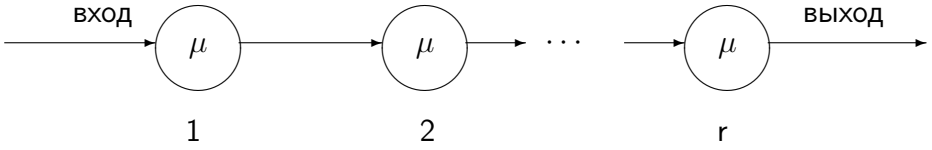

Рис. 1.1. Эрлангово распределение

Эрлангово распределение является частным случаем гамма-распределения при целом параметре r. Это определяет следующую схему подбора его параметров:

1. Найти $\tilde{r} = [f_1^2/D + 0.5]$ (ближайшее целое к указанному отношению).

2. Вычислить вещественный параметр $\mu = \tilde{r}/f_1$.

При этом строго выравнивается первый момент и приближенно — второй.

В [28, с. 180–181] приводятся формулы для комбинации распределений M и E_k, причем целочисленность k не оговаривается.

1.8.2. Обобщенное распределение Эрланга

Обобщенное распределение Эрланга обладает различными задержками в фазах. Для двухфазного распределения «моментные» уравнения

$$\begin{aligned} f_1 &= \frac{1}{\mu_1} + \frac{1}{\mu_2}, \\ f_2 &= \frac{2}{\mu_1^2} + \frac{2}{\mu_2^2} + \frac{2}{\mu_1\mu_2}. \end{aligned}$$

Полагая $x_i = 1/\mu_i, \quad i = \overline{1,2}$, приходим к системе уравнений

$$\begin{aligned} x_1 + x_2 &= f_1, \\ x_1^2 + x_2^2 + x_1x_2 &= f_2/2. \end{aligned}$$

Исключая из этой системы x_2, имеем квадратное уравнение

$$x_1^2 - x_1f_1 + f_1^2 - f_2/2 = 0$$

с решением

$$x_1 = \frac{f_1 \pm \sqrt{2f_2 - 3f_1^2}}{2}, \quad x_2 = f_1 - x_1.$$

Соответственно определяются параметры $\{\mu_i\}$.

Функцию распределения для этого закона мы получим с помощью преобразования Лапласа:

$$F(t) = L^{-1}\left[\frac{1}{s}\Big(\frac{\mu_1}{\mu_1+s}\frac{\mu_2}{\mu_2+s}\Big)\right] = 1 + \frac{1}{\mu_1-\mu_2}\Big(\mu_2e^{-\mu_1t} - \mu_1e^{-\mu_2t}\Big). \tag{1.8.4}$$

1.8.3. Гиперэкспоненциальное распределение

Представление ДФР в виде

$$\bar{F}(t) = \sum_{i=1}^{k} y_ie^{-\mu_it} \tag{1.8.5}$$

позволяет считать исходный процесс проходящим одну из n альтернативных фаз. Здесь $\{y_i\}, \quad 0 \le y_i \le 1, \quad \sum_{i=1}^{n} y_i = 1$, интерпретируются

как вероятности выбора i-й фазы, а $\{\mu_i\}$ — как параметры показательного распределения времени пребывания в ней. Диаграмма гиперэкспоненциального распределения приведена на рис. 1.2.

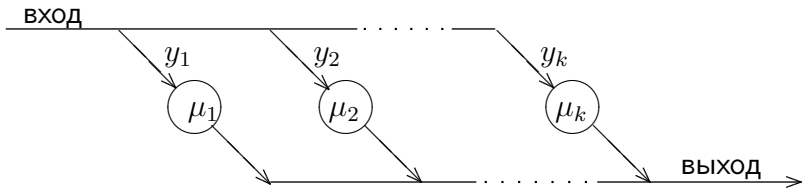

Рис. 1.2. Гиперэкспоненциальное распределение

Среди названных величин могут быть комплексные (попарно сопряженные), что подчеркивает *фиктивный* характер расщепления процесса на фазы. Допустимость таких параметров была впервые отмечена Д. Коксом в 1955 г. [188].

Применение формулы для моментов показательного распределения приводит к системе уравнений относительно параметров аппроксимации

$$\begin{array}{lllll} y_1 & + \; y_2 & +\ldots+ \; y_n & = 1, & \\ y_1/\mu_1 & + \; y_2/\mu_2 & +\ldots+ \; y_n/\mu_n & = \tilde{f}_1, & \\ y_1/\mu_1^2 & + \; y_2/\mu_2^2 & +\ldots+ \; y_n/\mu_n^2 & = \tilde{f}_2/2\,!, & \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ y_1/\mu_1^i & + \; y_2/\mu_2^i & +\ldots+ \; y_n/\mu_n^i & = \tilde{f}_i/i\,!, & i = \overline{0, 2n-1}. \end{array} \tag{1.8.6}$$

Здесь $\tilde{f}_i$ — i-й момент исходного распределения. Введем обозначения $x_i = \mu_i^{-1}$, $\quad f_i = \tilde{f}_i/i!$. Тогда система (1.8.6) переходит в

$$\sum_{j=1}^{n} y_j x_j^i = f_i, \qquad i = \overline{0, 2n-1}. \tag{1.8.7}$$

Для решения этой системы применим метод, описанный в книге Р. В. Хэмминга «Численные методы», с. 139–141. Введем многочлен

$$\pi(x) = \prod_{i=1}^{n} (x - x_i) = \sum_{k=0}^{n} C_k x^k. \tag{1.8.8}$$

Очевидно, для всех i имеет место $\pi(x_i) = 0$, а $C_n = 1$. Умножим первое из уравнений системы (1.8.7) на C_0, второе на C_1, ..., n-е —

на C_n и сложим результаты:

$$\sum_{i=0}^{n} C_i \sum_{j=1}^{n} y_j x_j^i = \sum_{i=0}^{n} C_i f_i = \sum_{j=1}^{n} y_j \sum_{i=0}^{n} C_i x_j^i = \sum_{j=1}^{n} y_j \pi(x_j) = 0.$$

Сдвинем теперь множители $\{C_i\}$ на одну строку вниз:

$$\sum_{i=0}^{n} C_i \sum_{j=1}^{n} y_j x_j^{i+1} = \sum_{i=0}^{n} C_i f_{i+1} = \sum_{j=1}^{n} y_j \sum_{i=0}^{n} C_i x_j^{i+1} = \sum_{j=1}^{n} y_j x_j \sum_{i=0}^{n} C_i x_j^i = 0.$$

Продолжая эти преобразования до сдвига на n строк, приходим к системе уравнений

$$\begin{array}{lcll} \sum\limits_{j=0}^{n} C_j f_{i+j} & = & 0, & i = \overline{0, n-1}, \\ C_n & = & 1, & \end{array} \tag{1.8.9}$$

решаемой одним из стандартных численных методов. Теперь согласно (1.8.8) имеем уравнение

$$\sum_{i=0}^{n} C_i x^i = 0$$

с известными коэффициентами. Решая его, находим корни $\{x_j\}$. Подстановка корней в первые n уравнений из (1.8.7) дает коэффициенты $\{y_j\}$. Наконец, вычисляем все $\mu_j = 1/x_j$.

Проверим это решение для H_2-аппроксимации гамма-распределения с моментами $f_i = \alpha(\alpha+1)\ldots(\alpha+i-1)/(i!\,\lambda^i)$. Система (1.8.9) теперь сводится к

$$\begin{array}{llcl} C_0 & +C_1 f_1 & = & -f_2, \\ C_0 f_1 & +C_1 f_2 & = & -f_3. \end{array}$$

Ее определитель

$$A = \begin{vmatrix} 1 & f_1 \\ f_1 & f_2 \end{vmatrix} = f_2 - f_1^2$$

обращается в нуль при $\alpha = 1$, что исключает стандартное применение метода для показательного распределения — точнее, для распределений с единичным коэффициентом вариации. Далее, при успешном прохождении этого этапа имеем

$$C_0 = \frac{f_1 f_3 - f_2^2}{f_2 - f_1^2}, \quad C_1 = \frac{f_1 f_2 - f_3}{f_2 - f_1^2}.$$

Соответственно корни квадратного уравнения $x^2+C_1x+C_0=0$ подсчитываются согласно

$$x_{1,2}=-\frac{f_1f_2-f_3}{2(f_2-f_1^2)}\pm\sqrt{D}.$$

Дискриминант уравнения

$$D=\left[\frac{f_1f_2-f_3}{2(f_2-f_1^2)}\right]^2-\frac{f_1f_3-f_2^2}{f_2-f_1^2}$$

равен нулю при $\alpha=2$, т. е. для распределения Эрланга второго порядка. В этом случае корни уравнения равны, и на заключительном этапе алгоритма определитель системы

$$\begin{array}{rcrcl} y_1 & + & y_2 & = & 1, \\ x_1y_1 & + & x_2y_2 & = & f_1. \end{array}$$

обращается в нуль. Таким образом, упомянутые типы распределений являются особыми случаями алгоритма и должны обрабатываться отдельно.

Для аппроксимации H_n любые распределения Эрланга порядка $k\le n$ являются особыми случаями. Это вполне естественно, поскольку последние предполагают последовательную систему фаз, тогда как H_n — параллельную. Для исключения аварийных завершений алгоритма моменты исходного эрланговского распределения E_k начиная с k-го имеет смысл умножить на $1+j\varepsilon$, где j — порядок момента и ε — допустимая относительная погрешность (например, порядка 0.02).

Более детальный анализ показывает, что при замене гамма-распределения с $1<\alpha<2$ на H_2-аппроксимацию одна из «вероятностей» $\{y_j\}$ будет отрицательной, а другая превысит единицу. Как показали вычислительные эксперименты (см. главу 13), эти парадоксальные промежуточные результаты не мешают успешному расчету СМО при соответствующих H_2-распределениях. На рис. 1.3 показана аппроксимация гамма-плотности, имеющей параметр $\alpha=1.5$, гиперэкспонентой с двумя составляющими.

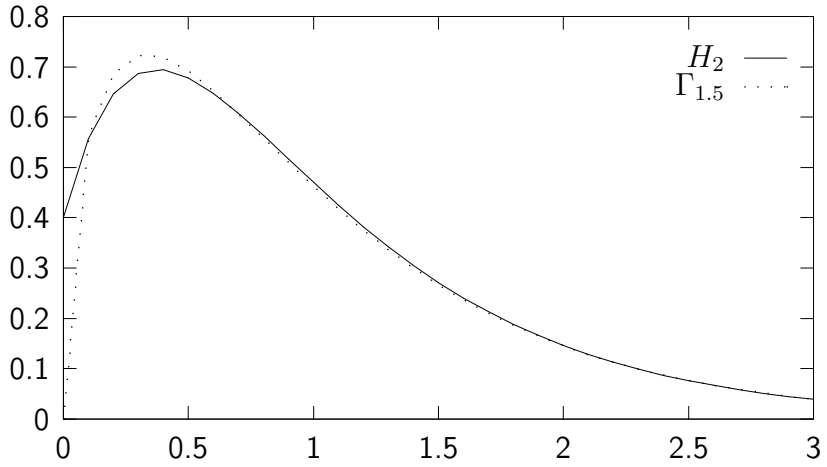

Рис. 1.3. Гиперэкспоненциальная аппроксимация плотности при $\alpha = 1.5$

Для некоторых задач ТМО отмеченная выше парадоксальная ситуация нежелательна. В таких случаях имеет смысл переходить к выравниванию H_2-распределений по *двум* моментам. Примем $y_1 = y_2 = 0.5$. Тогда условия выравнивания моментов $\{f_i\}$ принимают вид

$$\begin{array}{ll} 1/\mu_1 & +1/\mu_2 \\ 1/\mu_1^2 & +1/\mu_2^2 \end{array} \begin{array}{l} = 2f_1, \\ = f_2. \end{array}$$

Перейдя к обратным величинам $x_i = 1/\mu_i$, получаем из нее

$$\begin{array}{ll} x_1 & +x_2 \\ x_1^2 & +x_2^2 \end{array} \begin{array}{l} = 2f_1, \\ = f_2. \end{array}$$

Из первого уравнения следует $x_2 = 2f_1 - x_1$. Его подстановка во второе приводит к квадратному уравнению относительно x_1

$$2x_1^2 - 4f_1x_1 + 4f_1^2 - f_2 = 0$$

с решением

$$x_{1,2} = f_1 \pm \sqrt{f_2/2 - f_1^2}. \qquad (1.8.10)$$

Соответственно $\mu_i = 1/x_i, \quad i = 1, 2.$

Перепишем формулу (1.8.10) через коэффициент вариации v:

$$x_{1,2} = f_1(1 \pm \sqrt{(1+v^2)/2 - 1}\,).$$

Теперь ясно, что она может дать отрицательные значения параметров лишь при $(1+v^2)/2-1>1$, т. е. для распределений с коэффициентами вариации $v>\sqrt{3}\approx 1.732$, которые на практике встречаются редко.

Таким образом, рассмотренные варианты H_2-аппроксимации имеют непересекающиеся области «отказов» и в совокупности позволяют решить поставленную задачу. Известны и другие варианты подбора H_2-аппроксимации — см., например, [38].

Выполненные автором расчеты, связанные с теорией очередей (см. [130]), свидетельствуют об убывающем по индексу влиянии учета очередного момента. Представляется, однако, критически важным дополнительно оценить влияние отбора учтенных моментов при фиксированном их количестве на «хвост» распределения. Исследуем влияние выбора учтенных моментов на H_2-аппроксимацию дополнительной функции распределения Эрланга 3-го порядка с единичным средним. Эта пара распределение была выбрана из двух соображений:

1) распределение E_3 легко рассчитывается точно;

2) обсуждавшийся выше алгоритм подбора H_2-аппроксимации после очевидной замены обозначений позволяет выравнивать распределения по трем моментам с половинным и с двойным шагом индекса — см. таблицы 1.3 и 1.4.

Таблица 1.3 позволяет убедиться в том, что выравниваются именно моменты с затребованными целочисленными индексами, и тем самым верифицирует табл. 1.4.

Таблица 1.3. Результаты выравнивания моментов

t	E_3	H_2-аппроксимация по моментам		
		1,2,3	0.5,1.0,1.5	2,4,6
1	1.000e+0	1.000e+0	1.000e+0	1.020e+0
2	1.333e+0	1.333e+0	1.327e+0	1.333e+0
3	2.222e+0	2.222e+0	2.104e+0	2.212e+0
4	4.444e+0	4.346e+0	3.590e+0	4.444e+0
5	1.037e+1	9.438e+0	5.504e+0	1.041e+1
6	2.765e+1	2.136e+1	2.087e+0	2.765e+1

Теперь оценим качество аппроксимации ДФР (табл. 1.4):

Таблица 1.4. Влияние выбора моментов на аппроксимацию ДФР

t	E_3	H_2-аппр. по моментам			t	E_3	H_2-аппр. по моментам		
		1,2,3	0.5,1.0,1.5	2,4,6			1,2,3	0.5,1.0,1.5	2,4,6
0.0	1.00e-0	1.00e-0	1.00e-0	1.00e-0	2.0	6.20e-2	6.68e-2	6.88e-2	6.22e-2
0.1	9.96e-1	1.02e-0	1.01e-0	1.06e-0	2.1	4.98e-2	5.41e-2	5.49e-2	5.04e-2
0.2	9.77e-1	9.98e-1	9.78e-1	1.05e-0	2.2	4.00e-2	4.36e-2	4.33e-2	4.08e-2
0.3	9.37e-1	9.45e-1	9.27e-1	9.96e-1	2.3	3.20e-2	3.49e-2	3.38e-2	3.28e-2
0.4	8.79e-1	8.75e-1	8.62e-1	9.20e-1	2.4	2.55e-2	2.78e-2	2.60e-2	2.64e-2
0.5	8.09e-1	7.95e-1	7.89e-1	8.31e-1	2.5	2.03e-2	2.20e-2	1.97e-2	2.11e-2
0.6	7.31e-1	7.13e-1	7.13e-1	7.39e-1	2.6	1.61e-2	1.74e-2	1.46e-2	1.69e-2
0.7	6.50e-1	6.31e-1	6.36e-1	6.48e-1	2.7	1.27e-2	1.36e-2	1.06e-2	1.34e-2
0.8	5.70e-1	5.53e-1	5.62e-1	5.62e-1	2.8	1.00e-2	1.05e-2	7.43e-3	1.07e-2
0.9	4.94e-1	4.80e-1	4.92e-1	4.82e-1	2.9	7.92e-3	8.10e-3	4.95e-3	8.45e-3
1.0	4.23e-1	4.14e-1	4.27e-1	4.11e-1	3.0	6.23e-3	6.17e-3	3.05e-3	6.67e-3
1.1	3.59e-1	3.54e-1	3.68e-1	3.48e-1	3.1	4.90e-3	4.65e-3	1.61e-3	5.25e-3
1.2	3.03e-1	3.01e-1	3.14e-1	2.93e-1	3.2	3.84e-3	3.46e-3	5.48e-4	4.12e-3
1.3	2.53e-1	2.54e-1	2.67e-1	2.45e-1	3.3	3.01e-3	2.54e-3	-2.19e-4	3.23e-3
1.4	2.10e-1	2.13e-1	2.25e-1	2.04e-1	3.4	2.35e-3	1.82e-3	-7.53e-4	2.52e-3
1.5	1.74e-1	1.78e-1	1.88e-1	1.69e-1	3.5	1.83e-3	1.28e-3	-1.11e-3	1.96e-3
1.6	1.43e-1	1.48e-1	1.56e-1	1.40e-1	3.6	1.43e-3	8.64e-4	-1.32e-3	1.51e-3
1.7	1.16e-1	1.22e-1	1.29e-1	1.15e-1	3.7	1.11e-3	5.55e-4	-1.43e-3	1.17e-3
1.8	9.48e-2	1.00e-1	1.05e-1	9.38e-2	3.8	8.66e-4	3.27e-4	-1.46e-3	8.98e-4
1.9	7.68e-2	8.21e-2	8.54e-2	7.65e-2	3.9	6.73e-4	1.63e-4	-1.43e-3	6.88e-4

Таким образом, для построения ДФР в области малых значений вероятностей следует учитывать моменты максимально высокого порядка. Это особенно актуально при работе с распределениями, имеющими «толстые хвосты», т. е. значительные коэффициенты вариации.

1.8.4. Распределение Кокса

Диаграмма r-фазного распределения Кокса C_r представлена на рис. 1.4.

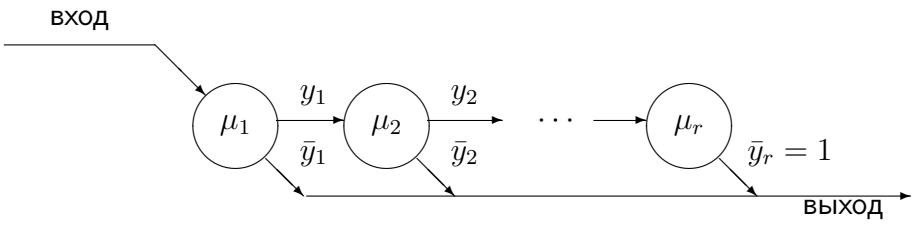

Рис. 1.4. Распределение Кокса

Распределение Эрланга и обсуждаемый ниже вариант гиперэрлангова распределения могут рассматриваться как частные случаи многофазного распределения Кокса.

Получим алгоритм расчета параметров распределения C_2. Прежде всего заметим, что его моменты с вероятностью y являются моментами свертки показательно распределенных задержек в фазах, а с дополнительной к ней суть моменты задержки в первой фазе. Таким образом, для расчета параметров C_2 имеем систему уравнений

$$\begin{aligned}
\frac{\bar{y}}{\mu_1} + y\left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right) &= \tilde{f}_1, \\
\bar{y}\frac{2}{\mu_1^2} + y\left(\frac{2}{\mu_1^2} + 2\frac{1}{\mu_1\mu_2} + \frac{2}{\mu_2^2}\right) &= \tilde{f}_2, \\
\bar{y}\frac{6}{\mu_1^3} + y\left(\frac{6}{\mu_1^3} + 3\frac{2}{\mu_1^2}\frac{1}{\mu_2} + 3\frac{1}{\mu_1}\frac{2}{\mu_2^2} + \frac{6}{\mu_2^3}\right) &= \tilde{f}_3.
\end{aligned}$$

Ее решение для общего случая приводится в [147, с. 35]. В наших обозначениях оно сводится к вычислению

$$\begin{aligned}
d &= (f_3 - f_1 f_2)^2 - 4(f_2 - f_1^2)(f_1 f_3 - f_2^2), \\
\mu_2 &= \frac{f_1 f_2 - f_3 + \sqrt{d}}{2(f_2^2 - f_1 f_3)}, \\
\mu_1 &= (\mu_2 f_1 - 1)/(\mu_2 f_2 - f_1), \\
y &= (\mu_1 f_1 - 1)\mu_2/\mu_1.
\end{aligned} \tag{1.8.11}$$

Особыми случаями являются показательное и E_2-распределения. Первый из них имеет место при $f_2 = f_1^2$ и требует выбора $\mu_1 = 1/f_1$ и $y = 0$ (значение μ_2 безразлично). Второй случай соответствует $f_2 = \frac{3}{4}f_1^2$. Здесь следует принять $\mu_1 = \mu_2 = 2/f_1$ и $y = 1$. Если f_2 меньше этой величины, аппроксимация должна иметь комплексные параметры.

В табл. 1.5 приведены параметры C_2-аппроксимации гамма-распределения при единичном первом моменте. Таблица подтверждает сделанный прогноз. Таблица содержит два варианта аппроксимации — в зависимости от знака перед корнем в выражении для μ_2 из системы (1.8.11). Оба варианта (проверено!) сохраняют три момента исходного распределения, причем второй при $\alpha < 1$ имеет парадоксальную (отрицательную) вероятность выхода во вторую фазу, что согласуется с рис. 1.4.

Обсчет моделей $M/H_2/n$ показал, что в парадоксальной зоне второй вариант работает при числе каналов не более 10.

Достоинством C_2-аппроксимации является естественное включение в нее как частных случаев M- и E_2-моделей, а недостатками — трудность обобщения на большее число составляющих и усложнение диаграмм переходов (см. главу 7).

Для приложений важна ДФР распределения Кокса. С учетом формулы (1.8.4) можно убедиться, что

$$\bar{F}(t) = \left(1 - \frac{y\mu_1}{\mu_1 - \mu_2}\right)e^{-\mu_1 t} + \frac{y\mu_1}{\mu_1 - \mu_2}e^{-\mu_2 t}. \qquad (1.8.12)$$

Легко проверить, что $\bar{F}(0) = 1$ и $\int\limits_0^\infty \bar{F}(t)\,dt = 1/\mu_1 + y/\mu_2$.

1.8.5. Гиперкоксово распределение

Отмеченные выше патологии с отрицательными вероятностями для H_2- и C_2-распределений имеют место в *различных* диапазонах коэффициентов вариации исходного распределения. Это дает надежду на существование их комбинации, свободной от упомянутого недостатка. Диаграмма процесса представлена на рис. 1.5.

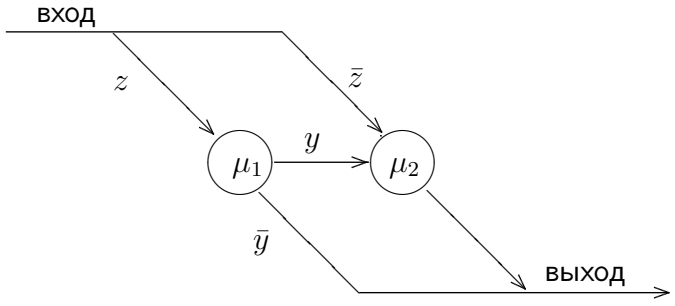

Рис. 1.5. Гиперкоксово распределение

Таблица 1.5. Параметры коксовой аппроксимации

α	$+\sqrt{D}$			$-\sqrt{D}$		
	y	μ_1	μ_2	y	μ_1	μ_2
0.2	0.20	3.73	0.27	-10.20	0.27	3.73
0.4	0.35	3.51	0.49	- 3.68	0.49	3.51
0.6	0.47	3.32	0.68	- 1.58	0.68	3.32
0.8	0.58	3.15	0.85	- 0.58	0.85	3.15
1.0	0.00	1.00	1.00	0.00	1.00	1.00
1.2	0.37	1.15	2.85	0.75	2.85	1.15
1.4	0.61	1.29	2.71	0.82	2.71	1.29
1.6	0.79	1.45	2.55	0.88	2.55	1.45
1.8	0.91	1.62	2.38	0.94	2.38	1.62
2.0	1.00	2.00	2.00	1.00	2.00	2.00
2.2	1.06+0.01i	2.00-0.35i	2.00+0.35i	1.06-0.01i	2.00+0.35i	2.00-0.35i
2.4	1.11+0.03i	2.00-0.49i	2.00+0.49i	1.11-0.03i	2.00+0.49i	2.00-0.49i
2.6	1.15+0.04i	2.00-0.58i	2.00+0.58i	1.15-0.04i	2.00+0.58i	2.00-0.58i
2.8	1.19+0.06i	2.00-0.65i	2.00+0.65i	1.19-0.06i	2.00+0.65i	2.00-0.65i
3.0	1.22+0.08i	2.00-0.71i	2.00+0.71i	1.22-0.08i	2.00+0.71i	2.00-0.71i
3.2	1.25+0.09i	2.00-0.76i	2.00+0.76i	1.25-0.09i	2.00+0.76i	2.00-0.76i
3.4	1.27+0.11i	2.00-0.80i	2.00+0.80i	1.27-0.11i	2.00+0.80i	2.00-0.80i
3.6	1.30+0.12i	2.00-0.83i	2.00+0.83i	1.30-0.12i	2.00+0.83i	2.00-0.83i
3.8	1.32+0.14i	2.00-0.87i	2.00+0.87i	1.32-0.14i	2.00+0.87i	2.00-0.87i
4.0	1.33+0.15i	2.00-0.89i	2.00+0.89i	1.33-0.15i	2.00+0.89i	2.00-0.89i

Очевидно, данное распределение с вероятностью z сводится к распределению Кокса C_2, а с дополнительной $\bar{z}$ — к показательному распределению с параметром μ_2. Это позволяет записать систему уравнений для его параметров в виде

$$\begin{aligned}
\tfrac{\bar{z}}{\mu_2} + z\left[\tfrac{\bar{y}}{\mu_1} + y\left(\tfrac{1}{\mu_1} + \tfrac{1}{\mu_2}\right)\right] &= \tilde{f}_1,\\
\tfrac{2\bar{z}}{\mu_2^2} + z\left[\bar{y}\tfrac{2}{\mu_1^2} + y\left(\tfrac{2}{\mu_1^2} + 2\tfrac{1}{\mu_1\mu_2} + \tfrac{2}{\mu_2^2}\right)\right] &= \tilde{f}_2,\\
\tfrac{6\bar{z}}{\mu_2^3} + z\left[\bar{y}\tfrac{6}{\mu_1^3} + y\left(\tfrac{6}{\mu_1^3} + 3\tfrac{2}{\mu_1^2}\tfrac{1}{\mu_2} + 3\tfrac{1}{\mu_1}\tfrac{2}{\mu_2^2} + \tfrac{6}{\mu_2^3}\right)\right] &= \tilde{f}_3,\\
\tfrac{24\bar{z}}{\mu_2^4} + z\left[\bar{y}\tfrac{24}{\mu_1^3} + y\left(\tfrac{24}{\mu_1^3} + 4\tfrac{6}{\mu_1^3}\tfrac{1}{\mu_2} + 6\tfrac{2}{\mu_1^2}\tfrac{2}{\mu_2^2} + 4\tfrac{1}{\mu_1}\tfrac{6}{\mu_2^3} + \tfrac{24}{\mu_2^4}\right)\right] &= \tilde{f}_4.
\end{aligned}$$

С помощью тех же подстановок, что в разделе 1.8.4, она приводится к виду

$$(1-z)x_2 + z[x_1 + yx_2] = f_1,$$

$$\begin{aligned}(1-z)x_2^2 + z[x_1^2 + y(x_1x_2 + x_2^2)] &= f_2,\\ (1-z)x_2^3 + z[x_1^3 + y(x_1^2x_2 + x_1x_2^2 + x_2^3)] &= f_3,\\ (1-z)x_2^4 + z[x_1^4 + y(x_1^3x_2 + x_1^2x_2^2 + x_1x_2^3 + x_2^4)] &= f_4.\end{aligned}$$

Ее явное решение получить не удается. Ограничившись выравниванием трех моментов, можно предложить следующий алгоритм:

1. Если $|f_2/f_1^2 - 2| < \varepsilon$ — считать распределение показательным: положить $z_1 = 1$, $z_2 = 0$, $y = 0$, $\mu_1 = \mu_2 = 1/f_1$ (значение μ_2 можно выбрать произвольно). Перейти к этапу 5.

2. Если $|f_2/f_1^2 - 1.5| < \varepsilon$ — считать распределение эрланговским второго порядка: положить $z_1 = 1$, $z_2 = 0$, $y = 1$, $\mu_1 = \mu_2 = 2/f_1$. Перейти к этапу 5.

3. Если $|f_2/f_1^2 - 1.75| < 0.25 - \varepsilon$, положить $z_1 = 1$, $z_2 = 0$ и применить аппроксимацию C_2. Перейти к этапу 5.

4. Положить $y = 0$ и найти параметры H_2-аппроксимации $\{z_i, \mu_i\}$.

5 Конец алгоритма.

Здесь для унификации обозначений принято $z_1 = z,\ z_2 = 1 - z_1$.

1.8.6. Гиперэрлангово распределение

По аналогии с гиперэкспоненциальным можно построить гиперэрлангово распределение — как взвешенную сумму эрланговых законов. На рис. 1.6 приведена одна из достаточно общих схем с тремя составляющими. Для такого закона сравнительно легко рассчитать моменты, однако решение обратной задачи (подбора параметров по заданным моментам) вызывает серьезные трудности. Кроме того, наличие большого числа фаз усложняет формирование матриц переходов, в особенности для многоканальных систем — см. главу 7.

Наиболее удобна для расчетов смесь двух эрланговых распределений со смежными значениями порядка и общим непрерывным параметром — см. рис. 1.7.

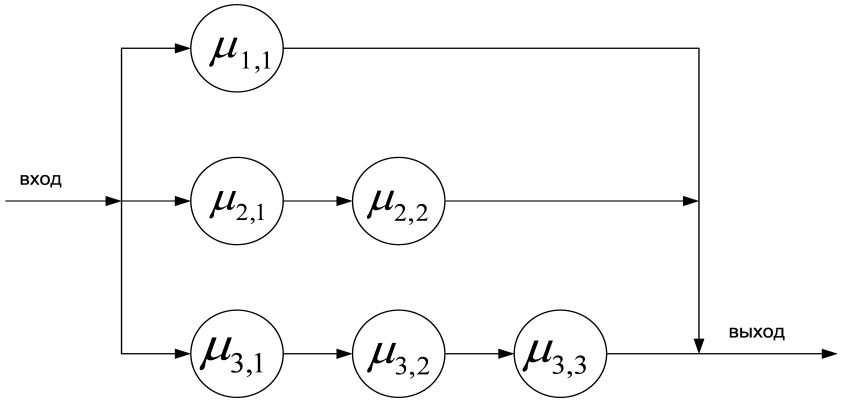

Рис. 1.6. Гиперэрлангово 3-распределение

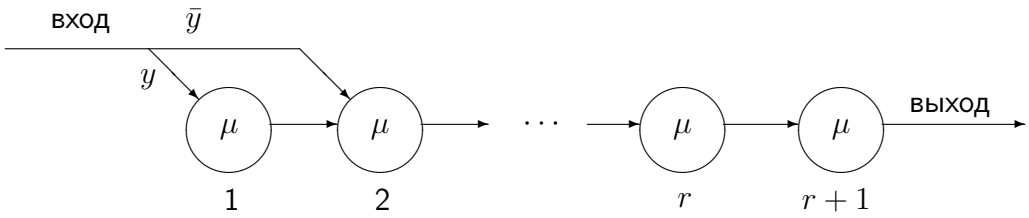

Рис. 1.7. Вариант гиперэрлангова распределения

При данном распределении состояние процесса полностью определяется одной дискретной переменной — номером текущей фазы. Его ДФР

$$\bar{F}(t) = \sum_{i=0}^{r-1} \frac{(\mu t)^i}{i\,!} e^{-\mu t} + y \frac{(\mu t)^r}{r\,!} e^{-\mu t}. \qquad (1.8.13)$$

Определим параметры аппроксимирующего гамма-распределения согласно (1.4.1). Тогда параметрами искомого гиперэрлангова распределения будут найденное из (1.4.1) значение μ, а также

$$r = [\alpha]; \quad y = \alpha - r.$$

Эта аппроксимация обеспечивает строгое выравнивание двух первых моментов и приближенное — третьего. В отладочном примере выравнивания гамма-распределения порядка $\alpha = 2.3$ с моментами f_1=2, f_2=5.739, f_3=21.302 третий момент гиперэрланговского распределения составил 21.459. Качество выравнивания высших моментов возрастает

с уменьшением коэффициента вариации; однако одновременно растет и число фаз, что увеличивает трудоемкость расчетов.

Можно показать, что в частном случае $\alpha < 1$

$$r = 1, \quad y = 2\alpha/(\alpha + 1), \quad \mu = y/f_1.$$

1.8.7. Распределение Ньютса

Наиболее общей формой представления распределений фазового типа (Ph-распределений) является схема М. Ньютса [28, 232]. Длительность каждой реализации процесса здесь соответствует случайному времени блуждания заявки по изображенной на рис. 1.8 сети с показательно распределенной (параметры $\{\nu_i\}, \quad i = \overline{1, M}$) задержкой в узлах и одним поглощающим состоянием.

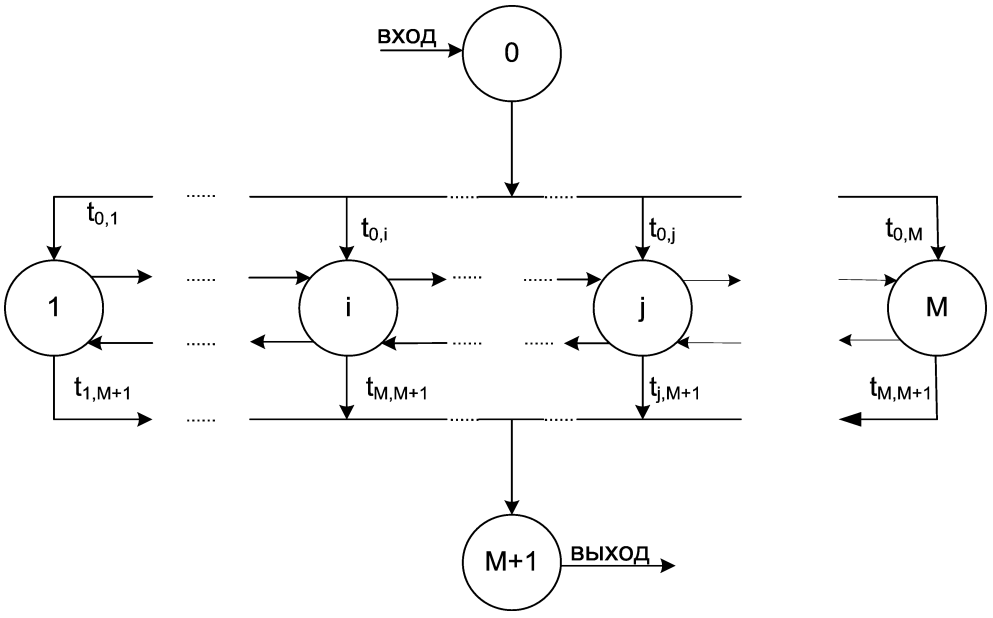

Рис. 1.8. Распределение Ньютса

Если мы запишем инфинитезимальный генератор Q такой марковской цепи в виде

$$Q = \begin{bmatrix} T & T^o \\ \mathbf{0} & 0 \end{bmatrix}$$

и обозначим его начальный вероятностный вектор $(\mathbf{a}, a_{M+1})$, то

$$F(x) = 1 - \mathbf{a} \cdot \exp(Tx) \cdot \mathbf{1},$$

где $\mathbf{1} = \{1, 1, \ldots, 1\}^T$.

Для распределения времени обслуживания фазового типа время обслуживания начинается с попадания в фазу i с вероятностью α_i, $i = \overline{1, M+1}$ (сумма вероятностей = 1). Изменения фаз управляются инфинитезимальным генератором

$$\tilde{Q} = \begin{bmatrix} T & \eta \\ \mathbf{0} & 0 \end{bmatrix}, \tag{1.8.14}$$

где T — квадратная матрица размера M, $\mathbf{0}$ и η — строка и столбец той же размерности. Состояние « $M+1$ » означает завершение процесса, m считается его порядком. Из приведенных выше определений следует

$$\eta = -T\mathbf{1}, \qquad \alpha_{m+1} = 1 - \alpha\mathbf{1}$$

(поскольку суммы строк генератора равны 0, а сумма компонент вероятностного вектора — 1). Вектор η называется вектором выхода PH-распределения.

Если матрица инфинитезимального генератора T неприводима, то результирующее распределение времени обслуживания будет

$$F(t) = 1 - \alpha e^{Tt}\mathbf{1}, \tag{1.8.15}$$

а моменты

$$M[t^k] = k!\alpha(-T)^{-k}\mathbf{1}. \tag{1.8.16}$$

Моменты могут быть найдены и численно — с помощью описанного в главе 11 численного алгоритма расчета распределения времени пребывания заявки в сети.

Приведем примеры матричного описания ранее рассмотренных нами распределений. Для распределения Эрланга

$$\alpha = (1, 0, \ldots, 0), \quad T = \begin{bmatrix} -\lambda & \lambda & & \\ & \ddots & \ddots & \\ & & -\lambda & \lambda \\ & & & -\lambda \end{bmatrix}, \qquad \eta = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \lambda \end{bmatrix}.$$

Для гиперэкспоненты

$$\alpha = (\pi_1, \ldots, \pi_n,), \quad T = \begin{bmatrix} -\lambda_1 & \lambda & \\ & \ddots & \ddots \\ & & -\lambda_n \end{bmatrix}, \qquad \eta = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix}.$$

Распределение Кокса имеет стартовый вектор $\alpha = (1, 0, \ldots, 0)$ и

$$T = \begin{bmatrix} -\lambda_1 & q_1\lambda_1 & & \\ & \ddots & \ddots & \\ & & -\lambda_{n-1} & q_{n-1}\lambda_{n-1} \\ & & & -\lambda_n \end{bmatrix}, \qquad \eta = \begin{bmatrix} (1-q_1)\lambda_1 \\ \vdots \\ (1-q_{n-1})\lambda_{n-1} \\ \lambda_n \end{bmatrix}.$$

В приведенных объектах все неспецифицированные элементы являются нулевыми. Элементы фундаментальной матрицы $(I - Q)^{-1}$ дают ожидаемое число посещений узлов — до поглощения [164].

Матрично-фазовые распределения замкнуты относительно операторов свертки, вероятностного взвешивания и взятия минимума (эти операции приводят к распределениям того же класса). Их результаты выражаются через параметры исходных распределений посредством кронекеровых сумм и произведений — см. [168, с. 186 и далее], а также [18, с. 74].

Распределение Ньютса соответствует классу плотностей, имеющих дробно-линейное преобразование Лапласа-Стилтьеса (ПЛС). Для последнего можно подобрать [190, с. 202–204] *аппроксимацию Паде* в виде отношения полиномов степеней m и l так, чтобы первые $m + l$ моментов совпадали. Теория выравнивания моментов с помощью аппроксимации Паде рассматривается в [155]. Однако неизбежно возникнет вопрос о ее интерпретации — *в расчете обязательно потребуется фазовое представление.*

Матрично-фазовые распределения позволяют элегантно описывать входящий поток (MAP — Markovian Arrival Process) и процесс обслуживания *в одноканальной системе*. Их применение для реального решения задач приводит к заметному увеличению потребностей в памяти и времени счета из-за использования сильно разреженных матриц и как следствие — к необходимости выполнения большого числа операций сложения и умножения с заведомо нулевыми результатами. Перспективы их

применения для расчета многоканальных систем представляются сомнительными. Во многих работах по теории очередей (к примеру, в [257]) применение Ph-аппроксимаций лишь декларируется и фактически подменяется использованием классических аппроксимаций. Ими в дальнейшем ограничимся и мы. Заметим кстати, что даже в весьма серьезных руководствах по теории очередей [18, 28, 165, 168] и в уже упомянутой статье [257] применение классических фазовых распределений ограничивается случаями вещественных параметров, что заметно обедняет арсенал исследователя.

1.8.8. Параметры фазовых распределений

Сопоставим типы значений параметров наиболее употребительных фазовых аппроксимаций в зависимости от параметра гамма-распределения $\alpha = f_1^2/D$. Эрланговы аппроксимации соответствуют целым положительным значениям α. Гиперэкспоненциальное распределение имеет положительные вещественные параметры при $0 < \alpha < 1$, коксово — при $0 < \alpha < 2$, P-распределение — при любых $\alpha > 1$. Гиперэкспонента и коксово распределение имеют комплексные параметры при $\alpha > 2$. Весьма популярна «фольклорная» рекомендация [28, 164, 165, 233]: если коэффициент вариации меньше 1 — выбирается Эрланг, иначе гиперэкспонента H_2. Таким образом, приходится отметить, что идея использования аппроксимирующих распределений с комплексными параметрами за полвека еще не овладела массами специалистов по ТМО.

Необходимо, однако, учитывать следующие дополнительные соображения:

1) Число выравниваемых моментов (обсуждалось ранее).

2) Расход оперативной памяти (для комплексных объектов он вдвое больше, чем для вещественных той же разрядности).

3) Трудоемкость вычислений (примерно такое же увеличение).

4) Наличие и возможности готовых программ вычислений (если готовая программа рассчитана на общий случай, то она будет обрабатывать вещественные объекты как комплексные с нулевой мнимой частью).

5) Множество возможных состояний системы (оно очень быстро разрастается с увеличением порядка эрланговых и P-распределений, что увеличивает требуемый объем памяти и время счета).

6) Наличные аппаратные возможности (увеличение требуемой памяти и объема вычислений до известных пределов может не иметь практического значения).

1.9. Гамма-распределение с поправочным многочленом

1.9.1. Выражение для плотности

В разд. 1.3 обсуждалось выравнивание распределения двухпараметрической гамма-плотностью. Для выравнивания большего числа моментов приходится применять гамма-плотность с поправочным многочленом:

$$f(t) \approx \frac{\mu(\mu t)^{\alpha-1}}{\Gamma(\alpha)} e^{-\mu t} \sum_{i=0}^{N} g_i t^i. \tag{1.9.1}$$

Параметры базовой плотности α и μ естественно подобрать по двум моментам — см. формулы (1.4.1). Коэффициенты поправочного многочлена определяются из системы линейных алгебраических уравнений, задающей выравнивание моментов до N-го порядка включительно:

$$\sum_{i=0}^{N} g_i \frac{\Gamma(\alpha+k+i)}{\mu^{k+i}\,\Gamma(\alpha)} = f_k, \quad k = \overline{0,N}. \tag{1.9.2}$$

В [92] с помощью теории ортогональных многочленов Лагерра были выведены явные выражения для коэффициентов $\{g_i\}$. Удобнее, однако, получать их из системы линейных уравнений

$$\begin{aligned} \int_0^\infty t^k \frac{\mu(\mu t)^{\alpha-1}}{\Gamma(\alpha)}\, e^{-\mu t} \sum_{i=0}^{N} g_i t^i\, dt &= \sum_{i=0}^{N} g_i \frac{\mu^\alpha}{\Gamma(\alpha)} \int_0^\infty t^{\alpha+k+i-1} e^{-\mu t}\, dt \\ = \sum_{i=0}^{N} g_i \frac{\mu^\alpha}{\Gamma(\alpha)\mu^{\alpha+k+i}} \int_0^\infty x^{\alpha+k+i-1} e^{-x}\, dx &= \sum_{i=0}^{N} \frac{g_i}{\mu^{k+i}} \frac{\Gamma(\alpha+k+i)}{\Gamma(\alpha)} = f_k, \\ k = \overline{0,N}. \end{aligned}$$

На рис. 1.9 показана аппроксимация формулой (1.9.1) плотности равномерного распределения (R), а на рис. 1.10 — треугольного распределения (T) при коэффициенте вариации $v = 0.3$ с учетом двух, четырех и шести моментов.

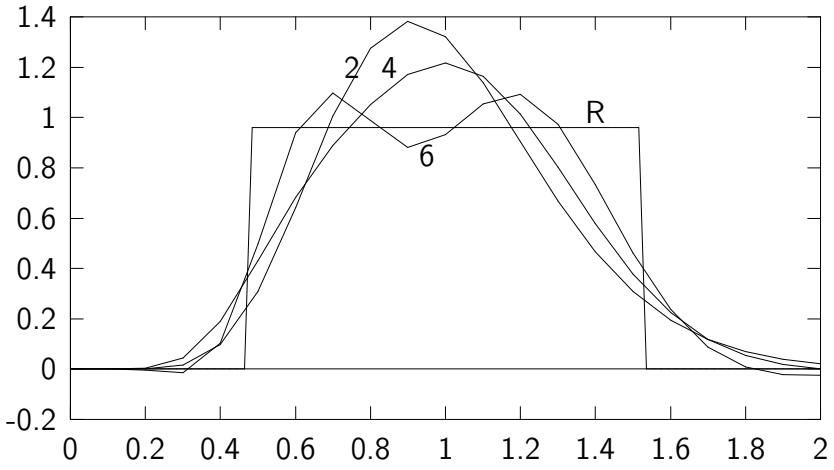

Рис. 1.9. Аппроксимации равномерного распределения

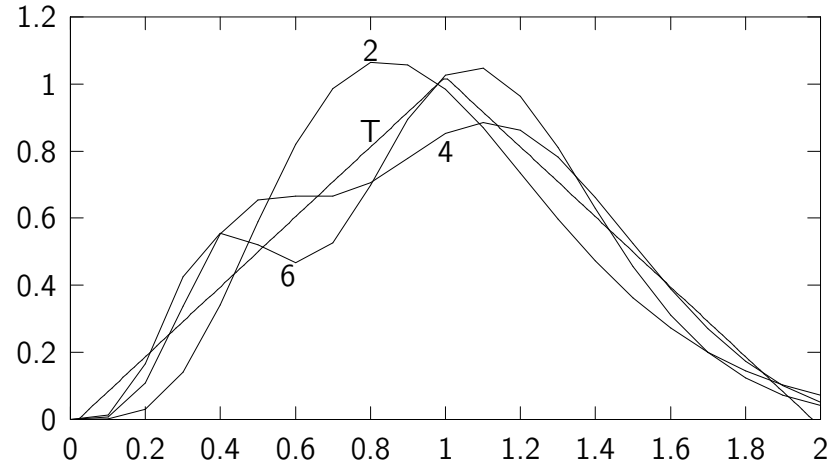

Рис. 1.10. Аппроксимации треугольного распределения

Заметен недостаток аппроксимации, проявляющийся при негладких исходных плотностях — возможность получения отрицательных значений

плотности. Кстати, такой же недостаток имеет давно и широко используемое в математической статистике [88] разложение Грама—Шарлье в виде плотности $нормального$ распределения с поправочным многочленом, зависящим от разностей одноименных моментов. В рассматриваемых ниже специфических приложениях обсуждаемой аппроксимации этот недостаток практического значения не имеет.

1.9.2. Вычисление и обращение преобразований Лапласа

Для плотности распределения, записанной в форме (1.9.1), легко вычисляется преобразование Лапласа:

$$\begin{aligned}\varphi(s) &= \int_0^\infty e^{-st}\frac{\mu(\mu t)^{\alpha-1}}{\Gamma(\alpha)}e^{-\mu t}\sum_{i=0}^{N} g_i t^i\,dt = \frac{\mu^\alpha}{\Gamma(\alpha)}\sum_{i=0}^{N} g_i \int_0^\infty t^{\alpha+i-1}e^{-(s+\mu)t}\,dt \\ &= \left(\frac{\mu}{s+\mu}\right)^\alpha \sum_{i=0}^{N}\frac{g_i}{(s+\mu)^i}\frac{\Gamma(\alpha+i)}{\Gamma(\alpha)},\end{aligned} \tag{1.9.3}$$

причем при рекуррентном расчете слагаемых вычисление гамма-функции не требуется.

Таблица 1.6 дает относительную погрешность вычисления ПЛС треугольного распределения со средним значением $f_1 = 1$ для двух коэффициентов вариации.

Таблица 1.6. Погрешность преобразований Лапласа

Число учитываемых моментов	Коэффициент вариации	
	0.3	0.4
2	-8.35e-4	-2.54e-3
3	-6.53e-5	-4.22e-4
4	-6.23e-7	-4.13e-5
5	9.81e-8	-4.17e-6
6	2.05e-9	-6.26e-7
7	4.83e-9	-9.32e-8
8	4.19e-9	-4.52e-9

Решение многих задач ТМО удается довести лишь до ПЛС $\nu(s)$ плотности $v(t)$ распределения времени пребывания заявки в системе:

$$\nu(s) = \int_0^\infty e^{-st} v(t)\, dt. \tag{1.9.4}$$

Последовательное дифференцирование левой и правой частей (1.9.4) по s приводит к определению начальных моментов $\{v_k\}$ согласно (1.6.3). Дифференцирование может быть выполнено численно; по найденным моментам можно построить сохраняющую их плотность вида (1.9.1).

Используемая в этой технологии идея получения оригинала преобразования Лапласа в виде разложения по ортогональным многочленам Лагерра была впервые выдвинута в работе А. Папулиса [236] (см. также [216]), но детальной разработки и применения не получила.

1.10. ДФР Вейбулла с поправочным многочленом

Для многих СМО основным показателем качества функционирования является ДФР времени пребывания заявки в системе, получаемая в виде таблицы для заданных значений аргумента. Кроме того, построение ДФР чистой длительности обслуживания является важным технологическим элементом расчета замкнутых СМО. Результат интегрирования правой части (1.9.1) через элементарные функции не выражается.

Эффективным методом вычисления ДФР по заданным моментам является представление ее в виде ДФР Вейбулла с поправочным многочленом

$$\bar{F}(t) = e^{-t^k/T} \sum_{i=0}^{N} g_i t^i. \tag{1.10.1}$$

Моменты этого распределения

$$f_j = (j/k) \sum_{i=0}^{N} g_i \Gamma[(i+j)/k] T^{(i+j)/k}, \qquad j = \overline{0, N}. \tag{1.10.2}$$

Как и в разд. 1.3, определим параметры опорного распределения выравниванием двух первых моментов. Тогда окажется, что

$$f_j = (j/k)\Gamma(j/k) T^{j/k} = T^{j/k} \Gamma(1 + j/k), \qquad j = 1, 2,$$

а отношение

$$\alpha = f_2/f_1^2 = 2k\Gamma(2/k)/\Gamma^2(1/k) = 2\Gamma(2u)/(u\Gamma^2(u)), \qquad (1.10.3)$$

где $u = 1/k$. Воспользовавшись формулой удвоения аргумента гамма-функции [152, с. 55], можно переписать (1.10.3) в виде

$$\alpha = \frac{2^{2u}\Gamma(u)\Gamma(u+1/2)}{u\sqrt{\pi}\,\Gamma^2(u)} = \frac{2^{2u}\Gamma(u+1/2)}{\sqrt{\pi}\,\Gamma(u+1)},$$

откуда следует обеспечивающая быстро сходящийся итерационный процесс уточнения переменной u формула

$$u_i = \frac{1}{2\ln 2}\ln\frac{\alpha\sqrt{\pi}\,\Gamma(u_{i-1}+1)}{\Gamma(u_{i-1}+1/2)}, \qquad i = 1, 2, \ldots \qquad (1.10.4)$$

Начальное приближение

$$u_0 = \ln 2\alpha/(2\ln 2). \qquad (1.10.5)$$

Теперь ясен алгоритм подбора параметров аппроксимации (1.10.1) по моментам $\{f_j\},\ j = \overline{1, N}$:

1. Вычислить $\alpha = f_2/f_1^2$.
2. Определить u_0 согласно (1.10.5).
3. Решить уравнение (1.10.4) методом итераций.
4. Вычислить $k = 1/u, \quad T = [f_1/\Gamma(u+1)]^k$.
5. Сформировать систему (1.10.2) линейных алгебраических уравнений относительно $\{g_i\}$.
6. Решить эту систему любым стандартным методом.

Подбор параметров этой аппроксимации по заданному множеству моментов и последующее вычисление ее моментов возвращают исходные моменты. Этот принцип тестирования (решение взаимообратных задач) в составе пакета применяется также к аппроксимациям распределений гамма-плотностью с поправочным многочленом и гиперэкспонентой.

На рис. 1.11 дана Вейбулл-аппроксимация треугольной плотности. Увеличение числа учтенных моментов улучшает качество приближения при больших значениях аргумента, но ухудшает при малых. Кроме того, как и в случае гамма-распределения с поправочным многочленом, плотность может принимать отрицательные значения.

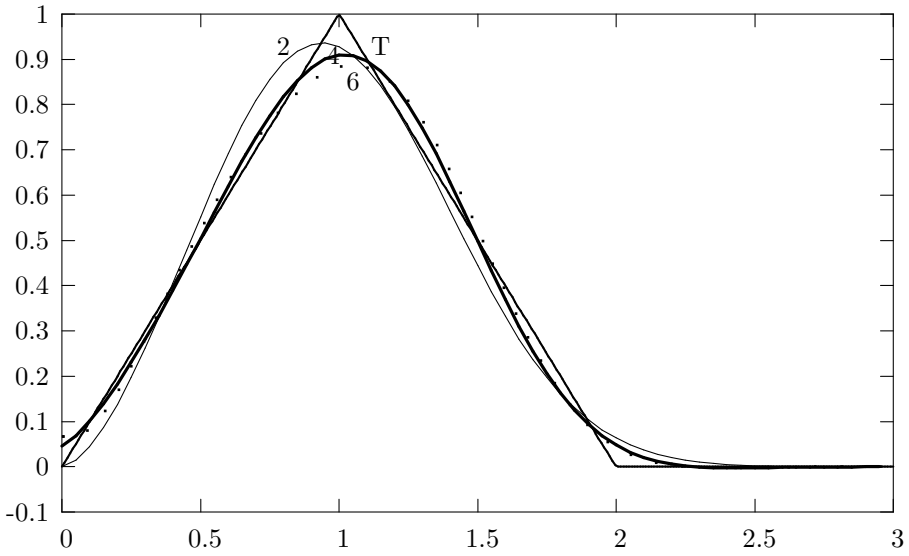

Рис. 1.11. Аппроксимация треугольной плотности по Вейбуллу

1.11. Распределение Парето

Распределение Парето (иначе — гиперболическое или степенное) описывается плотностью

$$f(t) = \alpha K^{\alpha}/t^{\alpha+1}, \qquad \alpha > 0, t \geq K,$$

где α — параметр формы, $K > 0$ — наименьшее значение случайной величины. Соответственно ДФР

$$\bar{F}(t) = (K/t)^{\alpha}, \quad t \geq K.$$

Можно показать, что *начальные моменты этого распределения порядка* m *существуют только для* $\alpha > m$. В случае $\alpha > 2$ математическое ожидание и дисперсия конечны и подсчитываются согласно

$$f_1 = \alpha K/(\alpha - 1), \qquad D = \alpha K^2/[(\alpha - 1)^2(\alpha - 2)].$$

Распределение Парето часто используется для тестирования различных аппроксимаций. В частности, в [37, с. 135] с его помощью эффектно иллюстрируется отсутствие сходимости среднего времени ожидания в процессе имитации системы M/M/1.

Глава 2

Математическая модель теории очередей

Модель задачи массового обслуживания включает в себя следующие элементы:

- поток заявок,
- каналы обслуживания,
- организацию очереди и дисциплину обслуживания,
- показатели эффективности.

Дадим содержательное описание и перечень возможных вариантов задания этих элементов и установим математическое содержание соответствующих понятий.

2.1. Поток заявок

2.1.1. Основные определения

Входящий поток задан, если для каждого номера n очередной заявки задано совместное распределение интервалов $\{z_1, z_2, \ldots, z_n\}$ между смежными заявками: $z_i = t_i - t_{i-1}, \quad t_0 = 0$. Если длины упомянутых интервалов независимы в совокупности, то зависимость момента прибытия очередной заявки t_n от предыстории процесса сводится к

фиксации t_{n-1} (поток с *ограниченным последействием*). Такой поток может быть задан функциями распределения $A_i(t) \stackrel{\text{def}}{=} P\{z_i < t\}$. Если $A_i(t) = A(t)$ при всех $i \geq 1$, поток называется *рекуррентным*. Это наиболее общий случай из реально используемых в расчетах.

Обозначим $Z_k(t,\tau)$ событие, состоящее в появлении ровно k заявок на полуинтервале $[t, t+\tau)$. Свойства потока заявок могут быть охарактеризованы через вероятности $\{p_k(t,\tau)\}$ таких событий. Поток называется *стационарным*, если эти вероятности определяются только длиной интервала τ и не зависят от его положения на оси времени (переменная t). Поток называется потоком *без последействия*, если события $Z_{k_1}(t_1,\tau_1)$ и $Z_{k_2}(t_2,\tau_2)$ для неперекрывающихся интервалов времени независимы[1]. Поток считается *ординарным*, если вероятность появления на элементарном участке $[t, t+\Delta t)$ более чем одного события имеет порядок малости $o(\Delta t)$, т. е. выше Δt. Поток, одновременно удовлетворяющий всем перечисленным требованиям, именуется *простейшим*.

Указанные свойства наблюдаются часто[2], но не всегда. Например, интенсивность потока посетителей в магазине или телефонных звонков в сети существенно зависит от времени суток. Один телефонный звонок с объявлением тревоги или сенсационной новости (дефолт) может вызвать целую лавину вторичных (типичная картина *последействия*). Заявки могут поступать группами постоянного или случайного объема. Характерный пример подобной «неординарной» ситуации — прибытие в арктический порт на разгрузку каравана судов после ледокольной проводки или пассажиров в аэропорт (семейными, деловыми и экскурсионными группами).

Для определения количественных характеристик потоков введем вероятность $\Pi_1(t,\tau)$ появления хотя бы одного требования за интервал длины τ, прилегающий справа к t, и математическое ожидание $M[k(t,\tau)]$ числа поступивших требований.

Интенсивность потока определяется согласно

$$\gamma(t) = \lim_{\tau\to 0} \frac{M[k(t,\tau)]}{\tau} = \lim_{\tau\to 0} \frac{\sum_{k=1}^{\infty} k p_k(t,\tau)}{\tau}. \qquad (2.1.1)$$

[1] В [2] потоком без последействия фактически именуется поток с ограниченным последействием.

[2] Есть мнение [17], что потоки депозитов и заявок на кредиты — простейшие.

Далее, $параметр$ потока вычисляется по формуле

$$\pi(t) = \lim_{\tau \to 0} \frac{\Pi_1(t,\tau)}{\tau}. \tag{2.1.2}$$

Поскольку $\Pi_1(t,\tau) = \sum_{k=1}^{\infty} p_k(t,\tau) \le \sum_{k=1}^{\infty} k p_k(t,\tau)$, то всегда $\pi(t) \le \gamma(t)$, причем равенство имеет место только для ординарного потока. Сразу же заметим, что в случае неординарного потока требований в виде «пачек» постоянного объема удобнее переходить к ординарному потоку групповых заявок.

Потоки заявок в сетях обслуживания подвергаются операциям суммирования, прореживания (случайного и регулярного) и преобразования в узлах. Хорошо изучены процессы просеивания и суммирования простейших потоков (см., например, [68]). Как будет показано ниже, вопреки мнению автора [33], можно сделать многое и для рекуррентных потоков общего вида.

2.1.2. Число событий рекуррентного потока на фиксированном интервале времени

Как уже отмечалось, интервалы между последовательными требованиями рекуррентного потока имеют одну и ту же функцию распределения $A(t)$. Таким образом, вероятность появления требования на полуинтервале $[t, t+\Delta t)$ с момента предыдущего требования есть

$$P(t,\Delta t) = \frac{A(t+\Delta t) - A(t)}{\bar{A}(t)} = \frac{\bar{A}(t) - \bar{A}(t+\Delta t)}{\bar{A}(t)}.$$

Параметр рекуррентного потока

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\bar{A}(t) - \bar{A}(t+\Delta t)}{\Delta t\, \bar{A}(t)} = -\frac{\bar{A}'(t)}{\bar{A}(t)}. \tag{2.1.3}$$

Отсюда следует, что

$$\bar{A}(t) = C \cdot \exp\Big(-\int_0^t \lambda(\tau)\, d\tau\Big).$$

Поскольку $\bar{A}(0) = 1, \quad C = 1$ и $\bar{A}(t) = \exp\Big(-\int_0^t \lambda(\tau)\, d\tau\Big)$.

Положим $a_1(t) = dA(t)/dt$, введем свертки исходных распределений по формуле

$$a_k(t) = \int_0^t a_1(t-\tau)a_{k-1}(\tau)\,d\tau, \qquad k = 2, 3, \ldots$$

и обозначим $A_k(t) = \int_0^t a_k(\tau)\,d\tau$. Очевидно, $A_k(t)$ есть функция распределения суммы k независимых случайных величин, подчиняющихся распределению $A_1(t) \equiv A(t)$. За полуинтервал $[0, t)$ придет *меньше* k требований, если сумма k интервалов между последовательными требованиями будет *больше* t. Переходя к вероятностям соответствующих событий, имеем

$$P_k(t) = 1 - A_k(t) = \bar{A}_k(t), \qquad k = 1, 2, \ldots$$

Очевидно, вероятность неприбытия заявок

$$q_0(t) = 1 - A(t) = \bar{A}(t). \tag{2.1.4}$$

Обозначим $T_k(t)$ событие, состоящее в том, что сумма k интервалов меньше t. Его вероятность равна $A_k(t)$. В это событие строго включается событие $T_{k+1}(t)$, имеющее вероятность $A_{k+1}(t)$. Очевидно, событие $T_k(t)\backslash T_{k+1}(t)$ соответствует случаям, когда сумма k интервалов меньше t, а $k+1$ интервалов — больше либо равна t. Вероятность такого события составляет $A_k(t) - A_{k+1}(t)$. Но подобное соотношение между t и суммами интервалов означает появление *ровно* k требований потока, так что

$$q_k(t) = A_k(t) - A_{k+1}(t). \tag{2.1.5}$$

Обозначим чере $\alpha(s)$ ПЛС от плотности распределения $A(t)$. Тогда производящая функция ПЛС от интересующих нас вероятностей

$$\begin{aligned} Q(s) &= \frac{1}{s}\sum_{k=0}^{\infty} z^k[\alpha^k(s) - \alpha^{k+1}(s)] \\ &= a\frac{1-\alpha(s)}{as}\sum_{k=0}^{\infty} z^k\alpha^k(s). \end{aligned}$$

Здесь a есть средний интервал между смежными заявками, а дробь перед суммой — ПЛС от *остаточного* распределения интервалов между ними. Таким образом, для рекуррентного потока

$$q_k(t) = aL^{-1}[\alpha^*(s)\cdot\alpha^k(s)], \qquad k = 0, 1, \ldots \tag{2.1.6}$$

(через L^{-1} обозначен оператор обратного ПЛС). Контрольное суммирование ПЛС дает $1/s$, откуда следует равенство суммы $\{q_k(t)\}$ единице.

Произведение сверток в формуле (2.1.6) имеет прозрачную вероятностную интерпретацию: чтобы за время t пришло ровно k заявок, в него должны уложиться k полных интервалов между заявками и один остаточный (последний). Напомним, что отсчет времени здесь начинается сразу после прибытия очередной заявки.

В случае рекуррентного потока *с запаздыванием* интервал времени до прибытия первой заявки является случайной модификацией типового. Это позволяет по аналогии с предыдущим вариантом сразу записать итоговые формулы в виде

$$\begin{aligned} q_0(t) &= \bar{A}^*(t), \\ q_k(t) &= aL^{-1}[(\alpha^*(s))^2 \cdot \alpha^{k-1}(s)], \qquad k = 1, 2, \ldots \end{aligned} \tag{2.1.7}$$

Условие нормировки $\{q_k(t)\}$ здесь проверяется (и выполняется) аналогично.

Рекуррентный поток с запаздыванием — это наиболее общий тип потока из реально используемых в расчетах.

2.1.3. Число событий простейшего потока на интервале фиксированной длины

Выясним, при каком распределении $A(t)$ параметр рекуррентного потока $\lambda = \mathrm{const}$, что эквивалентно полному отсутствию последействия. Из условия

$$\bar{A}(t) = \exp\Big(-\int_0^t \lambda\, d\tau\Big) = e^{-\lambda t}$$

находим, что

$$\begin{aligned} A(t) &= 1 - e^{-\lambda t}, \\ a(t) &= \lambda e^{-\lambda t}. \end{aligned} \tag{2.1.8}$$

Таким образом, интервалы между требованиями стационарного ординарного потока без последействия (простейшего потока) подчиняются *показательному* распределению. Его параметр $\lambda = 1/a_1$, где $a_1 = a$ — средний интервал между требованиями.

Для подсчета распределения числа требований простейшего потока за время t выполним свертку показательных распределений. Свертка

k-го порядка — распределение Эрланга того же порядка (см. разд. 1.8.1). Этому распределению, в частности, подчиняются интервалы между требованиями регулярно прореживаемого простейшего потока, в котором сохраняется каждое k-е требование. Подставляя формулы указанного раздела в (2.1.5), убеждаемся, что вероятность прихода за $[0,t)$ ровно k требований

$$p_k(t) = \frac{(\lambda t)^k}{k\,!} e^{-\lambda t}, \qquad k = 0, 1, \ldots \tag{2.1.9}$$

Эта формула задает распределение Пуассона с параметром λt, отчего простейший поток называют также $пуассоновым$.

Отметим важную для приложений $теорему\ Фрая$: при условии, что число событий простейшего потока на $[a,b)$ равно n, моменты этих событий независимы и равномерно распределены в упомянутом полуинтервале.

2.1.4. Особая роль простейшего потока

В теории массового обслуживания простейший поток занимает особое место по следующим причинам:

1) Сумма конечного числа независимых простейших потоков образует простейший поток с интенсивностью, равной сумме интенсивностей составляющих потоков.

2) Сумма n независимых стационарных потоков с ограниченным последействием при условии малой интенсивности составляющих в сравнении с суммарной интенсивностью при $n \to \infty$ сходится к простейшему потоку (это утверждение в точной формулировке доказывается теоремой Ососкова — Хинчина и ее обобщением — теоремой Григелиониса).

3) Случайное прореживание произвольного стационарного ординарного потока с ограниченным последействием, т. е. выбрасывание каждого очередного требования независимо с некоторой вероятностью, при увеличении вероятности выбрасывания приближает поток к простейшему. Точную формулировку этого результата дают теорема Реньи и ее обобщение — теорема Беляева о редеющих потоках.

4) Вероятность наступления события простейшего (и только простейшего) потока на малом интервале длины Δt пропорциональна длине этого интервала и не зависит от его положения на оси времени (см. разд. 1.2), что дает колоссальные расчетные преимущества.

Эти обстоятельства определяют подавляющее преобладание гипотезы о простейшем потоке в публикациях и решении задач ТМО — часто необоснованное, поскольку эффекты по пп. 2 и 3 вышеприведенного перечня являются лишь асимптотическими.

Пуассоновский поток заявок — может быть стационарным (постоянной интенсивности) или нестационарным; однородным или неоднородным; состоящим из единичных заявок или пачек фиксированного либо случайного объема.

Наряду с простейшим потоком часто рассматривают так называемый *примитивный* поток, связанный с понятием о *замкнутых* системах массового обслуживания. В таких системах имеется конечное число R источников заявок, причем суммарное количество действующих источников и необслуженных заявок постоянно. Если в системе находится k заявок, то входящий поток считается простейшим с мгновенной интенсивностью $\lambda(R-k)$, где λ — интенсивность простейшего потока в расчете на один источник. При $k=R$ поток заявок прерывается.

2.1.5. Распределение числа событий простейшего потока за случайный интервал времени

Пусть λ — параметр входящего потока. Тогда вероятность появления ровно j событий потока за случайное время, подчиненное распределению $B(t)$, должна вычисляться согласно

$$q_j = \int_0^\infty \frac{(\lambda t)^j}{j\,!} e^{-\lambda t}\, dB(t), \qquad j = 0, 1, \ldots \tag{2.1.10}$$

Указанные вероятности (иногда [18] их называют экспоненциальными моментами распределения) играют важную роль в расчете сложных СМО (см. главу 5), и необходимо уметь эффективно их вычислять для возможно более широкого круга распределений. Специальный интерес вызывает расчет вероятности q_0, которая может рассматриваться как преобразование Лапласа с параметром λ от распределения $B(t)$.

При равномерном распределении случайного времени на отрезке $[a-l, a+l]$ искомые вероятности

$$q_j = \frac{1}{2\lambda l}\Big[\sum_{i=0}^{j}\frac{[\lambda(a-l)]^i}{i\,!}e^{-\lambda(a-l)} - \sum_{i=0}^{j}\frac{[\lambda(a+l)]^i}{i\,!}e^{-\lambda(a+l)}\Big], \quad j = 0, 1, \ldots$$

В частности,

$$q_0 = (e^{-\lambda(a-l)} - e^{-\lambda(a+l)})/(2\lambda l).$$

Если t распределено на отрезке $[a-l, a+l]$ по треугольному закону (Симпсона), то

$$\begin{aligned} q_j &= \Big\{\Big[\frac{j+1}{\lambda} - (a-l)\Big]\sum_{i=0}^{j}\frac{[\lambda(a-l)]^i}{i\,!}e^{-\lambda(a-l)} \\ &+ \Big[\frac{j+1}{\lambda} - (a+l)\Big]\sum_{i=0}^{j}\frac{[\lambda(a+l)]^i}{i\,!}e^{-\lambda(a+l)} \\ &+ \frac{1}{\lambda}\frac{[\lambda(a-l)]^{j+1}}{j\,!}e^{-\lambda(a-l)} + \frac{1}{\lambda}\frac{[\lambda(a+l)]^{j+1}}{j\,!}e^{-\lambda(a+l)} \\ &+ 2\Big(a - \frac{j+1}{\lambda}\Big)\sum_{i=0}^{j}\frac{(\lambda a)^i}{i\,!}e^{-\lambda a} - \frac{2}{\lambda}\frac{(\lambda a)^{j+1}}{j\,!}e^{-\lambda a}\Big\}\Big/(\lambda l^2), \\ &\qquad\qquad j = 0, 1, \ldots \end{aligned} \tag{2.1.11}$$

Начальный коэффициент

$$q_0 = \frac{1}{(\lambda l)^2}\Big[e^{-\lambda(a-l)} + e^{-\lambda(a+l)} - 2e^{-\lambda a}\Big].$$

Для распределений с конечным размахом, отличающихся от равномерного и треугольного, может быть применено численное интегрирование по одной из формул интерполяционных квадратур. При этом цикл по j, т. е. перебор номеров коэффициентов, целесообразно делать самым внутренним, а интегральные суммы копить параллельно.

Для гамма-распределения с параметром формы r

$$\begin{aligned} q_j &= \int_0^\infty \frac{(\lambda t)^j}{j\,!} e^{-\lambda t} \frac{\mu(\mu t)^{r-1}}{\Gamma(r)} e^{-\mu t}\, dt \\ &= \frac{\lambda^j \mu^r}{j\,!\,\Gamma(r)} \int_0^\infty t^{r+j-1} e^{-(\lambda+\mu)t}\, dt \\ &= \frac{\lambda^j \mu^r}{j\,!\,(\lambda+\mu)^{r+j}\,\Gamma(r)} \int_0^\infty u^{r+j-1} e^{-u}\, du \\ &= \Big(\frac{\mu}{\lambda+\mu}\Big)^r \Big(\frac{\lambda}{\lambda+\mu}\Big)^j \frac{\Gamma(r+j)}{j\,!\,\Gamma(r)}, \qquad j = 0, 1, \ldots \end{aligned} \tag{2.1.12}$$

Получим рекуррентные формулы вычисления $\{q_j\}$ при времени обслуживания, подчиненном гамма-распределению. Прежде всего, из (2.1.12) следует $q_0 = (\mu/(\lambda+\mu))^r$. Далее,

$$\frac{q_j}{q_{j-1}} = \frac{\lambda}{\lambda+\mu} \frac{\Gamma(r+j)\cdot(j-1)\,!}{j\,!\,\Gamma(r+j-1)} = \frac{\lambda}{\lambda+\mu} \frac{r+j-1}{j}.$$

Итак, при гамма-распределении

$$\begin{aligned} q_0 &= \Big(\frac{\mu}{\lambda+\mu}\Big)^r, \\ q_j &= q_{j-1} \frac{\lambda}{\lambda+\mu} \frac{r+j-1}{j}, \qquad j = 1, 2, \ldots \end{aligned} \tag{2.1.13}$$

Частным случаем гамма-распределения при $r = 1$ является показательное распределение. При этом

$$q_j = \frac{\mu}{\lambda+\mu}\Big(\frac{\lambda}{\lambda+\mu}\Big)^j, \qquad j = 0, 1, \ldots \tag{2.1.14}$$

Наконец, для гамма-плотности с поправочным многочленом искомые вероятности

$$q_j = \Big(\frac{\mu}{\lambda+\mu}\Big)^\alpha \Big(\frac{\lambda}{\lambda+\mu}\Big)^j \frac{1}{j\,!} \sum_{i=0}^{N} \frac{g_i}{(\lambda+\mu)^i} \frac{\Gamma(\alpha+j+i)}{\Gamma(\alpha)}, \qquad j = 0, 1, \ldots \tag{2.1.15}$$

Расчет $\{q_j\}$ и здесь может быть организован рекуррентно.

В табл. 2.1 приведены первые 11 «треугольных» $\{q_j\}$ при $\lambda = 0.7$, $a = 1$ и коэффициенте вариации $v = 0.3$, вычисленные по (2.1.15)

с учетом различного числа N моментов. Точные значения, найденные согласно (2.1.10), совпадают с полученными по шести моментам по крайней мере в трех знаках.

Таблица 2.1. Сравнительный расчет коэффициентов $\{q_j\}$

j	N=2	N=3	N=4	N=6
0	5.07e-1	5.08e-1	5.08e-1	5.08e-1
1	3.34e-1	3.33e-1	3.33e-1	3.33e-1
2	1.20e-1	1.20e-1	1.20e-1	1.20e-1
3	3.10e-2	3.12e-2	3.12e-2	3.12e-2
4	6.49e-3	6.48e-3	6.46e-3	6.46e-3
5	1.16e-3	1.12e-3	1.12e-3	1.12e-3
6	1.85e-4	1.66e-4	1.68e-4	1.68e-4
7	2.68e-5	2.12e-5	2.24e-5	2.24e-5
8	3.60e-6	2.26e-6	2.69e-6	2.68e-6
9	4.52e-7	1.82e-7	2.96e-7	2.92e-7
10	5.40e-8	5.35e-9	3.08e-8	2.91e-8

С помощью процедуры-функции `laplag` вычисляется преобразование Лапласа с параметром s от распределения, аппроксимированного гамма-плотностью с поправочным многочленом.

2.1.6. Обобщенный пуассоновский поток

Здесь рассматривается пуассоновский поток $пачек$ требований, каждая из которых имеет одно и то же распределение $\{f_i\}$ числа заявок. Тогда распределение $\{f_i^{n*}\}$ суммарного числа заявок в n пачках будет n-кратной сверткой распределения $\{f_i^{1*}\} = \{f_i\}$. Для интервала времени с распределением $B(t)$ вероятность прибытия ровно i заявок

$$\begin{aligned} h_i &= \int_0^\infty e^{-\lambda t} \sum_{n=0}^\infty \frac{(\lambda t)^n}{n\,!} f_i^{n*}\, dB(t) \\ &= \sum_{n=0}^\infty f_i^{n*} \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n\,!}\, dB(t). \end{aligned} \tag{2.1.16}$$

Интегралы следует вычислить предварительно согласно рекомендациям предыдущего подраздела, а свертки (для конечного размаха) последовательно получать с помощью алгоритма дискретной свертки и выполнять подсуммирование раздельно для каждого i. Нужно иметь в виду

происходящее на каждом шаге свертки удлинение массива вероятностей $\{f_i^{n*}\}$.

Таблица 2.2. Обобщенный пуассоновский поток

n	Расчет	Модель	n	Расчет	Модель
0	.43143e-0	.43237e-0	20	.16770e-2	.17110e-2
1	.49974e-1	.49654e-1	21	.12536e-2	.12640e-2
2	.54396e-1	.53787e-1	22	.91291e-3	.88300e-3
3	.59148e-1	.59052e-1	23	.64986e-3	.67300e-3
4	.64250e-1	.64103e-1	24	.45644e-3	.50100e-3
5	.69724e-1	.69234e-1	25	.32116e-3	.32400e-3
6	.75594e-1	.75437e-1	26	.22857e-3	.24000e-3
7	.31910e-1	.32270e-1	27	.15886e-3	.18600e-3
8	.29800e-1	.30080e-1	28	.10808e-3	.12900e-3
9	.27174e-1	.27388e-1	29	.72346e-4	.91000e-4
10	.23976e-1	.24203e-1	30	.48001e-4	.55000e-4
11	.20146e-1	.20224e-1	31	.31759e-4	.46000e-4
12	.15622e-1	.15626e-1	32	.20907e-4	.30000e-4
13	.10335e-1	.10292e-1	33	.13506e-4	.19000e-4
14	.86349e-2	.86930e-2	34	.86336e-5	.12000e-4
15	.70107e-2	.68830e-2	35	.55360e-5	.80000e-5
16	.55038e-2	.53480e-2	36	.36252e-5	.60000e-5
17	.41612e-2	.40850e-2	37	.24694e-5	.30000e-5
18	.30353e-2	.29600e-2	38	.17744e-5	.10000e-5
19	.21854e-2	.21270e-2	39	.13579e-5	.00000e-0

В табл. 2.2 приводится сопоставление расчета по вышеописанной схеме и моделирования (1 млн. испытаний) числа заявок обобщенного пуассонова потока. Объем пачки предполагался равновероятным в диапазоне 1..6, интервал времени — равномерно распределенным на интервале [0,10].

2.1.7. Случайное прореживание потоков

Пусть в рекуррентном входящем потоке с ПЛС интервалов между заявками $\alpha(s)$ каждая заявка сохраняется в потоке с вероятностью z независимо от остальных заявок. Тогда для просеянного потока ПЛС

плотности распределения интервалов между заявками

$$\varphi(s) = \sum_{k=1}^{\infty} z(1-z)^{k-1}\alpha^k(s) = \frac{z\alpha(s)}{1-(1-z)\alpha(s)}. \tag{2.1.17}$$

Преобразование Лапласа от ДФР интервала между заявками

$$\bar{\Phi}_z(s) = \frac{1}{s}\left[1 - \frac{1-\alpha(s)}{1-(1-z)\alpha(s)}\right]. \tag{2.1.18}$$

Для получения моментов результирующего распределения через моменты исходного воспользуемся стандартной технологией разд. 1.7. Получим

$$\begin{aligned} f_1 &= a_1/z, \\ f_k &= [a_k + (1-z)\sum_{i=1}^{k-1}\binom{k}{i} f_i a_{k-i}]/z, \qquad k = 2,3,\ldots \end{aligned} \tag{2.1.19}$$

В частности,

$$f_2 = [a_2 + 2(1-z)f_1 a_1]/z.$$

Вычисляя коэффициент немарковости ξ_2 для распределения интервалов между заявками просеянного потока, убеждаемся, что

$$\xi_2^{(f)} = z\xi_2^{(a)}. \tag{2.1.20}$$

Эта неожиданно простая формула в усложненном виде (для квадрата коэффициента вариации) широко используется в расчете сетей обслуживания с потоками, аппроксимируемыми по двум моментам распределений [178]. При $z \to 0$ просеянный поток по ξ_2 стремится к простейшему (см. разд. 2.1.4), сохраняя знак $\xi_2^{(a)}$. Для высших коэффициентов немарковости $\xi_i^{(f)}$ соотношение вида (2.1.20) выполняется лишь приближенно — см. табл. 2.3. Однако получаемая погрешность мала в сравнении с $i!$ — см. формулу (1.4.4). Следовательно, можно считать

$$\xi_i^{(f)} \approx z\xi_i^{(a)}, \qquad i = 2,3,\ldots \tag{2.1.21}$$

Таблица 2.3. Просеивание регулярного потока

z	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\xi_3^{(f)}$	-1.16	-1.71	-2.24	-2.75	-3.24	-3.71	-4.16	-4.59	-5.00
$z\xi_3^{(a)}$	-1.00	-1.50	-2.00	-2.50	-3.00	-3.50	-4.00	-4.50	-5.00

В случае простейшего потока формула (2.1.18) сводится к

$$\bar{\Phi}(s) = \frac{1}{s}\,\frac{1-\frac{\lambda}{\lambda+s}}{1-(1-z)\frac{\lambda}{\lambda+s}} = \frac{1}{s+\lambda z},$$

что соответствует оригиналу $e^{-\lambda z t}$. Таким образом, простейший поток при случайном прореживании с любым z остается простейшим — меняется только его интенсивность.

2.1.8. Регулярный поток и регулярное прореживание

Процесс обслуживания любого рода легче организовать (но не рассчитать!) для *регулярного* потока — с постоянными интервалами между смежными заявками. Для такого потока коэффициенты немарковости

$$\xi_i^{(a)} = a^i/a^i - i! = 1 - i!, \quad i = 2, 3, \ldots$$

В частности, $\xi_2^{(a)} = -1, \quad \xi_3^{(a)} = -5$.

Справедливое (циклическое) распределение поступающих заявок между k обслуживающими устройствами порождает для каждого из них поток с регулярным прореживанием (остается k-я заявка исходного потока). Соответственно распределение интервалов между заявками оказывается n-кратной сверткой исходного распределения, а его ПЛС — k-й степенью исходного:

$$\begin{aligned}
\varphi_k(s) &= \left(1 - a_1 s + \frac{a_2}{2}s^2 - \frac{a_3}{6}s^3 + \ldots\right)^k + o(s^3) \\
&= 1 - k a_1 s + \left[\frac{k(k-1)}{2}a_1^2 + \frac{k}{2}a_2\right] s^2 \\
&+ \left[\frac{k(k-1)(k-2)}{6}a_1^3 - \frac{k(k-1)}{2}(a_1 a_2 - a_3/3)\right] s^3 + o(s^3).
\end{aligned}$$

Моменты просеянного потока

$$\begin{aligned}
f_1 &= k a_1, \\
f_2 &= k(k-1)a_1^2 + k a_2, \\
f_3 &= k(k-1)(k-2)a_1^3 - 3k(k-1)(a_1 a_2 - a_3/3).
\end{aligned}$$

По этим выражениям вычислим предельные при $k \to \infty$ коэффициенты немарковости:

$$\begin{aligned}
\xi_2^{(f)} &= \lim_{k\to\infty} \frac{k(k-1)a_1^2 + ka_2}{k^2 a_1^2} - 2 = 1 - 2 = -1, \\
\xi_3^{(f)} &= \lim_{k\to\infty} \frac{k(k-1)(k-2)a_1^3 - 3k(k-1)(a_1a_2 - a_3/3)}{k^3 a_1^3} - 6 \\
&= \lim_{k\to\infty} \left[\frac{(k-1)(k-2)}{k^2} - \frac{3(k-1)(a_1a_2 - a_3/3}{k^2 a_1^3} \right] - 6 = -5.
\end{aligned}$$

Таким образом, произвольный рекуррентный поток при увеличении интенсивности *регулярного* просеивания стремится к регулярному.

2.1.9. Суммирование потоков

Актуальность задачи суммирования потоков и явная недостаточность учета при этом только средних породила многочисленные попытки ее приближенного решения на уровне двух моментов.

Наибольшее распространение получили *эмпирические* формулы для квадратов коэффициентов вариации вида

$$C_\Sigma = \sum_i \left(\frac{\lambda_i}{\Lambda} \right)^k C_i \tag{2.1.22}$$

(в [177, 178, 275] предлагается $k = 2$, в [36] $k = 3$). Для N статистически одинаковых потоков эта формула сводится к

$$C_N = \sum_{i=1}^{N} C\, N^{-k} = C/N^{k-1}.$$

Но тогда отношение $C_N/C = N^{-(k-1)}$ не зависит от C и в логарифмическом масштабе должно давать прямую линию. Графики этого отношения на рис. 2.1, рассчитанные по описанной ниже более точной методике, эти утверждения опровергают. Таким образом, аппроксимации вида (2.1.22) оказываются несостоятельными.

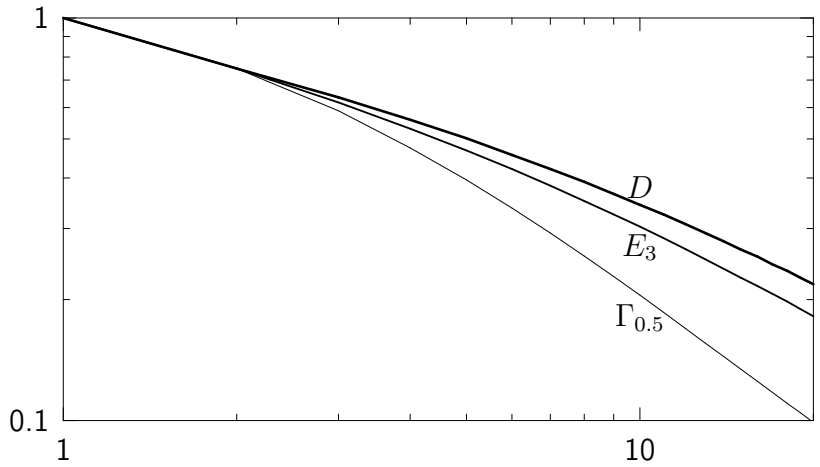

Рис. 2.1. Проверка эмпирических формул суммирования потоков

Попытки построения приближенных формул продолжаются и в наши дни [202, 258].

Смысл операции суммирования потоков иллюстрирует рис. 2.2.

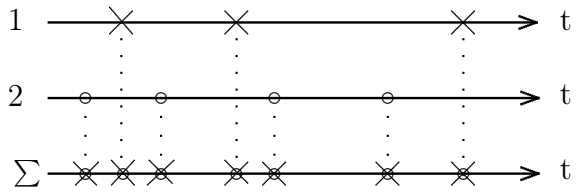

Рис. 2.2. Схема суммирования потоков

Момент появления очередной заявки суммарного потока — это минимум из моментов появления ближайших заявок составляющих. Если предыдущей была заявка первого потока, то распределение времени ожидания заявки второго заменяется на соответствующее остаточное распределение (см. разд. 1.7), и наоборот. Частота выбора вариантов определяется удельным весом заявок каждого типа в суммарном потоке, т. е. отношениями $\{\lambda_i/\Lambda\}$. Итак, ДФР интервалов в суммарном потоке

$$\bar{A}_\Sigma(t) = \frac{\lambda_1}{\Lambda}\bar{A}_1(t)\bar{A}_2^*(t) + \frac{\lambda_2}{\Lambda}\bar{A}_2(t)\bar{A}_1^*(t). \qquad (2.1.23)$$

Эта формула легко реализуется при гиперэкспоненциальной аппроксимации составляющих распределений. Для вывода моментов распределения интервалов суммарного потока воспользуемся формулой

$$f_k = \int_0^\infty t^k f(t)\,dt = k\int_0^\infty t^{k-1}\bar{F}(t)\,dt.$$

Теперь составляющие моментов распределения с ДФР типа (2.1.22) могут быть вычислены согласно

$$\begin{aligned} f_k &= k\int_0^\infty t^{k-1}\Big(\sum_{i=1}^2 y_i e^{-\mu_i t}\Big)\Big(\sum_{j=1}^2 u_j e^{-\lambda_j t}\Big)\,dt \\ &= k\sum_{i=1}^2 y_i \sum_{j=1}^2 u_j \int_0^\infty t^{k-1}e^{-(\mu_i+\lambda_j)t}\,dt = \sum_{i=1}^2\sum_{j=1}^2 y_i u_j\; k\,!/(\mu_i+\lambda_j)^k. \end{aligned} \tag{2.1.24}$$

Рассмотрим вопрос о случайной модификации H_2-распределения. Поскольку реализация этого закона состоит в выборе одной из экспонент, то его случайная модификация приведет к той же самой экспоненте с сохранением μ_i и исходных ее моментов. Вероятность выбора i-й экспоненты пропорциональна вероятности y_i и средней длительности задержки $b_i = 1/\mu_i$. После нормировки вероятностей к единице получаем новые «вероятностные коэффициенты»

$$y_1^* = y_1\mu_2/(y_1\mu_2 + y_2\mu_1), \quad y_2^* = y_2\mu_1/(y_1\mu_2 + y_2\mu_1). \tag{2.1.25}$$

Этот подход дает три верных момента остаточного распределения для аппроксимирующей гиперэкспоненты, но только два — по отношению к исходному распределению. Погрешность в расчете третьего растет при увеличении отклонения коэффициента вариации исходного распределения от единицы и может оказаться недопустимо большой. Приведем соответствующую таблицу для аппроксимации Г-распределения с единичным средним в зависимости от коэффициента вариации v:

Таблица 2.4. Третий момент остаточного распределения

Распределение	Коэффициент вариации v					
	0.2	0.3	0.5	0.7	2.0	3.0
Исходное	3.14e-1	4.08e-1	8.20e-1	1.82e-0	1.46e+2	1.33e+3
H_2-аппр.	2.38e-1	3.41e-1	7.81e-1	1.82e-0	1.38e+2	1.22e+3

Аппроксимируя H_2-законами распределения интервалов суммарного потока и очередного слагаемого, можно выполнить последовательное суммирование любого числа потоков. При большом числе составляющих имеет смысл организовать суммирование по схеме двоичного дерева. На рис. 2.3 показана зависимость второго и третьего нормированных (деленных на $k!$) коэффициентов немарковости суммарного потока от числа составляющих (первое число — параметр гамма-распределения слагаемых, второе — номер коэффициента). Эти графики (и формула (2.1.20)) показывают, что как при суммировании, так и при прореживании потоков, заметно отличающихся от простейшего, сходимость результирующего потока к простейшему может быть довольно медленной (из рисунка следует, что она очень близка к экспоненциальной по обоим коэффициентам немарковости).

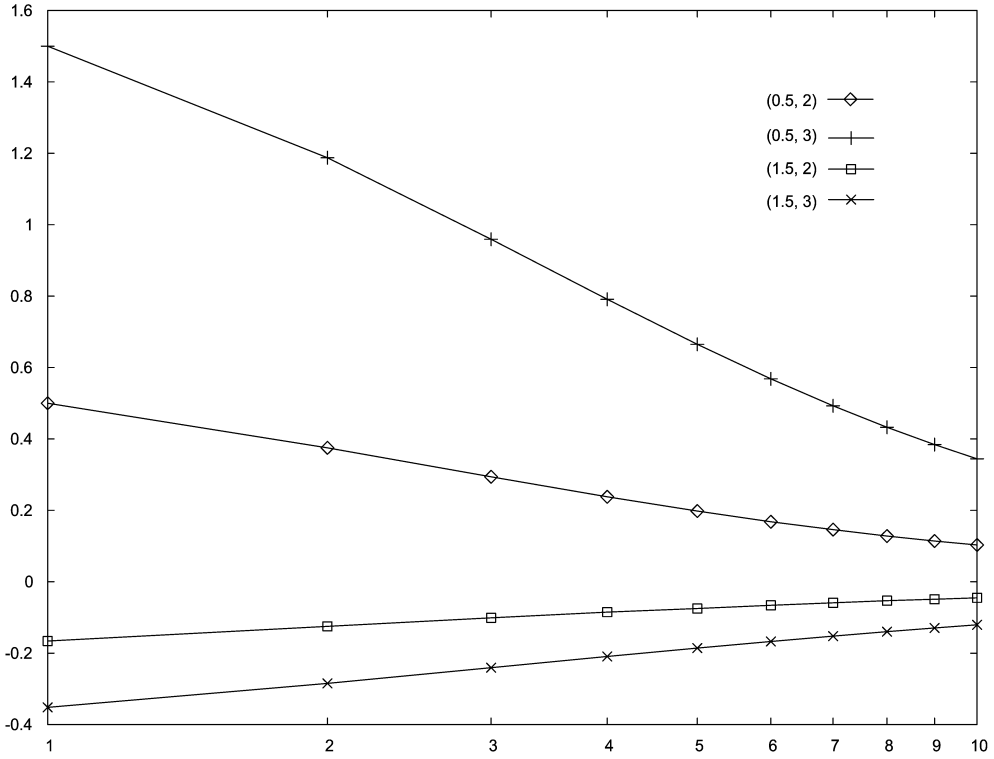

Рис. 2.3. Результаты суммирования потоков

Нетрудно привести пример ситуации, порождающей «неудобные» потоки: выходящий поток сильно загруженной одноканальной системы полностью определяется распределением длительности обслуживания. Поскольку вид распределения интервалов между заявками существенно влияет на характеристики СМО (см. главу 5), постулат о простейшем входящем потоке должен критически оцениваться в каждом конкретном случае.

Сумма пуассоновых потоков дает пуассонов поток суммарной интенсивности.

2.1.10. Суммирование в терминах семиинвариантов

Наряду с моментами важными числовыми характеристиками распределений случайных величин являются семиинварианты $\{\kappa_r\}$, для которых производящей функцией является логарифм функции характеристической. Это определение приводит к тождеству

$$\exp\Bigl(\kappa_1 s + \kappa_2 \frac{s^2}{2!} + \ldots \kappa_r \frac{s^r}{r!} + \ldots\Bigr) = 1 + a_1 s + a_2 \frac{s^2}{2!} + \ldots a_r \frac{s^r}{r!} + \ldots,$$

связывающему семиинварианты и начальные моменты $\{a_r\}$ распределения $A(t)$ [47]. Если существуют моменты порядка r и ниже, то существуют и семиинварианты соответствующих порядков, и обратно. Семиинварианты выражаются через моменты формулами

$$\begin{array}{rcl} \kappa_1 &=& a_1, \\ \kappa_2 &=& a_2 - a_1^2, \\ \kappa_3 &=& a_3 - 3a_2 a_1 + 2a_1^3, \\ \kappa_4 &=& a_4 - 4a_3 a_1 - 3a_2^2 + 12 a_2 a_1^2 - 6a_1^4, \\ \multicolumn{3}{c}{\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots} \end{array} \qquad (2.1.26)$$

Обратная задача решается по формулам

$$\begin{array}{rcl} a_1 &=& \kappa_1, \\ a_2 &=& \kappa_2 + \kappa_1^2, \\ a_3 &=& \kappa_3 + 3\kappa_1 \kappa_2 + \kappa_1^3, \\ a_4 &=& \kappa_4 + 4\kappa_3 \kappa_1 + 3\kappa_2^2 + 6\kappa_2 \kappa_1^2 + \kappa_1^4, \\ \multicolumn{3}{c}{\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots} \end{array} \qquad (2.1.27)$$

(в книге [47] эти системы доведены до $r = 10$).

В теории рекуррентных потоков (см. [53, с.71,166–168]) доказывается, что семиинварианты $\{\nu_r\}$ распределения числа событий за время t при $t \to \infty$ асимптотически равны

$$\begin{array}{rcl} \nu_1 & = & t/\kappa_1, \\ \nu_2 & = & t\kappa_2/\kappa_1^3, \\ \nu_3 & = & t(3\kappa_2^2/\kappa_1^5 - \kappa_3/\kappa_1^4), \\ \nu_4 & = & t(\kappa_4/\kappa_1^5 - 10\kappa_2\kappa_3/\kappa_1^6 + 15\kappa_2^3/\kappa_1^7), \\ \ldots\ldots & & \ldots\ldots \end{array} \tag{2.1.28}$$

(в формуле для ν_4 исправлена обнаруженная опечатка). Соответственно их нормированные (деленные на t) значения

$$\begin{array}{rcl} \bar\nu_1 & = & 1/\kappa_1, \\ \bar\nu_2 & = & \kappa_2/\kappa_1^3, \\ \bar\nu_3 & = & 3\kappa_2^2/\kappa_1^5 - \kappa_3/\kappa_1^4, \\ \bar\nu_4 & = & \kappa_4/\kappa_1^5 - 10\kappa_2\kappa_3/\kappa_1^6 + 15\kappa_2^3/\kappa_1^7, \\ \ldots\ldots & & \ldots\ldots \end{array} \tag{2.1.29}$$

Обратная задача восстановления семиинвариантов $\{\kappa_r\}$ по $\{\bar\nu_r\}$ решается системой

$$\begin{array}{rcl} \kappa_1 & = & 1/\bar\nu_1, \\ \kappa_2 & = & \bar\nu_2/\bar\nu_1^3, \\ \kappa_3 & = & (3\bar\nu_2^2 - \bar\nu_1\bar\nu_3)/\bar\nu_1^5, \\ \kappa_4 & = & [\bar\nu_4/\bar\nu_1^2 + 5\bar\nu_2(3\bar\nu_2^2 - 2\bar\nu_1\bar\nu_3)]/\bar\nu_1^7, \\ \ldots\ldots & & \ldots\ldots \end{array} \tag{2.1.30}$$

Семиинварианты обладают свойством *аддитивности*: семиинвариант суммы независимых случайных величин равен сумме соответствующих семиинвариантов составляющих. При суммировании независимых потоков благодаря свойству аддитивности будут суммироваться и их нормированные семиинварианты распределения числа событий. А. Н. Екимцов предложил следующую схему расчета моментов распределения интервалов между событиями суммарного потока:

1) Обнулить суммарные инварианты числа событий.

2) Для всех составляющих потоков $i = \overline{1, N}$

   - согласно (2.1.26) найти семиинварианты $\{\kappa_r^{(i)}\}$ распределения интервалов между заявками;

- вычислить нормированные семиинварианты числа событий $\{\bar{\nu}_r^{(i)}\}$ по формулам (2.1.29) и добавить их к соответствующим $\{\bar{\nu}_r^{(\Sigma)}\}$.

3) Согласно (2.1.30) вычислить семиинварианты $\{\kappa_r^{(\Sigma)}\}$ распределения интервалов между заявками суммарного потока.

4) С помощью (2.1.27) перейти к начальным моментам этого распределения.

2.1.11. «Патологические» потоки

Прежде всего отметим, что сравнительно часто встречаются рекуррентные потоки с конечным размахом интервалов — при жестких технологических ограничениях на минимальные и максимальные интервалы поступления. Их суммирование или просеивание приближают результирующий поток к простейшему.

Обсуждавшаяся выше теория относилась к потокам, в которых распределения интервалов между смежными заявками имеют требуемое число начальных моментов (при использовании гиперэкспоненциальных аппроксимаций — не менее трех). Имитационные эксперименты с распределениями Парето [37] показали *отсутствие сходимости* процессов суммирования и случайного прореживания потоков с исходными распределениями этого класса.

Недавно было обнаружено, что столь популярная гипотеза о близости реальных потоков к простейшим далека от истины в случае, касающемся практически всего человечества. В. Н. Задорожный [37] цитирует доклад В. Столлингса, в котором представлены результаты обширного анализа трафика в Интернете и показано, что трафик имеет фрактальную (самоподобную) структуру. Отсюда следует, что его статистические параметры не зависят от выбранного масштаба времени, и надежды на сглаживание трафика на длительных промежутках времени не имеют под собой никаких оснований. Вместо этого происходят не только уплотнение и разрежение потока данных, но и кластеризация самих этих уплотнений и разрежений.

2.2. Процесс обслуживания

2.2.1. Общие соображения

Системы обслуживания по числу установленных устройств делятся на одно- и многоканальные. Количество требований, одновременно могущих находиться на обслуживании, не превышает числа каналов n. При очень большом числе каналов можно считать $n = \infty$. Каналы могут быть однородными, специализированными по типам заявок, различающимися интенсивностями обслуживания и т.п. В случае однородных каналов они могут быть упорядоченными (тогда очередное требование занимает свободный канал обязательно с наименьшим номером) или неупорядоченными, полнодоступными и неполнодоступными. В дальнейшем (исключая разд. 7.11) мы не будем делать между каналами никакого различия.

Пусть распределение времени обслуживания задано функцией $B(t)$. Тогда аналогично проведенному выше исследованию потока заявок можно определить вероятность продолжительности обслуживания в полуинтервале $[t, t + \Delta t)$ согласно

$$P_{\text{обсл}}(t, \Delta t) = \frac{\bar{B}(t) - \bar{B}(t + \Delta t)}{\bar{B}(t)} + o(\Delta t).$$

Соответственно условная плотность окончания обслуживания

$$\mu(t) = -\frac{\bar{B}'(t)}{\bar{B}(t)} = \frac{B'(t)}{\bar{B}(t)}, \tag{2.2.1}$$

откуда следует

$$B(t) = 1 - \exp(-\int_0^t \mu(\tau)\, d\tau).$$

По смыслу параметра потока — см. формулу (2.1.2) — вероятность обслуживания хотя бы одного требования в полуинтервале $[t, t + \Delta t)$ с точностью $o(\Delta t)$ равна $\mu(t)\Delta t$. Полагая $\mu = \text{const}$, находим единственное распределение, для которого мгновенное значение параметра обслуживания постоянно: $B(t) = 1 - e^{-\mu t}$.

2.2.2. Показательное распределение обслуживания

В случае показательного распределения нетрудно найти вероятность окончания хотя бы одного обслуживания для многоканальной системы. Пусть обслуживанием заняты k каналов. Тогда каждый из них за малый интервал длины Δt закончит обслуживание с вероятностью $\mu\Delta t + o(\Delta t)$. По определению, параметр потока обслуживаний

$$\mu_k = \lim_{\Delta t \to 0} \frac{1 - (1 - \mu\Delta t)^k + o(\Delta t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{k\mu\Delta t + o(\Delta t)}{\Delta t} = k\mu,$$

а вероятность обслуживания за Δt двух и более заявок имеет порядок $o(\Delta t)$. Таким образом, в многоканальной системе при экспоненциально распределенной длительности обслуживания поток окончаний обслуживания при нахождении в каналах k требований является *ординарным*, а его параметр

$$\mu_k = \begin{cases} k\mu, & k = \overline{0, n-1}, \\ n\mu, & k = n, n+1, \ldots \end{cases} \tag{2.2.2}$$

(интенсивность обслуживания ограничивается числом каналов). Если допускается взаимопомощь между каналами (например, при распараллеливании сложной задачи на многопроцессорной вычислительной машине), то для $k \geq 1$ $\mu_k = n\mu$. В обоих случаях распределение интервалов между смежными завершениями обслуживания остается показательным с параметром μ_k.

2.2.3. Произвольное распределение

Сложнее обстоит дело при других распределениях длительности обслуживания. ДФР распределения интервалов между последовательными уходами из n-канальной полностью занятой системы обслуженных запросов можно (см. Саати [132, с. 253]) описать формулой

$$\bar{B}_n(t) = [\bar{B}^*(t)]^{n-1}\bar{B}(t), \tag{2.2.3}$$

которая имеет прозрачный вероятностный смысл: для одного из каналов (только что завершившего обслуживание) берется полное распределение длительности обслуживания, а для прочих — остаточное.

Возможность применения формулы (2.2.3) была проверена на имитационной модели, предварительно откалиброванной по задаче с показательным распределением длительности обслуживания (в этом случае

$\bar{B}_n(t) = e^{-n\mu t}$). Рабочее моделирование выполнялось для распределений длительности обслуживания Эрланга 3-го порядка (E_3), коэффициент вариации $v = 0.577$, и гиперэкспоненциального H_2, $v = 1.36$, при средней длительности обслуживания $b_1 = 5$. Результаты расчета моментов $\bar{B}_n(t)$ на основании формулы Саати (С) и посредством модели при 5000 испытаний (М) представлены в табл. 2.5.

Анализ таблицы с учетом роста относительных погрешностей моделирования по порядку вычисляемых моментов указывает на допустимость использования формулы (2.2.3).

Таблица 2.5. Моменты интервалов между обслуживаниями

Тип $B(t)$	Порядок момента	Число каналов n					
		2		3		4	
		М	С	М	С	М	С
E_3	1	2.50e0	2.50e0	1.62e0	1.67e0	1.22e0	1.25e0
	2	9.48e0	9.46e0	4.28e0	4.46e0	2.50e0	2.60e0
	3	4.64e1	4.57e1	1.47e1	1.57e1	6.83e0	7.21e0
	4	2.75e2	2.64e2	6.08e1	6.66e1	2.30e1	2.45e1
	5	1.90e3	1.77e3	2.90e2	3.28e2	9.13e1	9.72e1
	6	1.49e4	1.36e4	1.55e3	1.83e3	4.20e2	4.36e2
H_2	1	2.62e0	2.45e0	1.68e0	1.59e0	1.27e0	1.16e0
	2	1.70e1	1.58e1	6.48e0	6.57e0	3.62e0	3.48e0
	3	2.05e2	1.75e2	4.36e1	4.54e1	1.75e1	1.74e1
	4	3.95e3	2.94e3	4.41e2	4.55e2	1.25e2	1.25e2
	5	1.09e5	6.94e4	5.88e3	6.18e3	1.17e3	1.19e3
	6	3.80e6	2.13e6	9.37e4	1.09e5	1.30e4	1.43e4

2.2.4. Время до ближайшего обслуживания

При анализе работы многоканальных систем возникает вопрос о распределении времени ожидания вновь прибывшей заявкой ближайшего завершения обслуживания.

Аппроксимируем *остаточные* распределения длительности обслуживания в каналах законом Вейбулла. Здесь ДФР имеет вид $\bar{B}(t) = e^{-t^k/T}$. Вероятность того, что обслуживание не завершится ни в одном из n каналов, составит

$$\bar{B}_n(t) = e^{-nt^k/T}, \tag{2.2.4}$$

т. е. описывается тем же распределением с заменой T на T/n. Подставляя это значение в формулы для моментов распределения Вейбулла, убеждаемся, что его моменты порядка m получаются делением исходных на $n^{m/k}$.

Ту же аппроксимацию можно выполнить гиперэкспоненциальным законом с параметрами $\{y_k, \mu_k\}$. Теперь для ДФР имеем выражение

$$\bar{B}_n(t) = \left(\sum_{k=1}^{2} y_k e^{-\mu_k t} \right)^n = \sum_{i=0}^{n} \binom{n}{i} y_1^{n-i} y_2^i e^{[(n-i)\mu_1 + i\mu_2]t}. \qquad (2.2.5)$$

Соответственно среднее время до ближайшего завершения обслуживания составит

$$\bar{b}^* = \sum_{i=0}^{n} \binom{n}{i} \frac{y_1^{n-i} y_2^i}{(n-i)\mu_1 + i\mu(2)}. \qquad (2.2.6)$$

Результаты расчета с помощью вышеуказанных аппроксимаций при единичной средней длительности обслуживания представлены в табл. 2.6.

Поскольку при показательном законе ($v_B = 1$) ожидаемая остаточная длительность обслуживания совпадает с полной, среднее время до ближайшего обслуживания при $b_1 = 1$ должно быть равно $1/n$. Получение этого результата подтверждает правильность расчетных методик для обеих аппроксимаций. С другой стороны, при немарковских распределениях наблюдается заметное расхождение — в особенности при большом числе каналов. Предпочтение следует отдать H_2-аппроксимации, позволяющей учеть большее число моментов.

Отметим, наконец, возможность решения обсуждаемой задачи с помощью методики суммирования потоков.

Иногда приходится описывать *групповое обслуживание*. Возможны также варианты начала обслуживания с накоплением заявок:

T — за указанное время (например , автомобили у светофоров);

N — по числу заявок (в поликлиниках на рентген, на станциях микроавтобусов, на учебных курсах с началом занятий по мере комплектования групп);

X — по объему заявок (накопление грузов) у отправителя.

Таблица 2.6. Среднее время ожидания ближайшего обслуживания

n	$v_B = 2$		$v_B = 1$		$v_B = 1/\sqrt{3}$	
	H_2	Wb	H_2	Wb	H_2	Wb
1	2.5000	2.5000	1.0000	1.0000	0.6667	0.6667
2	1.1111	1.1048	0.5000	5.0000	0.3778	0.3797
3	0.7015	0.6852	0.3333	0.3333	0.2693	0.2732
4	0.5109	0.4883	0.2500	0.2500	0.2110	0.2163
5	0.4014	0.3754	0.2000	0.2000	0.1743	0.1804
6	0.3306	0.3028	0.1667	0.1667	0.1488	0.1556
7	0.2810	0.2525	0.1429	0.1429	0.1301	0.1373
8	0.2443	0.2158	0.1250	0.1250	0.1156	0.1232
9	0.2161	0.1878	0.1111	0.1111	0.1042	0.1195
10	0.1938	0.1659	0.1000	0.1000	0.0948	0.1028

2.3. Организация и продвижение очереди

Заявки, пришедшие в занятую систему, не могут быть обслужены немедленно и образуют $очередь$. Очередь может быть ограничена максимальной длиной R или максимальным временем $\hat{W}$ пребывания в ней. Примером задачи с временны́м ограничением является прибытие на стройку самосвала с бетонной смесью. При нарушении ограничения заявка получает отказ. Введение ограничения автоматически исключает очень большие задержки, но связано с дополнительными «штрафами» за отказ в обслуживании.

Вновь прибывшая заявка в зависимости от организации и назначения системы становится либо в конец очереди (дисциплина FCFS: First Come — First Served), либо в ее начало (LCFS: Last Come — First Served). Последний вариант иначе называется стековым («магазинным») принципом и широко применяется в различных областях вычислительной техники — в частности, при обработке быстро стареющей информации. По этому же принципу выбираются слитки для нагрева в колодцах перед прокатным станом [216]. Разумеется, самым популярным примером магазинного принципа остается выход патронов в ствол автомата Калашникова.

Возможен и случайный выбор из очереди (SIRO — Service In Random Order). Именно так обстоит дело при соединении абонентов АТС с одним из перегруженных номеров. Как другой пример такой дисципли-

ны в [216] приводится посадка на автобусы в Риме (по мнению автора этой книги, в Лондоне торжествует FCFS). Дисциплина IS (Immediate Service) присуща системе с бесконечным числом каналов.

При неоднородных заявках может вводиться *приоритетное* обслуживание. В этом случае заявки выстраиваются в несколько очередей, и в освободившийся канал поступает заявка из непустой очереди с наивысшим приоритетом. В некоторых ситуациях (абсолютный приоритет) срочная заявка может прервать уже начатое обслуживание. Снятая заявка поступает в одну из очередей или теряется. Примером заявки с абсолютным приоритетом является поломка обслуживающего устройства. Обслуживание подобной заявки состоит в отыскании и устранении неисправности. Поток поломок каналов обслуживания формируется *замкнутым* источником заявок, что не позволяет рассчитывать подобные СМО простой сменой интерпретации заявок с наивысшим абсолютным приоритетом. Варианты приоритетных дисциплин подробнее обсуждаются в главе 9.

Приоритет может предоставляться наикратчайшей заявке — дисциплина Shortest Remaining Processing Time (SRPT) с возможным прерыванием и дообслуживанием прерванной заявки. Эта дисциплина предполагает знание длины каждой поступающей в систему заявки и запоминание всех «остаточных» данных. SRPT минимизирует число заявок в системе в любой момент времени в классе консервативных (сохраняющих объем работы) дисциплин.

В вычислительных системах с разделением времени практикуется *циклическое* обслуживание — заявки с каждого абонентского пункта образуют отдельную очередь, и центральный процессор переходит от одной очереди к другой по кругу. Эта дисциплина (polling) характерна и для некоторых вариантов организации связи.

В ряде ситуаций целесообразно введение *порога обслуживания* — оно начинается при скоплении в системе $m > 1$ требований и заканчивается при полном рассасывании очереди. Этот подход может сочетаться с обслуживанием по кругу. К дисциплинам обслуживания можно отнести и такие факторы, как «разогрев» (при прибытии первой заявки в свободную систему); случайная длительность «прогулки» освободившегося устройства обслуживания; случайное ограничение времени ожидания заявки.

В многоканальных системах возникает новый аспект управления очередью — распределение накопившихся заявок между каналами.

Ясно, что если номер канала для k-й заявки определяется по правилу $k \pmod n + 1$, то система просто разделяется на n независимых с регулярно прореженными входными потоками. Рассматриваются также варианты со случайным распределением по очередям, зависящим и не зависящим от текущих длин очередей, а также с допущением переходов между очередями или без оных. С точки зрения загрузки системы и улучшения временны́х показателей оптимально наличие *общей* очереди. Обычно имется в виду именно этот вариант.

2.4. Классификация и обозначение моделей теории очередей

Для сокращенного наименования абстрактных СМО Д. Кендалл предложил использовать обозначения вида $A/B/n/R$, где A указывает распределение интервалов между требованиями, B — распределение времени обслуживания, n — число каналов, R — предельное число заявок в очереди. Типы распределений обозначаются следующим образом:

M — показательное (обладающее марковским свойством),

E_r — эрланговское порядка r,

D — детерминированное: постоянное время обслуживания или регулярный поток (D иногда интерпретируется как δ-функция Дирака),

H_k — гиперэкспоненциальное с k составляющими,

C_k — коксово с k составляющими,

Pa — распределение Парето,

Ph — общее распределение фазового типа,

G — произвольное распределение (для потока — GI).

При $R \to \infty$ четвертую позицию принято опускать. Таким образом, запись вида $M/G/1$ означает одноканальную СМО с простейшим входящим потоком, произвольным распределением времени обслуживания и неограниченной очередью.

Развитие ТМО привело к созданию модифицированных обозначений для новых классов исследованных систем: замкнутых, с неординарным потоком заявок, групповым обслуживанием, различными вариантами приоритетных дисциплин и т. д. — см., например, [86]. При анализе СМО, работающих в дискретном времени, используется $Geom$ — геометрическое распределение. Мы будем применять небольшое число модификаций:

$\rightarrow$ для обозначения неоднородных элементов системы,

$\hat{}$ для указания замкнутости по входящему потоку,

M^X для обозначения простейшего потока $пачек$ случайного объема;

$-T$ в позиции последнего элемента — для указания распределения предельной длительности пребывания заявки в очереди (вместо T подставляется обозначение одного из вышеприведенных законов).

Кроме того, в связи с обычно используемой идентификацией состояний СМО по полному числу заявок в ней будем использовать четвертую позицию схемы Кендалла для указания предельного числа заявок в $системе$. Именно так поступили и авторы [165].

Основная часть книги посвящена исследованию немарковских систем, в которых исходные распределения A и/или B не обладают марковским свойством.

2.5. Показатели эффективности

2.5.1. Перечень показателей

Показатели эффективности процессов обслуживания обычно устанавливаются для стационарного (предельного при $t \to \infty$) режима. Их можно разделить на две группы — счетные и временны́е. Счетные показатели связаны со стационарным распределением $\{p_j\}$ числа заявок в системе или финальным распределением $\{\pi_j\}$ числа заявок перед прибытием очередной заявки либо сразу после завершения обслуживания (эти распределения используются в разных целях и в общем случае различны — см. табл. 2.7, построенную по данным имитационного моделирования при 500000 испытаний). Из таблицы следует, что вероятности «Before» и «After» практически совпадают во всех случаях, чего

и следовало ожидать — из соображений симметрии. Со стационарными вероятностями они совпадают только в случае простейшего входящего потока. Этот факт был установлен в теореме PASTA — Poisson Arrivals See Time Average (ее доказательство приводится, например, в [256]). А. Я. Хинчин считал это свойство основным законом стационарной очереди [28, с. 40]. Принцип PASTA не относится к замкнутым системам и групповым потокам. В некоторых случаях принцип верен и для непуассонова потока.

Таблица 2.7. Сопоставление вероятностей состояний

j	Система M/D/3			Система D/M/3		
	Before	Stationary	After	Before	Stationary	After
0	2.155e-2	2.207e-2	2.163e-2	1.045e-1	4.384e-2	1.045e-1
1	6.708e-2	6.714e-2	6.731e-2	3.149e-1	2.190e-1	3.149e-1
2	1.083e-1	1.083e-1	1.086e-1	3.106e-1	3.313e-1	3.106e-1
3	1.260e-1	1.263e-1	1.264e-1	1.441e-1	2.168e-1	1.441e-1
4	1.220e-1	1.217e-1	1.224e-1	6.721e-2	1.009e-1	6.721e-2
5	1.054e-1	1.058e-1	1.057e-1	3.134e-2	4.704e-2	3.134e-2
6	8.794e-2	8.760e-2	8.825e-2	1.475e-2	2.204e-2	1.475e-2
7	7.220e-2	7.201e-2	7.245e-2	6.640e-3	1.021e-2	6.640e-3
8	5.913e-2	5.896e-2	5.934e-2	3.144e-3	4.708e-3	3.144e-3
9	4.795e-2	4.830e-2	4.812e-2	1.492e-3	2.187e-3	1.492e-3
10	3.916e-2	3.888e-2	3.930e-2	7.660e-4	1.100e-3	7.660e-4
11	3.182e-2	3.199e-2	3.193e-2	3.100e-4	4.849e-4	3.100e-4
12	2.594e-2	2.578e-2	2.603e-2	1.120e-4	2.058e-4	1.120e-4
13	2.165e-2	2.120e-2	2.172e-2	5.000e-5	7.365e-5	5.000e-5
14	1.753e-2	1.767e-2	1.760e-2	1.600e-5	3.168e-5	1.600e-5
15	1.426e-2	1.425e-2	1.431e-2	1.000e-5	1.112e-5	1.000e-5
16	1.195e-2	1.184e-2	1.199e-2	0.000e-0	4.099e-6	0.000e-0
17	9.859e-3	9.752e-3	9.894e-3	0.000e-0	0.000e-0	0.000e-0
18	6.864e-3	6.911e-3	6.888e-3	0.000e-0	0.000e-0	0.000e-0
19	3.532e-3	3.577e-3	0.000e-0	0.000e-0	0.000e-0	0.000e-0

К *счетным* показателям относятся:

1) сами упомянутые распределения (в частности, они нужны для расчета емкости буферных накопителей из условия размещения

очереди с вероятностью не менее заданной, распределения интервалов между обслуженными заявками и др.);

2) вероятность π_R отказа в приеме заявки на обслуживание;

3) вероятность нулевого ожидания $\Pi_0 = \sum_{j=0}^{n-1} \pi_j$;

4) среднее число заявок в системе $L = \sum_{j=1}^{R} j p_j$;

5) среднее число заявок в очереди $q = \sum_{j=n+1}^{R} (j-n) p_j$;

6) среднее число свободных каналов $f = \sum_{j=0}^{n-1} (n-j) p_j$;

7) среднее число занятых каналов $z = n - f$.

К показателям этой группы можно отнести также вероятность потери заявки в случае ограниченной очереди или «нетерпеливых» заявок.

Из *временны́х* показателей наиболее существенны

1) ДФР $\bar{V}(t)$ времени пребывания заявки в системе,

2) ДФР времени виртуального (ненулевого) ожидания,

3) моменты названных распределений,

Оперативность системы оценивают по характеристикам распределения времени пребывания (sojourn time = response time — [165]). Характеристики *ожидания* и, в частности, его средняя длительность отражают цену, которую клиент должен заплатить за совместное с другими клиентами (заявками) использование обслуживающей системы.

Как промежуточный показатель (обычно в связи с реализацией приоритетных дисциплин) применяется введенное Э. Борелем распределение периода непрерывной занятости (ПНЗ) и его моменты. При этом в случае n-канальной системы как правило имеется в виду период *полной* занятости. Период, «дополнительный» к ПНЗ, может использоваться для проведения регламентных и фоновых (менее важных) работ.

Имея распределения числа заявок в системе, можно получить большинство интересующих исследователя временных характеристик с помощью принципа сохранения стационарной очереди — см. разд. 3.2.

Показатель *пропускной способности системы* — интенсивность потока обслуженных заявок — в случае, когда система справляется с нагрузкой, совпадает с интенсивностью входящего потока. Поэтому его имеет смысл определять только для замкнутых систем и СМО с ограниченной очередью. При этом используются счетные характеристики и обсуждаемые в следующей главе балансовые соотношения.

2.5.2. Выбор показателей

Часть перечисленных показателей характеризует СМО с точки зрения потребителя, другие — с позиций эксплуатационного персонала. Улучшение показателей оперативности обслуживания, в котором заинтересованы клиенты системы, достигается путем увеличения мощности системы и ухудшает показатели загрузки. Поэтому говорить об оптимизации системы можно только при комплексном подходе к ней. В качестве обобщенного показателя эффективности обычно берется взвешенная сумма показателей разных групп — по одному от каждой.

Примером комплексного подхода и источником полезных аналогий может служить задача о выборе оптимального оборотного запаса агрегатов [90, 115]. Пусть

y — величина оборотного запаса,

s — цена хранения агрегата в единицу времени, включая убытки от омертвления денежных средств,

d — ущерб в единицу времени, возникающий при нехватке оборотного запаса,

p_j — вероятность того, что j агрегатов находятся в ремонте.

Оптимальный оборотный запас выбирается по минимуму целевой функции

$$L(y) = sy + d \sum_{j=y+1}^{\infty} p_j. \tag{2.5.1}$$

Он должен обеспечить одновременное выполнение неравенств

$$L(y^*+1)-L(y^*)\geq 0, \qquad L(y^*-1)-L(y^*)\geq 0,$$

которые после подстановки (2.5.1) сводятся к

$$p_{y^*+1}\leq s/d\leq p_{y^*}. \tag{2.5.2}$$

Распределение $\{p_j\}$ можно вычислить, рассматривая ремонтный орган как систему массового обслуживания. Вид условия (2.5.2) не зависит от параметров последней — числа каналов n, характеристик входящего потока, распределения обслуживания и т.п. Все это учитывается только при расчете $\{p_j\}$. На втором этапе оптимизации можно рассматривать минимизацию функции

$$M(n)=L(y^*(n))+cn.$$

Часто системы проектируются из условия обеспечения заданных вышестоящим органом показателей обслуживания при минимальных затратах. Однако обосновать требуемые показатели очень трудно.

Выбранный показатель эффективности должен быть достаточно чувствителен к варьируемым параметрам системы. Это требование, в частности, делает работу с ДФР предпочтительнее, чем с обычной функцией распределения, так как последняя в практически интересной области высоких вероятностей решения задачи меняется очень медленно.

Глава 3

Законы сохранения в теории очередей

В физике, химии, инженерном деле важную роль играют *законы сохранения*. Принципы сохранения массы, энергии, заряда, момента количества движения и др. часто позволяют непосредственно получить конечный результат или значительно сократить необходимые выкладки. Установление законов сохранения справедливо считается показателем зрелости соответствующей науки.

Существование стационарных режимов в системах массового обслуживания при стационарном входящем потоке возможно лишь при выполнении соотношений типа законов сохранения между некоторыми количествами, характеризующими состояние системы. Заслугой автора [211] М. Краковски является осознание ряда хорошо известных результатов — соотношения (3.1.1), его обобщения на неоднородный случай, (3.3.1) — как количественных следствий из формулируемых вербально законов сохранения. Эти законы и следствия из них были дополнены и развиты усилиями ряда других авторов — см. [50, 92, 138, 265]. Они имеют отчетливое физическое истолкование, а их применение упрощает анализ СМО. При их кажущейся очевидности из них удается извлечь далеко не тривиальные следствия.

Рассмотрим наиболее важные из этих законов.

3.1. Сохранение заявок

Закон сохранения заявок формулируется в следующем виде:

Частота поступления заявок в канал обслуживания в среднем равна частоте выходов из этого канала.

Покажем, как использовать этот принцип для расчета вероятности свободного состояния однолинейной системы $GI/G/1$. Средняя частота прибытия заявок в нее $\lambda = 1/a$, средняя длительность обслуживания — b, где a и b — средние соответствующих распределений. Частота обслуживаний равна вероятности занятости $1 - p_0$, деленной на b. Значит, в стационарном режиме должно быть $1/a = (1 - p_0)/b$, откуда следует

$$p_0 = 1 - b/a. \tag{3.1.1}$$

Достойно изумления, что аналогичный результат в [221, 233] выводится с применением метода производящих функций и теоремы Руше из условия $P(1) = 1$.

Поучителен вывод p_0 для неоднородного случая. Условие баланса заявок здесь имеет вид

$$\lambda_i = \frac{\lambda_i}{\Lambda} \frac{1 - p_0}{b_i},$$

где λ_i/Λ — вероятность обслуживания занятой системой заявки именно i-го типа. Умножив обе части этого равенства на b_i и просуммировав результаты по всем i, получаем

$$p_0 = 1 - \sum_i \lambda_i b_i.$$

Для n-канальной системы математическое ожидание числа занятых каналов составит

$$\sum_{j=0}^{n-1} j p_j + n(1 - \sum_{j=0}^{n-1} p_j) = n - \sum_{j=0}^{n-1} (n - j) p_j,$$

и условие баланса заявок сводится к

$$\sum_{j=0}^{n-1} (n - j) p_j = n - \lambda b = n(1 - \lambda b/n). \tag{3.1.2}$$

Отношение

$$\rho = \lambda b/n \tag{3.1.3}$$

называется *коэффициентом загрузки* системы и для однолинейной СМО совпадает с вероятностью ее занятости. Для разомкнутых СМО с неограниченной очередью условием существования стационарного режима является строгое неравенство $\rho < 1$.

Принцип сохранения заявок остается в силе и при анализе систем с отказами (ограниченной длиной очереди), а также с уходом из очереди «нетерпеливых» заявок. В этих случаях в уравнении баланса должны быть учтены все причины ухода заявок из системы. При неоднородных заявках этот принцип справедлив и для заявок каждого типа порознь. В случае превращающихся заявок действует его обобщенный вариант.

Принцип сохранения действует и при обработке потока *пачек* заявок интенсивности λ. В этом случае коэффициент загрузки

$$\rho = \lambda b\bar{f}/n, \tag{3.1.4}$$

где $\bar{f}$ — средний объем пачки.

3.2. Сохранение очереди

3.2.1. Вербальная формулировка

Зафиксируем число заявок в очереди перед прибытием очередной заявки и в момент приема ее на обслуживание. Очевидно, что при дисциплине очереди FCFS

> *распределение числа заявок, прибывших за время ожидания начала обслуживания, совпадает с распределением длины очереди перед прибытием заявки.*

Оказывается, поэтический образ Сасаки Микиро

> «. . . Голова Очереди
> Держит в зубах прошлое Очереди»

(Голоса вещей /Пер. с японского. — М.: Радуга, 1988) имеет четкое математическое содержание.

Сформулированный выше принцип верен для системы $GI/G/n$ и всех частных вариантов ее. Этот же принцип в некоторых случаях может быть применен и к системе в целом:

распределение числа заявок, прибывших за время ожидания окончания обслуживания, совпадает с распределением числа заявок в системе перед прибытием заявки.

Поскольку условие «первый пришел — первый обслужен» должно быть сохранено, вторая формулировка применима к более узкому классу систем: $GI/G/1$ и $GI/D/n$, в которых гарантируется совпадение порядка завершения обслуживания с порядком выборки из очереди.

Заметим, что упомянутые распределения при сделанных оговорках совпадают также с распределением числа заявок в очереди (в системе) *сразу после* завершения очередного обслуживания.

3.2.2. Общий случай

Сформулированные выше законы позволяют установить связь между распределением времени пребывания заявки в очереди и производящей функцией распределения числа заявок в ней перед прибытием очередной заявки.

Функция переменной z, $\quad 0 \le |z| \le 1$, получаемая из дискретного распределения вероятностей $\{\pi_k\}$ по формуле

$$\Pi(z) = \sum_{k=0}^{\infty} z^k \pi_k,$$

называется *производящей функцией* этого распределения. Имея производящую функцию, нетрудно получить вероятности $\{\pi_k\}$ и моменты $\{h_k\}$ распределения. В частности,

$$\pi_k \;=\; \frac{1}{k\,!}\,\Pi^{(k)}(z)\Big|_{z=0}, \qquad k = 0, 1, \ldots .$$

Можно показать, что

$$P^k(z)|_{z=1} = \sum_{n=k}^{\infty} \frac{n\,!}{(n-k)\,!}\pi_k = q_{[k]} -$$

k-му факториальному моменту рассматриваемого распределения. В частности, первый из них $q_{[1]}$ равен обычному начальному q_1.

Производящая функция является весьма компактной и потому удобной при расчетах формой задания распределения дискретной случайной величины. Производящей функции можно придать прямой

вероятностный смысл: если считать, что каждая заявка с вероятностью z независимо от остальных обладает некоторым свойством (например, является «красной»), то $\Pi(z)$ есть вероятность иметь все заявки «красными» [191, 204].

Пусть

$w(t)$ — плотность распределения времени ожидания начала обслуживания,

$\Pi(z)$ — производящая функция распределения числа заявок в очереди в момент прибытия новой заявки,

$\bar{A}_z(t)$ — ДФР интервалов между заявками исходного потока, просеянного с вероятностью сохранения заявки $\bar{z} = 1 - z$.

Поскольку $\Pi(z)$ можно истолковать как вероятность отсутствия в системе синих заявок в момент прибытия новой, из сформулированного выше принципа следует равенство

$$\Pi(z) = \int_0^\infty \bar{A}_z(t) w(t)\, dt. \tag{3.2.1}$$

3.2.3. Простейший входящий поток

При простейшем входящем потоке интенсивности λ просеянный поток также будет простейшим, и (3.2.1) переходит в

$$\Pi(z) = \int_0^\infty e^{-\lambda(1-z)t} w(t)\, dt. \tag{3.2.2}$$

Положим $\lambda(1-z) = s$. Тогда (3.2.2) сводится к

$$\Pi(1 - s/\lambda) = \int_0^\infty e^{-st} w(t)\, dt = \sum_{k=0}^\infty \frac{(-s)^k}{k\,!} w_k, \tag{3.2.3}$$

где $\{w_k\}$ — начальные моменты распределения $w(t)$. С другой стороны, по определению производящей функции

$$\Pi(1 - s/\lambda) = \sum_{n=0}^\infty \pi_n (1 - s/\lambda)^n.$$

Сгруппировав члены этой суммы, содержащие k-ю степень s, получаем

$$(-s/\lambda)^k \sum_{n=k}^{\infty} \left(\begin{array}{c} n \\ k \end{array} \right) \pi_n.$$

Поскольку равенство (3.2.3) справедливо в диапазоне $0 \leq s \leq \lambda$, коэффициенты при одинаковых степенях s в разложении левой и правой частей (3.2.3) должны быть равны. Таким образом,

$$\sum_{n=k}^{\infty} \left(\begin{array}{c} n \\ k \end{array} \right) \pi_n = \frac{\lambda^k}{k\,!} w_k.$$

Раскрывая биномиальный коэффициент, находим

$$\sum_{n=k}^{\infty} \frac{n\,!}{k\,!\,(n-k)\,!} \pi_n = \frac{\lambda^k}{k\,!} w_k,$$

или

$$\sum_{n=k}^{\infty} n(n-1)\ldots(n-k+1)\pi_n = \lambda_k w_k.$$

Выражение в левой части последнего равенства есть $q_{[k]}$ — k-й факториальный момент распределения числа заявок в очереди. Окончательно

$$w_k = q_{[k]}/\lambda^k, \qquad k = 1, 2, \ldots \tag{3.2.4}$$

Этот результат получен в статье [170]. В частности, среднее время ожидания в очереди

$$w_1 = q_{[1]}/\lambda = q_1/\lambda \tag{3.2.5}$$

(формула *Литтла*[1], в реальности общеизвестная с первых лет существования теории очередей и потому именуемая «фольклорной теоремой»). Варианты ее вывода приводятся в [68], [168, с. 17], [233, с. 48]; три варианта — в [246]. При соответствующем усреднении интенсивности входящего потока она верна и для замкнутых систем.

Как показал имитационный эксперимент (см. разд. 5.4.2), формула (3.2.4) применима и по отношению к *пачкам* заявок. При этом в очереди учитываются только «непочатые» пачки, а обсуждаемые формулы дадут моменты распределения ожидания для головной заявки пачки. Прочие

[1] Литтл первым доказал ее строго в 1961 г.

заявки получат дополнительные задержки, определяемые очередностью выбора заявки из пачки.

Аналогичную связь можно установить между факториальными моментами распределения числа заявок в системе и моментами распределения времени *пребывания* заявки в ней для систем $M/G/1$ и $M/D/n$. Для первого момента имеет место второй вариант формулы Литтла:

$$v_1 = L_1/\lambda, \tag{3.2.6}$$

где L_1 — среднее число заявок в системе.

Среднее число заявок в системе есть сумма средней длины очереди q_1 и среднего числа заявок в каналах. Последнее, очевидно, равно среднему числу занятых каналов $\bar{n}_{\text{З}}$. Итак, $L_1 = q_1 + \bar{n}_{\text{З}}$. Из условия баланса заявок имеем $\lambda = \bar{n}_{\text{З}}/b$, или $\bar{n}_{\text{З}} = \lambda b$. Но $q_1 = \lambda w_1$, так что

$$L_1 = \lambda w_1 + \lambda b = \lambda(w_1 + b) = \lambda v_1.$$

Следовательно, формула Литтла (3.2.6) справедлива независимо от сохранения принципа FCFS и типа распределения интервалов между смежными заявками. Более того, она верна и для *сетей обслуживания*.

Для системы $M/M/1$ распределеним числа заявок $p_k = (1-\rho)\rho^k$, $k = 0, 1, \ldots$, а производящая функция сводится к $\Pi_0(z) = (1-\rho)/(1-\rho z)$ и $\Pi_1(z) = \rho\Pi_0(z)$ для значений $n = 0$ и $n = 1$ соответственно. Эти результаты могут использоваться для тестирования процедуры расчета производящей функции.

Факториальные моменты вычисляются по формуле

$$f_{[k]} = \sum_{j=k}^{j_{\max}-n} j(j-1)\cdots(j-k+1)p_{j+n}.$$

Внешний цикл по k устанавливает начальное значение произведения $b = j(j-1)\cdots(j-k+1)$, при $j = k$ обращающегося в k. Во внутреннем цикле (по J) используется рекуррентный расчет произведения на основе соотношения

$$\frac{j(j-1)\cdots(j-k+1)}{(j-1)(j-2)\cdots(j-k)} = \frac{j}{j-k}, \qquad j = k+1, k+2, \ldots$$

После завершения расчета очередного момента готовится новый стартовый множитель.

В процедуру вычисления факториальных моментов можно ввести поправки, компенсирующие усечение бесконечной очереди. Полагая `x=p(j)/p(j-1)` и считая неучтенные вероятности образующими убывающую геометрическую прогрессию, имеем

$$\begin{aligned}
\Delta_k &= \sum_{j=J-n+1}^{\infty} j(j-1)\cdots(j-k+1)p_{j+n} \\
&= \sum_{j=J-n+1}^{\infty} j(j-1)\cdots(j-k+1)p_J x^{J-j+n} \\
&= \frac{p_J}{x^{J-n}}\left\{\sum_{j=k}^{\infty} j(j-1)\cdots(j-k+1)x^j - \sum_{j=k}^{J-n} j(j-1)\cdots(j-k+1)x^j\right\} \\
&= \frac{p_J}{x^{J-n}}\left\{\frac{k!}{1-x}\left(\frac{x}{1-x}\right)^k - \sum_{j=k}^{J-n} j(j-1)\cdots(j-k+1)x^j\right\}.
\end{aligned}$$

3.3. Сохранение вероятностей состояний

Рассмотрим процесс переходов через разрез AB в марковской системе, соответствующей диаграмме переходов рис. 3.1. Состояния системы характеризуются числом k находящихся в ней заявок. Переходы между состояниями размечаются условными интенсивностями. Очевидно,

> *в стационарном режиме средние частоты переходов через разрез в противоположных направлениях равны.*

Для диаграммы рис. 3.1 с переходами только между соседними состояниями (процесс «размножения и гибели»)

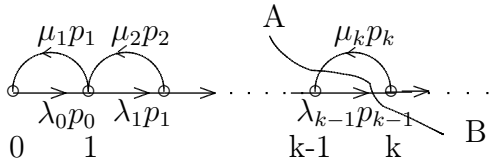

Рис. 3.1. К закону сохранения вероятностей для марковских систем

применение этого закона непосредственно приводит к равенству

$$p_{k-1}\lambda_{k-1} = p_k\mu_k, \tag{3.3.1}$$

так что необходимые вероятности можно определить рекуррентно:

$$p_k = \frac{\lambda_{k-1}}{\mu_k} p_{k-1}, \qquad k = 1, 2, \ldots, \tag{3.3.2}$$

отправляясь от начальной вероятности p_0.

Частным случаем разреза является выделяющий одно из состояний системы *кольцевой*. При этом противоположным направлениям соответствуют входящие и выходящие стрелки, а закон сохранения потока переходов для k-го узла диаграммы можно записать в виде

$$p_k \sum_{i \in \Gamma(x_k)} \nu_{k,i} = \sum_{j \in \Gamma^{-1}(x_k)} p_j \nu_{j,k}, \qquad k = 0, 1, \ldots \tag{3.3.3}$$

Здесь $\nu_{i,j}$ — параметр потока переходов из x_i в x_j, $\Gamma(x_k)$ означает множество вершин графа-диаграммы, в которое ведут стрелки из вершины x_k, а $\Gamma^{-1}(x_k)$ — множество вершин, стрелки *из которых* ведут в x_k.

Соотношения (3.3.3) образуют систему линейных алгебраических уравнений относительно искомых вероятностей $\{p_k\}$, решение которой при конечном числе узлов не вызывает принципиальных затруднений. Заметим, что предварительно нужно заменить одно из этих уравнений на условие нормировки или баланса заявок.

Формулы (3.3.3) определяют *глобальный* баланс вероятностей состояний, тогда как формулы (3.3.1) — *локальный* (рассматриваются переходы только внутри *пары состояний*). Из локального баланса для всех пар автоматически следует глобальный, тогда как обратное в общем случае неверно. Выполнение условий локального баланса является критическим условием применения многих методов расчета сетей обслуживания (см. главу 12).

3.4. Сохранение объема работы

Выделим класс так называемых «консервативных» дисциплин обслуживания, в котором:

1) все заявки остаются в системе до полного завершения обслуживания;

2) прерывание обслуживания не увеличивает загрузку системы.

Последнее условие означает, что затраты времени обслуживающего устройства на обеспечение прерывания пренебрежимо малы, а возобновление прерванного обслуживания производится без потери ранее затраченных ресурсов.

> *В указанном классе дисциплин распределение объема невыполненной работы, находящегося в СМО, постоянно и не зависит от выбора конкретной дисциплины.*

Если к этим требованиям добавить независимость выбора из очереди от длительности обслуживания, то можно утверждать, что

> *в указанном классе дисциплин распределение числа заявок в системе не зависит от выбора конкретной дисциплины.*

Дисциплины, нарушающие это правило (например, SPT — Shortest Processing Time) при случайной длительности обслуживания физически не реализуемы.

Покажем применение принципа сохранения объема работы к расчету среднего времени ожидания заявки в однолинейной системе. Прежде всего, в данном случае средний объем работы совпадает со средним значением w времени ожидания начала обслуживания вновь прибывшей заявкой. Это время будет складываться из времени завершения $\bar{f}_1$ начатого обслуживания и времени обслуживания ранее пришедших заявок. Средний остаток начатого обслуживания определяется формулой (1.7.4) и должен учитываться с вероятностью занятости системы, равной $\rho = \lambda b_1$:

$$\bar{f}_1 = \rho b_2/(2b_1) = \lambda b_2/2 = w^*.$$

Среднее число заявок в очереди на основании формулы Литтла составит λw , причем каждая из них в среднем обслуживается b_1 единиц времени. Итак, закон сохранения объема работы приводит к равенству

$$w = \lambda b_2/2 + \lambda w b_1,$$

откуда следует формула (Полячека—Хинчина)

$$w = \lambda b_2/[2(1-\rho)] \tag{3.4.1}$$

для среднего времени ожидания в $M/G/1$. Поскольку $b_2 = b_1^2(1+v_B^2)$, при одинаковых b_1 среднее время ожидания по отношению к показательно распределенному обслуживанию ($v_B = 1$) в «регулярном» случае

уменьшается вдвое, а при $v_B = 2$ возрастает в 2.5 раза. Это наиболее наглядная иллюстрация категорического императива учета высших моментов распределений обслуживания.

С другой стороны, из анализа знаменателя (3.4.1) следует необходимость $\rho < 1$ [2]. Отсюда следует вывод о *недопустимости выбора средней скорости обслуживания в одноканальной системе равной интенсивности потока заявок*.

Умножив обе части формулы (3.4.1) на ρ, получим

$$w\rho = \frac{\rho}{1-\rho}\,\frac{\lambda b_2}{2} = \frac{\rho}{1-\rho}w^*. \tag{3.4.2}$$

Равенство (3.4.2) в неоднородном случае принимает вид

$$\sum_i \rho_i w_i = \frac{R}{1-R}w^* = C. \tag{3.4.3}$$

Оно управляет всеми возможными перераспределениями среднего времени *ожидания* между заявками разных типов в случае введения приоритетного обслуживания. В случае прерывания с дообслуживанием оно выполняется строго только для экспоненциально распределенной трудоемкости. Его выполнение (с некоторыми оговорками — см. разд. 9.2.2) подтверждается как имитационными моделями, так и численными расчетами — в том числе для многоканальных систем. Напомним, что условие (3.4.3) не имеет места при прерываниях с *повторением* обслуживания.

На основе формулы (3.4.1) можно построить экономическую модель выбора оптимального числа n независимых каналов обслуживания [164, с. 204]. В этом случае на каждую систему M/G/1 придется интенсивность потока λ/n, и среднее время ожидания составит

$$w = \frac{\lambda b_2}{2(n-\lambda b_1)}.$$

Целевая функция может быть представлена в виде

$$Z = c_1 w + c_2 n.$$

Из условия $dZ/dn = 0$ находим оптимальное число каналов

$$n^* = \lambda b_1 + \sqrt{\frac{\lambda b_2}{2}\frac{c_1}{c_2}}.$$

[2] Это требование имеет место и для многоканальных систем.

3.5. Второй закон сохранения очереди

Приведем еще один инвариант (Клейнрока):

> *в классе дисциплин, не учитывающих длительность обслуживания, распределение числа заявок в системе не зависит от выбора конкретной дисциплины обслуживания.*

Это обстоятельство вместе с формулой Литтла позволяет при том же условии считать инвариантными и средние времена ожидания (пребывания). Однако на *высшие* моменты распределения времени ожидания дисциплина очереди влияет весьма существенно: FCFS гарантирует минимальную дисперсию времени ожидания, а случайный выбор из очереди и в особенности LCFS — увеличение долей заявок, ожидающих очень малые и очень большие промежутки времени.

Глава 4

Марковские системы и марковизация систем

Все методы анализа СМО в конечном счете применимы только к *марковским* процессам в них. Наиболее просто обстоит дело при простейшем входящем потоке и показательном распределении времени обслуживания. В этом случае интенсивности обслуживания одного из находящихся в каналах требований или прибытия еще одной заявки полностью определяются числом k находящихся в системе требований: $\mu = \mu(k), \quad \lambda = \lambda(k)$ независимо от уже истекших времени обслуживания и интервала с момента поступления последнего требования. Подобные системы называются *марковскими*. Для них вектор состояний сводится к скаляру k.

4.1. Метод Эрланга

Рассмотрим марковскую систему массового обслуживания с ограничением максимального количества требований в системе величиной R. Возможные переходы между состояниями указаны на рис. 3.1. Обслуживания в состоянии S_0 (система свободна) не происходит. Заявки, заставшие систему в состоянии S_R, получают отказ. Составим табл. 4.1 вероятностей переходов между состояниями за малый интервал времени Δt. Вероятности совершения двух и более скачков имеют порядок малости, высший в сравнении с Δt, и в таблицу не включены.

Таблица 4.1. Вероятности переходов за Δt

Исходное состояние на момент t	Конечное состояние на момент $t+\Delta t$	Вероятность перехода с точностью $o(\Delta t)$
S_0	S_1	$\lambda_0\Delta t$
S_0	S_0	$1-\lambda_0\Delta t$
$S_k, \quad k=\overline{1,R-1}$	S_{k-1} S_{k+1} S_k	$\mu_k\Delta t$ $\lambda_k\Delta t$ $1-(\lambda_k+\mu_k)\Delta t$
S_R	S_{R-1} S_R	$\mu_R\Delta t$ $1-\mu_R\Delta t$

Объединяя вероятности событий, приводящих в каждое из состояний на момент $t+\Delta t$, получаем конечно-разностные уравнения типа Чэпмена—Колмогорова

$$\begin{aligned}
p_0(t+\Delta t) &= p_0(t)(1-\lambda_0\Delta t)+\mu_1\Delta t\, p_1(t)+o(\Delta t);\\
p_k(t+\Delta t) &= p_k(t)[1-(\lambda_k+\mu_k)\Delta t]+\lambda_{k-1}\Delta t\, p_{k-1}(t)+\mu_{k+1}\Delta t\, p_{k+1}(t)\\
&\qquad +o(\Delta t), \qquad k=\overline{1,R-1};\\
p_R(t+\Delta t) &= p_R(t)(1-\mu_R\Delta t)+\lambda_{R-1}\Delta t\, p_{R-1}(t)+o(\Delta t).
\end{aligned}$$

Перегруппируем члены этих уравнений и разделим обе части каждого из них на Δt:

$$\begin{aligned}
(p_0(t+\Delta t)-p_0(t))/\Delta t &= -\lambda_0 p_0(t)+\mu_1 p_1(t)+o(\Delta t)/\Delta t;\\
(p_k(t+\Delta t)-p_k(t))/\Delta t &= -(\lambda_k+\mu_k)p_k(t)+\lambda_{k-1}p_{k-1}(t)+\mu_{k+1}p_{k+1}(t)\\
&\qquad +o(\Delta t)/\Delta t, \qquad k=\overline{1,R-1};\\
(p_R(t+\Delta t)-p_R(t))/\Delta t &= -\mu_R p_R(t)+\lambda_{R-1}p_{R-1}(t)+o(\Delta t)/\Delta t.
\end{aligned}$$

Устремив Δt к нулю, приходим к системе дифференциальных уравнений

$$\begin{aligned}
p_0'(t) &= -\lambda_0 p_0(t)+\mu_1 p_1(t);\\
p_k'(t) &= -(\lambda_k+\mu_k)p_k(t)+\lambda_{k-1}p_{k-1}(t)+\mu_{k+1}p_{k+1}(t), \qquad k=\overline{1,R-1};\\
p_R'(t) &= -\mu_R p_R(t)+\lambda_{R-1}p_{R-1}(t).
\end{aligned} \tag{4.1.1}$$

Задав начальные условия к системе (4.1.1) — например, в виде $p_0(0)=1$ и $p_k=0 \quad$ для $\quad k=\overline{1,R}$, — можно найти (численно) решение соответствующей задачи Коши для произвольного значения t. Другой возможный подход — это получение производящей функции переходных вероятностей $P(z,t)$.

При $t \to \infty$ производные по времени, если стационарный режим существует, обращаются в нуль. В итоге получаем систему линейных алгебраических уравнений для *стационарных* вероятностей

$$\begin{array}{lllll}
0 = & -\lambda_0 p_0 & +\mu_1 p_1, & & \\
0 = & \lambda_0 p_0 & -(\lambda_1+\mu_1)p_1 & +\mu_2 p_2, & \\
0 = & & \lambda_1 p_1 & -(\lambda_2+\mu_2)p_2 & +\mu_3 p_3, \\
\multicolumn{5}{l}{\dotfill} \\
0 = & \lambda_{k-1}p_{k-1} & -(\lambda_k+\mu_k)p_k & +\mu_{k+1}p_{k+1}, & \\
\multicolumn{5}{l}{\dotfill} \\
0 = & \lambda_{R-1}p_{R-1} & -\mu_R p_R. & &
\end{array}$$

Эти уравнения можно непосредственно получить из условий сохранения вероятностей для кольцевых разрезов (см., например, [35]).

Сложив первые $k+1$ уравнений этой системы, убеждаемся, что

$$-\lambda_k p_k + \mu_{k+1} p_{k+1} = 0. \tag{4.1.2}$$

В них мы узнаем уравнения, получаемые посредством вертикальных разрезов на диаграмме. Далее,

$$p_{k+1} = \frac{\lambda_k}{\mu_{k+1}} p_k = \frac{\lambda_k}{\mu_{k+1}} \frac{\lambda_{k-1}}{\mu_k} p_{k-1} = \frac{\lambda_k}{\mu_{k+1}} \frac{\lambda_{k-1}}{\mu_k} \dots \frac{\lambda_1}{\mu_2} p_1.$$

Но из первого уравнения вытекает $p_1 = \lambda_0 p_0/\mu_1$. Таким образом, для схемы «размножения и гибели» все вероятности выражаются через p_0 согласно $p_k = \alpha_k p_0$, $k = \overline{1,R}$, где

$$\alpha_k = \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}.$$

Вероятность p_0 определяется из условия нормировки (сумма всех вероятностей должна быть равна единице). Следовательно,

$$p_0 = \left(1 + \sum_{k=1}^{R} \alpha_k\right)^{-1}.$$

Из рассмотренной схемы могут быть получены порознь или в комбинациях различные частные случаи, например $\lambda_k = \lambda(R-k)$ (замкнутая система), $\lambda_k = \lambda$ (разомкнутая система), а также случай n-линейной системы, для которой интенсивности обслуживания вычисляются согласно (2.2.2). Для разомкнутой системы необходима сходимость $\sum_{k=1}^{\infty} \alpha_k$. Условием этой сходимости является существование

константы q, $0 < q < 1$, превышающей любое отношение λ_k/μ_{k+1}, начиная с некоторого номера k. Тогда последующие члены ряда мажорируются членами убывающей геометрической прогрессии со знаменателем q, сумма которой конечна.

Приведем окончательные формулы (Эрланга) для разомкнутой системы $M/M/n$:

$$\begin{aligned} p_0 &= \Big[1+\sum_{i=1}^{n-1}\Big(\frac{\lambda}{\mu}\Big)^i\frac{1}{i\,!}+\Big(\frac{\lambda}{\mu}\Big)^n\frac{n\mu}{n\,!\,(n\mu-\lambda)}\Big]^{-1};\\ p_k &= p_0\Big(\frac{\lambda}{\mu}\Big)^k\frac{1}{k\,!}, \qquad k=\overline{1,n};\\ p_k &= p_n\Big(\frac{\lambda}{n\mu}\Big)^{k-n}, \qquad k=n+1,n+2,\ldots \end{aligned} \tag{4.1.3}$$

Средняя длина очереди для этой системы

$$q = p_n\frac{\rho}{(1-\rho)^2}.$$

Она может быть полезна при расчете сетей обслуживания на основе инвариантов отношения.

Для систем *с отказами* ($R = n$) формулы Эрланга справедливы при произвольных распределениях длительности обслуживания $B(t)$ с заменой $1/\mu$ на $\int\limits_0^\infty \bar{B}(t)\,dt = b$ — см. [134, 270].

При $n \to \infty$ система (4.1.3) сводится к распределению Пуассона

$$p_k = \frac{(\lambda/\mu)^k}{k\,!}e^{-\lambda/\mu}, \qquad k=0,1,\ldots$$

Этот результат с вышеупомянутой заменой результат инвариантен к распределению обслуживания и определяется только средними значениями. Легко проверить выполнение формулы Литтла для среднего числа заявок в системе:

$$\bar{L} = \lambda/\mu = \lambda b.$$

Для систем с конечным числом состояний, в особенности замкнутых, практически удобно положить $p_0 = 1$ и с помощью формулы (4.1.2) вычислять $\{p_k\}$ для конкретных зависимостей $\mu(k)$ и $\lambda(k)$, одновременно подсчитывая сумму, а затем нормировать результаты. Поэтому

соответствующие аналитические зависимости (в частности, формулы Энгсета для замкнутых СМО) мы приводить не будем.

Сходным методом могут быть рассчитаны любые марковские СМО, в которых за малый интервал Δt возможны переходы только между соседними состояниями (*схема размножения и гибели*).

4.2. Расчет системы $M/M/1$ с применением законов сохранения

Конечным результатом разд. 4.1 для стационарного режима является соотношение (4.1.2), найденное в итоге длительных выкладок. Однако эта формула мгновенно получается из диаграммы рис. 3.1. Проделанный методом Эрланга расчет может рассматриваться как проверка упомянутого закона.

Применим законы сохранения к расчету основных показателей системы $M/M/1$. Прежде всего отметим, что для этой СМО $\lambda(k) = \lambda$ и $\mu(k) = \mu$ независимо от k. Следовательно, для всех k имеет место $\mu p_{k+1} = \lambda p_k$, или $p_{k+1} = (\lambda/\mu)p_k = \rho p_k$, и стационарные вероятности

$$p_k = (1-\rho)\rho^k, \qquad k = 0, 1, \ldots \tag{4.2.1}$$

образуют геометрическую прогрессию со знаменателем ρ. Условием существования стационарного режима является неравенство $\rho < 1$. Вероятность $p_0 = 1 - \rho$, найденная здесь из условия нормировки, совпадает с выражением (3.1.1) при соответствующей замене обозначений, что подтверждает закон сохранения требований.

Формула (4.2.1) позволяет получить явную форму условия (2.5.2) оптимальности оборотного запаса:

$$y^* = \left[\ln\Big(\frac{s/d}{1-\rho}\Big) \Big/ \ln\rho \right].$$

Среднее число заявок в системе

$$L(\rho) = \frac{\rho}{1-\rho}. \tag{4.2.2}$$

Применительно к модели M/M/1 легко решаются типичные задачи оптимизации: минимум стоимости при заданной нагрузке, минимизация

среднего времени ответа — при ограничении по стоимости. Рассмотрим, к примеру, выбор оптимального быстродействия. Пусть целевая функция имеет вид

$$Z = A\mu + B\bar{L} = A\mu + \frac{B\lambda}{\mu - \lambda}$$

(разыскивается оптимальный баланс между быстродействием канала и стоимостью пребывания заявок в системе). Тогда

$$\mu^* = \lambda + \sqrt{\lambda B/A}.$$

Распределение числа заявок перед прибытием очередной заявки

$$\pi'_j = \lambda_j p_j / \sum_{i=0}^{\infty} \lambda_i p_i = p_j, \qquad j = 0, 1, \ldots,$$

т. е. совпадает со стационарным распределением. Этот результат имеет место для любых разомкнутых СМО с простейшим входящим потоком и неограниченной очередью (теорема PASTA).

Распределение числа заявок после завершения обслуживания

$$\pi''_j = \mu_{j+1} p_{j+1} / \sum_{i=0}^{\infty} \mu_{i+1} p_{i+1} = \mu\rho^{j+1}(1-\rho)/(\mu\rho) = \pi'_j = p_j,$$

что подтверждает исходную формулировку закона сохранения стационарной очереди.

Производящая функция стационарных вероятностей

$$P(z) = \sum_{k=0}^{\infty} (1-\rho)\rho^k z^k = \frac{1-\rho}{1-\rho z}.$$

На основании формулы (3.2.3) ПЛС распределения времени пребывания заявки в системе

$$\nu(s) = P(1 - s/\lambda) = \frac{1-\rho}{1-\rho(1-s/\lambda)} = \frac{\mu - \lambda}{\mu - \lambda + s}. \tag{4.2.3}$$

Следовательно, время пребывания заявки в системе подчинено показательному закону с параметром $\mu - \lambda$. Среднее время пребывания

$$v = 1/(\mu - \lambda) \tag{4.2.4}$$

при $\lambda \to \mu$ неограниченно возрастает из-за роста времени ожидания. С помощью (4.2.4) легко проверяется формула Литтла $L = \lambda v$.

Сильно загруженные СМО очень чувствительны к дальнейшему возрастанию загрузки:

$$L'(\rho) = \frac{1}{(1-\rho)^2}.$$

СМО с *конечным* источником заявок обладают свойством саморегулируемости, присущим системам с отрицательной обратной связью. Неограниченный рост очереди в них исключен.

Простота выражения для среднего времени пребывания заявки в системе $M/M/1$ позволяет продемонстрировать несколько неожиданную возможность уменьшения v пропорциональным увеличением интенсивности входящего потока и обслуживания (с сохранением коэффициента загрузки). В самом деле, при m-кратном их увеличении

$$v(m) = 1/(m\mu - m\lambda) = v(1)/m. \tag{4.2.5}$$

Можно показать, что распределение времени ожидания (*при условии ненулевого ожидания и дисциплине* $FCFS$) также описывается ПЛС (4.2.3) и имеет моменты

$$\tilde{w}_k = k\,!/(\mu - \lambda)^k.$$

Следовательно, тот же масштабный эффект наблюдается и по отношению к среднему времени ожидания.

Можно ожидать, что вышеупомянутый масштабный эффект имеет место независимо от числа каналов и вида распределений интервалов между смежными заявками и длительностей обслуживания. В книге [50] этот тезис обосновывается с помощью *оценок* среднего времени ожидания, полученных разными авторами для систем $GI/G/1$, $GI/M/n$, $GI/D/n$, $GI/E_k/n$. Применив численные методы последующих глав, удалось установить, что соотношения (4.2.5) выполняются *строго* независимо от вида исходных распределений и числа каналов обслуживания.

Среднее число заявок в системе

$$\begin{aligned} L &= (1-\rho)\sum_{k=1}^{\infty} k\rho^k &&= \rho(1-\rho)\sum_{k=1}^{\infty} k\rho^{k-1} \\ &= \rho(1-\rho)\frac{d}{d\rho}\,\frac{1}{1-\rho} &&= \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}. \end{aligned}$$

Соотношение типа (3.2.5) между L и v оказалось верным, что согласуется с законом сохранения очереди.

Учитывая очевидность законов сохранения, нельзя не изумляться фактам продолжающегося применения вековой давности технологии Эрланга в учебниках по ТМО — даже такого высокого уровня, как [18].

4.3. Обзор методов марковизации СМО

4.3.1. Метод линейчатых марковских процессов

В разд. 4.1 рассматривалась наиболее удобная для расчета модель СМО вида $M/M/n/R$, в которой оба исходных распределения — показательные. Если одно из них отличается от показательного, то процесс, характеризуемый числом требований, уже не является марковским. Для вероятностного прогнозирования его поведения можно ввести в вектор состояния дополнительную непрерывную переменную — значение немарковской компоненты (в случае $M/G/1$ — истекшее время очередного обслуживания)[1].

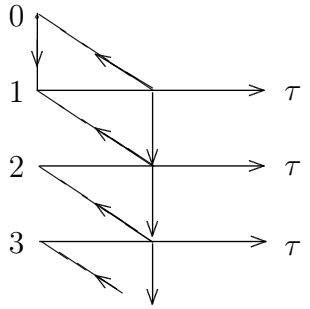

Рис. 4.1. Линейчатая диаграмма переходов для системы $M/G/1$

Полученный марковский процесс в связи с графическим методом его представления (рис. 4.1) называется *линейчатым*. Здесь номер каждой полупрямой указывает число заявок в системе. Завершение обслуживания уменьшает число заявок на единицу и возобновляет отсчет времени обслуживания τ с нуля. Прибытие заявки приводит к мгновенному переходу вниз на следующую полупрямую при сохранении τ.

[1] В [18, с. 278] рассматривается *оставшееся* время обслуживания.

Для сравнения и лучшего уяснения возможных переходов приведем еще несколько диаграмм. В системе $GI/M/1/R$ (см. рис. 4.2) непрерывный параметр θ представляет собой интервал, истекший с момента поступления очередного требования.

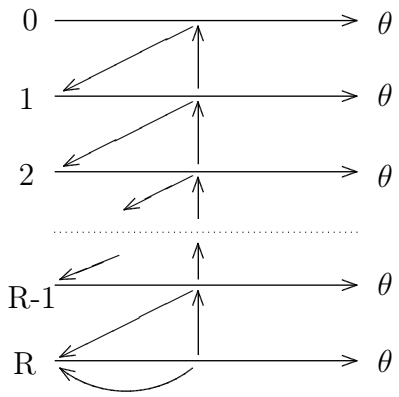

Рис. 4.2. Линейчатый процесс для $GI/M/1/R$

Таким образом, прибытие очередного требования переводит изображающую точку в *начало* нижележащей оси θ. При $j = R$ новая заявка получает отказ, и отсчет времени с момента прибытия последней заявки вновь начинается с нуля. Для этой модели ограничение числа заявок в системе необходимо *принципиально*, поскольку расчет плотностей вероятностей здесь приходится начинать с самого нижнего яруса (см. [92]).

Рис. 4.3 описывает систему $M/G/1$ *с порогом включения* — обслуживание начинается при скоплении в ней $n > 1$ заявок и выключается при полном рассасывании очереди. Поэтому система при нахождении в ней $k = \overline{1, n-1}$ заявок может пребывать в режимах как с отключенным, так и со включенным обслуживанием. Такие модели соответствуют ситуации со значительными затратами на *запуски* процесса обслуживания, частоту которых желательно уменьшить.

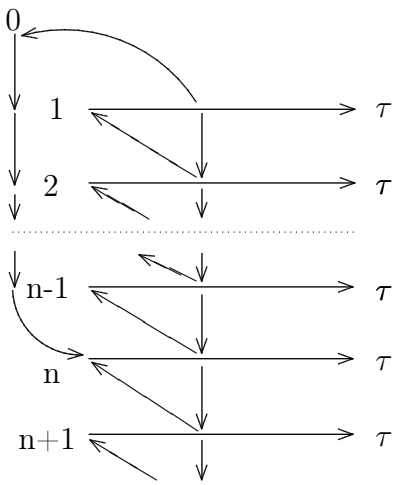

Рис. 4.3. Линейчатый процесс для $M/G/1$ с порогом включения

Наконец, значительный интерес представляет система $M/G/1$ с «разогревом» (см. рис. 4.4).

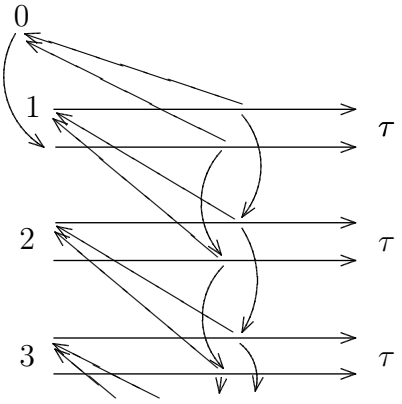

Рис. 4.4. Линейчатый процесс для $M/G/1$ с «разогревом»

Здесь первая заявка, прибывшая в свободную систему, получает обслуживание с иной функцией распределения, чем все последующие (в частности, это может быть задержка на приведение системы в рабочее состояние). Соответственно каждая временна́я ось расщепляется на две: нижняя представляет обслуживание головной заявки периода занятости системы, а верхняя — всех прочих. Завершение обслуживания головной заявки переводит изображающую точку в начало вышележащей верхней оси. Прибытие новой заявки сохраняет выбор элемента пары.

4.3.2. Законы сохранения для линейчатых процессов

Сформулированный в предыдущей главе закон сохранения вероятностей непосредственно применим только для анализа марковских систем. В стационарных немарковских задачах, исследуемых методом линейчатых процессов, имеют место *дифференциальные* аналоги этого закона:

- для скользящей точки оси: производная от плотности вероятностей равна разности интенсивностей переходов по входящим и выходящим стрелкам;
- для начальных точек осей: плотность равна интегралу от взвешенных исходными плотностями интенсивностей переходов по входящим стрелкам;
- для произвольного разреза на диаграмме переходов: средние интенсивности противоположных переходов через разрез равны.

Приведенные формулировки обладают достаточной общностью и, в частности, допускают зависимость λ и μ от номера состояния и от фиксируемого на диаграмме непрерывного параметра, что позволяет легко получить начальные условия и дифференциальные уравнения для плотностей применительно к следующим случаям:

- $\lambda_k = \lambda(R-k)$ — замкнутая система с R источниками заявок;
- $\lambda = \lambda(\tau)$ — интенсивность входящего потока зависит от истекшего времени времени обслуживания очередной заявки;
- $\mu(\tau) = \mu_k(\tau)$ — распределение времени обслуживания меняется в зависимости от числа требований, находящихся в системе.

Разумеется, остается открытым вопрос о возможности получения *решений* этих уравнений в удобной для практики форме.

4.3.3. «Линейчатый» расчет системы $M/G/1$

Для простейшей системы $M/G/1$ с диаграммой переходов, показанной на рис. 4.1, интенсивность потока переходов вниз $\lambda = \mathrm{const}$,

а мгновенная интенсивность обслуживания $\mu(\tau)$ определяется согласно (2.2.1). Обозначим $\rho_k(\tau)$ стационарную плотность вероятности нахождения системы в состоянии $E_k(\tau)$. Применение к диаграмме законов сохранения приводит к системе дифференциальных уравнений

$$\begin{aligned} \frac{d\rho_1(\tau)}{d\tau} &= -[\mu(\tau)+\lambda]\rho_1(\tau), \\ \frac{d\rho_k(\tau)}{d\tau} &= -[\mu(\tau)+\lambda]\rho_k(\tau)+\lambda\rho_{k-1}(\tau), \qquad k=2,3,\ldots \end{aligned} \tag{4.3.1}$$

с начальными условиями

$$\begin{aligned} \rho_1(0) &= \lambda p_0 + \int\limits_0^\infty \mu(\tau)\rho_2(\tau)\,d\tau, \\ \rho_k(0) &= \int\limits_0^\infty \mu(\tau)\rho_{k+1}(\tau)\,d\tau, \qquad k=2,3,\ldots \end{aligned} \tag{4.3.2}$$

Ход решения этой задачи представлен в [92]. Читателю будет интересно сопоставить его с серией аналогичных задач, рассмотренных автором в [89] *без использования законов сохранения* применительно к задачам управления запасами со случайной задержкой поставок.

4.3.4. Идея метода вложенных цепей Маркова

Процессу с *одной* немарковской составляющей можно придать марковское свойство и другим путем. Если рассматривать, например, модель типа $M/G/1$ только в моменты окончания обслуживания очередного требования, то состояния системы достаточно характеризовать числом требований. Этот второй процесс является дискретным марковским и «вложен» в исходный линейчатый. Соответствующим подбором моментов фиксации состояний («точек регенерации») можно построить *вложенную цепь Маркова* для тех же ситуаций, в которых применим метод линейчатых процессов. Точки регенерации для исключения неоднозначности количества находящихся в системе требований должны быть сдвинуты на бесконечно малый интервал относительно моментов выбранных событий.

Расчет вероятностей состояний вложенной цепи выполняется более простыми средствами. Однако эти вероятности, вообще говоря, не тождественны стационарным вероятностям состояний системы на

произвольный момент времени, и переход к последним обычно требует дополнительных вычислений. Применение метода вложенных цепей Маркова будет описано в следующей главе.

4.3.5. Понятие о методе фиктивных фаз

Немарковскую компоненту процесса (например, прибытие заявки) можно представить одним из распределений фазового типа — см. разд. 1.8. Тогда состояние процесса задается полным числом заявок в ней и фазой поступления очередной заявки. Аналогично обстоит дело при марковизации процесса обслуживания (в многоканальном случае указывается количество заявок, проходящих каждую из фаз обслуживания). Поскольку время пребывания заявки в каждой фазе распределено экспоненциально, фиксировать его необходимости нет. Таким образом, из описания динамики очереди непрерывная составляющая исключается.

Метод фиктивных фаз является единственным инструментом расчета многоканальных систем с немарковским распределением времени обслуживания (исключая детерминированное). Реализация его связана с разрастанием размерности пространства состояний и с некоторой потерей точности из-за учета ограниченного числа моментов и приближенного их выравнивания. Она рассмотрена в главах 5–7.

Отметим, что возможно и комбинированное использование перечисленных методов — например, в разд. 5.8 и 5.9 используются и метод фаз, и метод вложенных цепей Маркова.

Глава 5

Метод вложенных цепей Маркова

5.1. Основные этапы расчета

Расчет немарковских СМО методом вложенных цепей Маркова проходит следующие этапы:

1) выбор моментов регенерации процесса;

2) расчет финальных вероятностей состояний вложенной цепи Маркова;

3) переход к стационарным вероятностям состояний;

4) расчет временны́х характеристик.

По первому и третьему этапам из рассматриваемого в главе множества систем можно «вынести за скобки» два общих соображения:

- моменты регенерации выбираются по событиям немарковской компоненты процесса с бесконечно малым сдвигом, позволяющим исключить неоднозначность числа заявок в СМО;

- переход к стационарным вероятностям осуществляется на основе закона сохранения вероятностей (см. разд. 3.3).

Классифицируем переходы между состояниями на марковские и немарковские. Свяжем марковские переходы, параметры которых постоянны для любого момента времени, со стационарными вероятностями состояний $\{p_k\}$, а немарковские переходы — с финальными вероятностями вложенной цепи $\{\pi_k\}$. Обозначим через α суммарную интенсивность немарковских переходов (она обратна среднему интервалу времени между смежными моментами регенерации). Тогда интенсивность переходов типа k составит $\alpha\pi_k$.

Составляя уравнения баланса противоположно направленных переходов для узлов или через разрезы на диаграмме, можно получить искомые соотношения между стационарными вероятностями состояний и финальными вероятностями вложенной цепи [62, 92].

5.2. Стандартная система $M/G/1$

5.2.1. Вероятности состояний

Выделим смежные моменты $\{\eta_n\}$, $n = 1, 2, \ldots$, окончания обслуживания очередного требования в системе $M/G/1$ и рассмотрим состояния системы в моменты $\{\eta_n + 0\}$. В такие моменты система либо свободна, либо только начинает обслуживание, так что фиксировать значение истекшего времени обслуживания не нужно.

Обозначим q_j вероятность прибытия j новых требований за время обслуживания. Нетрудно видеть, что $\{q_j\}$ вычисляются согласно (2.1.10). Для установившегося режима вероятность застать систему свободной в последующий момент регенерации связана с вероятностями состояний предыдущего формулой

$$\pi_0 = \pi_0 q_0 + \pi_1 q_0$$

(система была свободной, прибыло одно требование, оно обслужилось и за время его обслуживания новых заявок не прибыло, или было одно требование, оно обслужилось и новых не прибыло). Вообще в системе будет k требований в следующих случаях:

а) если в ней в предыдущий момент регенерации находилась $j+1$ заявка $(j \le k)$, одна обслужилась и пришло еще $k-j \ge 0$ заявок;

б) если за время обслуживания первой заявки, прибывшей в свободную систему, пришло еще k требований.

Следовательно, финальные (предельные при устремлении номера момента регенерации к бесконечности) вероятности состояний вложенной цепи связаны системой уравнений

$$\pi_k = \sum_{j=0}^{k} \pi_{j+1} q_{k-j} + \pi_0 q_k = \sum_{j=1}^{k+1} \pi_j q_{k+1-j} + \pi_0 q_k, \qquad k = 0, 1, \ldots \quad (5.2.1)$$

Из этих формул легко получить рекуррентные выражения для расчета вероятностей:

$$\pi_k = (\pi_{k-1} - \pi_0 q_{k-1} - \sum_{j=1}^{k-1} \pi_j q_{k-j})/q_0, \qquad k = 1, 2, \ldots \quad (5.2.2)$$

В разомкнутой системе $M/G/1$ ожидаемое время до первого немарковского перехода из свободного состояния составит среднее время до поступления заявки λ^{-1} (независимо от момента прибытия предыдущей) плюс среднее время b ее обслуживания. Для всех прочих состояний имеется только одна составляющая b. Таким образом,

$$\alpha = [(\lambda^{-1} + b)\pi_0 + b(1 - \pi_0)]^{-1} = (b + \pi_0/\lambda)^{-1}. \quad (5.2.3)$$

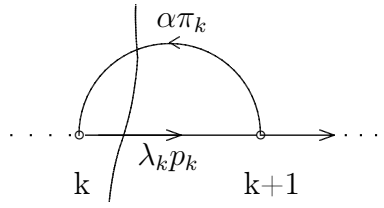

Рис. 5.1. Диаграмма переходов в $\hat{M}/G/1/R$

При анализе диаграммы переходов рис. 5.1, составленной для более общего случая (замкнутая система), учтем выбор моментов регенерации: сразу $после$ окончания обслуживания. Следовательно, произведение $\alpha\pi_k$ имеет смысл интенсивности перехода из $(k+1)$-го в k-е состояние. Уравнение баланса переходов через разрез для разомкнутой системы имеет вид $\lambda p_k = \alpha\pi_k$, откуда

$$p_k = \alpha\pi_k/\lambda, \qquad k = 0, 1, \ldots \quad (5.2.4)$$

Подставив в (5.2.4) при $k = 0$ известное выражение $p_0 = 1 - \lambda b$, находим $\pi_0 = \lambda(1 - \lambda b)/\alpha$. Далее с учетом выражения (5.2.3) имеем $\pi_0 = 1 - \lambda b$, откуда следует $\alpha = \lambda$. Итак, в разомкнутой системе $M/G/1$

$$p_k = \pi_k, \qquad k = 0, 1, \ldots,$$

т. е. стационарные вероятности при выбранном направлении сдвига совпадают с финальными вероятностями вложенной цепи Маркова, что согласуется с теоремой PASTA. Говорят, что система $M/G/1$ *представима* вложенной цепью Маркова. Это единственный случай представимости немарковских систем — и лишь при указанном направлении бесконечно малого сдвига моментов регенерации.

5.2.2. Распределения времени пребывания и ожидания

Умножим уравнения системы (5.2.1) на z в степени, соответствующей индексу стоящей в левой части вероятности, и сложим результаты. Заменяя финальные вероятности вложенной цепи стационарными вероятностями, получаем

$$\begin{aligned} P(z) = & \sum_{k=0}^{\infty} z^k \Big(\sum_{j=1}^{k+1} p_j q_{k+1-j} + p_0 q_k \Big) \\ = & \sum_{j=1}^{\infty} \sum_{k=j-1}^{\infty} z^k p_j q_{k+1-j} + p_0 \sum_{k=0}^{\infty} z^k q_k \\ = & z^{-1} \sum_{j=1}^{\infty} p_j z^j \sum_{k=j-1}^{\infty} z^{k+1-j} q_{k+1-j} + p_0 Q(z) \\ = & z^{-1} Q(z) \Big[P(z) - p_0 \Big] + p_0 Q(z), \end{aligned}$$

где $P(z)$, $Q(z)$ — производящие функции $\{p_k\}$ и $\{q_k\}$ соответственно. Приведя подобные члены, получаем производящую функцию распределения числа заявок в системе $M/G/1$

$$P(z) = p_0 Q(z) \frac{1-z}{Q(z)-z}. \tag{5.2.5}$$

Воспользовавшись законом сохранения очереди — формула (3.2.3), для ПЛС распределения времени пребывания заявки в системе $M/G/1$ имеем выражение

$$\nu(s) = P(1-s/\lambda) = p_0 Q(1-s/\lambda) \frac{s/\lambda}{Q(1-s/\lambda) - 1 + s/\lambda}.$$

Но

$$Q(z) = \sum_{k=0}^{\infty} z^k q_k = \sum_{k=0}^{\infty} z^k \int_0^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t}\, dB(t)$$
$$= \int_0^{\infty} \Big(\sum_{k=0}^{\infty} \frac{(\lambda t z)^k}{k!} \Big) e^{-\lambda t}\, dB(t) = \beta(\lambda - \lambda z),$$

где $\beta(s)$ — ПЛС распределения времени обслуживания с параметром s. Значит,

$$Q(1 - s/\lambda) = \beta(\lambda - \lambda + s) = \beta(s).$$

Окончательно

$$\nu(s) = \frac{p_0 \beta(s)}{1 - \frac{\lambda}{s}[1 - \beta(s)]}. \tag{5.2.6}$$

Пример. Пусть время обслуживания имеет показательное распределение с плотностью $b(t) = \mu e^{-\mu t}$, причем $\mu = 1/b$. Найдем распределение времени пребывания заявки в системе. Прежде всего,

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

Значит,

$$\nu(s) = \frac{p_0 \mu/(\mu + s)}{1 - \frac{\lambda}{s}\left(1 - \frac{\mu}{\mu + s}\right)} = \frac{\mu - \lambda}{\mu - \lambda + s},$$

что совпадает с непосредственно выведенной формулой (4.2.3).

Время ожидания окончания обслуживания складывается из времени ожидания и чистой длительности обслуживания. Следовательно, ПЛС плотности распределения $w(t)$ времени ожидания начала обслуживания

$$\omega(s) = \frac{p_0}{1 - \frac{\lambda}{s}[1 - \beta(s)]} \tag{5.2.7}$$

(формула Полячека—Хинчина). Отметим, что распределение $w(t)$ является смешанным (дискретно-непрерывным): нулевое время ожидания имеет конечную вероятность p_0.

5.2.3. Моменты распределения времени ожидания

ПЛС плотности распределения $w(t)$ случайной длительности ожидания может быть выражено через ее начальные моменты $\{w_k\}$. Разложим ПЛС в формуле (5.2.7) по моментам соответствующих распределений согласно (1.6.2):

$$\begin{aligned} & w_0 - sw_1 + \frac{s^2}{2!}w_2 - \cdots \\ = \; & p_0 \Big/ \Big[1 - \tfrac{\lambda}{s}(1 - 1 + sb_1 - \tfrac{s^2}{2!}b_2 + \tfrac{s^3}{3!}b_3 - \cdots)\Big] \\ = \; & p_0 \Big/ \Big(1 - \lambda b_1 + \tfrac{s}{2!}\lambda b_2 - \tfrac{s^2}{3!}\lambda b_3 + \cdots\Big). \end{aligned}$$

Избавляясь от знаменателя в правой части, приходим к равенству

$$\Big(w_0 - sw_1 + \frac{s^2}{2!}w_2 - \cdots\Big)\Big(1 - \lambda b_1 + \frac{s}{2!}\lambda b_2 - \frac{s^2}{3!}\lambda b_3 + \cdots\Big) = p_0.$$

Раскрывая скобки в его левой части и группируя слагаемые с одинаковыми степенями s, получаем равенство

$$(-1)^k \frac{w_k}{k!}(1 - \lambda b_1) + \lambda \sum_{j=0}^{k} (-1)^{k+1} \frac{w_j b_{k+1-j}}{j!\,(k+1-j)!} = 0,$$

откуда следует формула для рекуррентного расчета $\{w_k\}$, отправляясь от очевидного $w_0 = 1$:

$$w_k = \frac{\lambda}{1 - \lambda b_1} \sum_{j=0}^{k-1} \frac{k!}{j!\,(k+1-j)!}\, b_{k+1-j} w_j, \qquad k = 1, 2, \ldots \tag{5.2.8}$$

В частности,

$$w_1 = \frac{\lambda b_2}{2(1 - \lambda b_1)}.$$

Эта формула совпадает с выведенной в разд. 3.4 элементарным методом на основе принципа сохранения объема работы и тем подтверждает справедливость последнего.

Итак, среднее время ожидания пропорционально *второму* моменту b_2 распределения чистого времени обслуживания и при фиксированном b_1 увеличивается с ростом дисперсии времени обслуживания. Минимальное среднее время ожидания достигается при нулевой дисперсии (постоянной длительности обслуживания), когда $b_2 = b_1^2$. В случае показательного распределения времени обслуживания с тем же значением b_1 второй момент $b_2 = 2b_1^2$, и среднее время ожидания возрастает ровно *вдвое*.

Отмеченный факт является наиболее наглядным свидетельством недостаточности учета только средних значений исходных временны́х распределений, хотя бы одно из которых заметно отличается от показательного, даже для получения усредненных характеристик систем обслуживания. С увеличением числа каналов этот эффект ослабевает, а при бесконечном числе каналов исчезает вообще.

Среднее время пребывания заявки в системе

$$v_1 = w_1 + b_1.$$

О влиянии коэффициента вариации распределения времени обслуживания на распределения числа заявок и времени пребывания в системе $M/G/1$ можно судить по рис. 5.2, 5.3 соответственно.

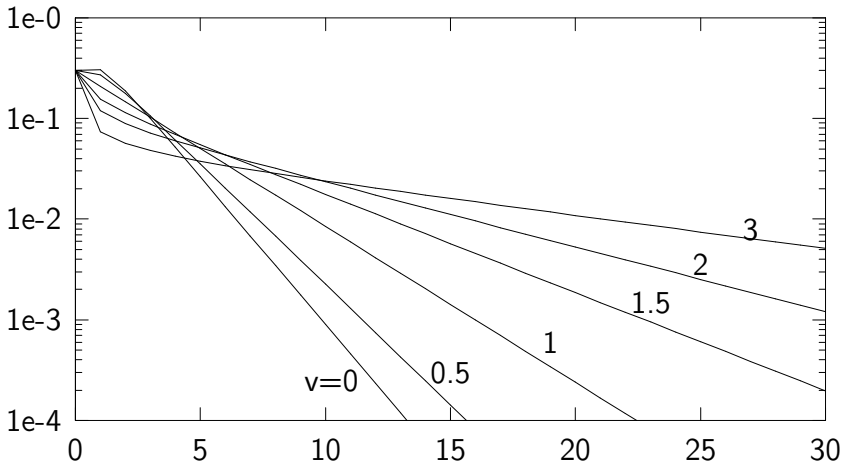

Рис. 5.2. Распределение числа заявок в системе $M/G/1$

Сопоставление рис. 5.2 с формулой (2.5.2) убедительно свидетельствует о сильном влиянии второго и высших моментов распределения времени обслуживания на принимаемое решение (оптимальный запас) — в особенности при малых отношениях s/d и большой вариации.

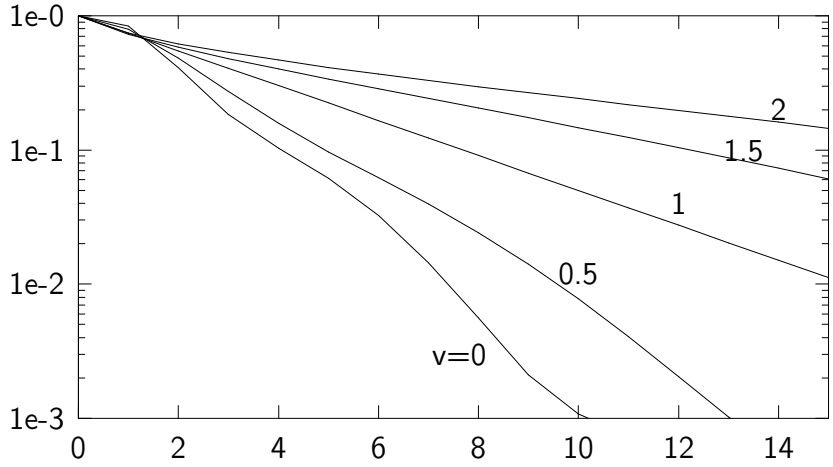

Рис. 5.3. ДФР времени пребывания заявки в системе $M/G/1$

Таблица 5.1 демонстрирует влияние на ДФР времени пребывания высших моментов распределения времени обслуживания.

Таблица 5.1. ДФР времени пребывания заявки в системе $M/G/1$

t	Распред. времени обслуж.			t	Распред. времени обслуж.		
	равномерн.	треугольн.	гамма-		равномерн.	треугольн.	гамма-
0.1	0.998318	0.998318	0.998609	3.1	0.231264	0.231266	0.228396
0.3	0.982587	0.982595	0.983796	3.3	0.204550	0.204497	0.202151
0.5	0.948678	0.948700	0.950386	3.5	0.181533	0.181467	0.179644
0.7	0.899583	0.899625	0.901267	3.7	0.161692	0.161618	0.160327
0.9	0.836379	0.836440	0.837324	3.9	0.144550	0.144469	0.143700
1.1	0.767659	0.767734	0.767714	4.1	0.129680	0.129595	0.129318
1.3	0.696074	0.696158	0.695133	4.3	0.116708	0.116623	0.117940
1.5	0.624911	0.624998	0.623009	4.5	0.105318	0.105235	0.105803
1.7	0.556594	0.556676	0.553861	4.7	0.095243	0.951642	0.096072
1.9	0.492754	0.492825	0.489381	4.9	0.086265	0.086191	0.087380
2.1	0.438207	0.438264	0.434725	5.1	0.078206	0.078138	0.079547
2.3	0.385648	0.385687	0.381953	5.3	0.070921	0.070861	0.072435
2.5	0.338969	0.338987	0.335256	5.5	0.064299	0.064247	0.065932
2.7	0.297956	0.297955	0.294398	5.7	0.058251	0.058207	0.059954
2.9	0.262220	0.262199	0.258954	5.9	0.052707	0.052671	0.054438

Она построена для трех различных распределений по шести моментам, полученным численным дифференцированием (5.2.6) при $b_1 = 1$, $\lambda = 0.7$ и коэффициенте вариации $v = 0.4$. Отметим, что первые два распределения как симметричные имеют одинаковые третьи моменты, и результаты счета для них согласуются лучше. Подчеркнем также, что расхождение на «хвостах» распределений (скажем, для значений ДФР, меньших 10^{-2}) неизбежно при любом разумном числе учтенных моментов.

Практически те же результаты были получены при вычислении моментов распределения времени ожидания согласно (5.2.8) и свертке их с моментами распределения времени обслуживания.

5.3. Обслуживание с порогом включения и разогревом

В ряде систем массового обслуживания, в особенности при относительно малой загрузке, оказывается целесообразным введение «порога включения» — обслуживание начинается при скоплении в системе $r > 1$ заявок и заканчивается при полном освобождении системы. Такой режим увеличивает как период непрерывной занятости, так и время, в течение которого обслуживание не ведется. Это на относительно длительные периоды позволяет переводить автоматическую аппаратуру — в целях экономии ресурса и электроэнергии — в облегченный (дежурный) режим. В системах с участием человека появляется возможность полностью выключать значительную часть техники и переводить оператора на решение других задач.

Очевидно, что обслуживание первой заявки в этих условиях (а практически даже при $r = 1$) сопряжено с выполнением некоторых дополнительных операций («разогревом» системы в прямом и/или переносном смысле) и в среднем будет продолжаться дольше, чем обслуживание прочих заявок, причем изменение показателей оперативности обслуживания в общем случае не может быть сведено к постоянной задержке. Ниже предлагается методика расчета системы $M/G/1$, позволяющая учитывать обе (при желании — порознь) упомянутые особенности алгоритма ее функционирования.

5.3.1. Вложенная цепь Маркова

Аналогично разд. 5.2 выберем моменты регенерации процесса. Снова обозначим q_j вероятность прибытия j новых требований за время обслуживания с распределением $B(t)$. Эти вероятности определяются формулой (2.1.10), и в главе 3 описан способ их эффективного вычисления. Аналогично определим вероятности $\{q_j^*\}$ и для распределения $B^*(t)$ обслуживания с разогревом. Для установившегося режима застать в системе $k < r-1$ требований можно, если в ней в предыдущий момент регенерации находилось $j = \overline{1, k+1}$ заявок, одна обслужилась и пришло еще $k+1-j \geq 0$ заявок. После разогрева вложенная цепь Маркова может оказаться только в состояниях $r-1, r, \ldots$, причем до начала разогрева следует дождаться прибытия r заявок. Следовательно, финальные (предельные при устремлении номера момента регенерации к бесконечности) вероятности состояний вложенной цепи связаны уравнениями

$$\begin{aligned} \pi_k &= \sum_{j=1}^{k+1} \pi_j q_{k+1-j}, && k = \overline{0, r-2}, \\ \pi_k &= \sum_{j=1}^{k+1} \pi_j q_{k+1-j} + \pi_0 q^*_{k-r+1}, && k = r-1, r, r+1, \ldots. \end{aligned}$$

Из них следуют рекуррентные выражения для расчета вероятностей:

$$\begin{aligned} \pi_k &= \left(\pi_{k-1} - \sum_{j=1}^{k-1} \pi_j q_{k-j}\right) \Big/ q_0, && k = \overline{1, r-1}, \\ \pi_k &= \left(\pi_{k-1} - \pi_0 q^*_{k-r} - \sum_{j=1}^{k-1} \pi_j q_{k-j}\right) \Big/ q_{k,0}, && k = r, r+1, \ldots \end{aligned} \tag{5.3.1}$$

5.3.2. Переход к стационарным вероятностям

Обозначим

α — среднюю частоту немарковских (по завершению обслуживания) переходов между состояниями системы,

b — среднюю длительность обслуживания,

b^* — среднюю длительность обслуживания с разогревом.

Средний интервал между немарковскими переходами

$$\alpha^{-1} = (r/\lambda + b^*)\pi_0 + b(1 - \pi_0). \tag{5.3.2}$$

Свяжем немарковские переходы с моментами регенерации вложенной цепи, для которых рассчитываются $\{\pi_k\}$. Параметр λ_k марковских переходов (прибытие новых заявок) при наличии в системе ровно k заявок может быть отнесен к произвольному моменту и должен связываться со стационарной вероятностью p_k. Баланс переходов между смежными состояниями системы в стационарном режиме требует равенства

$$\lambda p_k = \alpha\pi_k, \qquad k = 1, 2, \ldots \tag{5.3.3}$$

Частоты включения и выключения режима обслуживания совпадают:

$$\lambda p^*_{r-1} = \alpha\pi_0.$$

Поскольку вероятности $\{p^*_k\}$ всех r состояний с отключенным обслуживанием равны между собой, это уравнение можно заменить на

$$\lambda p^*_0 = \alpha\pi_0. \tag{5.3.4}$$

Определим суммарную вероятность свободного состояния p^*. Она равна отношению средней длительности свободного состояния r/λ к сумме этой длительности и средней длины T_B периода непрерывной занятости (ПНЗ). В среднюю длину ПНЗ входят:

- средняя длительность разогрева b^*;
- длительность обслуживания остальных $r - 1$ заявок, формирующих порог включения — в среднем $(r - 1)b$;
- длительность обслуживания вновь прибывших заявок — в среднем $b\lambda T_B$.

Итак, $T_B = b^* + b(r - 1 + \lambda T_B)$, откуда

$$T_B = \frac{b^* + (r - 1)b}{1 - \lambda b}. \tag{5.3.5}$$

Общая вероятность свободного состояния

$$p^* = \frac{r/\lambda}{r/\lambda + T_B} = \frac{r}{r + \lambda T_B},$$

а p_0^* — в r раз меньше.

Подставляя полученные выражения в (5.3.4), приходим к уравнению

$$\pi_0 = \frac{\lambda}{r + \lambda \dfrac{b^* + (r-1)b}{1 - \lambda b}} \left[\left(\frac{r}{\lambda} + b^* \right) \pi_0 + (1 - \pi_0) b \right]$$

с решением

$$\pi_0 = \frac{1 - \lambda b}{r + \lambda (b^* - b)}.$$

Подстановка его в (5.3.2) дает $\alpha = \lambda$. Таким образом, рассматриваемая нами система, как и ее простейший аналог — обычная $M/G/1$, при сделанном выборе моментов регенерации процесса представима вложенной цепью Маркова.

Итак, расчет стационарных вероятностей состояний для нашей системы должен выполняться в следующем порядке:

1. Вычислить среднюю длину T_B периода непрерывной занятости согласно (5.3.5).
2. Положить $p_0^* = \pi_0 = 1/(r + \lambda T_B)$.
3. По формулам (5.3.2) найти вероятности состояний для режима со включенным обслуживанием.
4. К найденным вероятностям для $k = \overline{1, r-1}$ добавить p_0^*.

5.3.3. Распределение времени пребывания

Используя закон сохранения стационарной очереди, можно показать (разд. 3.2), что ПЛС $\nu(s)$ распределения времени пребывания заявки в системе с дисциплиной обслуживания FCFS связано с производящей функцией $P(z)$ распределения числа заявок в системе формулой

$$\nu(s) = P(1 - s/\lambda). \tag{5.3.6}$$

Умножая строки системы уравнений (5.3.2) на соответствующие степени z и выполняя суммирование, для производящей функции вероятностей состояний со включенным обслуживанием получаем уравнение

$$\tilde{P}(z) = p_0^* \frac{z^r Q^*(z) - Q(z)}{z - Q(z)}.$$

Здесь $Q(z)$ и $Q^*(z)$ — производящие функции вероятностей $\{q_k\}$ для обычного режима обслуживания и разогрева соответственно. Заметим, что

$$\begin{aligned} Q(z) &= \sum_{k=0}^{\infty} z^k \int_0^{\infty} \frac{(\lambda t)^k}{k\,!} e^{-\lambda t}\, dB(t) = \int_0^{\infty} \Big[\sum_{k=0}^{\infty} \frac{(z\lambda t)^k}{k\,!}\Big] e^{-\lambda t}\, dB(t) \\ &= \int_0^{\infty} e^{-\lambda t(1-z)}\, dB(t) = \beta(\lambda(1-z)), \end{aligned}$$

т. е. сводится к ПЛС с указанным параметром от распределения длительности обслуживания.

С учетом вероятностей дополнительных состояний с отключенным обслуживанием производящая функция распределения числа заявок

$$P(z) = \tilde{P}(z) + p_0^* \sum_{j=1}^{r-1} z^j = \tilde{P}(z) + p_0^* \frac{z^r - z}{z-1}.$$

Подставляя эти результаты в (5.3.6), получаем обобщение формулы Полячека—Хинчина:

$$\nu(s) = p_0^* \Big\{ \frac{(1-s/\lambda)^r \beta^*(s) - \beta(s)}{1 - s/\lambda - \beta(s)} - \frac{\lambda}{s}\Big[(1-s/\lambda)^r - 1 + s/\lambda\Big]\Big\}.$$

Моменты соответствующего распределения можно найти дифференцированием $\nu(s)$ в нуле — аналитическим или численным.

Самостоятельный интерес представляют частные случаи — порог включения и разогрев порознь. В первом случае $b = b^*$, $\beta^*(\cdot) = \beta(\cdot)$ и подставляемая в выражение для $p*$ средняя длина ПНЗ

$$T_B = rb/(1-\lambda b).$$

Во втором случае ($r = 1$) ПЛС распределения пребывания сводится к

$$\nu(s) = p_0^* \frac{(1-s/\lambda)\beta^*(s) - \beta(s)}{1 - s/\lambda - \beta(s)},$$

причем $T_B = b^*/(1-\lambda b)$, $p_0^* = 1/(1+\lambda T_B)$.

5.4. Система с пачками заявок

Задачи этого типа возникают как непосредственно (к примеру, обслуживание семейных и иных коллективов), так и в расчете сетей обслуживания с процессами расщепления заявок. Для их решения необходимо уметь рассчитывать распределения:

- трудоемкости пачки;
- времени ожидания пачки в целом;
- дополнительных задержек заявок внутри пачки;
- общего числа заявок в системе.

5.4.1. Трудоемкость пачки

ПЛС распределения трудоемкости пачки, объем которой задается дискретным распределением $\{f_m\}$, определяется формулой

$$B(s) = \sum_m f_m[\beta(s)]^m. \tag{5.4.1}$$

Ее моменты $\{B_i\}$ могут быть вычислены с помощью численного дифференцирования таблицы ПЛС в окрестности нуля или последовательным накоплением сверток распределения длительности обслуживания заявки и их взвешенным суммированием. Средняя трудоемкость пачки B_1 равна произведению b_1 на средний объем пачки.

5.4.2. Распределения числа пачек и времени ожидания

Упомянутые распределения рассчитываются по тем же формулам, что в стандартной модели $M/G/1$, с заменой трудоемкости заявки на трудоемкость пачки. Приведем сравнительные результаты моделирования (1 млн. пачек равновероятного объема от 1 до 6) и расчета системы с равномерной на [0,10] длительностью обслуживания заявки для коэффициента загрузки 0.8 (табл. 5.2).

Таблица 5.2. Распределение числа пачек в очереди по MG1

j	Расчет	Модель	j	Расчет	Модель
0	.40341e-0	.40224e-0	10	.85699e-2	.88636e-2
1	.16340e-0	.16312e-0	11	.61429e-2	.63665e-2
2	.12174e-0	.12099e-0	12	.44032e-2	.44870e-2
3	.88167e-1	.88313e-1	13	.31563e-2	.31766e-2
4	.63288e-1	.63167e-1	14	.22624e-2	.21522e-2
5	.45342e-1	.45830e-1	15	.16217e-2	.16070e-2
6	.32483e-1	.32488e-1	16	.11625e-2	.11304e-2
7	.23276e-1	.23682e-1	17	.83326e-3	.78494e-3
8	.16681e-1	.17140e-1	18	.59728e-3	.60351e-3
9	.11956e-1	.12123e-1	19		.50317e-3

Кроме того, были получены моменты длительности ожидания для пачки (табл. 5.3).

Таблица 5.3. Моменты распределения ожидания для пачки

Способ расчета	Порядок момента		
	1	2	3
по ФПХ	.46594e+2	.51851e+4	.86156e+6
через MFACT	.46594e+2	.51851e+4	.86156e+6
в модели	.46938e+2	.52437e+4	.86378e+6
в модели через MFACT	.46943e+2	.52475e+4	.86595e+6

Здесь вычисление «по ФПХ» (формула Полячека—Хинчина) подразумевает расчет высших моментов по рекуррентным формулам (5.2.8), а по MFACT — через факториальные моменты длины очереди пачек.

5.4.3. Задержки внутри пачки

ПЛС средней дополнительной задержки внутри пачки должна определяться по формуле

$$\delta(s) = \sum_{m>0} \frac{f_m}{m} \sum_{i=1}^{m} \beta^{i-1}(s) = \frac{1}{1-\beta(s)} \sum_{m>0} \frac{f_m}{m} [1 - \beta^m(s)]. \qquad (5.4.2)$$

Средневзвешенная задержка заявки получается как свертка задержки пачки и средней задержки внутри пачки.

5.4.4. Распределение числа заявок в системе

Распределение числа заявок в системе существенно для решения проблем размещения очереди и задач типа определения восстанавливаемого ЗИП. Это распределение (см. табл. 5.4) можно получить по алгоритму для стандартной системы $M/G/1$ после замены $\{q_j\}$ на их аналоги для обобщенного пуассонова потока из разд. 2.1.6.

Таблица 5.4. Распределение числа заявок в системе $M^X/G/1$

j	Расчет	Модель	j	Расчет	Модель
0	.20025e-0	.19985e-0	21	.11847e-1	.11988e-1
1	.49231e-1	.49081e-1	22	.10809e-1	.10988e-1
2	.52554e-1	.52130e-1	23	.98625e-2	.99888e-2
3	.54630e-1	.54487e-1	24	.89987e-2	.91151e-2
4	.55031e-1	.54938e-1	25	.82106e-2	.83720e-2
5	.53278e-1	.52955e-1	26	.74914e-2	.76478e-2
6	.48852e-1	.48722e-1	27	.68353e-2	.69639e-2
7	.41197e-1	.41141e-1	28	.62366e-2	.63042e-2
8	.38521e-1	.38470e-1	29	.56904e-2	.57371e-2
9	.35595e-1	.35358e-1	30	.51920e-2	.52330e-2
10	.32594e-1	.32395e-1	31	.47372e-2	.47071e-2
11	.29685e-1	.29406e-1	32	.43223e-2	.43543e-2
12	.27006e-1	.26686e-1	33	.39438e-2	.39469e-2
13	.24627e-1	.24475e-1	34	.35984e-2	.36217e-2
14	.22507e-1	.22442e-1	35	.32832e-2	.33337e-2
15	.20541e-1	.20569e-1	36	.29956e-2	.30517e-2
16	.18737e-1	.18869e-1	37	.27333e-2	.28815e-2
17	.17093e-1	.17107e-1	38	.24939e-2	.26155e-2
18	.15595e-1	.15729e-1	39	.22755e-2	.24088e-2
19	.14231e-1	.14347e-1	40	.20762e-2	.22400e-2
20	.12985e-1	.13017e-1	41	.18944e-2	.20520e-2

Обращает на себя внимание вызванное группировкой заявок в пачки сильнейшее затягивание хвостов распределения.

5.5. Замкнутая система $\hat{M}/G/1$

5.5.1. Распределение числа заявок

Рассмотрим замкнутую СМО с R источниками заявок, каждый из которых в своей активной фазе формирует пуассоновский поток, и построим вложенную цепь Маркова на моменты, непосредственно следующие за окончанием обслуживания. Финальные вероятности $\{\pi_k\}$ вложенной цепи оказываются связанными системой уравнений

$$\pi_k = \sum_{j=0}^{k+1} \pi_j q_{j,k+1-j}, \qquad k = \overline{0, R-1}. \tag{5.5.1}$$

Коэффициенты $\{q_{j,r}\}$, как и определяемые (2.1.10), имеют смысл вероятности прибытия ровно r новых заявок за случайное время обслуживания, но *с учетом начального числа заявок в системе* $j = \overline{0, R-1}$; $r = \overline{0, R-1-j}$.

Каждый потенциальный источник заявок независимо от остальных за время t формирует заявку с вероятностью $1 - e^{-\lambda t}$. Распределение числа поступивших заявок подчиняется биномиальному закону, так что

$$\begin{aligned} q_{j,r} = & \binom{R-j}{r} \int_0^\infty (1-e^{-\lambda t})^r e^{-\lambda(R-j-r)t}\, dB(t) \\ = & \binom{R-j}{r} \sum_{m=0}^{r} \binom{r}{m} (-1)^m \int_0^\infty e^{-\lambda(R+m-j-r)t}\, dB(t) \\ = & \binom{R-j}{r} \sum_{m=0}^{r} \binom{r}{m} (-1)^m a_0^{(R+m-j-r)}\, dB(t), \end{aligned}$$

где коэффициенты $\{a_0\}$ вычисляются по формуле типа (2.1.10) для $j = 0$ с дополнительным множителем при λ, равным верхнему индексу. Заметим, что по построению вложенной цепи Маркова $q_{0,r} = q_{1,r}$ при всех r (отсчет времени обслуживания в свободной системе начинается с момента прибытия первой заявки).

Имея все $\{q_{j,r}\}$, можно, задавшись $\pi_0 = 1$, получить остальные вероятности вложенной цепи по формуле

$$\pi_k = \left(\pi_{k-1} - \pi_0 q_{1,k-1} - \sum_{j=1}^{k-1} \pi_j q_{j,k-j}\right) \Big/ q_{k,0}, \qquad k = \overline{1, R-1}, \tag{5.5.2}$$

после чего нормировать результаты.

Переход к стационарным вероятностям состояний здесь может быть осуществлен тем же способом, что и в разомкнутой системе $M/G/1$, но с учетом зависимости λ от числа заявок в системе (см. рис. 5.1):

$$\alpha = (b + \pi_0/\lambda_0)^{-1} = [b + \pi_0/(\lambda R)]^{-1}, \tag{5.5.3}$$

$$\begin{aligned} p_k &= \frac{\alpha\pi_k}{\lambda(R-k)}, \qquad k = \overline{0, R-1}, \\ p_R &= 1 - \sum_{k=0}^{R-1} p_k. \end{aligned} \tag{5.5.4}$$

Благодаря отмеченному выше свойству саморегулируемости замкнутые системы по своим характеристикам менее чувствительны к виду исходных распределений, чем разомкнутые.

5.5.2. Моменты распределения времени ожидания

Обозначим

π_i — вероятность того, что вновь прибывший запрос застает в системе i ранее пришедших, $i = \overline{0, R-1}$;

$B(t)$ — распределение чистой длительности обслуживания запроса ($\bar{B}(t)$ — ДФР);

$\tilde{B}(t) = b_1^{-1} \int_0^t \bar{B}(u)\, du$ — распределение остатка обслуживания.

Тогда распределение длительности ненулевого ожидания

$$W(t) = \sum_{i=1}^{R-1} \pi_i \tilde{B}(t) * B^{(i-1)*}(t); \tag{5.5.5}$$

здесь знаком $*$ обозначен оператор свертки.

Входящие в эту формулу вероятности $\{\pi_i\}$ застать в системе ровно i заявок подсчитываются согласно

$$\pi_i = \lambda_i p_i \Big/ \sum_{j=0}^{R-1} \lambda_j p_j, \qquad i = \overline{0, R-1}, \tag{5.5.6}$$

где $\{p_j\}$ — известное стационарное распределение числа запросов в системе, а λ_j есть интенсивность входящего потока при наличии в системе j заявок. Обычно $\lambda_j = \lambda(R-j)$. Свертки, необходимые для реализации формулы (5.5.5), удобно выполнять в моментах согласно (1.6.11).

Для замкнутой одноканальной системы с R источниками заявок среднее время v реакции системы может быть найдено из условия равенства частоты вошедших в систему и обслуженных заявок:

$$\frac{R}{v+1/\lambda} = \frac{1-p_0}{b},$$

откуда

$$v = Rb/(1-p_0) - 1/\lambda.$$

5.5.3. «Разомкнутая» аппроксимация

При большом числе источников заявок разумно считать интенсивность потока заявок меняющейся мало. Это позволяет представить упомянутую интенсивность формулой

$$\lambda = \frac{R}{t+w+b_1}.$$

Знаменатель этой формулы есть средняя длительность цикла заявки, в которую входят среднее время t «обдумывания» заявки в источнике, средняя длительность обслуживания b_1 и средняя длительность ожидания w. Последняя может быть найдена по формуле Полячека—Хинчина (3.4.1), так что

$$\lambda = \frac{R}{t+[\lambda b_2/2(1-\lambda b_1)]+b_1}.$$

Это уравнение можно привести к квадратному

$$\lambda^2[b_2 - 2b_1(t+b_1)] + 2\lambda[t+(R+1)b_1] - 2R = 0$$

с решением

$$\lambda = \frac{t+(R+1)b_1 \pm \sqrt{t^2+(R-1)^2b_1^2 - 2tb_1(R-1) + 2Rb_2}}{2b_1(t+b_1)-b_2}.$$

В случае $R=30$, $b_1=4$, $b_2=32$ (показательное распределение) и $t=15$ уравнение имеет корни 0.240872 и 2.075795, причем только первый удовлетворяет условию $\lambda b_1 < 1$. При этом среднее время ожидания $w = 105.553$.

5.6. Система $GI/M/1$

5.6.1. Вложенная цепь Маркова

Будем контролировать число заявок в системе $GI/M/1$ на моменты, непосредственно *предшествующие* поступлению очередной заявки. Обозначим

$$q_j = \int_0^\infty \frac{(\mu t)^j}{j\,!} e^{-\mu t}\, dA(t), \qquad j = 0, 1, \ldots, \tag{5.6.1}$$

вероятности обслуживания ровно j требований за интервал между прибытиями смежных заявок; эти вероятности также могут быть вычислены согласно (2.1.15). Для перехода вложенной цепи Маркова из состояния S_j в состояние S_k, $k \geq 1$, за этот интервал должно быть завершено обслуживание $j-k+1$ заявки, $j = 0, 1, \ldots$. Для перехода из S_j в S_0 за такой же интервал должно успеть завершиться обслуживание $j+1$ заявки. Это событие имеет вероятность $1 - \sum_{i=0}^{j} q_i$. Таким образом, здесь финальные вероятности $\{\pi_j\}$ вложенной цепи Маркова связаны системой линейных алгебраических уравнений

$$\begin{aligned} \pi_0 &= \sum_{j=0}^{\infty} \pi_j (1 - \sum_{i=0}^{j} q_i), \\ \pi_k &= \sum_{j=k-1}^{\infty} \pi_j q_{j-k+1}, \qquad k = 1, 2, \ldots \end{aligned} \tag{5.6.2}$$

Форма уравнений (5.6.2), в отличие от (5.2.1), не допускает последовательного расчета вероятностей состояний вложенной цепи. Можно показать, однако, что эти вероятности имеют вид

$$\pi_k = (1-\omega)\omega^k, \qquad k = 0, 1, \ldots, \tag{5.6.3}$$

где ω — единственный корень уравнения

$$\omega = \int_0^\infty e^{-\mu t(1-\omega)}\, dA(t) \tag{5.6.4}$$

на интервале (0,1).

Проверим подстановкой уравнения (5.6.2). Первое из них

$$\begin{aligned} \pi_0 = \; & (1-\omega) \sum_{j=0}^{\infty} \omega^j (1 - \sum_{i=0}^{j} q_i) \;\; = \;\; 1 - (1-\omega) \sum_{j=0}^{\infty} \omega^j \sum_{i=0}^{j} q_i \\ = \; & 1 - (1-\omega) \sum_{i=0}^{\infty} q_i \sum_{j=i}^{\infty} \omega^j \;\; = \;\; 1 - \sum_{i=0}^{\infty} q_i \omega^i. \end{aligned}$$

Поскольку

$$\sum_{i=0}^{\infty} q_i \omega^i = \int_0^{\infty} \Big[\sum_{i=0}^{\infty} \frac{(\mu\omega t)^i}{i\,!} e^{-\mu t} \Big]\, dA(t) = \int_0^{\infty} e^{-\mu t(1-\omega)}\, dA(t) = \omega,$$

правильность формулы (5.6.3) при $k = 0$ подтвердилась. Далее, для $k \geq 1$ имеем

$$\begin{aligned} \pi_k = & \sum_{j=k-1}^{\infty} \pi_j q_{j-k+1} = \sum_{j=0}^{\infty} q_j \pi_{k+j-1} \\ = & (1-\omega) \sum_{j=0}^{\infty} \omega^{k+j-1} \int_0^{\infty} \frac{(\mu t)^j}{j\,!} e^{-\mu t}\, dA(t) \\ = & (1-\omega)\omega^{k-1} \int_0^{\infty} e^{-\mu t} \Big[\sum_{j=0}^{\infty} \frac{(\mu\omega t)^j}{j\,!} \Big]\, dA(t) \\ = & (1-\omega)\omega^{k-1} \int_0^{\infty} e^{-\mu t(1-\omega)}\, dA(t) = (1-\omega)\omega^k, \end{aligned}$$

что окончательно доказывает (5.6.3).

Укажем способ нахождения ω. Представив плотность распределения интервалов между заявками входящего потока $a(t) = dA(t)/dt$ с помощью разложения (1.9.1), убеждаемся, что правая часть (5.6.4) может быть получена по формуле (1.9.3) для q_0 при соответствующей замене обозначений. Таким образом, (5.6.4) сводится к уравнению

$$\omega = \Big[\frac{\nu}{\mu(1-\omega)+\nu} \Big]^{\alpha} \sum_{i=0}^{N} \frac{g_i}{[\mu(1-\omega)+\nu]^i} \frac{\Gamma(\alpha+i)}{\Gamma(\alpha)}, \qquad (5.6.5)$$

в котором μ — интенсивность обслуживания, а ν — параметр аппроксимирующего гамма-распределения. Это уравнение можно решить, например, методом итераций при начальном значении $\omega = \rho^{2/(v_A^2+1)}$.

Из уравнения (5.6.4) следует, что параметр ω определяется распределением $A(t)$ *в целом*, т. е. *всеми его моментами*. Соответственно так же обстоит дело со всеми показателями системы GI/G/1, выражаемыми через ω — в частности, со средним временем ожидания. Напомним, что в случае M/G/1 оно зависит от *двух* моментов распределения времени обслуживания.

5.6.2. Распределение времени пребывания

Время пребывания в системе заявки, застающей в одноканальной СМО k ранее прибывших требований (назовем ее k-заявкой), будет

состоять из времени завершения начатого обслуживания и полного времени обслуживания еще k заявок. Для показательного закона распределение времени завершения начатого обслуживания совпадает с исходным распределением. Таким образом, распределение времени пребывания в системе k-заявки есть $(k+1)$-кратная свертка показательного закона, а его ПЛС есть $[\mu/(\mu+s)]^{k+1}$. С учетом формулы (5.6.3) ПЛС распределения времени пребывания

$$\begin{aligned}\nu(s) = & \sum_{k=0}^{\infty}(1-\omega)\omega^k[\mu/(\mu+s)]^{k+1} \\ = & \frac{\mu(1-\omega)}{\mu+s}\sum_{k=0}^{\infty}\Big(\frac{\mu\omega}{\mu+s}\Big)^k = \frac{\mu(1-\omega)}{\mu(1-\omega)+s}.\end{aligned}$$

Выполнив обратное преобразование, убеждаемся, что время пребывания заявки в системе $GI/M/1$ подчинено $показательному$ закону с параметром $\mu(1-\omega)$, причем тип распределения интервалов между заявками влияет только на численное значение параметра ω — см. уравнение (5.6.4). Соответственно моменты распределения времени пребывания

$$v_k = \frac{k\,!}{[\mu(1-\omega)]^k}, \qquad k = 1, 2, \ldots$$

Аналогичным образом устанавливаем, что плотность распределения времени ожидания начала обслуживания для $t \geq 0$

$$w(t) = \mu\omega(1-\omega)e^{-\mu(1-\omega)t}.$$

Поскольку вероятность ненулевого ожидания равна ω, $условное$ распределение его длительности подчинено тому же показательному закону, что и распределение времени пребывания заявки в системе.

Проверим полученные результаты для простейшего входящего потока. Подстановка $A(t) = 1 - e^{-\lambda t}$ в уравнение (5.6.4) дает

$$\omega = \int_0^{\infty} e^{-\mu(1-\omega)t}\lambda e^{-\lambda t}\,dt = \lambda/(\lambda+\mu(1-\omega)).$$

Квадратное уравнение

$$\mu\omega^2 - (\lambda+\mu)\omega + \lambda = 0$$

имеет на промежутке (0,1) только один корень $\omega = \lambda/\mu = \rho$. Следовательно, параметр распределения времени пребывания заявки в системе $\mu(1-\omega) = \mu - \lambda$, что согласуется с результатами разд. 4.2.

5.6.3. Стационарное распределение числа заявок

В системе $GI/M/1$ параметр потока немарковских переходов $\alpha = a^{-1}$, где a — средний интервал между заявками входящего потока. Так как моменты регенерации $предшествуют$ поступлению очередной заявки, то $\alpha\pi_k$ имеет смысл параметра потока переходов из k-го в $(k+1)$-е состояние. Диаграмма переходов рис. 5.4, которой мы воспользуемся для $n = 1$ и $\mu_k = \mu$, позволяет записать уравнения баланса $\mu p_k = \alpha\pi_{k-1}$, так что

$$p_k = \frac{\alpha}{\mu}\pi_{k-1} = \frac{\pi_{k-1}}{\mu a}, \qquad k = 1, 2, \ldots \tag{5.6.6}$$

Вероятность свободного состояния $p_0 = 1 - 1/(\mu a)$ находим из условия баланса (3.1.1).

Подстановка коэффициента загрузки $\rho = 1/(\mu a)$ и выражение $\{\pi_k\}$ через ω позволяют получить формулы

$$\begin{array}{ll} p_0 = & 1 - \rho, \\ p_k = & \rho(1-\omega)\omega^{k-1}, \qquad k = 1, 2, \ldots \end{array} \tag{5.6.7}$$

Проверим для системы $GI/M/1$ закон сохранения стационарной очереди. Ожидаемое число заявок в системе

$$\begin{array}{rl} \bar{k} = & \sum\limits_{k=1}^{\infty} k\rho(1-\omega)\omega^{k-1} = \rho(1-\omega)\sum\limits_{k=1}^{\infty} k\omega^{k-1} \\ = & \rho(1-\omega)\dfrac{d}{d\omega}\dfrac{\omega}{1-\omega} = \rho(1-\omega)\dfrac{1}{(1-\omega)^2} = \dfrac{\rho}{1-\omega}. \end{array}$$

Подставляя значение ρ, убеждаемся, что

$$L_1 = \bar{k} = 1/[a\mu(1-\omega)] = v_1/a = \lambda v_1,$$

т. е. формула Литтла справедлива и для рекуррентного потока. Проделав аналогичные выкладки, видим, что второй факториальный момент распределения числа заявок в системе

$$f_{[2]} = 2\rho\omega/(1-\omega)^2,$$

и соотношение $v_2 = f_{[2]}/\lambda^2$ не имеет места. Таким образом, формулы (3.2.4) для высших (начиная со второго) моментов справедливы только при простейшем входящем потоке.

5.7. Система $GI/M/n$

5.7.1. Вложенная цепь Маркова

Метод расчета системы вида $GI/M/n$ был предложен Л. Такачем [264] в 1960 г. В этой системе предельные значения вероятностей состояний на моменты, непосредственно предшествующие поступлению очередного требования, отвечают системе уравнений

$$\pi_k = \sum_{j=k-1}^{\infty} q_{j,k}\pi_j, \qquad k = 1, 2, \ldots \tag{5.7.1}$$

Вероятности $\{q_{j,k}\}$ скачков за интервал между моментами прибытия смежных требований в данном случае зависят не только от разности между исходным и конечным индексами, но и от значения начального индекса, который определяет количество каналов, фактически производящих обслуживание. Получим выражения для вероятностей переходов.

Прежде всего отметим, что для всех j и k при $0 \le j \le k-2$ $q_{j,k} = 0$. При $j = \overline{0, n-1}$ требование, прибывшее в момент регенерации процесса, переводит его в состояние $j+1 \le n$; из этого числа требований ровно k до следующего момента регенерации должны остаться необслуженными. При показательном распределении времени обслуживания распределение времени ожидания его окончания не зависит от уже истекшей длительности. Значит,

$$\begin{aligned} q_{j,k} &= \binom{j+1}{k} \int_0^{\infty} (e^{-\mu t})^k (1 - e^{-\mu t})^{j+1-k}\, dA(t) \\ &= \binom{j+1}{k} \int_0^{\infty} e^{-k\mu t} (1 - e^{-\mu t})^{j+1-k}\, dA(t), \\ &\qquad j = \overline{0, n-1}; \quad k = \overline{0, j+1}. \end{aligned} \tag{5.7.2}$$

В случае $j \ge n$ и $k \ge n$ в течение всего интервала времени между последовательными заявками имеет место простейший поток обслуживаний с параметром $n\mu$. Таким образом,

$$q_{j,k} = \int_0^{\infty} \frac{(n\mu t)^{j+1-k}}{(j+1-k)!} e^{-n\mu t}\, dA(t), \qquad j, k \ge n. \tag{5.7.3}$$

Более сложных рассуждений требует случай $j \ge n, \quad k < n$ — здесь система, приняв очередное требование, в процессе обслуживания

переходит от режима полной занятости к работе с недогрузкой. Для перехода между состояниями $S_j \to S_k$ необходимо последовательное выполнение двух условий:

1) n каналами обслуживаются $j+1-n$ требований, после чего в системе исчезает очередь и остается ровно n требований;

2) за оставшееся до момента регенерации время обслуживаются ровно $n-k>0$ требований.

Распределение длительности первой фазы при параметре обслуживания $n\mu$ есть распределение Эрланга с плотностью

$$\psi(\tau) = \frac{n\mu(n\mu\tau)^{j-n}}{(j-n)\,!}e^{-n\mu\tau}.$$

При полной длительности интервала между требованиями t вероятность обслуживания $n-k$ требований за оставшееся время $t-\tau$

$$\varphi_k(t-\tau) = \left(\begin{array}{c} n \\ k \end{array}\right)\Big[e^{-\mu(t-\tau)}\Big]^k\Big[1-e^{-\mu(t-\tau)}\Big]^{n-k} = \left(\begin{array}{c} n \\ k \end{array}\right)e^{-k\mu(t-\tau)}\Big[1-e^{-\mu(t-\tau)}\Big]^{n-k}.$$

Окончательно для этого диапазона

$$\begin{aligned} q_{j,k} = & \int_0^\infty\Big[\int_0^t \varphi_k(t-\tau)\psi(\tau)\,d\tau\Big]\,dA(t) \\ = & \binom{n}{k}\int_0^\infty\Big\{\int_0^t e^{-k\mu(t-\tau)}\Big[1-e^{-\mu(t-\tau)}\Big]^{n-k}\,\frac{n\mu(n\mu\tau)^{j-n}}{(j-n)\,!}e^{-n\mu\tau}\,d\tau\Big\}\,dA(t) \\ = & \binom{n}{k}\int_0^\infty e^{-k\mu t}\Big\{\int_0^t\Big[1-e^{-\mu(t-\tau)}\Big]^{n-k}e^{k\mu\tau-n\mu\tau}\,\frac{n\mu(n\mu\tau)^{j-n}}{(j-n)\,!}\,d\tau\Big\}\,dA(t) \\ = & \binom{n}{k}\int_0^\infty e^{-k\mu t}\Big[\int_0^t(e^{-\mu\tau}-e^{-\mu t})^{n-k}\,\frac{(n\mu\tau)^{j-n}}{(j-n)\,!}n\mu\,d\tau\Big]\,dA(t), \\ & \qquad\qquad j\ge n,\quad k<n. \end{aligned} \tag{5.7.4}$$

Рассмотрим теперь уравнения (5.7.1) для состояний, соответствующих полностью загруженной системе $(k\ge n)$. Эти уравнения могут быть преобразованы к виду

$$\pi_k = \sum_{j=0}^\infty \pi_{j+k-1}\int_0^\infty \frac{(n\mu t)^j}{j\,!}e^{-n\mu t}\,dA(t), \qquad k=n,n+1,\ldots \tag{5.7.5}$$

Покажем, что в указанном диапазоне справедлива формула

$$\pi_k = C\omega^{k-n}, \qquad k = n-1, n, \ldots, \tag{5.7.6}$$

где ω — корень уравнения

$$\omega = \int_0^\infty e^{-n\mu t(1-\omega)}\, dA(t) \tag{5.7.7}$$

из (0,1), который существует при $n\mu a > 1$. Подставим $\{\pi_k\}$, определяемые формулой (5.7.6), в правую часть (5.7.5):

$$\begin{aligned}
\pi_k = &\sum_{j=0}^{\infty} C\omega^{j+k-1-n} \int_0^\infty \frac{(n\mu t)^j}{j\,!} e^{-n\mu t}\, dA(t) \\
= &\; C\omega^{k-1-n} \int_0^\infty \Big[\sum_{j=0}^{\infty} \frac{(n\mu\omega t)^j}{j\,!}\Big] e^{-n\mu t}\, dA(t) \\
= &\; C\omega^{k-1-n} \int_0^\infty e^{-n\mu t(1-\omega)}\, dA(t).
\end{aligned}$$

С помощью (5.7.7) убеждаемся, что $\pi_k = C\omega^{k-n}$, т. е. формулы (5.7.6) действительно имеют место. Из (5.7.6) при $k = n$ следует $C = \pi_n$, так что

$$\pi_k = \pi_n \omega^{k-n}, \qquad k = n-1, n, \ldots.$$

Таким образом, для $k \geq n$ значения $\{\pi_k\}$ могут быть найдены рекуррентно через π_n, и задача сводится к определению начальных вероятностей $\{\pi_k\}$.

Перепишем систему (5.7.1) для $k = \overline{1,n}$, выделив отдельно вероятности, которые выражаются через π_n и ω:

$$\pi_k = \sum_{j=k-1}^{n-1} q_{j,k}\pi_j + \sum_{j=n}^{\infty} q_{j,k}\pi_n \omega^{j-n}. \tag{5.7.8}$$

Вторая сумма после подстановки $\{q_{j,k}\}$ из (5.7.4) сводится к

$$\begin{aligned}
& n\mu\pi_n \int_0^\infty \int_0^t \binom{n}{k} e^{-k\mu t}(e^{-\mu\tau}-e^{-\mu t})^{n-k} \sum_{j=n}^\infty \frac{(n\mu\omega\tau)^{j-n}}{(j-n)\,!}\, d\tau\, dA(t) \\
= \;& n\mu\pi_n \int_0^\infty \left\{ \binom{n}{k} e^{-k\mu t} \sum_{i=0}^{n-k} (-1)^i \binom{n-k}{i} \int_0^t e^{-(n-k-i)\mu\tau} e^{-i\mu t} e^{n\mu\omega\tau}\, d\tau \right\} dA(t) \\
= \;& n\mu\pi_n \binom{n}{k} \sum_{i=0}^{n-k} (-1)^i \binom{n-k}{i} \int_0^\infty e^{-(k+i)\mu t} \left\{ \int_0^t e^{-\mu\tau(n-k-i-n\omega)}\, d\tau \right\} dA(t) \\
= \;& n\mu\pi_n \binom{n}{k} \sum_{i=0}^{n-k} (-1)^i \binom{n-k}{i} \int_0^\infty e^{-(k+i)\mu t} \frac{1-e^{-\mu t(n-k-i-n\omega)}}{\mu(n-k-i-n\omega)}\, dA(t) \\
= \;& n\pi_n \sum_{i=0}^{n-k} (-1)^i \frac{n\,!}{k\,!\,i\,!\,(n-k-i)\,!} \int_0^\infty \frac{e^{-(k+i)\mu t} - e^{-\mu t n(1-\omega)}}{n-k-i-n\omega}\, dA(t).
\end{aligned}$$

Положим

$$\beta_0^{(j)} = \int_0^\infty e^{-j\mu t}\, dA(t). \tag{5.7.9}$$

Расчет этого интеграла можно выполнить по формуле (1.9.3) при соответствующем пересчете параметра экспоненты. С учетом (5.7.7) и (5.7.9)

$$\sum_{j=k-1}^{n-1} q_{j,k}\pi_j = n\pi_n \sum_{i=0}^{n-k} (-1)^i \frac{n\,!}{k!\,i!\,(n-k-i)!}\, \frac{\beta_0^{(k+i)} - \omega}{n(1-\omega)-(k+i)}. \tag{5.7.10}$$

Условие нормировки вероятностей $\{\pi_j\}$ имеет вид

$$\sum_{j=0}^{n-1} \pi_j + \sum_{j=n}^\infty \pi_n \omega^{j-n} = \sum_{j=0}^{n-1} \pi_j + \pi_n/(1-\omega) = 1.$$

Упростим выражение для $q_{j,k}$ из первой суммы в (5.7.8). На основании (5.7.2) и (5.7.9)

$$\begin{aligned}
q_{j,k} = \;& \binom{j+1}{k} \int_0^\infty e^{-k\mu t}(1-e^{-\mu t})^{j+1-k}\, dA(t) \\
= \;& \sum_{i=0}^{j+1-k} (-1)^i \binom{j+1}{k} \binom{j+1-k}{i} \int_0^\infty e^{-\mu t(k+i)}\, dA(t) \\
= \;& \sum_{i=0}^{j+1-k} (-1)^i \frac{(j+1)\,!}{k\,!\,i\,!\,[j+1-(k+i)]\,!} \beta_0^{(k+i)}.
\end{aligned}$$

Теперь можно записать систему линейных алгебраических уравнений для нахождения начальных $\{\pi_j\}$:

$$\begin{aligned}
&\sum_{j=0}^{n-1} \pi_j + \pi_n/(1-\omega) &&= 1,\\
&\sum_{j=k-1}^{n-1} \pi_j \sum_{i=0}^{j+1-k} (-1)^i \frac{(j+1)!\,\beta_0^{(k+i)}}{k!\,i!\,[j+1-(k+i)]!} - \pi_k &&\\
&+n\pi_n \sum_{i=0}^{n-k} (-1)^i \frac{n!}{k!\,i!(n-k-i)!}\,\frac{\beta_0^{(k+i)}-\omega}{n(1-\omega)-(k+i)} &&= 0,\\
& && k=\overline{1,n}.
\end{aligned} \tag{5.7.11}$$

Отметим особо частный случай, когда $n\omega$ — целое число. Тогда при $k+i = n(1-\omega)$ вычислявшийся при выводе формулы (5.7.10) интеграл

$$\begin{aligned}
&\int_0^\infty e^{-(k+i)\mu t}\left\{\int_0^t e^{-\mu\tau(n-k-i-n\omega)}\,d\tau\right\} dA(t)\\
=\ &\int_0^\infty te^{-n\mu(1-\omega)t}\,dA(t) = \gamma/[n\mu(1-\omega)],
\end{aligned} \tag{5.7.12}$$

где γ можно рассматривать как коэффициент типа q_1 — см. формулу (2.1.10) — при параметре экспоненты $n\mu(1-\omega)$. Соответственно при $k+i = n(1-\omega)$ множитель $(\beta_0^{(k+i)}-\omega)/[n(1-\omega)-(k+i)]$ в последней строке формул (5.7.11) заменяется на $\gamma/[n(1-\omega)]$. Указанный случай встретился при отладке программы, реализующей метод Такача, на модели $M/M/n$ и потребовал доработки основного алгоритма.

Последующие финальные вероятности состояний вложенной цепи могут быть вычислены по формуле (5.7.6).

5.7.2. Распределение времени ожидания

Найдем распределение времени ожидания начала обслуживания. Прибывающая в систему $GI/M/n$ заявка с вероятностью $\Pi_0 = \sum_{i=0}^{n-1}\pi_i$ немедленно принимается на обслуживание. Если же она застает в системе $k \geq n$ заявок, то до ее выборки из очереди должны обслужиться $k-n+1$ заявок, причем параметр потока обслуживаний равен $n\mu$. Тогда ПЛС

искомого распределения

$$\varphi(s) = \sum_{k=n}^{\infty}\pi_k\Big(\frac{n\mu}{n\mu+s}\Big)^{k-n+1} = \pi_n\Big(\frac{n\mu}{n\mu+s}\Big)\sum_{k=0}^{\infty}\Big(\frac{n\mu\omega}{n\mu+s}\Big)^{k}$$
$$= \pi_n\frac{n\mu}{n\mu+s}\,\frac{1}{1-n\mu\omega/(n\mu\omega+s)} = \frac{n\mu\pi_n}{n\mu(1-\omega)+s}.$$

Выполняя обратное преобразование для имеющего дополнительный множитель $1/s$ изображения *функции* распределения, получаем

$$\Phi(t) = \frac{n\mu\pi_n}{-n\mu(1-\omega)}[e^{-n\mu(1-\omega)t}-1] = \frac{\pi_n}{1-\omega}[1-e^{-n\mu(1-\omega)t}].$$

Заметим, что множитель $\pi_n/(1-\omega)$ равен вероятности новой заявке застать систему занятой. Условная функция распределения ненулевого ожидания

$$\tilde{\Phi}(t) = 1-e^{-n\mu(1-\omega)t}.$$

Следовательно, упомянутое распределение является экспоненциальным независимо от вида распределения интервалов между смежными заявками и числа каналов (эти характеристики влияют только на значение параметра ω).

Начальные моменты безусловного распределения ожидания

$$w_k = \frac{\pi_n}{1-\omega}\,\frac{k\,!}{[n\mu(1-\omega)]^k}, \qquad k=1,2,\ldots \tag{5.7.13}$$

5.7.3. Стационарные вероятности состояний

Применение принципа баланса марковских и немарковских переходов между состояниями (см. рис. 5.4) к различным диапазонам значе-

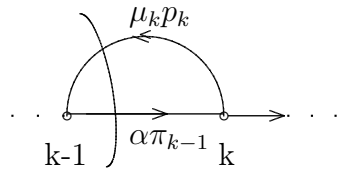

Рис. 5.4. Сохранение вероятностей в $GI/M/n$

ний k числа заявок в системе (с полной и неполной ее занятостью) позволяет вычислить стационарное распределение числа заявок в системе по формулам

$$p_k = \begin{cases} \pi_{k-1}/(k\mu a), & k = \overline{1,n}, \\ \pi_{k-1}/(n\mu a), & k = n, n+1, \ldots \end{cases} \tag{5.7.14}$$

Стационарную вероятность p_0 застать систему свободной можно определить из условия сохранения требований типа (3.1.2). Подставляя в него $\lambda = 1/a$ и $b = 1/\mu$, находим

$$p_0 = 1 - \Big[\sum_{k=1}^{n-1}(n-k)p_k + b/a\Big]\Big/n. \tag{5.7.15}$$

5.8. Система $GI/E_k/1$

Опишем несколько упрощенный вариант предложенного в [277] алгоритма расчета данной системы.

5.8.1. Вложенная цепь Маркова

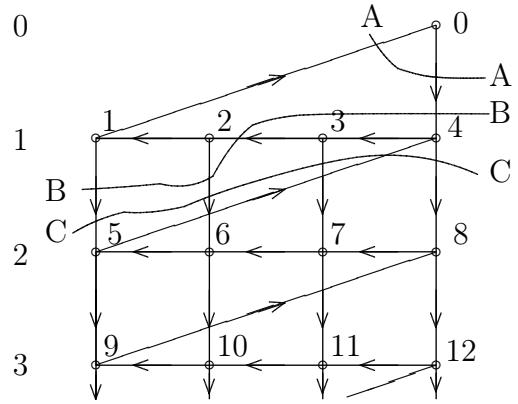

Рис. 5.5. Диаграмма переходов в системе $GI/E_4/1$

Построим вложенную цепь Маркова, описывающую число r непройденных фаз обслуживания в СМО вида $GI/E_k/1$ на моменты $\{\eta_n - 0\}$, непосредственно предшествующие поступлению очередной заявки

(см. рис. 5.5), и введем финальные вероятности состояний этой вложенной цепи $\pi_j = \lim_{n\to\infty} P\{r(\eta_n - 0) = j\}, \quad j = 0, 1, \ldots$ Обозначим вероятность обслуживания ровно j *фаз* за время между смежными моментами прибытия заявок

$$q_j = \int_0^\infty \frac{(\mu t)^j}{j\,!} e^{-\mu t}\, dA(t), \qquad j = 0, 1, \ldots \tag{5.8.1}$$

Здесь μ — параметр показательного распределения времени обслуживания каждой из k фаз.

Проводя рассуждения, аналогичные разд. 5.6, с учетом привнесения в систему каждой новой заявкой $k \geq 1$ дополнительных фаз, получаем систему уравнений, связывающих финальные вероятности вложенной цепи:

$$\begin{aligned} \pi_0 &= 1 - \sum_{j=0}^{\infty} \pi_j \sum_{m=0}^{j+k-1} q_m, \\ \pi_i &= \sum_{j=\max\{0,i-k\}}^{\infty} \pi_j q_{k-i+j}, \qquad i = 1, 2, \ldots \end{aligned} \tag{5.8.2}$$

Пусть $\omega_1, \omega_2, \ldots, \omega_k$ — корни уравнения

$$\omega^k = \int_0^\infty e^{-\mu\theta(1-\omega)t}\, dA(t), \tag{5.8.3}$$

лежащие внутри и на границе единичного круга (можно доказать, что таких корней при условии докритической загрузки системы имеется ровно k). Представим $\{\pi_i\}$ как линейные комбинации соответствующих степеней этих корней:

$$\pi_i = \sum_{m=1}^{k} \beta_m \omega_m^i, \qquad i = 0, 1, \ldots \tag{5.8.4}$$

Для $i \geq k$ подстановка (5.8.4) в уравнения системы (5.8.2) дает

$$\pi_i = \sum_{j=i-k}^{\infty} \sum_{m=1}^{k} \beta_m \omega_m^j q_{k-i+j} = \sum_{m=1}^{k} \beta_m \omega_m^{i-k} \sum_{j=0}^{\infty} \omega_m^j q_j = \sum_{m=1}^{k} \beta_m \omega_m^i,$$

так что в этом диапазоне (5.8.4) выполняется для произвольных наборов $\{\beta_m\}$.

Из системы (5.8.2) следует, что для $i = \overline{1, k-1}$ имеет место равенство

$$\pi_i = \sum_{j=0}^{\infty} \pi_j q_{k+j-1}. \tag{5.8.5}$$

С учетом (5.8.4) его правую часть можно преобразовать к виду

$$\begin{aligned} \sum_{j=0}^{\infty}\sum_{m=1}^{k} \beta_m \omega_m^j q_{k+j-i} &= \sum_{m=1}^{k} \beta_m \omega_m^{i-k} \sum_{j=0}^{\infty} \omega_m^{k+j-i} q_{k+j-i} \\ &= \sum_{m=1}^{k} \beta_m \omega_m^{i-k} \Big(\sum_{j=0}^{\infty} \omega_m^j q_j - \sum_{j=0}^{k-i-1} \omega_m^j q_j \Big). \end{aligned}$$

Но

$$\begin{aligned} \sum_{j=0}^{\infty} \omega_m^j q_j &= \sum_{j=0}^{\infty} \omega_m^j \int_0^{\infty} \frac{(\mu t)^j}{j!} e^{-\mu t}\, dA(t) \\ &= \int_0^{\infty} \sum_{j=0}^{\infty} \frac{(\mu \omega_m t)^j}{j!} e^{-\mu t}\, dA(t) = \int_0^{\infty} e^{-\mu t(1-\omega_m)t}\, dA(t) = \omega_m^k. \end{aligned} \tag{5.8.6}$$

Теперь на основании формул (5.8.4) – (5.8.6) можно записать

$$\sum_{m=1}^{k} \beta_m \omega_m^i = \sum_{m=1}^{k} \beta_m (\omega_m^i - \omega_m^{i-k} \sum_{j=0}^{k-1-i} \omega_m^j q_j), \qquad i = \overline{1, k-1},$$

откуда после приведения подобных членов получаем уравнения

$$\sum_{m=1}^{k} \beta_m \omega_m^{i-k} \sum_{j=0}^{k-1-i} \omega_m^j q_j = 0, \qquad i = \overline{1, k-1},$$

для определения весовых коэффициентов $\{\beta_m\}$. Еще одно уравнение

$$\sum_{m=1}^{k} \beta_m / (1 - \omega_m) = 1$$

следует из условия нормировки $\{\pi_i\}$. Таким образом, $\{\beta_m\}$ могут быть найдены решением системы линейных алгебраических уравнений

$$\begin{aligned} \sum_{m=1}^{k} \beta_m \omega_m^{i-k} \sum_{j=0}^{k-1-i} \omega_m^j q_j &= 0, \qquad i = \overline{1, k-1}, \\ \sum_{m=1}^{k} \beta_m / (1 - \omega_m) &= 1. \end{aligned} \tag{5.8.7}$$

5.8.2. Распределение времени ожидания

Распределение времени ожидания начала обслуживания заявкой, заставшей в системе на момент своего прибытия j фаз, будет j-кратной сверткой показательного закона. Следовательно, ПЛС плотности распределения времени ожидания

$$\begin{aligned}\varphi(s) &= \sum_{j=0}^{\infty}\Big(\frac{\mu}{\mu+s}\Big)^j\pi_j = \sum_{j=0}^{\infty}\Big(\frac{\mu}{\mu+s}\Big)^j\sum_{m=1}^{k}\beta_m\omega_m^j \\ &= \sum_{m=1}^{k}\beta_m\sum_{j=0}^{\infty}\Big(\frac{\mu\omega_m}{\mu+s}\Big)^j = \sum_{m=1}^{k}\beta_m\Big/\Big(1-\frac{\mu\omega_m}{\mu+s}\Big) \\ &= \sum_{m=1}^{k}\beta_m\Big[1+\frac{\omega_m}{1-\omega_m}\,\frac{\mu(1-\omega_m)}{\mu(1-\omega_m)+s}\Big].\end{aligned}$$

Вероятность нулевого ожидания $\Pi_0 = \sum_{m=1}^{k}\beta_m$. Распределение ненулевого ожидания является взвешенной суммой показательных законов, так что i-й начальный момент безусловного распределения времени ожидания

$$w_i = \sum_{m=1}^{k}\frac{\beta_m\omega_m}{1-\omega_m}\,\frac{i\,!}{[\mu(1-\omega_m)]^i}. \tag{5.8.8}$$

5.8.3. Стационарные вероятности состояний

Диаграмма рис. 5.5 переходов между состояниями системы позволяет записать соотношения баланса между $\{\pi_k\}$ и стационарными вероятностями состояний $\{\tilde{p}_k\}$ в форме

$$\mu\tilde{p}_i = \alpha\sum_{j=\max\{0,i-k\}}^{i-1}\pi_j, \qquad i=1,2,\ldots,$$

где интенсивность немарковских переходов $\alpha = 1/a$, т. е. обратна среднему интервалу между смежными заявками. Таким образом,

$$\tilde{p}_i = \frac{1}{\mu a}\sum_{j=\max\{0,i-k\}}^{i-1}\pi_j, \qquad i=1,2,\ldots. \tag{5.8.9}$$

Вероятность $\tilde{p}_0$ отсутствия фаз в системе совпадает с вероятностью ее свободного состояния и может быть вычислена по формуле типа (3.1.1):

$$p_0 = \tilde{p}_0 = 1 - k/(\mu a).$$

Вероятности $\{p_i\}$ нахождения в системе $i > 1$ заявок получим суммированием вероятностей фаз по соответствующим ярусам диаграммы рис. 5.5:

$$p_i = \sum_{j=(i-1)k+1}^{ik} \tilde{p}_j, \qquad i = 1, 2, \ldots$$

Подстановка в эту формулу значений $\{\tilde{p}_i\}$, найденных согласно (5.8.9), позволяет исключить из расчетной схемы распределение $\{\tilde{p}_i\}$ числа фаз. Окончательно система уравнений для $\{p_i\}$ принимает вид

$$\begin{aligned} p_0 &= \tilde{p}_0 = 1 - k/(\mu a), \\ p_1 &= (\mu a)^{-1} \sum_{i=0}^{k-1} (k-i)\pi_i, \\ p_j &= (\mu a)^{-1} \sum_{i=(j-2)k+1}^{jk-1} (k - |(j-1)k - i|)\pi_i, \qquad j = 2, 3, \ldots \end{aligned} \tag{5.8.10}$$

Рассмотренный алгоритм может быть применен для анализа системы с рекуррентным потоком $пачек$ заявок объема k и показательно распределенной длительностью обслуживания.

5.9. Система $E_k/G/1$

5.9.1. Вложенная цепь Маркова

Будем характеризовать состояние системы числом i находящихся в ней необслуженных фаз прибытия, $i = 0, 1, \ldots$. Поскольку немарковской компонентой процесса здесь является обслуживание, аналогично модели $M/G/1$ построим вложенную цепь Маркова по моментам, непосредственно следующим за окончанием обслуживания. Диаграмма переходов между состояниями системы представлена на рис. 5.6.

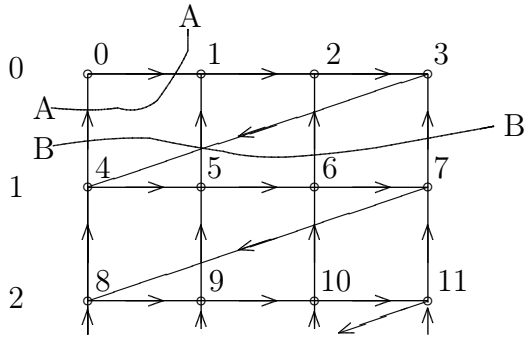

Рис. 5.6. Переходы для системы $E_4/G/1$

В очередной момент регенерации в системе будет i фаз, если в предыдущий такой момент

а) было $j < k$ фаз, пришло еще $k - j > 0$ (после чего началось обслуживание) и за время обслуживания прибыло ровно i фаз;

б) было $j \geq k$ фаз и прибыло $i - j + k$.

Финальные вероятности $\{\pi_i\}$ вложенной цепи связаны системой линейных уравнений

$$\pi_i = \sum_{j=k}^{i+k} \pi_j q_{i-j+k} + q_i \sum_{j=0}^{k-1} \pi_j, \quad i = 0, 1, \ldots, \tag{5.9.1}$$

где коэффициенты $\{q_i\}$ вновь подсчитываются согласно (2.1.10), но λ является параметром простейшего потока *фаз*.

Система (5.9.1) позволяет рекуррентно определить «старшие» $\{\pi_i\}$ через начальные:

$$\pi_i = (\pi_{i-k} - q_{i-k} \sum_{j=0}^{k-1} \pi_j - \sum_{j=k}^{i-1} \pi_j q_{i-j}) \Big/ q_0, \qquad i = k, k+1, \ldots \tag{5.9.2}$$

Умножая уравнения (5.9.1) на z^i и суммируя результаты, находим производящую функцию

$$\Pi(z) = \sum_{i=0}^{\infty} z^i \sum_{j=k}^{i+k} \pi_j q_{i+k-j} + \sum_{i=0}^{\infty} z^i q_i \sum_{j=0}^{k-1} \pi_j = S_1 + S_2. \tag{5.9.3}$$

Первая двойная сумма

$$\begin{aligned} S_1 &= \sum_{i=0}^{\infty} z^i \sum_{j=0}^{i} q_j \pi_{i+k-j} &&= \sum_{j=0}^{\infty} q_j \sum_{i=j}^{\infty} z^i \pi_{i+k-j} \\ &= z^{-k} \sum_{j=0}^{\infty} q_j z^j \sum_{i=k}^{\infty} z^i \pi_i &&= [\Pi(z) - \sum_{j=0}^{k-1} z^j \pi_j] Q(z)/z^k, \end{aligned}$$

а вторая $S_2 = Q(z) \sum_{j=0}^{k-1} \pi_j$. В обоих случаях через $Q(z)$ обозначена производящая функция $\{q_i\}$. Подставляя результаты в (5.9.3), находим

$$\Pi(z) = \frac{\sum_{j=0}^{k-1} \pi_j (z^j - z^k)}{Q(z) - z^k}. \tag{5.9.4}$$

Поскольку при $|z| \le 1$ производящая функция $\Pi(z)$ имеет вероятностный смысл (см. разд. 3.2), при всех таких z она должна оставаться ограниченной. Следовательно, лежащие в единичном круге корни знаменателя должны совпадать с корнями числителя. Тогда k неизвестных начальных вероятностей могут быть найдены из уравнений

$$\sum_{j=0}^{k-1} \pi_j (z_i^j - z_i^k) = 0, \tag{5.9.5}$$

где $\{z_i\}$ — корни знаменателя правой части (5.9.4), лежащие строго внутри единичного круга. Корень $z = 1$ должен быть исключен из рассмотрения, так как обращает (5.9.5) в тождество. Поскольку $Q(z) = q_0(\lambda(1-z))$, для расчета $\{z_i\}$ имеем уравнение

$$z^k - q_0(\lambda(1-z)) = 0, \tag{5.9.6}$$

причем q_0 следует вычислять согласно (1.9.3).

5.9.2. Стационарные вероятности состояний

Связь между финальным распределением числа фаз для вложенной цепи Маркова и соответствующим стационарным распределением следует из баланса переходов через разрезы, показанные на диаграмме рис. 5.6:

$$\begin{aligned} \alpha \sum_{i=0}^{j} \pi_i &= \lambda \tilde{p}_j, \qquad j = \overline{0, k-1}, \\ \alpha \sum_{i=j-k+1}^{j} \pi_i &= \lambda \tilde{p}_j, \qquad j = k, k+1, \ldots \end{aligned} \tag{5.9.7}$$

Здесь суммарная интенсивность потока немарковских переходов

$$\alpha = \Big(\sum_{i=0}^{k-1} \frac{k-i}{\lambda} \pi_i + b \Big)^{-1}. \tag{5.9.8}$$

Вероятность свободного состояния непосредственно получается из условия баланса заявок:

$$p_0 = \sum_{j=0}^{k-1} \tilde{p}_j = 1 - \lambda b/k. \tag{5.9.9}$$

Стационарное распределение числа *заявок* в системе устанавливается суммированием $\{\tilde{p}_j\}$ по ярусам диаграммы:

$$p_j = \frac{\alpha}{\lambda} \sum_{i=kj}^{k(j+1)-1} \sum_{m=i-k+1}^{i} \pi_m. \tag{5.9.10}$$

Подставляя в (5.9.9) выражения для $\{\tilde{p}_j\}$ из первой группы уравнений (5.9.7), находим

$$\sum_{j=0}^{k-1} \tilde{p}_j = \frac{\alpha}{\lambda} \sum_{j=0}^{k-1} \sum_{i=0}^{j} \pi_j = \frac{\alpha}{\lambda} \sum_{j=0}^{k-1} (k-j)\pi_j.$$

Теперь с учетом (5.9.8) можно переписать (5.9.9) в виде

$$\sum_{j=0}^{k-1} (k-j)\pi_j = k - \lambda b. \tag{5.9.11}$$

Уравнение (5.9.11) замыкает систему (5.9.5) для расчета начальных значений $\{\pi_j\}$. Итак, вероятности вложенной цепи с начальными индексами должны определяться из системы линейных алгебраических уравнений

$$\begin{aligned} \sum_{j=0}^{k-1} \pi_j (z_i^j - z_i^k) &= 0, \qquad i = \overline{0, k-2}, \\ \sum_{j=0}^{k-1} (k-j)\pi_j &= k - \lambda b. \end{aligned} \tag{5.9.12}$$

5.9.3. Временные характеристики

Для перехода к временны́м характеристикам системы нужно знать вероятности $\{\varphi_j\}$ застать в системе j заявок перед прибытием

новой. Эти вероятности пропорциональны стационарным вероятностям нахождения изображающей точки в последних микросостояниях ярусов. Из системы (5.9.7) следует, что

$$\varphi_j = C \sum_{i=jk}^{(j+1)k-1} \pi_i. \qquad (5.9.13)$$

Постоянная C определяется из условия нормировки.

Описанный алгоритм можно использовать для расчета системы с простейшим потоком заявок и произвольно распределенной длительностью обслуживания $пачек$ заявок фиксированного объема k.

5.10. Новые задачи

Отметим два важных направления в развитии $инструментария$ теории очередей: системы «с прогулками» и с повторными вызовами.

Под «прогулкой» обслуживающего устройства понимается его отключение по завершении обслуживания всех находящихся в системе заявок. Во время «прогулки» каналом может пройти профилактическое обслуживание или переключиться на выполнение других работ, для которых «прогулкой» является обслуживание первоначально обсуждавшейся очереди. Здесь приходится рассматривать системы «с разогревом», в который входит обслуживание накопившихся за время прогулки заявок. Очевидна возможность итерационного расчета основной и дополняющей подсистем данной задачи и использование в нем ранее обсуждавшейся модели с порогом включения и разогревом.

Системы $с\ повторными\ вызовами$ [28, 233] адекватно описывают процессы передачи информации в интенсивно развиваемых мобильных сотовых сетях связи и локальных вычислительных сетях. Здесь заявки, получившие отказ, «остаются на орбите» системы и время от времени повторяют запрос на обслуживание. В типичном варианте успех (выбор в канал) достигается с вероятностью q; с дополнительной вероятностью $1-q$ заявка возвращается в очередь — бернуллиева обратная связь. Можно показать, что среднее число потребных попыток равно $1/v$.

Глава 6

Многоканальные системы с детерминированным обслуживанием

6.1. Система $M/D/n$

Положим

$$q_k = \frac{(\lambda b)^k}{k\,!} e^{-\lambda b}, \qquad k = 0, 1, \ldots,$$

где b — время обслуживания одной заявки, а λ — интенсивность простейшего входящего потока, и обозначим через

$$u = \sum_{i=0}^{n} p_i$$

вероятность отсутствия очереди. Следуя Кроммелину (см. [189]), будем рассматривать вероятности состояний системы в моменты времени, отстоящие друг от друга на b. Очевидно, для стационарных вероятностей должны выполняться равенства

$$\begin{array}{rcl} p_0 & = & uq_0, \\ p_1 & = & uq_1 + p_{n+1}q_0, \\ \ldots\ldots & \ldots\ldots & \ldots\ldots\ldots\ldots \\ p_k & = & uq_k + p_{n+1}q_{k-1} + \cdots + p_{n+k}q_0, \\ \ldots\ldots & \ldots\ldots & \ldots\ldots\ldots\ldots \end{array} \tag{6.1.1}$$

Из записи общего вида уравнений системы (6.1.1)

$$p_k = uq_k + \sum_{j=1}^{k} p_{n+j}q_{k-j} = \sum_{j=0}^{k-1} p_{n+k-j}q_j, \qquad k = 0, 1, \ldots,$$

следует формула рекуррентного расчета $\{p_k\}$:

$$p_k = (p_{k-n} - uq_{k-n} - \sum_{j=1}^{k-1-n} q_j p_{k-j})\Big/ q_0, \qquad k = n+1, n+2, \ldots \quad (6.1.2)$$

Первое из уравнений (6.1.1) с учетом выражения для u дает

$$p_n = p_0/q_0 - \sum_{j=1}^{n-1} p_j. \quad (6.1.3)$$

Таким образом, определение $\{p_k\}$, $k = \overline{0, n-1}$, является необходимым условием рекуррентного расчета старших вероятностей.

Умножим уравнения системы (6.1.1) для p_k на z^k, $k = 0, 1, \ldots$, и просуммируем результаты. Обозначая $P(z)$ и $Q(z)$ производящие функции $\{p_k\}$ и $\{q_k\}$ соответственно, получаем

$$P(z) = uQ(z) + \sum_{k=1}^{\infty} z^k \sum_{j=1}^{k} p_{n+j}q_{k-j}. \quad (6.1.4)$$

Обозначим через S двойную сумму в уравнении (6.1.4) и изменим порядок суммирования. Теперь

$$S = \sum_{j=1}^{\infty}\sum_{k=j}^{\infty} z^k p_{n+j}q_{k-j} = \sum_{j=1}^{\infty} p_{n+j}z^j \sum_{k=0}^{\infty} z^k q^k = \frac{Q(z)}{z^n}\sum_{j=1}^{\infty} p_{n+j}z^{n+j}.$$

Последняя сумма равна $P(z) - \sum\limits_{j=0}^{n} p_j z^j$, и мы можем переписать (6.1.4) в виде

$$P(z) = Q(z)\Big\{u + z^{-n}\Big[P(z) - \sum_{j=0}^{n} p_j z^j\Big]\Big\},$$

откуда следует выражение для производящей функции

$$P(z) = \frac{\sum\limits_{j=0}^{n}(z^j - z^n)p_j}{1 - z^n/Q(z)}. \quad (6.1.5)$$

Последнее слагаемое в числителе может быть опущено, а

$$Q(z) = \sum_{k=0}^{\infty} \frac{(\lambda b z)^k}{k\,!} e^{-\lambda b} = e^{-\lambda b(1-z)}.$$

Итак,

$$P(z) = \frac{\sum\limits_{j=0}^{n-1}(z^j - z^n)p_j}{1 - z^n e^{\lambda b\,(1-z)}}. \tag{6.1.6}$$

Неизвестные n начальных вероятностей должны определяться аналогично разд. 5.9. Применяемый ниже метод основан на известной теореме Руше (подробная аргументация приводится в [18, с. 321]). Приравняем знаменатель (6.1.6) нулю и перепишем соответствующее уравнение в форме

$$z^n e^{\lambda b(1-z)} = e^{2m\pi i}, \qquad m = 0, 1, \ldots \tag{6.1.7}$$

Тогда корень его, соответствующий конкретному значению m, может быть найден по итерационной формуле

$$z = \exp\Big\{\frac{2m\pi i}{n} - \frac{\lambda b}{n}(1 - z)\Big\}, \tag{6.1.8}$$

обеспечивающей сходящийся итерационный процесс. В качестве начального приближения можно выбирать, например, $z_m^{(0)} = 0.5e^{2m\pi i/n}$.

Уравнение (6.1.7) имеет ровно n различных корней. В частности, $m = 0$ дает корень $z = 1$, подстановка которого в числитель правой части (6.1.6) приводит к тождественному нулю. Можно показать, что если z_m — корень уравнения (6.1.7), то комплексно сопряженное с ним число $\bar{z}_m = z_{n-m}$ также является корнем этого уравнения. Следовательно, практически уравнение (6.1.8) нужно решать для $m = \overline{1, [(n-1)/2]}$. Подстановка каждого комплексного корня в числитель правой части (6.1.6) дает два уравнения:

$$\begin{aligned} \sum_{j=0}^{n-1} p_j(\Re(z_m^j - \Re(z_m^n)) &= 0, \\ \sum_{j=0}^{n-1} p_j(\Im(z_m^j - \Im(z_m^n)) &= 0. \end{aligned} \tag{6.1.9}$$

При четных n корень, соответствующий $m = n/2$, лежит на действительной оси и имеет фазу $\varphi = \pi$. Его подстановка в числитель

(6.1.6) дает одно уравнение типа первого уравнения системы (6.1.9). Еще одно уравнение (единственное с ненулевой правой частью) можно получить из условия (3.1.2) баланса пришедших и обслуженных заявок.

После решения полученной системы относительно начальных $\{p_j\}$ применением формулы (6.1.3) находим p_n и далее через (6.1.2) — требуемое число старших вероятностей.

6.2. Система $E_k/D/n$

А. В. Быкадоров в статье [22] применил подход Кроммелина к системе $E_k/D/n$ — с эрланговским входящим потоком. Развивая его результаты, можно получить для этой задачи расчетную схему, обобщающую описанную в предыдущем разделе.

Введем вероятности

$$q_j = \frac{(\lambda b)^j}{j\,!}e^{-\lambda b}, \qquad j = 0, 1, \ldots, \tag{6.2.1}$$

прибытия ровно j *фаз* поступления заявки за время b обслуживания одной заявки, где λ — параметр распределения Эрланга

$$a(t) = \frac{\lambda(\lambda t)^{k-1}}{(k-1)\,!}e^{-\lambda t} \tag{6.2.2}$$

интервалов между смежными заявками. Обозначим w_i вероятность иметь в системе ровно i фаз заявок и положим

$$W_{n,i} = \sum_{j=0}^{n} w_{jk+i}, \qquad i = \overline{0, k-1}. \tag{6.2.3}$$

Это вероятности иметь в системе не более n заявок и i фаз поступления очередной. Тогда вероятности $\{w_i\}$ для отстоящих на b один от другого моментов времени оказываются связанными системой линейных алгебраических уравнений

$$w_i = \sum_{m=0}^{i} W_{n,m}q_{i-m}, \qquad i = \overline{0, k-1}, \tag{6.2.4}$$

$$w_i = \sum_{m=0}^{k-1} W_{n,m}q_{i-m} + \sum_{m=0}^{i-k} q_m w_{nk+i-m}, \qquad i \geq k. \tag{6.2.5}$$

Перепишем формулы (6.2.4) с учетом (6.2.3), выделив отдельно слагаемые $\{w_{nk+i}\}$:

$$w_i = \sum_{m=0}^{i}\Big(\sum_{j=0}^{n-1} w_{jk+m} + w_{nk+m}\Big)q_{i-m}.$$

Тогда

$$w_i = \sum_{m=0}^{i} q_{i-m}\sum_{j=0}^{n-1} w_{jk+m} + \sum_{m=0}^{i} q_{i-m}w_{nk+m}, \qquad i = \overline{0,k-1}, \qquad (6.2.6)$$

откуда следует возможность рекуррентного определения

$$w_{nk+i} = \Big(w_i - \sum_{m=0}^{i} q_{i-m}\sum_{j=0}^{n-1} w_{jk+m} - \sum_{m=0}^{i-1} q_{i-m}w_{nk+m}\Big)\Big/ q_0, \qquad i = \overline{0,k-1}. \qquad (6.2.7)$$

Уравнение (6.2.5) также позволяет рекуррентно рассчитывать $\{w_{nk+i}\}$ через вероятности с меньшими значениями индексов. Выделим в правой части (6.2.5) слагаемое второй суммы для $m = 0$:

$$w_i = \sum_{m=0}^{k-1} W_{n,m}q_{i-m} + \sum_{m=1}^{i-k} q_m w_{nk+i-m} + w_{nk+i}q_0.$$

Тогда

$$w_{nk+i} = \Big(w_i - \sum_{m=0}^{k-1} W_{n,m}q_{i-m} - \sum_{m=1}^{i-k} q_m w_{nk+i-m}\Big)\Big/ q_0, \qquad i = k, k+1, \ldots \qquad (6.2.8)$$

Таким образом, последующие вероятности состояний могут быть выражены через $\{w_j\}$ для $j = \overline{0, nk-1}$, и задача расчета распределения числа фаз сводится к нахождению начальных вероятностей.

Получим производящую функцию $W(z)$ распределения числа фаз в системе. Умножив уравнения (6.2.4) и (6.2.5) на соответствующие степени z и сложив результаты, имеем

$$\begin{aligned} W(z) &= \sum_{i=0}^{k-1} z^i \sum_{m=0}^{i} W_{n,m}q_{i-m} \\ &+ \sum_{i=k}^{\infty} z^i\Big(\sum_{m=0}^{k-1} W_{n,m}q_{i-m} + \sum_{m=0}^{i-k} q_m w_{nk+i-m}\Big). \end{aligned}$$

Меняя порядок суммирования, выводим

$$\begin{aligned} W(z) &= \sum_{m=0}^{k-1} W_{n,m} z^m \sum_{i=m}^{\infty} z^{i-m} q_{i-m} + \sum_{m=0}^{\infty} q_m \sum_{i=k+m}^{\infty} z^i w_{nk+i-m} \\ &= Q(z) \sum_{m=0}^{k-1} W_{n,m} z^m + \sum_{m=0}^{\infty} q_m z^{m-nk} \sum_{i=k+m}^{\infty} z^{nk+i-m} w_{nk+i-m}, \end{aligned} \tag{6.2.9}$$

где снова

$$Q(z) = \sum_{i=0}^{\infty} z^i q_i = e^{-\lambda b(1-z)}.$$

Входящая в правую часть (6.2.9) двойная сумма

$$S_2 = \frac{1}{z^{nk}} \sum_{m=0}^{\infty} q_m z^m \sum_{i=(n+1)k}^{\infty} z^i w_i = z^{-nk} Q(z) \Big[W(z) - \sum_{i=0}^{(n+1)k-1} z^i w_i \Big].$$

Первая сумма из того же равенства (6.2.9)

$$S_1 = \sum_{m=0}^{k-1} W_{n,m} z^m = \sum_{m=0}^{k-1} z^m \sum_{j=0}^{n} w_{jk+m}.$$

Подставив эти результаты в (6.2.9) и приведя подобные члены, приходим к выражению для производящей функции распределения числа фаз

$$W(z) = \frac{\sum_{j=0}^{n-1} \sum_{m=0}^{k-1} (z^{nk} - z^{jk}) z^m w_{jk+m}}{1 - z^{nk} e^{-\lambda b(1-z)}}. \tag{6.2.10}$$

Как и в разд. 6.1, подстановка каждого комплексного корня знаменателя в числитель даст два уравнения для отыскания начальных вероятностей $\{w_j\}$:

$$\begin{aligned} \sum_{j=0}^{n-1} \sum_{m=0}^{k-1} (\Re(z^{nk+m}) - \Re(z^{jk+m})) w_{jk+m} &= 0, \\ \sum_{j=0}^{n-1} \sum_{m=0}^{k-1} (\Im(z^{nk+m}) - \Im(z^{jk+m})) w_{jk+m} &= 0, \end{aligned} \tag{6.2.11}$$

а подстановка действительного корня — одно. Сами корни вычисляются аналогично разд. 6.1 с заменой n на произведение nk. Замыкающее систему условие нормировки типа (3.1.2) после подстановки в левую

часть вместо распределения числа заявок распределения числа фаз прибытия на основе

$$p_j = \sum_{m=jk}^{(j+1)k-1} w_m \tag{6.2.12}$$

и в правую часть — интенсивности потока заявок λ/k принимает вид

$$\sum_{j=0}^{n-1}(n-j)\sum_{m=jk}^{(j+1)k-1} w_m = n - \lambda b/k. \tag{6.2.13}$$

Совместное решение систем уравнений (6.2.11) и (6.2.13) дает исходные вероятности $\{w_j\}$, $j = \overline{0, kn-1}$. Последующие k вероятностей получим согласно (6.2.7), после чего по формуле (6.2.8) досчитаем требуемое число вероятностей $\{w_j\}$ для $j \geq k(n-1)$. Распределение $\{p_j\}$ числа заявок в системе можно найти с помощью (6.2.12).

На рис. 6.1 показана зависимость распределения числа заявок в системе $E_k/D/n$ от порядка k входящего потока Эрланга.

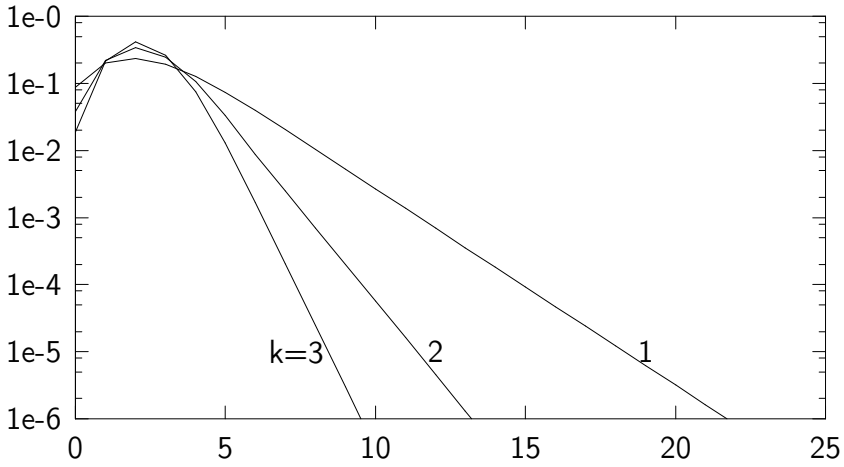

Рис. 6.1. Распределение заявок в $E_k/D/3$

6.3. Бесконечное число каналов

Увеличение числа каналов существенно увеличивает трудоемкость и снижает устойчивость методов, обсуждаемых в главах 5–7. В таких случаях разумной аппроксимацией является модель $M/G/\infty$. Этот

материал по своей логической значимости заслуживает отдельной главы, но ввиду крайне малого объема замыкает собой данную.

В [68] показано, что распределение числа заявок в системе $M/G/\infty$ подчинено закону Пуассона

$$p_k = \frac{\rho^k}{k!} e^{-\rho}, \qquad k = 0, 1, \ldots, \tag{6.3.1}$$

где $\rho = \lambda b$. Время ожидания в такой системе тождественно равно нулю (дисциплина очереди IS — Immediate Servicing). Распределение времени пребывания заявки в системе совпадает с распределением времени чистой длительности обслуживания.

Глава 7

Многоканальные системы с распределениями фазового типа

Вступление человечества в эру информационных технологий определяет растущий интерес к методам проектирования и оценки эффективности систем обработки и передачи данных. Особенность текущего момента — приближение технологии производства СБИС, составляющих основу современных ЭВМ, к фундаментальным физическим ограничениям. По этой причине требуемые показатели производительности достигаются созданием многопроцессорных и многомашинных систем, возможности расчета производительности которых современными методами ТМО исследованы недостаточно. Весьма широкий круг задач этого класса может быть решен на основе *фазовых аппроксимаций* исходных распределений.

7.1. Многофазное представление сложных СМО

Эффективным методом марковизации сложных СМО является многофазное представление составляющих распределений. Каждая комбинация распределений (и их порядков) порождает специфическую диаграмму, которая может быть представлена различными способами. Разумеется, при изображении диаграмм порядки распределений и число

каналов приходится конкретизировать.

На рис. 7.1 приведена простейшая диаграмма переходов — для системы $H_2/M/3$.

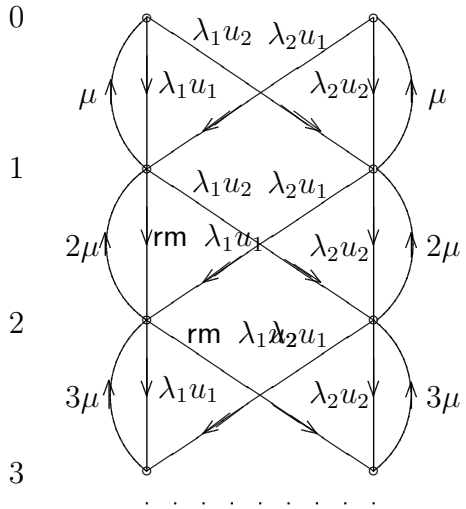

Рис. 7.1. Переходы в системе $H_2/M/3$

Эта диаграмма состоит из двух ветвей, каждая со своим значением λ_i интенсивности простейшего входящего потока. Вероятности $\{u_i\}$ выбора ветвей не зависят от исходной ветви. Множитель при интенсивности обслуживания μ не зависит от номера ветви и ограничивается числом каналов (нижележащая часть диаграммы на рисунке не показана).

На диаграмме переходов для $M/E_3/2$ (рис. 7.2) состояние системы идентифицируется полным количеством заявок в ней (номер яруса) и распределением проходящих обслуживание заявок по его фазам («ключ» соответствующего микросостояния). Здесь и на последующих диаграммах сумма значений позиций ключа равна числу задействованных каналов обслуживания.

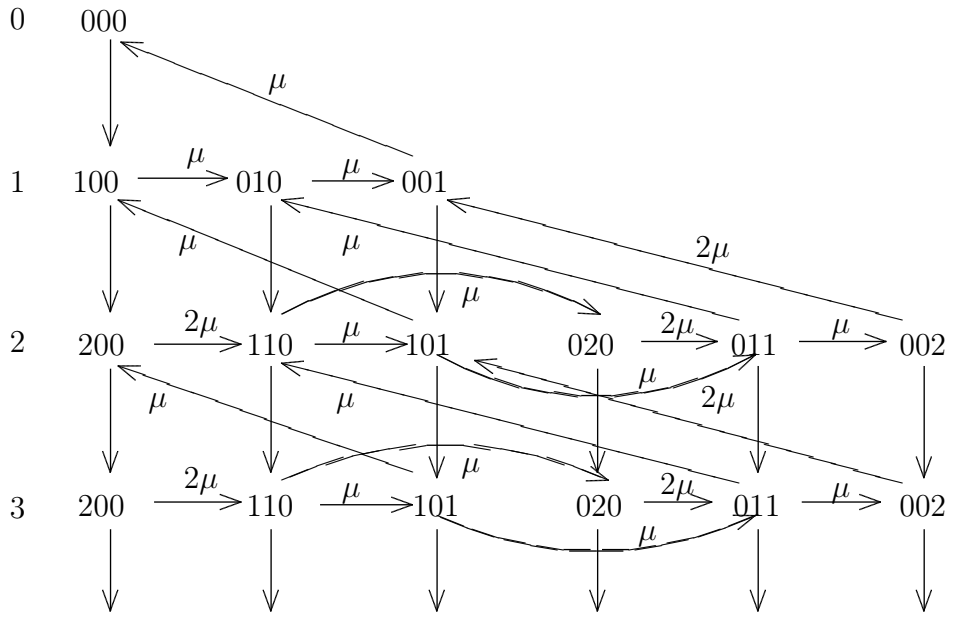

Рис. 7.2. Переходы в системе $M/E_3/2$

При гиперэрланговой аппроксимации распределения обслуживания обсуждавшегося в разд. 1.8.6 типа вновь принимаемая на обслуживание (извне или из ранее накопившейся очереди) заявка с вероятностью y попадает в первую фазу, а с вероятностью $\bar{y}$ — сразу во вторую. Это обстоятельство существенно усложняет диаграммы и вынуждает изобразить их в более крупном масштабе (рис. 7.3, 7.4). Диаграммы переходов по прибытию заявки и по завершению фазы обслуживания для наглядности разделены.

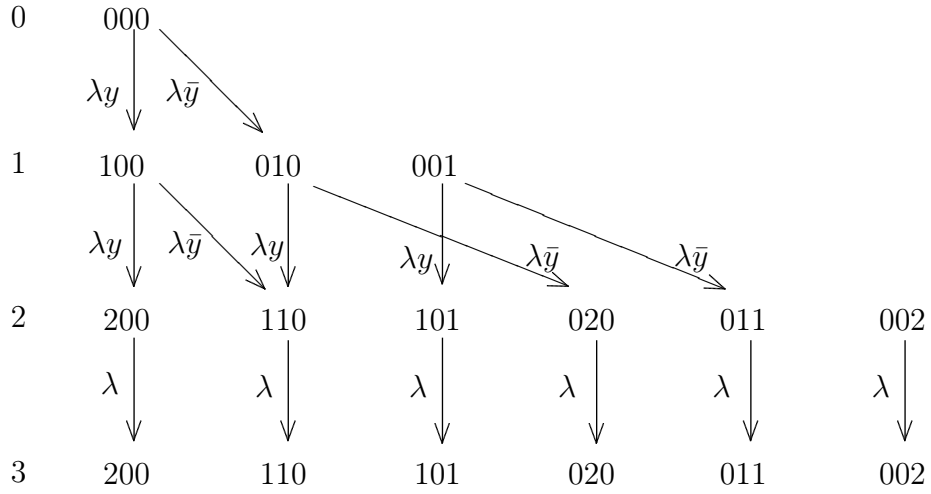

Рис. 7.3. Переходы по прибытию заявки в системе $M/P_3/2$

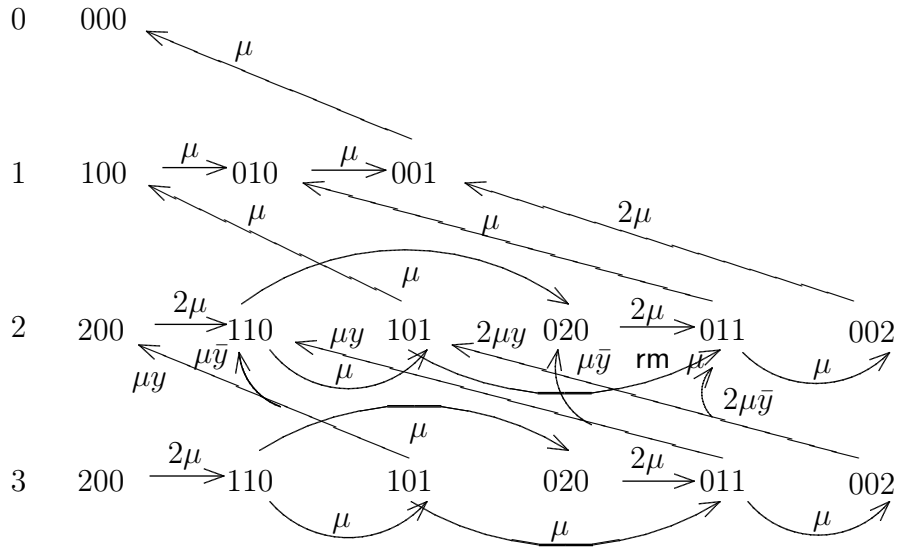

Рис. 7.4. Переходы по завершению обслуживания в системе $M/P_3/2$

Работа системы $M/H_2/3$ может быть интерпретирована как процесс обслуживания неоднородного потока заявок [94], причем выбор типа заявки определяет параметр показательно распределенного обслуживания. Теперь ключ микросостояния указывает количество находящихся в каналах обслуживания заявок каждого типа (рис. 7.5, 7.6). Суммарный входящий поток имеет интенсивность λ; прибывающая (или выбираемая из очереди) заявка с вероятностью y_i относится к i-му типу.

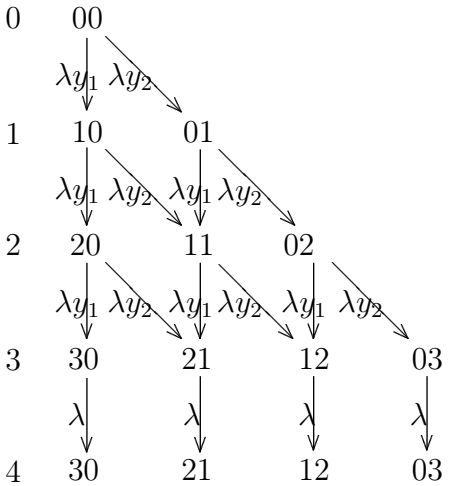

Рис. 7.5. По прибытию

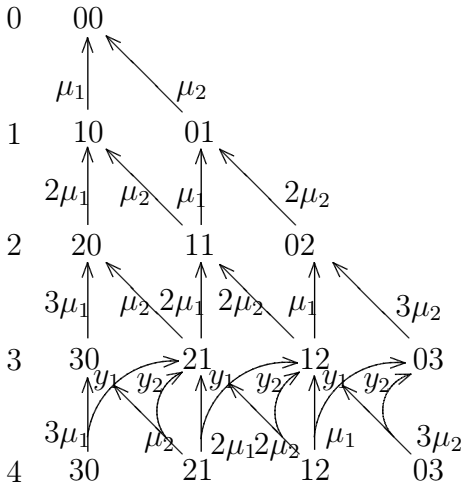

Рис. 7.6. По обслуживанию

На правом рисунке при $j > n$ параметр потока обслуживаний заявок i-го типа равен $m_i\mu_i$, где m_i — содержимое i-й позиции ключа. Завершение обслуживания с вероятностями $\{y_i\}$ в зависимости от типа выбранной из очереди заявки приводит в одно из микросостояний вышележащего яруса.

С учетом этих диаграмм легко представить (но трудно нарисовать с сохранением наглядности) диаграмму переходов для более общей системы $H_k/H_k/n$, в которой ДФР составляющих распределений $\bar{A}(t) = \sum\limits_{i=1}^{k} u_i e^{-\lambda_i t}$, $\bar{B}(t) = \sum\limits_{i=1}^{k} y_i e^{-\mu_i t}$.

Диаграмма может, например, состоять из слоев, каждый из которых соответствует определенному числу заявок в системе. Каждый слой включает k ярусов. Все ярусы одного слоя имеют одинаковый набор «ключей», аналогичный случаю $M/H_k/n$. Переходы по завершению обслуживания связывают пары ключей смежных слоев так же, как и

в модели $M/H_k/n$ — см. рис. 7.6, но переводят изображающую точку ровно на k ярусов вверх. Переходы по прибытию заявки строятся аналогично рис. 7.5, но интенсивность ухода изображающей точки вниз с i-го яруса любого слоя, $i = \overline{1,k}$, равна λ_i, а вероятность ее попадания на j-й ярус нижележащего слоя, $j = \overline{1,k}$, равна u_j.

Диаграмма переходов в модели $H_k/E_q/n$ будет отличаться от вышеописанной лишь тем, что в ней переходы вверх будут строиться по аналогии с рис. 7.2.

Далее, в случае $E_q/E_k/n$ количество ярусов в слое определяется числом q фаз поступления заявки. Здесь переходы вверх также аналогичны рис. 7.2, но ведут на q ярусов вверх. Прохождение фазы прибытия заявки перемещает изображающую точку строго вниз на один ярус, причем окончательное прибытие заявки (и изменение ключа микросостояния при номере исходного слоя $j < n$) фиксируется только при переходе в следующий слой.

Наконец, в модели $E_q/H_k/n$ переходы вверх идут аналогично рис. 7.6, но на q ярусов вверх. Переходы вниз в пределах слоя ключ микросостояния не меняют, а переход в следующий слой аналогичен рис. 7.5 с интерпретацией λ как интенсивности потока фаз прибытия заявок.

Подобные рассуждения могут быть проведены и для других комбинаций фазовых распределений. Достоинством *коксовых* аппроксимаций является естественное включение в них как частных случаев всех эрланговых (в том числе и показательных) моделей, а недостатками — трудность обобщения на большее число составляющих и усложнение диаграмм переходов. Для практических целей применение распределения C_2, обеспечивающего выравнивание трех начальных моментов, представляется вполне достаточным.

Будем представлять состояние системы $C_2/C_2/n$ фазой прибытия очередной заявки и распределением находящихся в каналах заявок по фазам обслуживания («ключами» соответствующих микросостояний). Диаграмма состояний в зависимости от фазы прибытия разделяется на две ветви, а по числу заявок в системе — на ярусы.

На рис. 7.7 приведена диаграмма переходов в системе $M/C_2/3$ по завершению фазы обслуживания.

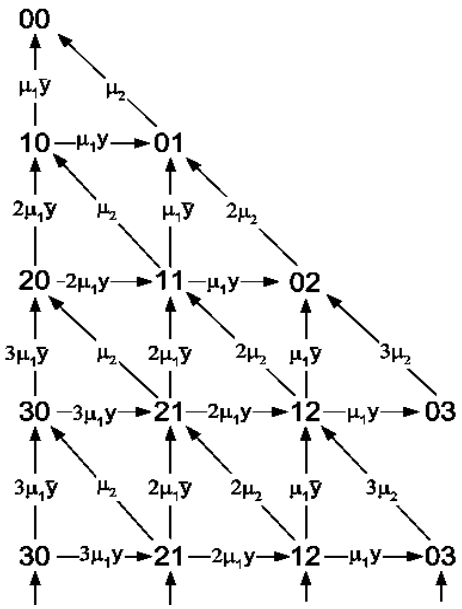

Рис. 7.7. Переходы по завершению фазы C_2-обслуживания

Здесь окончание первой фазы с интенсивностью $y\mu_1$ переводит заявку во вторую фазу обслуживания, а с интенсивностью $\bar{y}\mu_1$ — завершает обслуживание и поднимает изображающую точку на вышележащий ярус. Последнее всегда происходит при завершении второй фазы (интенсивность μ_2). На рис. 7.7 интенсивности переходов проставлены с учетом числа заявок, находящихся в каждой фазе исходного микросостояния. В случае C_2-потока эти переходы $не\ меняют$ фазы прибытия, так что диаграммы переходов данного вида для обеих ветвей идентичны и совпадают с только что рассмотренной.

Диаграмма переходов по фазам прибытия заявок оказывается весьма запутанной; поэтому мы ограничимся описанием логики ее построения. Если интервалы между заявками подчинены C_2-распределению с параметрами $\{u, \lambda_1, \lambda_2\}$, то все переходы из первой фазы (ветви) с интенсивностями $u\lambda_1$ приводят в аналогичное микросостояние второй фазы прибытия того же яруса, а с интенсивностями $\bar{u}\lambda_1$ — в первую фазу нижележащего яруса. Переходы по завершению второй (всегда последней) фазы имеют интенсивность λ_2 и также приводят в первую фазу нижележащего яруса.

7.2. К расчету переходных матриц

Расчет многофазных систем состоит в решении уравнений баланса переходов для вероятностей микросостояний, в которые входят матрицы интенсивностей переходов между последними. Зависимость структуры этих матриц от типа и порядка аппроксимирующих фазовых распределений ставит ценность программной реализации расчетной схемы в прямую зависимость от возможности автоматического построения матриц переходов. Эту проблему можно решить следующим образом:

а) для каждого j-го яруса системы, $j = \overline{0,n}$, автоматически сформировать последовательность ключей микросостояний;

б) формировать матрицы переходов, сопоставляя «ключи-источники» j-го яруса и совместимые с ними по выбранному типу переходов ключи-«преемники»: для матрицы B на $(j-1)$-м ярусе, для матрицы C — на j-м, для A — на $(j+1)$-м ярусе.

Для уменьшения объема занимаемой памяти целесообразно располагать ключи в такой последовательности, чтобы матрица C была верхней треугольной. Это достигается размещением ключей в пределах яруса по их убыванию. В частных случаях возможны дополнительные средства экономии памяти, применение которых расширяет возможности соответствующих программ [92].

Реализация метода фиктивных фаз оказалась значительно сложнее, чем это представляется при ознакомлении с идеей метода. Дело в том, что метод предполагает работу с *переменным числом* матриц перехода *изменяемой* размерности, определяемой в процессе счета. Его первая реализация отрабатывалась на ЭВМ М-220 с объемом оперативной памяти всего в 4096 слов. Соответственно пришлось:

- упаковывать матрицы каждого вида (и их половинки для треугольных матриц) в *линейные* массивы;
- выделять под эти объекты динамическую (`[ALLOCATABLE]`) память;
- фиксировать отдельно их координаты в линейных структурах и размерности;
- явно переадресовывать операнды матричных операций.

Описанные трудности многократно усугубляются для моделей со входящим потоком, отличным от простейшего. При нынешних технических возможностях можно избежать упомянутой упаковки, задав каждую матрицу как сечение трехмерного массива с максимально необходимым размером.

Основой автоматического построения переходных матриц является процедура KEY, которая генерирует лексикографически убывающую последовательность ключей микросостояний систем. Первый ключ j-го яруса для $j \leq n$ имеет вид {J,0,...,0}. Для получения остальных ключей:

- снимается копия с предыдущего;
- отыскивается крайняя справа (кроме первой) ненулевая позиция M;
- значение ключа в ней уменьшается на единицу;
- подсчитывается сумма в позициях от 1 до M; весь остаток заносится в (M+1)-ю позицию, а стоящие правее нее обнуляются.

Сами матрицы интенсивностей инфинитезимальных переходов формируются на основе сопоставления ключей исходного и результирующего ярусов в соответствии с логикой переходов данного вида (прибытие заявки, завершение промежуточной или конечной фазы обслуживания).

7.3. Эрланг и гиперэкспонента

В ряде источников [28, 164, 233] фигурирует «фольклорная» рекомендация: при коэффициенте вариации распределения $v > 1$ применять H_2-аппроксимацию, а в противном случае — эрлангову. Заметим, что ширина диаграмм (максимальное количество микросостояний на ярусе) для моделей с эрланговым обслуживанием быстро растет по числу каналов n и порядку k распределения обслуживания согласно формуле $\binom{n+k-1}{n}$ [265] — см. табл. 7.1.

Таблица 7.1. Количество микросостояний на ярусах системы $M/E_k/n$

Число каналов n	Число фаз обслуживания k				
	2	3	4	5	6
2	3	6	10	15	21
3	4	10	20	35	56
5	6	21	56	126	252
10	11	66	286	1001	3003
20	21	231	1771	10626	53130
30	31	496	5456	46376	324632

Эрланговы распределения позволяют строго выравнивать первый и лишь приближенно — второй моменты распределения обслуживания. Коэффициент вариации эрлангова распределения порядка k составляет $v = 1/\sqrt{k} < 1$. Для наибольшего из включенных в таблицу 7.1 значения $k = 6$ он равен 0.408. При обсчете моделей с мéньшими коэффициентами вариации могут потребоваться значительно большие значения k. С другой стороны, аппроксимация H_2 позволяет выравнивать три момента произвольного (исключая E_2) распределения, что представляется необходимым и достаточным. Даже «патологический характер» H_2-аппроксимации при $1/\sqrt{k} < v < 1$ (одна из вероятностей выбора фазы отрицательна, а другая больше единицы) и комплекснозначные параметры при $v < 1/\sqrt{2}$ на завершающих этапах расчета приводят к вполне осмысленным результатам, согласующимся с найденными другими методами. Диаграмма переходов для $M/H_2/n$ имеет ширину $n + 1$, что позволяет обсчитывать системы с достаточно большим числом каналов. По указанным причинам распределения времени обслуживания целесообразно представлять гиперэкспонентой H_2.

7.4. Уравнения глобального баланса при простейшем входящем потоке

Основная задача данной главы — получение стационарных вероятностей микросостояний, представленных на соответствующих диаграммах разд. 7.1. Для удобства сопоставления различных методов ее решения дадим формальную постановку задачи ограниченной общности — с простейшим входящим потоком, когда все микросостояния с фиксированным числом заявок находятся на одном ярусе диаграммы.

Обозначим через S_j множество всех возможных микросостояний системы, при которых на обслуживании находится ровно j заявок, а через σ_j — количество элементов в S_j. Далее в соответствии с диаграммой переходов для выбранной модели построим матрицы интенсивностей инфинитезимальных переходов:

$A_j[\sigma_j \times \sigma_{j+1}]$ — в S_{j+1} (прибытие заявки),
$C_j[\sigma_j \times \sigma_j]$ — в S_j (конец промежуточной фазы обслуживания),
$B_j[\sigma_j \times \sigma_{j-1}]$ — в S_{j-1} (полное завершение обслуживания заявки),
$D_j[\sigma_j \times \sigma_j]$ — ухода из состояний яруса j

(в квадратных скобках здесь и далее указывается размер матриц).

Введем векторы-строки $\gamma_j = \{\gamma_{j,1}, \gamma_{j,2}, \ldots, \gamma_{j,\sigma_j}\}$ нахождения СМО в состоянии (j,i), $j = 0, 1, \ldots$ Теперь можно записать векторно-матричные уравнения баланса переходов между состояниями

$$\begin{array}{rcll} \gamma_0 D_0 & = & \gamma_0 C_0 + \gamma_1 B_1, & \\ \gamma_j D_j & = & \gamma_{j-1} A_{j-1} + \gamma_j C_j + \gamma_{j+1} B_{j+1}, & j = 1, 2, \ldots \end{array} \qquad (7.4.1)$$

Систему (7.4.1), дополненную условием нормировки, практически приходится расписывать покомпонентно. Даже для моделей с ограниченной очередью она характеризуется чрезвычайно высокой размерностью, и стандартные методы решения систем линейных алгебраических уравнений применительно к ней оказываются малоэффективными. Мы рассмотрим два класса методов ее решения — итерационный и матрично-геометрической прогрессии.

7.5. Итерационный метод для простейшего потока

Под основной схемой мы будем понимать ее базовый вариант, ориентированный на простейший входящий поток (каждый слой диаграммы состоит из одного яруса). *Идея* этой схемы была впервые предложена Такахаси и Таками [265]. Мы опишем ее в более общей форме, с детальной проработкой расчетных зависимостей и с частными вариантами [92, 93, 94].

7.5.1. Основная схема

Положим $t_j = \gamma_j / p_j$, где p_j — суммарная вероятность наличия в системе ровно j заявок, и обозначим

$$x_j = p_{j+1}/p_j, \qquad z_j = p_{j-1}/p_j. \tag{7.5.1}$$

Тогда систему (7.4.1) можно переписать относительно векторов условных вероятностей $\{t_j\}$, нормированных к единице в пределах яруса:

$$\begin{array}{rcl} t_0 D_0 & = & t_0 C_0 + x_0 t_1 B_1, \\ t_j D_j & = & z_j t_{j-1} A_{j-1} + t_j C_j + x_j t_{j+1} B_{j+1}, \qquad j = 1, 2, \ldots \end{array} \tag{7.5.2}$$

С помощью векторов-столбцов $\mathbf{1}_j = \{1, 1, \ldots, 1\}^T$ размера σ_j для всех j могут быть записаны дополняющие систему (7.5.2) условия нормировки

$$t_j \mathbf{1}_j = 1 \tag{7.5.3}$$

и баланса суммарных интенсивностей переходов между смежными ярусами

$$z_j t_{j-1} A_{j-1} \mathbf{1}_j = t_j B_j \mathbf{1}_{j-1}. \tag{7.5.4}$$

Алгоритм расчета набора векторов $\{t_j\}$ и чисел $\{x_j\}$ и $\{z_j\}$, удовлетворяющих соотношениям (7.5.2)–(7.5.4), в случае разомкнутой системы с неограниченной очередью опирается на существование предельного вектора условных вероятностей $t_\infty = \lim_{j\to\infty} t_j$, которое является следствием стабилизации переходных матриц при $j > n$. Алгоритм основан на последовательном приближении к искомым характеристикам для ограниченного множества индексов $j = \overline{0, N}$ и по существу является блочным вариантом известного итерационного метода Гаусса — Зейделя.

Перепишем уравнения системы (7.5.2) для $j \geq 1$ в виде

$$t_j^{(m)}(D_j - C_j) = z_j^{(m)} t_{j-1}^{(m)} A_{j-1} + x_j^{(m)} t_{j+1}^{(m-1)} B_{j+1}, \qquad j = 1, 2, \ldots,$$

где верхний индекс указывает номер итерации. Теперь ясно, что

$$t_j^{(m)} = z_j^{(m)} \beta'_j + x_j^{(m)} \beta''_j, \tag{7.5.5}$$

где

$$\begin{array}{l} \beta'_j = t_{j-1}^{(m)} \quad A_{j-1}(D_j - C_j)^{-1}, \\ \beta''_j = t_{j+1}^{(m-1)}\, B_{j+1}(D_j - C_j)^{-1}. \end{array} \tag{7.5.6}$$

При $j = N$ считается, что

$$\beta''_N = t^{(m)}_{N-1} B_N (D_N - C_N)^{-1}. \tag{7.5.7}$$

Осталось указать способ расчета $\{z^{(m)}_j\}$ и $\{x^{(m)}_j\}$. Перепишем (7.5.4) с учетом (7.5.5):

$$(z^{(m)}_j \beta'_j + x^{(m)}_j \beta''_j) B_j \mathbf{1}_{j-1} = z^{(m)}_j t^{(m)}_{j-1} A_{j-1} \mathbf{1}_j.$$

Отсюда следует пропорциональность

$$z^{(m)}_j = c x^{(m)}_j \tag{7.5.8}$$

с коэффициентом

$$c = \frac{\beta''_j B_j \mathbf{1}_{j-1}}{t^{(m)}_{j-1} A_{j-1} \mathbf{1}_j - \beta'_j B_j \mathbf{1}_{j-1}}. \tag{7.5.9}$$

В этой и последующих формулах произведения матриц перехода на вектор $\mathbf{1}_j$ равны суммам строк соответствующих матриц и могут быть вычислены до начала итераций.

Подстановка (7.5.7) в (7.5.5) и умножение обеих частей результата на $\mathbf{1}_j$ дают

$$1 = t^{(m)}_j \mathbf{1}_j = x^{(m)}_j (c\beta'_j + \beta''_j) \mathbf{1}_j.$$

Итак,

$$x^{(m)}_j = 1/[(c\beta'_j + \beta''_j)\mathbf{1}_j]. \tag{7.5.10}$$

Удобным критерием прекращения итераций является условие

$$\max_j |x^{(m)}_j - x^{(m-1)}_j| \le \varepsilon_1.$$

Описанный алгоритм в связи с усечением числа ярусов — формула (7.5.7) — будет давать надежные результаты лишь при условии

$$\|t_N - t_\infty\| \le \varepsilon_2. \tag{7.5.11}$$

Но $\{t_j\}$ можно получить лишь в процессе реализации алгоритма, трудоемкость итерационной части которого возрастает по N. К тому же, увеличение N приводит к ухудшению сходимости итераций и росту требуемого объема памяти из расчета по крайней мере σ_n слов на каждый дополнительный ярус диаграммы (расход удвоится при двойной точности и учетверится — при использовании гиперэкспоненциальной аппроксимации, имеющей в общем случае комплексные параметры). Поэтому целесообразно выполнить итерации для умеренного значения N, после чего проверить выполнение условия (7.5.11).

7.5.2. Безусловные вероятности

После прекращения итераций можно переходить к нахождению абсолютных значений вероятностей. Прежде всего отметим, что из определения чисел $\{x_j\}$ следуют равенства

$$p_{j+1} = p_j x_j, \qquad j = \overline{0, N-1}. \tag{7.5.12}$$

Подставив их в условие баланса прибытия и ухода заявок

$$\sum_{j=0}^{n-1} (n-j) p_j = n - \lambda b,$$

получаем формулу для вероятности свободного состояния системы

$$p_0 = \frac{n - \lambda b}{n + \sum\limits_{j=1}^{n-1} (n-j) \prod\limits_{i=0}^{j-1} x_i}. \tag{7.5.13}$$

Последующие вероятности для $j = \overline{1, N}$ определяются рекуррентно с помощью (7.5.12). При необходимости та же формула может быть применена для бóльших значений j с использованием $x_j = x_\infty$.

С точки зрения «философии вычислений» отметим, что в методе Такахаси—Таками чередуются фазы декомпозиции и агрегации. Его достоинства вытекают из общих свойств итеративных методов: уплотнение информации в случае разреженных матриц и возможность использования дополнительной информации (например, для улучшения начальных приближений).

7.5.3. Начальные приближения

Простейший способ задания начальных векторов вероятностей микросостояний базируется на равновероятном распределении в пределах яруса.

При гиперэкспоненциальном обслуживаниии применим более продвинутый подход, восходящий к уже упоминавшемуся аналогу модели $M/H_2/n$ — процессу обслуживания неоднородных заявок двух типов. Условное распределение числа обслуживаемых заявок каждого типа естественно принять биномиальным с вероятностями, пропорциональными $\{y_i/\mu_i\}$.

Можно, наконец, опереться на сделанное допущение о существовании предельного вектора вероятностей микросостояний при $j \to \infty$. Упомянутое условие указывает естественный путь задания начального приближения: $t_N^{(0)} = t_\infty$. Рассмотрим способ нахождения такого приближения. Будем обозначать предельные при $j \to \infty$ матрицы, векторы и отношения смежных вероятностей прежними символами, но без индексов. Если существует $t_\infty = t$, то существуют и предельные значения отношений смежных вероятностей x и $z = 1/x$, причем из (7.5.4) следует

$$x = tA\mathbf{1}_n / tB\mathbf{1}_n \tag{7.5.14}$$

и стационарные вероятности состояний с большими индексами образуют геометрическую прогрессию. Расчеты свидетельствуют о хорошей аппроксимации x формулой

$$x = \rho^{2/(v_A^2+v_B^2)}. \tag{7.5.15}$$

Полагая x известным, как следствие (7.5.5) имеем предельное равенство

$$t = (x^{-1}tA + xtB)(D-C)^{-1} = t(x^{-1}A + xB)(D-C)^{-1} = tQ,$$

Обозначим $(Q-I)_1$ матрицу, полученную из $Q-I$ заменой ее первой строки на единичную, и положим $\delta_1 = \{1,0,0,\ldots,0\}$. Тогда заведомо $\det(Q-I) \neq 0$, и искомый вектор получается как решение системы линейных алгебраических уравнений

$$t(Q-I)_1 = \delta_1. \tag{7.5.16}$$

В качестве начальных приближений к векторам $\{t_j\}$ в ходе описанных выше итераций, охватывающих все ярусы диаграммы, для $j \geq n$ следует воспользоваться решением (7.5.16) при значении x, найденном по формуле (7.5.15). На вышележащих ярусах можно принять все состояния равновероятными. Опыт расчетов свидетельствует о слабом влиянии выбора начальных приближений (в рамках рассмотренных подходов) на требуемое число итераций.

7.5.4. Направление прогонки

Вариант с «предельным» стартовым вектором вероятностей микросостояний порождает естественную идею смены направления прогонки

— от бóльших индексов к меньшим (от «хороших» приближений к менее удачным). Тогда естественно проводить горизонтальный разрез между j-м и $(j+1)$-м ярусами, что приводит к условию баланса переходов

$$(z_j^{(m)}\beta_j' + x_j^{(m)}\beta_j'')A_j\mathbf{1}_{j+1} = x_j^{(m)}t_{j+1}^{(m)}B_{j+1}\mathbf{1}_j.$$

Отсюда следует единственное изменение в основной расчетной схеме алгоритма:

$$c = \frac{t_{j+1}^{(m)}B_{j+1}\mathbf{1}_j - \beta_j''A_j\mathbf{1}_{j+1}}{\beta_j'A_j\mathbf{1}_{j+1}}. \tag{7.5.17}$$

7.5.5. Метод сверхрелаксации

Расчеты для каждого яруса диаграммы весьма нетривиальны, количество ярусов в одном прогоне обычно измеряется десятками, а число итераций для получения достаточно точных значений $\{x_j\}$ достигает нескольких сотен. Одним из популярных подходов к ускорению сходимости итерационных процессов является метод *последовательной сверхрелаксации*. Его применение к расчету многофазных систем обслуживания описано в [255] и сводится к получению результирующих векторов условных вероятностей по правилу

$$t_j^{(m)} = \omega\tilde{t}_j^{(m)} + (1-\omega)t_j^{(m-1)}, \tag{7.5.18}$$

где $\tilde{t}_j^{(m)}$ — приближение, получаемое в стандартной версии метода. Скорость сходимости определяется вторым по модулю собственным числом преобразования. Если все собственные значения известны, можно получить оптимальное для сверхрелаксации значение ω. Это значение коэффициента релаксации ω, $1 < \omega < 2$, чрезвычайно чувствительно к параметрам задачи и зависит от них очень сложным образом. В связи с этим в [255] был предложен адаптивный алгоритм: для текущего значения параметра выполнялось 10 итераций, после чего определялось отношение норм невязок

$$\xi = \frac{||t^{(10)} - t^{(9)}||}{||t^{(9)} - t^{(8)}||}. \tag{7.5.19}$$

Если предыдущее изменение ω приводило к увеличению ξ, то предлагалось принять новое значение $\omega = 1 + 0.75 \cdot (\omega - 1)$, иначе

$\omega = 1 + 1.25 \cdot (\omega - 1)$. Декларировалось, что при комплексных векторах $\{t_j\}$ процесс не сходится.

При применении метода последовательной сверхрелаксации возникают дополнительные вопросы:

- обязательно ли контролировать нормы невязок векторов условных вероятностей;
- действительно ли в комплексном случае имеет место расходимость;
- как применение метода сверхрелаксации влияет на общее количество итераций;
- как влияет количество шагов при фиксированном ω на скорость сходимости;
- насколько энергично следует корректировать ω.

Ниже обсуждаются варианты ответа на них и результаты численных экспериментов.

7.5.6. Численные результаты

Приведем (табл. 7.2, 7.3) сведения о количестве итераций и (через слэш) трудоемкости в секундах процессорного времени Pentium 4, 2.4 Ггц, для двух моделей систем обслуживания (первая — с комплексными параметрами распределения времени обслуживания, вторая — с вещественными). Сопоставляются обсуждавшиеся выше варианты задания начальных условных векторов вероятностей микросостояний и различные направления прогонки. Точность оценки времени счета определялась разрешающей способностью системных часов (порядка 0.01 с). Обсчитывалось 70 ярусов диаграммы состояний. Точность стабилизации отношений смежных вероятностей назначалась $\varepsilon = 10^{-6}$.

Более подробное описание и обоснование описанных выше методов приведено в [126]. Они были запрограммированы и дали результаты, совпадающие с высокой степенью точности и подтвержденные перекрестным тестированием на ряде моделей систем обслуживания — см. [129].

Таблица 7.2. Итерационный обсчет модели $M/E_3/n$

Метод	Число каналов n					
	2	3	5	10	20	30
Равновер., ↓	8/0	13/0	23/0	50/0.031	84/0.110	108/0.312
Равновер., ↑	28/0	37/0	55/0	100/0.047	154/0.218	208/0.594
Бином., ↓	7/0	12/0	20/0	38/0.031	70/0.109	94/0.266
Бином., ↑	18/0	23/0	33/0	55/0.031	96/0.141	120/0.359
Lim, равновер., ↓	8/0	14/0	24/0	49/0.016	78/0.109	170/0.485
Lim, бином., ↓	8/0	14/0	24/0	52/0.032	78/0.109	89/0.250
Lim, равновер., ↑	21/0	28/0	42/0	79/0.047	146/0.187	188/0.516
Lim, бином., ↑	21/0	28/0	42/0	75/0.031	151/0.235	167/0.468

Таблица 7.3. Итерационный обсчет модели $M/H_2/n$, $v_B = 2$

Метод	Число каналов n					
	2	3	5	10	20	30
Равновер., ↓	27/0	37/0	52/0	91/0.031	171/0.250	249/0.688
Равновер., ↑	24/0	34/0	55/0	106/0.046	205/0.297	299/0.812
Бином., ↓	23/0	31/0	42/0	62/0.031	89/0.140	85/0.235
Бином., ↑	15/0	20/0	30/0	52/0.015	74/0.125	87/0.250
Lim, равновер., ↓	27/0	36/0	51/0	89/0.047	169/0.234	248/0.688
Lim, бином., ↓	27/0	36/0	51/0	83/0.031	136/0.203	168/0.469
Lim, равновер., ↑	18/0	26/0	41/0	80/0.031	153/0.234	223/0.610
Lim, бином., ↑	17/0	23/0	35/0	62/0.031	98/0.141	120/0.328

Из этих таблиц можно сделать следующие выводы:

1) Все рассмотренные варианты позволяют обсчитывать модели с несколькими десятками каналов быстрее, чем за секунду.

2) Требуемое число итераций возрастает несколько медленнее, чем число каналов.

3) Наименее устойчив в работе метод, использующий предельный вектор микросостояний.

4) Наилучшим вариантом выбора начальных приближений является биномиальный.

5) Для модели $M/E_3/n$ предпочтительна прогонка сверху вниз, а для $M/H_2/n$ — снизу вверх.

Теперь приведем результаты исследования сверхрелаксации. В нижеприведенной таблице под «коррекцией» подразумевается модуль добавки к единице при множителе $(1-\omega)$. Показатель ξ, на основе которого корректировалась ω — см. формулу (7.5.19) — определялся по максимальным уточнениям $\{x_j\}$ для переменного числе шагов при соответственно измененных индексах. Прочерками отмечены случаи, где невязка 10^{-6} не была достигнута за 500 шагов.

Таблица 7.4. Количество просчетов для модели $M/H_2/10$ при $\omega = \text{const}$

Коррекция	Просчетов при									
	$v_B = 2$					$v_B = 1/\sqrt{3}$				
	3	4	5	7	10	3	4	5	7	10
0.05	55	57	56	57	61	34	33	36	36	41
0.07	52	57	56	57	61	34	33	36	36	41
0.10	175	53	56	57	61	34	33	36	36	41
0.15	-	-	51	57	61	31	33	36	36	41
0.20	253	-	-	57	61	31	33	36	36	41
0.25	-	297	171	134	61	31	33	36	36	41

Таким образом,

1) Экономичный вариант вычисления ξ по максимальным уточнениям $\{x_j\}$ оказался вполне работоспособным.

2) Из таблицы 7.4 для модели с коэффициентом вариации обслуживания $v_B = 1/\sqrt{3}$ следует отрицавшаяся автором [255] возможность применения метода сверхрелаксации к задачам с комплексными параметрами.

3) «Слишком энергичная» коррекция как правило не окупается; по-видимому, целесообразно ограничиться поправкой ± 0.10.

4) При разумной организации программы (циркуляция данных) потребность в памяти не зависит от числа k просчетов при неизменном ω. Малое их количество, вообще говоря, позволяет быст-

рее реагировать на ситуацию, но снижает обоснованность коррекции. По данным нашего эксперимента, разумно остановиться на $k = 7$.

Сравним по числу итераций метод сверхрелаксации с наилучшей версией итерационного метода («биномиальные» начальные приближения, прогонка сверху вниз):

Таблица 7.5. Модель $M/H_2/n, \quad v_B = 1/\sqrt{3}$

Модель	Метод	Число каналов					
		2	3	5	10	20	30
$v_B = 1/\sqrt{3}$	Зейделя	12	22	38	74	138	186
	Сверхрелаксации	15	15	22	36	64	141
$v_B = 2.0$	Зейделя	44	60	82	122	176	169
	Сверхрелаксации	22	29	36	57	78	71

Несмотря на бесспорное преимущество в рассмотренном случае, метод сверхрелаксации сложен в реализации, требует дополнительной оперативной памяти и весьма капризен в связи с трудностью выбора параметра сверхрелаксации ω (достаточно сослаться на отмеченные выше весьма противоречивые — и даже абсурдные — рекомендации по выбору его начального значения и последующей коррекции). Для массового использования следует рекомендовать вариант с биномиальными начальными приближениями и прогонкой от меньших индексов к бóльшим.

При распределениях времени обслуживания с коэффициентом вариации, превышающим единицу, фактические параметры H_2-аппроксимации этого распределения оказываются вещественными, и проведение с ними «комплексных» вычислений приводит к избыточной трудоемкости. Соответственно имеет смысл работать с двумя вариантами процедуры обсчета упомянутой модели: с комплексными и вещественными параметрами. В табл. 7.6 приведены результаты обсчета $M/H_2/n$ для $J_{\max} = 120$ и $\varepsilon = 10^{-8}$ (i — число итераций, t — время счета). Помимо четко прослеживаемого «эффекта овеществления», отметим иллюстрируемую этой таблицей возможность расчета систем с очень большим числом каналов.

Таблица 7.6. Эффект «овеществления» параметров

Представление данных		Число каналов							
		5	10	20	30	40	50	70	100
Комплексные	i	114	184	246	384	360	416	526	662
	t	0.031	0.109	0.485	1.484	2.422	4.203	12.672	46.656
Вещественные	i	50	79	112	117	137	159	199	250
	t	0	0.046	0.172	0.391	0.750	1.422	4.640	19.954

7.5.7. Ограниченная очередь

Описанная расчетная схема легко модифицируется применительно к системам с ограничением числа заявок в системе величиной R — естественно принять $N = R$. Кроме того, отсутствие ярусов для $j > N$ приводит к замене формулы (7.5.5) при $j = N$ на

$$t_N^{(m)} = z_N^{(m)} t_{N-1}^{(m-1)} A_{N-1}(D_N - C_N)^{-1}. \tag{7.5.20}$$

Умножив обе части (7.5.20) справа на $\mathbf{1}_N$, получаем явную формулу

$$z_N^{(m)} = [t_{N-1}^{(m-1)} A_{N-1}(D_N - C_N)^{-1}\mathbf{1}_N]^{-1}. \tag{7.5.21}$$

Теперь с помощью (7.5.20) можно найти $t_N^{(m)}$. Остальные элементы итерационной части алгоритма сохраняются. После окончания итераций для расчета p_0 вместо (3.1.2) используется условие баланса

$$(1 - p_N)/a = \Big[\sum_{j=0}^{n-1} j p_j + n\Big(1 - \sum_{j=0}^{n-1} p_j\Big)\Big]\Big/ b,$$

откуда следует

$$p_0 = \frac{n - b/a}{n + \sum\limits_{j=1}^{n-1}(n-j)\prod\limits_{i=0}^{j-1} x_i - (b/a)\prod\limits_{i=0}^{N-1} x_i}. \tag{7.5.22}$$

7.5.8. Замкнутые системы

Введение зависимости λ от номера яруса позволяет применить алгоритм и для расчета замкнутых систем (с кусочно-постоянным простейшим потоком). Здесь в целях определения p_0 используется условие баланса

$$\sum_{j=0}^{N-1} \lambda_j p_j = \Big[\sum_{j=0}^{n-1} j p_j + n\Big(1 - \sum_{j=0}^{n-1} p_j\Big)\Big]\Big/ b,$$

которое после выражения $\{p_j\}$ через p_0 и $\{x_j\}$ дает формулу

$$p_0 = n \Big/ \Big\{ n + \sum_{j=1}^{n-1} (n-j) \prod_{i=0}^{j-1} x_i + b\Big(\lambda_0 + \sum_{j=1}^{N-1} \lambda_j \prod_{i=0}^{j-1} x_i \Big) \Big\}. \qquad (7.5.23)$$

Расчет t_N и z_N в итерациях для этой модели также выполняется по формулам (7.5.20) и (7.5.21).

7.5.9. Условия сходимости метода

Получим условия сходимости метода для случая неограниченной очереди и выполнения итераций снизу вверх. Для согласования схемы рассуждений с традиционно используемой в линейной алгебре [31] транспонируем все введенные в разд. 7.4 матричные объекты, сохранив за ними прежние обозначения. Тогда (7.4.1) можно переписать в виде

$$\begin{array}{rcll} D_0\gamma_0 & = & C_0\gamma_0 + B_1\gamma_1, & \\ D_j\gamma_j & = & A_{j-1}\gamma_{j-1} + C_j\gamma_j + B_{j+1}\gamma_{j+1}, & j = 1, 2, \ldots \end{array} \qquad (7.5.24)$$

Сходимость описанной выше расчетной схемы эквивалентна сходимости процесса

$$\begin{array}{rcll} \gamma_N^{(k)} & = & (D_N - C_N)^{-1}(A_{N-1} + xB_N)\gamma_{N-1}^{(k-1)}, & \\ \gamma_j^{(k)} & = & (D_j - C_j)^{-1}A_{j-1}\gamma_{j-1}^{(k-1)} + (D_j - C_j)^{-1}B_{j+1}\gamma_{j+1}^{(k)}, & \\ & & & j = N-1, N-2, \ldots, 1, \\ \gamma_0^{(k)} & = & (D_0 - C_0)^{-1}B_1\gamma_1^{(k)}, & \end{array} \qquad (7.5.25)$$

где верхние индексы указывают номера приближений.

Запишем уравнения системы (7.5.25) в стандартной форме

$$\gamma_j^{(k)} = \alpha_{j,j-1}\gamma_{j-1}^{(k-1)} + \alpha_{j,j+1}\gamma_{j+1}^{(k)}, \qquad j = \overline{1, N}, \qquad (7.5.26)$$

где

$$\begin{array}{rcll} \alpha_{0,-1} & = & 0, & \\ \alpha_{j,j-1} & = & (D_j - C_j)^{-1}A_{j-1}, & j = \overline{1, N-1}, \\ \alpha_{N,N-1} & = & (D_N - C_N)^{-1}(A_{N-1} + xB_N), & \\ \alpha_{j,j+1} & = & (D_j - C_j)^{-1}B_{j+1}, & j = \overline{0, N-1}, \\ \alpha_{N,N+1} & = & 0. & \end{array} \qquad (7.5.27)$$

Если итерации сходятся, то предельные значения векторов вероятностей микросостояний

$$\gamma_j = \alpha_{j,j-1}\gamma_{j-1} + \alpha_{j,j+1}\gamma_{j+1}. \tag{7.5.28}$$

Вычитая из (7.5.28) равенство (7.5.26), получим

$$\gamma_j - \gamma_j^{(k)} = \alpha_{j,j-1}(\gamma_{j-1} - \gamma_{j-1}^{(k-1)}) + \alpha_{j,j+1}(\gamma_{j+1} - \gamma_{j+1}^{(k)}). \tag{7.5.29}$$

Найдем достаточные условия сходимости процесса по m-норме (максимальному из модулей элементов вектора $\gamma - \gamma^{(k)}$). Из (7.5.29) и свойств канонических норм матриц следует, что

$$\|\gamma_j - \gamma_j^{(k)}\|_m \le \|\alpha_{j,j-1}\|_m \cdot \|\gamma_{j-1} - \gamma_{j-1}^{(k-1)}\|_m + \|\alpha_{j,j+1}\|_m \cdot \|\gamma_{j+1} - \gamma_{j+1}^{(k)}\|_m. \tag{7.5.30}$$

Пусть i — значение индекса j, при котором достигается $\max\limits_j \|\gamma_j - \gamma_j^{(k)}\|_m$. Тогда $\|\gamma - \gamma^{(k)}\|_m = \|\gamma_i - \gamma_i^{(k)}\|_m$, и можно переписать (7.5.30) в виде

$$\|\gamma - \gamma^{(k)}\|_m \le \|\alpha_{i,i-1}\|_m \cdot \|\gamma - \gamma^{(k-1)}\|_m + \|\alpha_{i,i+1}\|_m \cdot \|\gamma - \gamma^{(k)}\|_m,$$

откуда выводим

$$\|\gamma - \gamma^{(k)}\|_m \le \frac{\|\alpha_{i,i-1}\|_m}{1 - \|\alpha_{i,i+1}\|_m}\|\gamma - \gamma^{(k-1)}\|_m.$$

Сходимость по m-норме будет обеспечена при любых начальных приближениях, если

$$y_m = \max_i \frac{\|\alpha_{i,i-1}\|_m}{1 - \|\alpha_{i,i+1}\|_m} < 1. \tag{7.5.31}$$

Теперь получим достаточные условия сходимости процесса (7.5.25) по l-норме (сумме модулей элементов вектора $\gamma - \gamma^{(k)}$). Из равенства (7.5.29) вновь следует

$$\|\gamma_j - \gamma_j^{(k)}\|_l \le \|\alpha_{j,j-1}\|_l \cdot \|\gamma_{j-1} - \gamma_{j-1}^{(k-1)}\|_l + \|\alpha_{j,j+1}\|_l \cdot \|\gamma_{j+1} - \gamma_{j+1}^{(k)}\|_l.$$

Суммируя эти неравенства по всем j, получим

$$\sum_j \|\gamma_j - \gamma_j^{(k)}\|_l \le \sum_j \|\alpha_{j,j-1}\|_l \cdot \|\gamma_{j-1} - \gamma_{j-1}^{(k-1)}\|_l + \sum_j \|\alpha_{j,j+1}\|_l \cdot \|\gamma_{j+1} - \gamma_{j+1}^{(k)}\|_l. \tag{7.5.32}$$

Положим

$$\begin{aligned} \varphi_- &= \max_j \|\alpha_{j,j-1}\|_l, \\ \varphi_+ &= \max_j \|\alpha_{j,j+1}\|_l. \end{aligned} \tag{7.5.33}$$

Учитывая, что

$$\sum_j \|\gamma_j - \gamma_j^{(k)}\|_l = \|\gamma - \gamma^{(k)}\|_l,$$

можно переписать (7.5.32) в виде

$$\|\gamma - \gamma^{(k)}\|_l \le \varphi_- \|\gamma - \gamma^{(k-1)}\|_l + \varphi_+ \|\gamma - \gamma^{(k)}\|_l.$$

Таким образом,

$$\|\gamma - \gamma^{(k)}\|_l \le \frac{\varphi_-}{1-\varphi_+} \|\gamma - \gamma^{(k)}\|_l,$$

и сходимость итераций по l-норме гарантируется в случае

$$y_l = \frac{\varphi_-}{1-\varphi_+} < 1. \tag{7.5.34}$$

Проверка условий (7.5.31) и (7.5.34) требует выполнения трудоемких вычислений. Кроме того, оба эти условия являются достаточными, но не необходимыми, и отрицательный результат проверки не обязательно влечет за собой расходимость итераций. Практически *при наличии программ*, реализующих итерационный метод, удобнее непосредственно проверить его сходимость на ЭЦВМ. Возможные в случае расходимости непроизводительные затраты машинного времени легко ограничить указанием предельно допустимых времени счета или числа шагов. По названным причинам мы воздержимся от теоретического исследования сходимости итераций для более общих случаев.

Реальные случаи расходимости были зафиксированы при коэффициентах загрузки $\rho \le 0.3$ — в подобных ситуациях влиянием очередей на работу системы можно пренебречь.

Сходимость заметно ухудшается при увеличении числа рассматриваемых ярусов. Поэтому целесообразно ограничиваться 15–20 ярусами, при необходимости постулируя постоянство отношений вероятностей смежных ярусов для бóльших индексов.

7.6. Прикладная задача

Определим оптимальное число аварийных линейных бригад для кругового энергорайона диаметром 600 км. Все данные условны.

Район потребляет мощность 200 тыс. кВт, подводимую от внешнего источника по линиям напряжением 110 кВ. Это напряжение подводится к 10 зонам, где понижается до 6 кВ и распределяется между 10 потребителями в каждой зоне. Состав обслуживаемого бригадами электрооборудования, интенсивности $\{\lambda_i\}$ отказов в год и средние времена устранения аварии, а также суммарные частоты отказов приведены в табл. 7.7. Данные по удельным интенсивностям отказов (для линий электропередач из расчета на 100 км длины) и средним временам восстановления заимствованы из «Справочника по проектированию электроснабжения, линий электропередач и сетей» под редакцией Я. М. Большама и В. И. Круповича. — М.: изд-во «Энергия», 1974, с. 88. «Деятельность» охотников за цветными металлами не учитывается.

В дальнейшем времена восстановления для каждого вида оборудования считаем постоянными. Учитывая малую относительную частоту отказов трансформаторов и распределительных устройств, дальнейший расчет делаем только для линий электропередач.

Таблица 7.7. «Электрические» исходные данные

Наименование	Мера	Частота аварий в год		Ср. время восст., час.
		удельная	суммарно	
1. Линии передач				
- воздушные 110 кВ	4500 км	1.2	54	4
- воздушные 6 кВ	2500 км	1.5	37.5	8
- кабельные 6 кВ	2500 км	8	200	16
2. Трансформаторы				
110 Квольт – 6 кВ по 50000 кВА	5	0.04	0.2	90–120
6 Квольт – 380 В по 2800 кВА	100	0.005	0.5	90–120
3. Распределительные устройства				
- 110 кВ	11	0.03	0.33	25
- 6 кВ	10	0.02	0.20	15

Время занятости бригады на конкретной аварии складывается из чистого времени ремонта и времени проезда в оба конца. Пусть авария с равной плотностью вероятностей происходит в любой точке района. Тогда дифференциал плотности распределения расстояния от центра района до места аварии

$$\varphi(r)dr = \frac{2\pi r\,dr}{\pi R^2} = \frac{2r\,dr}{R^2}, \qquad 0 \le r \le R,$$

а k-й момент времени проезда в оба конца при средней скорости v

$$t_k = \int_0^R (2r/v)^k \frac{2r\,dr}{R^2} = \frac{2}{k+2}\left(\frac{2R}{v}\right)^k.$$

При R=300 км и v=60 км/час

$$t_k = \frac{2}{k+2}10^k \text{час}^k, \qquad k = 1, 2, \ldots$$

В частности, среднее время

$$t_1 = 10 \cdot 2/3 = 6.67 \quad \text{час.}$$

Моменты чистой длительности ремонта

$$\tau_k = \sum_{i=1}^{3} \frac{\lambda_i}{\Lambda} \tau_k^{(i)},$$

где $\{\lambda_i\}$ и $\{\tau_k^{(i)}\}$ данные из 4-й и 5-й граф табл. 7.7 для соответствующих типов линий передач, а $\Lambda = \sum\limits_i \lambda_i$. Моменты распределения суммарного времени занятости бригады на выезде можно получить сверткой моментов распределений длительности проезда и ремонта с помощью процедуры `CONV`.

Вычислим загрузку «одноканального» устройства обслуживания $\hat{\rho} = \Lambda(t_1 + \tau_1)$. В нашем случае

Λ = 54+37.5+200=291.5 1/год,

τ_1 = (4 · 54+8 · 37.5+16 · 200)/291.5 = 12.75 час.

В году 8760 часов; следовательно, загрузка

$$\hat{\rho} = 291.5 \cdot (6.67 + 12.75)/8760 = 0.646.$$

Поскольку $\hat{\rho} < 1$, с расчетным объемом работ может справиться и одна бригада. Вопрос об оптимальном числе бригад следует решать с учетом экономических факторов.

В стоимость содержания бригад включим:

- оплату персонала (4 смены по три человека со средним окладом 200 долл. в месяц), т. е. $200 \cdot 12 \cdot 12$=28800 долл.;
- оплату специальной машины (исходная стоимость C=12 тыс. долл., годовые эксплуатационные расходы и амортизационные отчисления около $0.2 \cdot$ — итого 2400 долл).

Всего годовая стоимость содержания бригады $C_{\text{бр}}$=31.2 тыс. долл.

Ущерб от недостаточного числа бригад связан с дополнительной задержкой в восстановлении энергоснабжения из-за ожидания освобождения занятой бригады. Для укрупненных расчетов этот ущерб можно принять равным 0.8 долл. на недопоставленный киловаттчас (см. Л. А. Солдаткина. Электрические сети и системы. — М.: Энергия, 1972, с. 183)[1]. Примем, что авария на линии 110 кВ приводит к отключению всех потребителей зоны (20 МВт), а на линии 6 кВ — одного потребителя (2 МВт). Тогда средневзвешенные потери составят

$$\frac{54 \cdot 20 + 237.5 \cdot 2}{291.5} \cdot 0.8 = 4.27 \text{ тыс. долл.}$$

за час *дополнительной* задержки. Среднее число случаев задержки равно ожидаемому числу аварий в год (291.5), а средняя задержка — функция числа бригад $w(n)$, определяемая как среднее время ожидания начала обслуживания в соответствующей n-канальной системе обслуживания. Нам предстоит выбрать такое значение количества бригад n, при котором функция

$$L(n) = 31.2n + 4.27 \cdot 291.5w(n) = 31.2n + 1244.7w(n)$$

[1] В упомянутой книге эта цифра названа в рублях, но в те годы покупательная способность рубля была не меньше, чем у доллара.

имеет минимальное значение. Отметим, что в эту сумму не входит ущерб от задержек, *не связанных с числом бригад* (например, на время проезда бригады и собственно ремонта).

Табулирование этой функции было выполнено в цикле по n с помощью процедуры расчета СМО вида $M/H_k/n$, вычисления средней длины очереди и последующего применения формулы Литтла. Результаты представлены в табл. 7.8.

Таблица 7.8. Зависимость затрат от числа бригад

n	$w(n)$, час.	$L(n)$, тыс. долл.
1	19.120	23833.073
2	1.330	1719.642
3	0.160	292.903
4	0.018	147.549
5	0.002	158.242

Итак, наивыгоднейшим является довольно неожиданое решение (иметь 4 бригады), что обусловлено высокой ценой «штрафа» за перерывы в энергоснабжении.

Результаты аналогичного расчета в простейших допущениях (показательное распределение длительности занятости бригады при том же среднем) приведены в табл. 7.9.

Таблица 7.9. Результаты расчета по марковской модели

n	$w(n)$, час.	$L(n)$, тыс. долл.
1	35.440	44137.580
2	2.260	2877.592
3	0.250	401.073
4	0.026	157.433
5	0.002	159.070

В данном случае оптимальный выбор (4 бригады) совпадает с точным решением, а ожидаемые затраты завышены на 7%. Это согласие объясняется исключительно высокой ценой штрафа, при которой оптимум соответствует очень малому среднему ожиданию ($\approx$ 1.5 минуты).

Разумеется, при принятии окончательного решения могут быть учтены и другие соображения.

В качестве альтернативного подхода рассмотрим разбивку района на n зон с четырехсменной бригадой, закрепленной за каждой зоной. В этой модели будут следующие отличия:

1. Суммарная интенсивность отказов Λ делится на n.

2. Площадь зоны составит S/n, а ее радиус — $R/\sqrt{n}$. Соответственно будут изменяться моменты распределения времени пребывания бригады в пути:

$$t^k = \frac{2}{k+2}(10/\sqrt{n})^k, \qquad k = 1, 2, \ldots$$

3. К оплате бригады следует добавить оплату диспетчера в расчете на четырехсменный режим, что составит $200 \cdot 12 \cdot 4$=9600 долл. в год. Для сравнимости с результатами исходного расчета эту сумму надо умножать на $n-1$ (в централизованной системе тоже есть диспетчер).

Таким образом, в данном варианте

$$\begin{aligned} L(n) &= 31.2n + 9.6(n-1) + 1244.7w'(n) \\ &= 40.8n + 1244.7w'(n) - 9.6, \end{aligned}$$

где $w'(n)$ — среднее время ожидания начала обслуживания в «одноканальной» системе с интенсивностью входящего потока Λ/n и уменьшенными с учетом сокращения расстояний моментами распределения времени занятости бригады на одной аварии. Заметим, что коэффициент при $w'(n)$ должен быть разделен на n (ожидаемое число аварий уменьшается в n раз) и умножен на n для суммирования затрат по району в целом, т. е. останется без изменения.

Поскольку в данной модели нетривиальная часть расчета состоит в определении среднего времени ожидания для системы $M/G/1$, здесь достаточно воспользоваться формулой Полячека — Хинчина. Результаты этого расчета представлены в табл 7.10.

Таблица 7.10. Эффект разбивки на зоны

n	$w(n)$, час.	$L(n)$, тыс. долл.
2	3.89	4917.890
3	2.05	2664.745
4	1.37	1852.809
5	1.01	1456.420
.....		
11	0.38	916.799
12	0.35	911.289
13	0.32	913.725

Значения $w'(n)$ убывают по n значительно медленнее, чем в предыдущих случаях. Минимум $L(n)$ достигается при $n = 12$ и равен 911.289 тыс. долл. Таким образом, дробление энергорайона на зоны обслуживания в рамках сделанных допущений невыгодно.

7.7. Итерационный метод для рекуррентного потока

Здесь мы рассмотрим расчетные схемы, получаемые при H_k- и E_k- аппроксимации распределений интервалов между смежными заявками. Обе названные аппроксимации приводят к диаграммам с переходами в микросостояния несмежных ярусов (см. разд. 7.1), в связи с чем формулы разд. 7.4 нуждаются в модификации. Этапы алгоритмов, совпадающие с базовой схемой, дополнительно не обсуждаются.

7.7.1. Модель $H_k/M/n$

В этом частном случае благодаря особой простоте и симметрии диаграммы переходов (рис. 7.1) удается получить чрезвычайно простой и эффективный алгоритм. Для произвольного яруса $j > 0$ уравнения баланса могут быть расписаны покомпонентно в форме

$$t_{j,i}(\lambda_i + \mu_j) = z_j s u_i + x_j \mu_{j+1} t_{j+1,i}, \qquad i = \overline{1,k}, \tag{7.7.1}$$

где принято $s = \sum_{m=1}^{k} \lambda_m t_{j-1,m}$, а интенсивности обслуживания $\{\mu_j\}$ вычисляются с учетом числа задействованных каналов. Баланс перехо-

дов для разреза $(j-1,j)$ требует равенства $z_j s = \mu_j$, что позволяет переписать (7.7.1) с исключенным z_j:

$$t_{j,i} = \mu_j \frac{u_i}{\lambda_i + \mu_j} + x_j \mu_{j+1} \frac{t_{j+1,i}}{\lambda_i + \mu_j}. \tag{7.7.2}$$

Суммируя по всем i, находим

$$x_j = \Bigl(1 - \mu_j \sum_{i=1}^{k} \frac{u_i}{\lambda_i + \mu_j}\Bigr) \Big/ \Bigl(\mu_{j+1} \sum_{i=1}^{k} \frac{t_{j+1,i}}{\lambda_i + \mu_j}\Bigr) \tag{7.7.3}$$

и подставляем его в (7.7.2) для вычисления компонент $\{t_{j,i}\}$. Заметим, что составляющие (7.7.2) уже вычислялись при накоплении сумм из (7.7.3) — осталось только получить их взвешенные комбинации. Система уравнений для компонент предельного при $j \to \infty$ вектора преобразуется в

$$\begin{aligned} (u_i/x) \sum_{m=1}^{k} \lambda_m t_m + [n\mu(x-1) - \lambda_i] t_i &= 0, \qquad i = \overline{1, k-1}, \\ \sum_{m=1}^{k} t_m &= 1. \end{aligned}$$

Для нулевого яруса уравнения (7.7.1) сводятся к $t_{0,i}\lambda_i = x_0 \mu t_{1,i}$, откуда $t_{0,i} = G t_{1,i}/\lambda_i$. Здесь G — нормирующий множитель.

При расчете системы с ограниченной очередью для последнего яруса R имеем

$$n\mu t_{R,i} = z_R \Bigl(\sum_{m=1}^{k} \lambda_m t_{R-1,m}\Bigr) u_i,$$

и соображения нормировки приводят к равенствам $t_{R,i} = u_i$ независимо от R и хода итераций на остальных ярусах.

7.7.2. Модель $H_k/H_k/n$

В данном случае уравнения баланса принимают вид

$$\begin{aligned} \gamma_{0,i} D_{0,i} &= \gamma_{1,i} B_1, \qquad i = \overline{1, k}, \\ \gamma_{j,i} D_{j,i} &= u_i \Bigl(\sum_{l=1}^{k} \lambda_l \gamma_{j-1,l}\Bigr) A_{j-1} + \gamma_{j+1,i} B_{j+1}, \\ &\qquad i = \overline{1, k}, \qquad j = 1, 2, \ldots \end{aligned} \tag{7.7.4}$$

Для этой модели все обозначения имеют прежний смысл за исключением матриц $\{A_j\}$, которые выгоднее сформировать для единичной интенсивности входящего потока. Диагональные матрицы $\{D_{j,i}\}$ интенсивностей ухода из микросостояний ярусов получают дополнительный индекс номера яруса i в связи с тем, что интенсивности потоков заявок $\{\lambda_i\}$ зависят от i.

Переходя к условным векторам вероятностей микросостояний, перепишем (7.7.4) в виде

$$\begin{aligned} t_{0,i}D_{0,i} &= x_0 t_{1,i}B_1, \qquad i=\overline{1,k}, \\ t_{j,i}D_{j,i} &= u_i z_j\Big(\sum_{l=1}^{k}\lambda_l t_{j-1,l}\Big)A_{j-1} + x_j t_{j+1,i}B_{j+1}, \\ & \qquad\qquad i=\overline{1,k}, \qquad j=1,2,\ldots \end{aligned} \tag{7.7.5}$$

Зафиксируем направление итераций снизу вверх (опыт расчетов показал, что в системах с немарковским входящим потоком это обстоятельство существенно для обеспечения сходимости). Положим

$$V_j = \sum_{l=1}^{k}\lambda_l t_{j,l}^{(m-1)}. \tag{7.7.6}$$

Тогда из (7.7.5) следуют формулы типа (7.5.5):

$$t_{j,i}^{(m)} = z_j\beta'_{j,i} + x_j\beta''_{j,i}, \tag{7.7.7}$$

где

$$\begin{aligned} \beta'_{j,i} &= u_i V_{j-1}\, A_{j-1}D_{j,i}^{-1}, \\ \beta''_{j,i} &= t_{j+1,i}^{(m)} \quad B_{j+1}D_{j,i}^{-1}. \end{aligned} \tag{7.7.8}$$

Условие баланса суммарных интенсивностей переходов между j-м и $(j-1)$-м слоями в нашем случае принимает вид

$$z_j V_{j-1}A_{j-1}\mathbf{1}_j = \Big(\sum_{i=1}^{k} t_{j,i}^{(m)}\Big)B_j\mathbf{1}_{j-1},$$

причем $A_{j-1}\mathbf{1}_j = \mathbf{1}_{j-1}$. Подставляя в его правую часть (7.7.7), убеждаемся, что вновь

$$z_j = cx_j, \tag{7.7.9}$$

где

$$\begin{aligned} c &= \Big(\sum_{i=1}^{k}\beta''_{j,i}\Big)B_j\mathbf{1}_{j-1}\Big/\Big[V_{j-1}A_{j-1}\mathbf{1}_j-\Big(\sum_{i=1}^{k}\beta'_{j,i}\Big)B_j\mathbf{1}_{j-1}\Big] \\ &= \mathcal{B}''_jB_j\mathbf{1}_{j-1}/\Big[V_{j-1}\mathbf{1}_{j-1}-\mathcal{B}'_jB_j\mathbf{1}_{j-1}\Big]. \end{aligned} \tag{7.7.10}$$

Здесь

$$\mathcal{B}'_j=\sum_{i=1}^{k}\beta'_{j,i},\qquad \mathcal{B}''_j=\sum_{i=1}^{k}\beta''_{j,i}. \tag{7.7.11}$$

Наконец, из условия нормировки вероятностей микросостояний в пределах слоя

$$\Big(\sum_{i=1}^{k}t_{j,i}^{(m)}\Big)\mathbf{1}_j=1$$

после подстановки в него правой части (7.7.7) с учетом (7.7.9) и (7.7.11) получаем

$$x_j=1/(c\mathcal{B}'_j+\mathcal{B}''_j)\mathbf{1}_j. \tag{7.7.12}$$

Общий алгоритм большой итерации для слоев, соответствующих числу заявок $j=N-1,N-2,\ldots,1$, состоит в следующем:

1. Рассчитать V_{j-1} по формуле (7.7.6).
2. Для всех $i=\overline{1,k}$ вычислить $\{\beta'_{j,i}\}$ и $\{\beta''_{j,i}\}$ согласно (7.7.8); одновременно
   копить их суммы — см. формулы (7.7.11).
3. Вычислить коэффициент c согласно (7.7.10); при этом матричные умножения следует выполнять справа налево, т. е. умножать первый сомножитель на вычисленную однократно (до начала итераций) сумму строк второго.
4. Вычислить x_j согласно (7.7.12).
5. Вычислить z_j согласно (7.7.9).
6. Сформировать очередные приближения $\{t_{j,i}^{(m)}\}$ согласно (7.7.7).

При расчете N-го (нижнего) слоя аналогично случаю простейшего потока $\{t^{(m)}_{N+1,i}\}$ заменяются на $\{t^{(m-1)}_{N-1,i}\}$.

Для нулевого слоя в каждом ярусе будет только одно микросостояние, матрица $D_{0,i}$ сводится к скаляру λ_i, а матрица B_1 — к вектору-столбцу $\{\mu_1, \mu_2, \ldots, \mu_k\}^T$. Поэтому первое уравнение системы (7.7.5) дает

$$t_{0,i,1} = x_0 \sum_{l=1}^{k} t^{(m)}_{1,i,l}\mu_l/\lambda_i, \qquad i = \overline{1,k}. \tag{7.7.13}$$

Множитель x_0 определяется нормировкой компонент $\{t_{0,i}\}$.

Начальные значения компонент векторов $\{t^{(0)}_{j,i}\}$ для $j < n$ проще всего принять равными в пределах слоя. Можно также считать, что вероятности между ярусами распределяются пропорционально отношениям $\{u_i/\lambda_i\}$, а в пределах яруса — обратно пропорционально интенсивностям ухода по завершению обслуживания, т. е. суммам строк матриц $\{B_j\}$.

При $j \geq n$ в качестве начальных приближений целесообразно принять предельные векторы $\{t_i\}$, получаемые решением системы линейных алгебраических уравнений

$$t_i D_i = x^{-1} u_i \sum_{l=1}^{k} \lambda_l t_l + x t_i B \tag{7.7.14}$$

после замены одного из уравнений условием нормировки.

7.7.3. Модель $H_k/E_q/n$

Данный случай отличается от предыдущего ненулевыми матрицами $\{C_j\}$ интенсивностей переходов между микросостояниями в пределах яруса. Это обстоятельство требует замены в формулах (7.5.1)–(7.5.2) матриц $\{D_{j,i}\}$ на разности $\{D_{j,i} - C_j\}$ (обратных значений в формулах (7.7.8)). Различие в структуре переходных матриц отражается только на уровне программной реализации. Исключение составляет формула (7.7.13), которая здесь должна быть записана в виде

$$t_{0,i,1} = x_0 t_{1,i,q} \cdot \mu/\lambda_i. \tag{7.7.15}$$

7.7.4. Модель $E_q/H_k/n$

В этой модели исходные уравнения баланса имеют вид

$$\begin{array}{rcll}
\gamma_{0,1}D_0 &=& \gamma_{1,1}B_1, & \\
\gamma_{0,i}D_0 &=& \lambda\gamma_{0,i-1} + \gamma_{1,i}B_1, & i = \overline{2,q}; \\
\gamma_{j,1}D_j &=& \gamma_{j-1,q}A_{j-1} + \gamma_{j+1,1}B_{j+1}, & \\
\gamma_{j,i}D_j &=& \lambda\gamma_{j,i-1} + \gamma_{j+1,i}B_{j+1}, & i = \overline{2,q}; \qquad j = 1,2,\ldots
\end{array} \tag{7.7.16}$$

После деления уравнений для j-го слоя на суммарную вероятность p_j его микросостояний система (7.7.16) преобразуется в

$$\begin{array}{rcll}
t_{0,1}D_0 &=& x_0 t_{1,1}B_1, & \\
t_{0,i}D_0 &=& \lambda t_{0,i-1} + x_0 t_{1,i}B_1, & i = \overline{2,q}; \\
t_{j,1}D_j &=& z_j t_{j-1,q}A_{j-1} + x_j t_{j+1,1}B_{j+1}, & \\
t_{j,i}D_j &=& \lambda t_{j,i-1} + x_j t_{j+1,i}B_{j+1}, & i = \overline{2,q}; \qquad j = 1,2,\ldots
\end{array} \tag{7.7.17}$$

Положим

$$\begin{array}{rcl}
\beta' &=& t_{j-1,q}^{(m-1)}A_{j-1}D_j^{-1}, \\
\beta_i'' &=& t_{j+1,i}^{(m)}B_{j+1}D_j^{-1}, \quad i = \overline{1,q}.
\end{array} \tag{7.7.18}$$

Из уравнения для первого яруса j-го слоя следует

$$t_{j,1}^{(m)} = z_j\beta' + x_j\beta_1''. \tag{7.7.19}$$

Последующие уравнения дают

$$t_{j,i}^{(m)} = t_{j,i-1}^{(m)}\lambda D_j^{-1} + x_j\beta_i'', \quad i = \overline{2,q}. \tag{7.7.20}$$

Обозначим

$$F = \lambda D_j^{-1} \tag{7.7.21}$$

и запишем общее выражение для $t_{j,i}^{(m)}$ в виде

$$t_{j,i}^{(m)} = z_j h_i + x_j g_i, \qquad i = \overline{1,q}. \tag{7.7.22}$$

Из (7.7.19) следуют начальные значения

$$h_1 = \beta', \qquad g_1 = \beta_1'', \tag{7.7.23}$$

а из (7.7.20) и (7.7.21) — возможность рекуррентного определения

$$h_i = h_{i-1}\cdot F, \qquad g_i = g_{i-1}\cdot F + \beta_i'', \qquad i = \overline{2,q}. \tag{7.7.24}$$

Условие баланса переходов между j-м и $(j-1)$-м ярусами

$$z_j t_{j-1,q} C_{j-1} \mathbf{1}_j = \sum_{i=1}^{q} t_{j,i} B_j \mathbf{1}_{j-1}$$

с учетом (7.7.22) и структуры C_{j-1} сводится к

$$z_j \lambda t_{j-1,q} \mathbf{1}_{j-1} = z_j \Big(\sum_{i=1}^{q} h_i \Big) B_j \mathbf{1}_{j-1} + x_j \Big(\sum_{i=1}^{q} g_i \Big) B_j \mathbf{1}_{j-1}.$$

Полагая $z_j = c x_j$, находим

$$c = \Big(\sum_{i=1}^{q} g_i \Big) B_j \mathbf{1}_{j-1} \Big/ \Big(\lambda t_{j-1,q} \mathbf{1}_{j-1} - \Big(\sum_{i=1}^{q} h_i \Big) B_j \mathbf{1}_{j-1} \Big).$$

Нормировка в пределах слоя векторов вероятностей, представленных в форме (7.7.22), дает последнее уравнение:

$$x_j = 1 \Big/ \Big(c \sum_{i=1}^{q} h_i + \sum_{i=1}^{q} g_i \Big) \mathbf{1}_j.$$

Теперь можно описать шаг большой итерации в этой модели при $j = N-1,\ N-2, \ldots, 1$:

1. Вычислить β' и $\{\beta''_i\}$ согласно (7.7.18).
2. Вычислить F по формуле (7.7.21).
3. Вычислить векторы $\{h_i\}$ и $\{g_i\}$ согласно (7.7.23)–(7.7.24); одновременно копить их суммы по i и суммы компонент s_g и s_h.
4. Найти коэффициент c.
5. С помощью s_g и s_h вычислить x_j и затем $z_j = c x_j$.
6. Согласно (7.7.22) вычислить $\{t^{(m)}_{j,i}\}$.

Для нулевого слоя $\beta' = 0$, $F = \lambda D_0^{-1} = I$ и формула (7.7.22) сводится к

$$t_{0,i} = x_0 \sum_{m=1}^{i} \beta''_m.$$

Все эти векторы при $j = 0$ фактически являются скалярами,

$$\sum_{i=1}^{q} t_{0,i} = x_0 \sum_{i=1}^{q} \sum_{m=1}^{i} \beta''_m = x_0 \sum_{m=1}^{q} (q - m + 1)\beta''_m$$

и потому

$$x_0 = 1 \Big/ \sum_{m=1}^{q} (q - m + 1)\beta''_m. \tag{7.7.25}$$

Начальные значения вероятностей при $j < n$ можно равномерно распределить между ярусами слоя. Предельные вектора вероятностей связаны системой уравнений

$$\begin{array}{rcll} t_1 D & = & \lambda t_q/x + x t_1 B, & \\ t_i D & = & \lambda t_{i-1} + x t_i B, & i = \overline{2, q}. \end{array} \tag{7.7.26}$$

7.7.5. Модель $E_q/E_k/n$

Отличие этой модели от $E_q/H_k/n$ аналогично установленному в разд. 7.7.2.

7.7.6. Модель $P/H_k/n$

Данная модель отличается от $E_q/H_k/n$ возможностью прибытия заявки сразу во вторую фазу поступления с вероятностью $\bar{y} = 1 - y$. Соответственно исходные уравнения баланса принимают вид

$$\begin{array}{rcll} \gamma_{0,1} D_0 & = & \gamma_{1,1} B_1, & \\ \gamma_{0,i} D_0 & = & \lambda \gamma_{0,i-1} + \gamma_{1,i} B_1, & i = \overline{2, q}; \\ \gamma_{j,1} D_j & = & y \gamma_{j-1,q} A_{j-1} + \gamma_{j+1,1} B_{j+1}, & \\ \gamma_{j,2} D_j & = & \bar{y} \gamma_{j-1,q} A_{j-1} + \lambda \gamma_{j,1} + \gamma_{j+1,2} B_{j+1}, & \\ \gamma_{j,i} D_j & = & \lambda \gamma_{j,i-1} + \gamma_{j+1,i} B_{j+1}, & i = \overline{3, q}; \qquad j = 1, 2, \ldots \end{array}$$

Переходя к условным векторам вероятностей микросостояний, имеем

$$\begin{array}{rcll} t_{0,1} D_0 & = & x_0 t_{1,1} B_1, & \\ t_{0,i} D_0 & = & \lambda t_{0,i-1} + x_0 t_{1,i} B_1, & i = \overline{2, q}; \\ t_{j,1} D_j & = & z_j y t_{j-1,q} A_{j-1} + x_j t_{j+1,1} B_{j+1}, & \\ t_{j,2} D_j & = & z_j \bar{y} t_{j-1,q} A_{j-1} + \lambda t_{j,1} + x_j t_{j+1,2} B_{j+1}, & \\ t_{j,i} D_j & = & \lambda t_{j,i-1} + x_j t_{j+1,i} B_{j+1}, & i = \overline{3, q}; \qquad j = 1, 2, \ldots \end{array}$$

Воспользовавшись обозначениями (7.7.18) и (7.7.21), можно переписать уравнения для первого яруса j-го слоя как

$$\begin{array}{rcll} t_{j,1} & = & z_j y\beta' + x_j\beta''_1, & \\ t_{j,2} & = & z_j\bar{y}\beta' + t_{j,1}F + x_j\beta''_i, & \\ t_{j,i} & = & t_{j,i-1}F + x_j\beta''_i, & i = \overline{3,q}. \end{array}$$

Легко убедиться, что

$$t_{j,i} = z_j h_i + x_j g_i, \qquad i = \overline{1,q},$$

где

$$\begin{array}{rcll} h_1 & = & \beta' y, & \\ h_2 & = & \beta'(yF + \bar{y}), & \\ h_i & = & h_{i-1}F, & i = \overline{3,q}; \\ g_1 & = & \beta''_1, & \\ g_i & = & g_{i-1}F + \beta''_i, & i = \overline{2,q}. \end{array}$$

Условие баланса переходов через разрез между j-м и $(j+1)$-м слоями диаграммы имеет вид

$$x_j\Big(\sum_{i=1}^{q} t_{j+1,i}\Big)B_{j+1}\mathbf{1}_j = t_{j,q}A_j\mathbf{1}_{j+1}$$

и после подстановки выражения для $t_{j,q}$ позволяет найти связь между z_j и x_j:

$$x_j = \frac{h_q A_j\mathbf{1}_{j+1}}{(\sum\limits_{i=1}^{q} t_{j+1,i})B_{j+1}\mathbf{1}_j - g_q A_j\mathbf{1}_{j+1}} z_j = cz_j.$$

Коэффициент z_j выводим из условия нормировки:

$$z_j = 1\Big/\sum_{i=1}^{q}(h_i + cg_i)\mathbf{1}_j.$$

7.7.7. «Двойной Кокс»

Названная задача имеет особую актуальность и соответственно заслуживает более подробного разбора. Эффективная реализация алгоритма требует учета специфической структуры матриц интенсивностей перехода:

A — состоит из двух диагональных блоков (один над другим) с элементами $u\lambda_1$ и λ_2 соответственно, и ее можно не запоминать вообще;

B — блочно-диагональная с одинаковыми блоками (достаточно запомнить один блок);

C — верхняя треугольная.

Поэтому вычисление необходимых для итерационного метода стандартных произведений $A_{j-1}(D_j - C_j)^{-1}$ и $B_{j+1}(D_j - C_j)^{-1}$ удобно реализовать как действия над *клеточными* матрицами. Действия над таковыми формально реализуются аналогично обычным матричным операциям при условии согласованности структуры матриц и клеток. В интересующем нас случае клетка (i, j) матрицы-произведения вычисляется как сумма произведений блоков i-й строки первого сомножителя на блоки j-го столбца второго. Кроме того, будем иметь в виду, что умножение матрицы слева на диагональную матрицу равносильно умножению ее строк на соответствующие элементы диагонали. Поскольку B_j блочно-диагональна и состоит из двух одинаковых блоков, для получения $B_{j+1}(D_j - C_j)^{-1}$ достаточно умножить блоки второго сомножителя на стандартный блок первого. Наконец, $D_j - C_j$ и обратные им (в приводимом ниже тексте процедуры `C` и `C1` соответственно) являются верхними треугольными и могут быть разбиты на две верхних треугольных подматрицы (1-й и 3-й квадранты), полную квадратную (2-й) и нулевую (4-й). Это учитывается при перемножении *клеток* надлежащим ограничением циклов.

Выбор распределений с *двумя* фазами сделал алгоритмы формирования матриц очевидными и позволил встроить их в основной текст процедуры.

7.7.8. Сравнение результатов счета

В табл. 7.11 приведены сравнительные результаты обсчета модели $E_3/E_4/3$ при $\rho = 0.7$ с помощью четырех процедур, использующих различные комбинации фазовых аппроксимаций.

Таблица 7.11. Стационарное распределение заявок в системе $E_3/E_4/3$

j	Вариант расчета			
	$E_3/E_4/3$	$E_3/H_2/3$	$H_2/E_4/3$	$H_2/H_2/3$
0	4.03895e-2	4.05880e-2	4.09605e-2	4.12129e-2
1	2.20380e-1	2.21283e-1	2.20442e-1	2.21150e-1
2	3.38071e-1	3.35670e-1	3.36234e-1	3.34061e-1
3	2.40631e-1	2.41643e-1	2.41759e-1	2.42813e-1
4	1.06676e-1	1.06252e-1	1.07643e-1	1.07088e-1
5	3.71713e-2	3.75305e-2	3.69732e-2	3.73156e-2
6	1.16652e-2	1.19561e-2	1.13100e-2	1.16217e-2
7	3.52145e-3	3.60327e-3	3.32030e-3	3.40555e-3
8	1.05058e-3	1.05247e-3	9.64173e-4	9.63910e-4
9	3.12554e-4	3.01943e-4	2.79448e-4	2.67411e-4
10	9.29465e-5	8.57381e-5	8.09858e-5	7.33322e-5
11	2.76403e-5	2.42035e-5	2.34720e-5	1.99772e-5
12	8.21993e-6	6.81007e-6	6.80312e-6	5.42187e-6
13	2.44455e-6	1.91261e-6	1.97183e-6	1.46844e-6
14	7.26997e-7	5.36617e-7	5.71518e-7	3.97250e-7
15	2.16205e-7	1.50475e-7	1.65650e-7	1.07399e-7
16	6.42983e-8	4.21825e-8	4.80123e-8	2.90261e-8
17	1.91220e-8	1.18232e-8	1.39160e-8	7.84333e-9
18	5.68678e-9	3.31358e-9	4.03344e-9	2.11920e-9
$\bar{q}$	0.23821	0.23955	0.23614	0.23737

Хорошее согласие практически значимых вероятностей и средней длины очереди $\bar{q}$ свидетельствует о правильности как расчетных соотношений, так и их программной реализации.

Сопоставление времени счета обнаруживает заметное возрастание трудоемкости при эрланговской аппроксимации — в особенности для распределения времени обслуживания, что связано с увеличением количества микросостояний на ярусах диаграммы переходов.

7.7.9. Масштабный эффект и дробление производительности

Здесь исследовалось влияние обоих упомянутых эффектов на средние времена ожидания и пребывания в системе при сохранении коэффи-

циента загрузки $\rho = 0.8$. Первый эффект создавался одинаковым увеличением интенсивности входящего потока и интенсивности обслуживания в каждом из каналов, второй — изменением числа каналов.

Принято считать, что средние времена ожидания начала обслуживания и пребывания заявки в системе при умножении на m интенсивностей потока заявок и их обслуживания должны оставаться неизменными. Этот тезис элементарно опровергается для простейшей системы M/M/1 (см. раздел 4.2). Для исследования масштабного эффекта в *немарковских многоканальных* системах была рассчитана система $M/E_3/n$ (через факториальные моменты) и на имитационной модели (1 млн испытаний). Результаты приведены в табл. 7.12. Как видно, реакция средних времен ожидания и пребывания на масштабирование интенсивностей поступления и обслуживания заявок аналогична той, что имеет место в простейшем случае.

Таблица 7.12. Масштабный эффект

Показатель		n=1	n=2	n=3
w	Расчет	2.3944	1.1972	0.7981
	Имитация	2.3689	1.1845	0.7896
v	Расчет	4.3944	2.1972	1.4648
	Имитация	4.3689	2.1845	1.4563

На той же системе изучался эффект дробления производительности (табл. 7.13).

Таблица 7.13. Эффект дробления производительности

Показатель		n=1	n=2	n=3
Ожидание	Расчет	2.6667	2.3944	2.1941
	Имитация	2.6419	2.3689	2.1241
Пребывание	Расчет	3.6667	4.3944	5.1941
	Имитация	3.6419	4.3689	5.1241

Из табл. 7.13 следует:

- с ростом числа каналов среднее время ожидания начала обслуживания уменьшается, но незначительно;

- среднее время обслуживания изменяется обратно пропорционально интенсивности обслуживания в одном канале, т. е. возрастает, причем быстрее, чем уменьшается среднее время ожидания;
- среднее время пребывания является суммой средних времен ожидания и обслуживания и по числу каналов возрастает.

Таким образом, с точки зрения оперативности обработки информации целесообразнее иметь одно обслуживающее устройство, чем несколько с тем же суммарным быстродействием. Но, поскольку в процессе эксплуатации технических устройств необходимо проводить профилактику и ремонт, оптимальное число каналов равно двум. Такой принцип (спарки) использовался при эксплуатации машин ЕС. Другой аргумент в пользу многоканальных систем — проблема переменной (в частности, пиковой) загрузки.

7.7.10. Эффект общей очереди

Часто возникает вопрос, как при наличии нескольких обслуживающих устройств организовать очередь — отдельно к каждому устройству или общую.

Исследовались системы с простейшим входящим потоком и различными коэффициентами вариации v времени обслуживания при сложении n очередей. Расчет проводился при коэффициентах загрузки $\rho = 0.7$ и 0.9. Случай с одним каналом $n = 1$ соответствует изолированным очередям.

Таблица 7.14. Среднее время ожидания в системе с общей очередью

n	$v = 0.5$		$v = 1.0$		$v = 2.0$	
	0.7	0.9	0.7	0.9	0.7	0.9
1	1.458	5.625	2.333	9.000	5.833	22.500
2	0.611	2.679	0.961	4.263	2.334	10.574
3	0.353	1.717	0.547	2.724	1.305	6.718
4	0.232	1.245	0.357	1.969	0.839	4.828
5	0.166	0.967	0.252	1.525	0.583	3.696

Результаты табл. 7.14 показывают, что при наличии нескольких устройств гораздо выгоднее организовать общую очередь: ввиду случайного

поступления заявок при разделении очередей могут возникать ситуации, когда к одному каналу возникла очередь, а другой канал свободен. При объединении очередей система становится многоканальной, и подобные ситуации исключены. При увеличении количества «суммируемых» очередей рассматриваемый эффект затухает.

7.8. Метод матрично-геометрической прогрессии

7.8.1. Сущность метода

Для расчета *разомкнутых* систем обслуживания весьма многообещающим выглядит предложенный Ивэнсом [193] и развиваемый М. Ньютсом и его последователями (например, в [18]) метод *матрично-геометрической прогрессии* (МГП). Этот метод принципиально применим только к процессам типа QBD (quasi birth and death), в которых переходы совершаются лишь между соседними ярусами диаграммы. К типу QBD можно свести любые процессы с ординарными потоками заявок и обслуживаний, если все микросостояния с фиксированным числом заявок объединять в один ярус диаграммы.

Идея метода матрично-геометрической прогрессии заключается в переносе подхода, применявшегося при обсчете модели M/M/n, на векторно-матричные объекты. Здесь векторы вероятностей микросостояний полностью занятой системы представляются соотношением типа

$$\gamma_j = \gamma_n R^{j-n}, \qquad j = n, n+1, \ldots, \tag{7.8.1}$$

где R — *матричный* знаменатель прогрессии, спектральный радиус которого предполагается строго меньшим единицы. Выпишем одно из уравнений системы (7.4.1) для $j > n$, опуская индексы у стабилизировавшихся к этому ярусу матриц переходов:

$$\gamma_j D = \gamma_{j-1} A + \gamma_{j+1} B + \gamma_j C. \tag{7.8.2}$$

Далее в этом разделе мы для упрощения обозначений заменим $D - C$ на D. Подставив выражения векторов вероятностей состояний согласно (7.8.1), можно переписать (7.8.2) в виде

$$\gamma_{j-1} R D = \gamma_{j-1} A + \gamma_{j-1} R^2 B, \tag{7.8.3}$$

откуда следует, что искомый знаменатель прогрессии должен удовлетворять матричному квадратному уравнению

$$R^2B - RD + A = 0. \tag{7.8.4}$$

Найдя этот знаменатель и имея вектор p_n вероятностей микросостояний n-го яруса, можно согласно (7.8.1) вычислить вероятности микросостояний для $j > n$.

Ожидаемая длина очереди выражается как сумма членов геометрической прогрессии:

$$\begin{aligned}\bar{q} &= \Big[\sum_{j=n+1}^{\infty}(n-j)\gamma_n R^{j-n}\Big]\mathbf{1}_n \\ &= \gamma_n(R+R^2+R^3+\ldots)\mathbf{1}_n = \gamma_n R\Big(\frac{d}{dR}\sum_{j=1}^{\infty}R^j\Big)\mathbf{1}_n.\end{aligned}$$

Последняя сумма равна $R(I-R)^{-1}$. Окончательно

$$\bar{q} = \gamma_n R(I-R)^{-2}\mathbf{1}_n.$$

Среднее время ожидания легко получить по формуле Литтла.

Основная доля трудоемкости метода МГП приходится на расчет знаменателя. Сразу же отметим, что сходимость и трудоемкость процесса его вычисления не зависят от количества обсчитываемых ярусов диаграммы. Здесь критическим фактором является *ширина диаграммы*. Последняя для моделей с эрланговым обслуживанием очень быстро растет по числу каналов n и порядку k распределения обслуживания — см. табл. 7.1. По указанным причинам все последующие рассуждения проводятся применительно к модели $M/H_2/n$.

Варианты реализации МГП различаются способами расчета знаменателя прогрессии R и техникой вычисления начальных векторов вероятностей.

7.8.2. Расчет знаменателя прогрессии

Основная проблема реализации МГП-метода заключается в определении знаменателя прогрессии R. Уравнение вида (7.8.4) может быть решено лишь численными методами. Возможные подходы к решению полиномиальных матричных уравнений обсуждаются в [44]. Общей их

чертой является необходимость составления и частичного решения вспомогательной проблемы собственных значений — обычной или обобщенной. Затем из полученной спектральной информации конструируется решение исходного уравнения. Все они имеют очень сложную теорию и соответственно — программную реализацию, что вынудило искать решение в классе итерационных подходов.

Применительно к решению уравнения вида (7.8.4) имеет смысл рассматривать следующие методы:

- простые итерации;
- метод касательных;
- метод Эйткена;
- метод Ньютона.

Простые итерации. Методы этого типа могут быть применены в вариантах

$$R = R^2BD^{-1} + AD^{-1} \tag{7.8.5}$$

или

$$R = A(D - RB)^{-1}. \tag{7.8.6}$$

Разумеется, постоянные части матричных произведений следует вычислять однократно. С целью избежать трудоемкого обращения во втором варианте матриц дополнительно рассматривалось двух- и трехшаговое итерационное уточнение матрицы, обратной к $[I - RBD^{-1}]$, согласно [31, с. 310].

Применение **метода сверхрелаксации** восходит к работе [214], см. также [18, 190]. После прибавления к обеим частям (7.8.4) слагаемого $R\omega$ можно получить формулу

$$R = \frac{1}{\omega}[R^2B - R(D - I\omega) + A]. \tag{7.8.7}$$

Параметр ω в [190] предлагалось выбирать равным трем, что противоречит общепризнанному требованию $1 < \omega < 2$, а в [18] — как минимальный модуль диагонального элемента матрицы D_n. В ходе специально проведенного эксперимента обнаружилось, что оба этих подхода приводили к расходящемуся вычислительному процессу. Приемлемым

решением оказался выбор среднего геометрического из крайних элементов D_n. В связи с чрезвычайной чувствительностью метода к выбору ω имеет смысл уже обсуждавшаяся нами адаптивная стратегия, предложенная в статье [255][2] применительно к итеративной схеме Такахаси—Таками. В нашем случае пересчет знаменателя прогресии идет по формуле

$$R^{(m)} = \omega\tilde{R}^{(m)} + (1-\omega)\tilde{R}^{(m-1)}.$$

Здесь верхним индексом помечен номер итерации, а $\tilde{R}^m$ — приближение, полученное по схеме базового итерационного метода (например, одного из вышеперечисленных). Начальное значение $\omega = 1.15$ в дальнейшем через каждые k итераций перевычисляется на основе показателя

$$\psi = \frac{\|R^{(m)} - R^{(m-1)}\|}{\|R^{(m-1)} - R^{(m-2)}\|}.$$

Если предыдущее изменение ω привело к увеличению ψ, для последующих k шагов принимается $\omega = 1+0.75(\omega-1)$, иначе $\omega = 1+1.25(\omega-1)$. Эксперименты показали, что вышеприведенная коррекция как правило оказывается чрезмерно энергичной: лучшие результаты получались при множителях 0.9 и 1.1 соответственно. Кроме того, выявилась целесообразность более частой коррекции — через 7 шагов вместо 10.

Метод Эйткена ([60, с. 192]) широко применяется для ускорения итерационного решения *скалярных* нелинейных уравнений. Пусть результатами некоторого итерационного процесса являются

$$\begin{aligned} R_0 &= R + E, \\ R_1 &= R + EF, \\ R_2 &= R + EF^2, \\ &\cdots\cdots\cdots\cdots \end{aligned}$$

где R — искомое решение; E — погрешность; F, $|F| < 1$ — скорость ее убывания по *закону геометрической прогрессии*. Из этих уравнений можно найти поправку E и решение R.

[2] Сходная идея Л. А. Люстерника описана в [31, с. 444–445].

Разовьем этот подход применительно к решению *матричного* уравнения (7.8.4). В данном случае матричные конечные разности

$$\begin{aligned}\Delta_0 &= R_1 - R_0 = EF - E = E(F-I),\\ \Delta_1 &= R_2 - R_1 = EF^2 - EF = EF(F-I),\\ \Delta_0^{(2)} &= \Delta_1 - \Delta_0 = E(F-I)^2.\end{aligned}$$

Можно показать, что $\Delta_0^{(2)} = \Delta_0(F-I) = \Delta_0 F - \Delta_0$, откуда следует

$$F = \Delta_0^{-1}\Delta_0^{(2)} + I.$$

Теперь $\Delta_0 = E(F-I) = E\Delta_0^{-1}\Delta_0^{(2)}$. Окончательное выражение для поправки

$$E = \Delta_0(\Delta_0^{-1}\Delta_0^{(2)})^{-1} = \Delta_0(\Delta_0^{(2)})^{-1}\Delta_0. \tag{7.8.8}$$

Соответственно

$$R = R_0 - E. \tag{7.8.9}$$

Реализация процесса, начавшись со стартового приближения, должна проходить «пакетами» по три итерации. В конце каждого пакета определяются по описанной схеме поправка и новое приближение. Если норма поправки меньше указанного допуска, результат считается окончательным; в противном случае R, найденное согласно (7.8.9), становится стартовым значением для следующей тройки шагов.

Ньютоново (в линейном приближении) вычисление матричных поправок Δ к знаменателю прогрессии проводится на основе записи (7.8.4) в виде

$$(R+\Delta)(R+\Delta)B - (R+\Delta)D + A = 0. \tag{7.8.10}$$

Пренебрегая членом, содержащим квадрат поправки, имеем

$$R^2B + \Delta\cdot RB + R\Delta\cdot B - RD - \Delta\cdot D + A \approx 0,$$

или

$$\Delta\cdot(RB-D) + R\cdot\Delta\cdot B \approx RD - R^2B - A.$$

Здесь и далее точки используются как знак умножения во избежание интерпретации ΔR как приращения R.

Полагая $R^2B - RD + A = F$ — левая часть (7.8.4), можно переписать предыдущее уравнение как обобщенное матричное уравнение Риккати

$$F + \Delta \cdot R \cdot B + R \cdot \Delta \cdot B - \Delta \cdot D = 0. \qquad (7.8.11)$$

Оно является *линейным* относительно поправочной матрицы. Однако вследствие «разностороннего» вхождения в слагаемые поправочной матрицы Δ применение к решению этой системы стандартных методов исключено. Приходится расписывать (7.8.11) как систему линейных уравнений относительно *компонент* матрицы Δ:

$$\sum_k d_{ik} f_{kj} + \sum_m r_{im} \sum_k d_{mk} b_{kj} = g_{ij}.$$

Число неизвестных системы для модели $M/H_2/n$ составляет $(n+1)^2$. Трудоемкость стандартных методов решения системы линейных алгебраических уравнений пропорциональна кубу ее размера. Значит, трудоемкость ньютонова метода расчета знаменателя МГП имеет порядок $O(n^6)$, что при большом n предъявляет значительные требования к оперативной памяти и делает трудоемким каждый шаг итерации.

По указанной причине имеет смысл определять поправки в методе Ньютона с помощью итерационной формы уравнения (7.8.11):

$$\Delta = (F + R \cdot \Delta \cdot B)(D - RB)^{-1}. \qquad (7.8.12)$$

Увеличение количества «внутренних» итераций уменьшает количество «внешних». Численные эксперименты показали, что целесообразно проведение *одной* внутренней итерации.

Для получения начального приближения к поправке в правой части (7.8.12) имеет смысл предположить матрицы R и B коммутирующими. Тогда уравнение (7.8.11) можно записать в виде

$$F + \Delta(2RB - D) = 0,$$

откуда следует

$$\Delta_0 = F(D - 2RB)^{-1}.$$

Известно, что для ньютонова варианта область сходимости мала. По этой причине могут потребоваться предварительные шаги по второму итерационному варианту — (7.8.8). Количество шагов уточнения с его помощью начального приближения в целях сокращения числа ньтоновых итераций при больших n должно возрастать по n (оно задавалось равным $n/2$).

7.8.3. Начальные приближения

Для всех обсуждавшихся методов существенна проблема начальных приближений к знаменателю МГП. Обычно рекомендуется применять

$$R_0 = xI, \tag{7.8.13}$$

где x — обсуждавшееся выше предельное при $j \to \infty$ отношение суммарных вероятностей смежных ярусов, а I — единичная матрица. Значительно лучшие результаты дает

$$R_0 = A(D - A)^{-1}. \tag{7.8.14}$$

7.8.4. Расчет вероятностей микросостояний

При известном R векторы $\{\gamma_j\}$, $j = \overline{0, n}$, находим решением системы уравнений глобального баланса для микросостояний соответствующих ярусов, дополненной условием баланса заявок:

$$\begin{array}{lcll} \sum_{j=0}^{n-1}(n-j)\gamma_j \mathbf{1}_j & = & n - b/a, & \\ \gamma_0(D_0 - C_0) & = & \gamma_1 B_1, & \\ \gamma_j(D_j - C_j) & = & \gamma_{j-1}A_{j-1} + \gamma_{j+1}B_{j+1}, & j = \overline{1, n-1}, \\ \gamma_n(D_n - C_n) & = & \gamma_{n-1}A_{n-1} + \gamma_n R B_{n+1}. & \end{array} \tag{7.8.15}$$

Практически эту систему перед решением необходимо расписать по компонентам упомянутых в ней векторов — в модели $M/H_2/n$ их будет $(n+1)(n+2)/2$. Решение может быть проведено как прямым методом типа гауссова исключения, так и методом итераций. Гарантирующие получение точного решения прямые методы выглядят предпочтительнее. Размер системы для модели $M/H_2/n$ составит

$$N = \sum_{j=0}^{n}(j+1) = (n+1)(n+2)/2.$$

Следовательно, трудоемкость расчета по Гауссу этого этапа также имеет порядок $O(n^6)$.

Последующие векторы вероятностей определяются на основе (7.8.1) рекуррентно — домножением предыдущего вектора на матрицу R.

Трудоемкость определения начальных векторов вероятностей может быть снижена, если, воспользовавшись результатами [98], свести эту задачу к определению вектора γ_n. Перепишем систему (7.4.1) в виде

$$\begin{array}{rcccll} \gamma_j A_j & = & \gamma_{j+1} D_{j+1} & - & \gamma_{j+2} B_{j+2}, & j = \overline{0, n-2}, \\ \gamma_{n-1} A_{n-1} & = & \gamma_n D_n & - & \gamma_n R B_{n+1}. & \end{array} \tag{7.8.16}$$

Обозначим через A_j^* матрицу, комплексно сопряженную с A_j и транспонированную. Домножим уравнения (7.8.16) справа на A_j^*:

$$\begin{array}{rcccll} \gamma_j A_j A_j^* & = & \gamma_{j+1} D_{j+1} A_j^* & - & \gamma_{j+2} B_{j+2} A_j^*, & j = \overline{0, n-2}, \\ \gamma_{n-1} A_{n-1} A_{n-1}^* & = & \gamma_n D_n A_{n-1}^* & - & \gamma_n R B_{n+1} A_{n-1}^*. & \end{array} \tag{7.8.17}$$

С помощью (7.8.17) рекуррентно выразим начальные векторы вероятностей микросостояний через γ_n:

$$\begin{array}{rcl} \gamma_{n-1} & = & \gamma_n (D_n - R B_{n+1}) A_{n-1}^* (A_{n-1} A_{n-1}^*)^{-1}, \\ \gamma_j & = & \gamma_{j+1} D_{j+1} A_j^* (A_j A_j^*)^{-1} - \gamma_{j+2} B_{j+2} A_j^* (A_j A_j^*)^{-1}, \\ & & j = n-2, n-3, \ldots, 0. \end{array} \tag{7.8.18}$$

Для упрощения обозначений положим

$$\tilde{A}_j = A_j^* (A_j A_j^*)^{-1}$$

и введем пересчетные матрицы $\{V_J\}$, такие, что

$$\gamma_j = \gamma_n V_j, \qquad j = \overline{0, n}. \tag{7.8.19}$$

Очевидно, $V_n = I$ (единичная матрица соответствующего размера). Теперь можно переписать систему (7.8.18) в виде

$$\begin{array}{rcll} \gamma_n V_{n-1} & = & \gamma_n (D_n - R B_{n+1})\ \tilde{A}_{n-1}, & \\ \gamma_n V_j & = & \gamma_n (V_{j+1} D_{j+1} - V_{j+2} B_{j+2}) \tilde{A}_j, & j = n-2, n-3, \ldots, 0. \end{array}$$

Отметим (проверено экспериментально), что все $\{\tilde{A}_j\}$ являются подматрицами $\tilde{A}_{n-1}$. Итак, пересчетные матрицы можно определить рекуррентно с помощью системы

$$\begin{array}{rcll} V_n & = & I, & \\ V_{n-1} & = & (D_n - R B_{n+1})\ \tilde{A}_{n-1}, & \\ V_j & = & (V_{j+1} D_{j+1} - V_{j+2} B_{j+2}) \tilde{A}_j, & j = n-2, n-3, \ldots, 0. \end{array} \tag{7.8.20}$$

Для замыкания расчетной схемы осталось указать способ вычисления γ_n. Запишем условие баланса заявок, полагая суммарные вероятности j-го яруса $p_j = \gamma_j \mathbf{1}_j$:

$$\gamma_n \sum_{j=0}^{n-1} (n-j) V_j \mathbf{1}_j = n - \lambda b. \tag{7.8.21}$$

Его можно переписать в виде

$$\begin{aligned} \sum_{j=0}^{n-1} (n-j) p_j &= \sum_{j=0}^{n-1} (n-j) \sum_{i=1}^{j+1} \sum_{m=1}^{n+1} \gamma_{n,m} v_{j,m,i} \\ &= \sum_{m=1}^{n+1} \sum_{j=0}^{n-1} (n-j) \sum_{i=1}^{j+1} v_{j,m,i}. \end{aligned}$$

Еще n недостающих уравнений получим из условия баланса переходов через вертикальные разрезы на диаграмме переходов рис. 7.8,

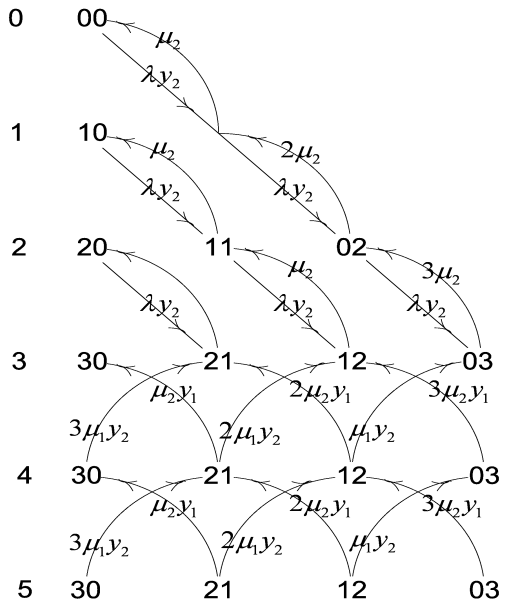

Рис. 7.8. «Диагональные» переходы

полученной из первоначальных диаграмм их объединением и удалением «строго вертикальных» переходов. Для k-го вертикального разреза

имеем условие баланса

$$\begin{aligned} & \lambda y_2 \sum_{j=k-1}^{n-1} \gamma_{j,k} + (n+1-k)\mu_1 y_2 \sum_{j=n+1}^{\infty} \gamma_{j,k} \\ = \quad & k\mu_2 \sum_{j=k}^{n} \gamma_{j,k+1} + k\mu_2 y_1 \sum_{j=n+1}^{\infty} \gamma_{j,k+1}, \qquad k = \overline{1,n}. \end{aligned}$$

Вычислим входящие в это равенство суммы с помощью МГП:

$$\sum_{j=n+1}^{\infty} \gamma_{j,k} = \sum_{j=n+1}^{\infty} \gamma_n R_k^{j-n} = \gamma_n \sum_{j=n+1}^{\infty} (R)_k^{j-n} = \gamma_n [R(I-R)^{-1}]_k,$$

(нижний индекс у матричного множителя указывает номер столбца). Соответственно уравнение для k-го разреза можно переписать в виде

$$\begin{aligned} & \gamma_n \left\{ \lambda y_2 \Big(\sum_{j=k-1}^{n-1} V_j \Big)_k + (n+1-k)\mu_1 y_2 [R(I-R)^{-1}]_k \right. \\ - \quad & \left. k\mu_2 \Big[\Big(\sum_{j=k}^{n} V_j \Big)_{k+1} + y_1 (R(I-R)^{-1})_{k+1} \Big] \right\} = 0. \end{aligned} \tag{7.8.22}$$

Решая систему уравнений (7.8.21)–(7.8.22), находим компоненты вектора γ_n. Теперь векторы $\{\gamma_j\}$ начальных вероятностей можно вычислить согласно (7.8.19), а для $j > n$ — как члены МГП.

7.8.5. «Овеществление» расчетной схемы

Как и для итерационного метода, универсальная процедура обсчета модели $M/H_2/n$ методом МГП должна быть рассчитана на работу с виртуально комплексными элементами векторов и матриц. Ее применение к задаче с *вещественными* элементами (для превышающего единицу коэффициента вариации распределения времени обслуживания) избыточно увеличивает трудоемкость вычислений, которые для МГП и без того намного более сложны. Здесь также целесообразно иметь два варианта процедуры МГП-расчета системы $M/H_2/n$: для вещественных и для комплексных параметров аппроксимации распределения длительности обслуживания. Эффективность реализации этого предложения оценивается ниже.

7.8.6. Численные эксперименты

Правильность программ контролировалась прежде всего по близости суммы вероятностей состояний системы к единице (условие нормировки вероятностей в расчетной схеме не использовалось). Результаты счета по описанным выше версиям МГП совпали с «итерационными» с высокой степенью точности.

Приведем сведения о количестве итераций и (через слэш) трудоемкости в секундах процессорного времени Pentium 4, 2.4 ГГц, для двух МГП-моделей систем обслуживания ($M/E_3/n$ — с комплексными параметрами распределения времени обслуживания, $M/H_2/n$ — с вещественными для коэффициента вариации $v = 2.0$). Точность оценки времени счета определялась разрешающей способностью системных часов (порядка 0.01 с). Обсчитывалось 70 ярусов диаграммы состояний. Точность стабилизации отношений смежных вероятностей назначалась $\varepsilon = 10^{-7}$. На пределе возможностей метода несколько начальных вероятностей порядка 10^{-12} оказывались отрицательными. Этот эффект может быть устранен повышением требований к точности расчета знаменателя МГП (соответственно — ростом трудоемкости этого расчета).

Прежде всего сопоставим обсуждавшиеся выше варианты задания начальных приближений к знаменателю МГП — табл. 7.15. В этой таблице вариант a) соответствует начальному приближению xI , a b) — $A(D-A)^{-1}$. Под «составным Ньютоном» понималось решение матричного уравнения Риккати с помощью $внутренних$ итераций.

Опыт расчетов показал, что система $M/E_3/n$ (с комплексными параметрами аппроксимации распределения обслуживания) при $n > 20$ расчету обсуждаемыми методами недоступна. Применительно к модели с вещественными параметрами удается показать эффективность предложенных в данной статье рекомендаций для бóльших значений n — табл. 7.16. Предполагалось, что начальные приближения вычислялись лучшим из представленных в табл. 7.15 способов.

В этих таблицах использованы следующие обозначения:

S — объединенная система для компонент векторов начальных вероятностей,

N — метод прогонки для расчета вектора γ_n,

I — внутренняя итерация для решения уравнения Риккати,

D — выбор параметра сверхрелаксации как квадратного корня из произведения крайних элементов диагональной матрицы D.

Таблица 7.15. Влияние начальных приближений

Расчет знаменателя	n	E_3		H_2	
		a)	b)	a)	b)
$A(D-RB)^{-1})$	3	19/0	8/0	52/0	26/0
	5	21/0	9/0	52/0	31/0
	10	26/0	17/0	54/0.016	39/0
	20	36/0.031	28/0.031	57/0.062	48/0.063
$R^2BD^{-1}+AD^{-1}$	3	36/0	54/0	76/0	74/0
	5	40/0	56/0	79/0	146/0
	10	49/0	63/0	84/0.015	139/0.015
	20	66/0.062	83/0.062	91/0.079	138/0.110
Ньютон	3	4/0	3/0	4/0	4/0
	5	3/0	2/0	4/0	4/0
	10	3/0.109	2/0.078	3/0.110	3/0.109
	20	2/4.672	4/9.328	3/6.969	3/6.985
Ньютон составной	3	9/0	3/0	28/0	12/0
	5	9/0	3/0	30/0	12/0
	10	9/0.016	5/0	30/0.016	17/0.015
	20	10/0.047	6/0.031	29/0.094	20/0.063

Таблица 7.16. Сравнение методов для модели $M/H_2/40$

Метод	Вариант	Трудоемкость
Ньютона	S	3/434.20
	N	3/414.62
	I	7/0.515
Сверхрелаксация	S	106/18.578
	N	106/0.641
	D	9/0.172

Таблица 7.17. Эффективность «овеществления» уравнения Риккати

Подпрограмма	Факторы	Число каналов n			
		5	10	15	20
Riccati	Число итераций	4	5	5	10
	Время счета, с	0.015	0.156	1.969	23.016
HalfRicc	Число итераций	4	5	5	21
	Время счета, с	0	0.125	1.297	34.187

Анализ результатов счета (в том числе не включенных в книгу из-за ограниченности ее объема) по итерационным и МГП-версиям программ приводит к следующим выводам:

1) Согласие результатов существенно различных методов подтверждают работоспособность исходной концепции, расчетных зависимостей и их программной реализации.

2) Все обсуждавшиеся версии метода МГП применимы к случаю как вещественных, так и комплексных элементов знаменателя прогрессии, хотя в последнем случае уже при $n = 20$ возникают проблемы со сходимостью или некорректностью результатов. Метод Эйткена при таких n расходится и для вещественных матриц.

3) Опыт расчетов подтвердил:

   - предпочтительность начальных приближений вида (7.8.14);
   - полезность ньютоновой схемы, в десятки раз сокращающей число итераций для решения (7.8.4); целесообразность предваряющего ее некоторого числа обычных итераций; необходимость $итерационного$ решения матричного уравнения Риккати вместо решения системы из $(n + 1)^2$ уравнений относительно компонент матричной поправки;
   - значительную экономию от применения метода прогонки при расчете начальных векторов вероятностей микросостояний (в примере для сверхрелаксации из табл. 7.16 — почти в 30 раз);
   - необходимость повышения требований к точности расчета знаменателя МГП при увеличении числа каналов (ввиду накопления погрешностей при реализации последующей прогонки);

- целесообразность применения в методе сверхрелаксации начального значения параметра ω среднего геометрического из элементов диагональной матрицы D (экономия по числу итераций в десять раз, а по трудоемкости — более чем на два порядка).

Важным результатом является также определение границ применимости метода МГП. П. Швейцер [219, с. 418] еще в 1984 г. в дискуссии на конференции по расчету характеристик вычислительных систем сказал: «Мы затратили слишком много времени на такие модели, как product-form networks and matrix-geometric solutions». Метод МГП в случае комплексных параметров, т. е. для умеренных и малых коэффициентов вариации, применим при числе каналов до 15. В случае с действительными параметрами с учетом обсуждавшихся выше предложений его можно использовать для значительно бóльших n. Отметим еще раз, что он $принципиально$ применим только к QBD (quasi birth and death) процессам — с переходами лишь между смежными ярусами диаграммы состояний.

7.9. Сопоставление итерационного и МГП-методов

Из логического анализа вычислительных схем и сопоставления результатов расчета (в том числе не представленных в этой книге из-за ограниченности места) вытекают следующие $выводы$:

1. Итерационный метод, как показали эксперименты, работает по крайней мере до $n = 100$. Его трудоемкость для комплекснозначной аппроксимации H_2-распределения обслуживания, как и для МГП, может быть уменьшена на основе симметрии (мнимые части векторов условных вероятностей микросостояний яруса должны компенсировать друг друга). Метод легко модифицируется применительно к системам с интенсивностью входящего потока, зависящей от состояния системы (в частности, замкнутым), и к системам с ограниченной очередью. Он сравнительно легко обобщается на системы, где допустимы переходы между несмежными ярусами диаграммы (например, при потоке групповых заявок). Благодаря работе с векторами относительных вероятностей и наличию этапа агрегации на каждом слое его точность практически не зависит от

числа обсчитываемых ярусов. К сожалению, его сходимость (в особенности при малых загрузке и коэффициентах вариации) не гарантируется. Впрочем, при таких данных проблема ожидания снимается сама собой. Сходимость существенно ухудшается при большом числе ярусов. В таком случае имеет смысл ограничить количество обсчитываемых ярусов диаграммы. Вероятности состояний со старшими индексами можно получить последовательным домножением их на предельное значение x, найденное по формуле (7.5.15).

Вследствие своей универсальности итерационный метод предпочтителен в стандартных процедурах расчета сетей обслуживания (прежде всего при немарковских входящих потоках с итерационным уточнением последних).

2. Неумеренно пропагандируемый метод матрично-геометрической прогрессии уступает ему по универсальности — он принципиально применим только для QBD-процессов и при умеренном числе каналов. При увеличении числа каналов необходимо повышение требований к точности расчета знаменателя прогрессии. Его сходимость не зависит от количества ярусов диаграммы. Как наиболее быстродействующий, он в области своей применимости предпочтителен при высоких требованиях к точности; в массовых расчетах; при работе с бесконечными суммами вероятностей.

7.10. Групповое прибытие заявок

Здесь мы рассмотрим ситуацию с пачками заявок *ограниченного объема* и гиперэкспоненциальной (H_2) аппроксимацией распределения обслуживания. Такая аппроксимация обеспечивает относительную наглядность рассуждений и позволяет сохранить три момента исходного распределения, что можно считать необходимым и достаточным.

Распределение количества находящихся в системе $M^X/H_2/n$ *пачек* заявок заманчиво рассчитывать аналогично одноканальному случаю, т. е. применяя H_2-аппроксимацию к распределению трудоемкости *пачки*. Далее моменты распределения ожидания пачки могут быть получены по распределению числа пачек в очереди — формула (3.2.4). Имитационные эксперименты подтвердили хорошее согласие полученных результатов с непосредственным определением ожидания головной заявки пачки.

Этот подход неявно предполагает, что все заявки каждой пачки последовательно обслуживаются в *одном и том же* канале. Тем самым исключается разделение пачки между несколькими каналами в недогруженной системе и снижается реальная производительность последней. Можно предполагать, что возникающая погрешность будет возрастать при увеличении числа каналов и уменьшении номинального коэффициента загрузки системы

$$\rho = \lambda \bar{f} b_1 / n. \tag{7.10.1}$$

Проиллюстрируем упоминавшиеся ожидаемые тенденции сопоставлением результатов расчета (Р) и имитации (И):

Таблица 7.18. Моменты распределения ожидания пачек

	$\rho = 0.8$				$\rho = 0.6$			
	$n = 3$		$n = 5$		$n = 3$		$n = 5$	
	И	Р	И	Р	И	Р	И	Р
w_1	.145e+2	.128e+2	.787e+1	.667e+1	.477e+1	.362e+1	.244e+1	.149e+1
w_2	.550e+3	.483e+3	.170e+3	.152e+3	.838e+2	.674e+2	.265e+2	.173e+2
w_3	.307e+5	.268e+5	.531e+4	.510e+4	.217e+4	.180e+4	.428e+3	.284e+3

Таким образом, описанный подход вполне может найти практическое применение, но в ограниченном диапазоне условий.

7.10.1. Дополнительные задержки в пачках

Эти задержки можно рассчитать аналогично одноканальному случаю с заменой распределения времени обслуживания распределением интервалов между выбираемыми из очереди заявками — см. разд. 2.2.3.

7.10.2. Итерационный метод для модели $M^X/H_2/n$

Запишем условие баланса для векторов-строк $\{\gamma_j\}$ вероятностей микросостояний системы с учетом прибытия заявок пачками случайного объема не свыше M:

$$\gamma_j D_j = \lambda \sum_{m=1}^{\hat{M}} f_m \gamma_{j-m} \prod_{i=0}^{m-1} A_{j-m+i} + \gamma_{j+1} B_{j+1}. \tag{7.10.2}$$

Здесь все матрицы $\{A_j\}$ разделены на интенсивность λ потока пачек. Предельный индекс суммирования $\hat{M} = \min\{j, M\}$, так что для нулевого яруса правая часть (7.10.2) сводится к $\gamma_1 B_1$.

Переходя к векторам $\{t_j\}$ условных вероятностей микросостояний, с использованием отношений вероятностей смежных ярусов имеем

$$t_j D_j = \lambda \sum_{m=1}^{\hat{M}} f_m t_{j-m} \Big(\prod_{i=0}^{m-1} z_{j-i} \Big) \Big(\prod_{i=0}^{m-1} A_{j-m+i} \Big) + x_j t_{j+1} B_{j+1}. \quad (7.10.3)$$

Таким образом, у нас снова

$$t_j = z_j \beta'_j + x_j \beta''_j, \quad (7.10.4)$$

где

$$\beta'_j = \lambda \Big[\sum_{m=1}^{\hat{M}} f_m t_{j-m}^{(k-1)} \Big(\prod_{i=1}^{m-1} z_{j-i} \Big) \Big(\prod_{i=0}^{m-1} A_{j-m+i} \Big) \Big] D_j^{-1}, \quad (7.10.5)$$

$$\beta''_j = t_{j+1}^{(k)} B_{j+1} D_j^{-1}. \quad (7.10.6)$$

Переход к условным вероятностям связан с введением для каждого яруса дополнительных неизвестных $\{x_j\}$ и $\{z_j\}$. Соответственно нужны дополнительные уравнения для их определения. В качестве первого выберем баланс переходов между j-м и $(j+1)$-м ярусами. Интенсивность переходов сверху должна учитывать $\hat{M} = \min\{M-1, j\}$ ярусов выше j-го; она составит

$$\begin{aligned} \Lambda_j &= \lambda \Big[p_j + p_{j-1}(1-f_1) + p_{j-2}(1-f_1-f_2) + \ldots + p_{j-\hat{M}+1}(1 - \sum_{i=1}^{\hat{M}-1} f_i) \Big] \\ &= p_j \lambda \Big[1 + z_j \sum_{i=1}^{\hat{M}} \Big(1 - \sum_{m=1}^{i} f_m \Big) \Big(\prod_{m=1}^{i-1} z_{j-m} \Big) \Big]. \end{aligned}$$

Интенсивность переходов снизу есть $p_{j+1} t_{j+1}^{(k)} B_{j+1} \mathbf{1}_j$. Приравнивая эти количества и разделив обе части равенства на p_j, получаем уравнение

$$\lambda + \lambda z_j \Big[\sum_{i=1}^{\hat{M}} \Big(1 - \sum_{m=1}^{i} f_m \Big) \Big(\prod_{m=1}^{i-1} z_{j-m} \Big) \Big] = x_j t_{j+1}^{(k)} B_{j+1} \mathbf{1}_j. \quad (7.10.7)$$

Второе дополнительное уравнение дает условие нормировки компонент t_j, которое можно записать как

$$(z_j \beta'_j + x_j \beta''_j) \mathbf{1}_j = 1.$$

Из последнего равенства следует, что

$$x_j = \frac{1}{\beta''_j} - z_j \frac{\beta'_j \mathbf{1}_j}{\beta''_j \mathbf{1}_j}. \tag{7.10.8}$$

Подставляя (7.10.8) в (7.10.7), убеждаемся, что

$$z_j = \frac{B_{j+1}\mathbf{1}_j/\beta''_j\mathbf{1}_j - \lambda}{\lambda\Big[\sum\limits_{i=1}^{\hat{M}}\Big(1 - \sum\limits_{m=1}^{i} f_m\Big)\Big(\prod\limits_{m=1}^{i-1} z_{j-m}\Big)\Big] + B_{j+1}\mathbf{1}_j\beta'_j\mathbf{1}_j/\beta''_j\mathbf{1}_j} \tag{7.10.9}$$

Выведенные соотношения для $j = j_{\max}, j_{\max}-1, \ldots, 1$ применяются в следующей очередности:

1) получить β'_j и β''_j согласно (7.10.5) и (7.10.6);

2) найти z_j по формуле (7.10.9);

3) вычислить x_j согласно (7.10.8);

4) получить очередное приближение $t_j^{(k)}$ по формуле (7.10.4).

Поскольку на нулевом ярусе имеется всего одно микросостояние, $t_0 \equiv 1$ и в пересчете не нуждается.

Итерации можно считать законченными, когда стабилизируются значения $\{z_j\}$, т. е. будет выполнено условие $\max_j\{|z_j^{(k)} - z_j^{(k-1)}|\} \le \varepsilon$.

7.10.3. Начальные приближения

Для расчета начальных приближений вновь воспользуемся стабилизацией векторов $\{t_j\}$ и отношений смежных вероятностей при $j \to \infty$. Для предельного x в разд. 7.4 обсуждалась формула (7.5.15). В нашем случае коэффициент загрузки вычисляется согласно (7.10.1), а интервалы между заявками для второй и последующих заявок каждой пачки равны нулю, что требует коррекции выражения и для v_A^2. При простейшем потоке пачек интенсивности λ и объеме пачки m средний интервал между заявками

$$a_1(m) = \frac{1}{m}\cdot\frac{1}{\lambda} + \frac{m-1}{m}\cdot 0 = 1/(m\lambda),$$

а второй момент $a_2(m) = 2/(m\lambda^2)$. Подставляя в выражение для квадрата коэффициента вариации $v_A^2 = a_2/a_1^2 - 1$ моменты интервалов между заявками, усредненные по распределению объема пачки, находим

$$v_A^2 = 2\Big/\Big(\sum_{m=1}^{M} f_m/m\Big) - 1. \qquad (7.10.10)$$

Для определения предельного вектора t индексы при векторах, матрицах, $\{x_j\}$ и $\{z_j\}$ можно опустить, а все матрицы A, упоминаемые в уравнении (7.10.3), считать единичными и из рассмотрения исключить. Теперь уравнение (7.10.3) запишется в виде

$$tD = \lambda\Big(\sum_{m=1}^{M} f_m z^m\Big)t + xtB, \qquad (7.10.11)$$

где $z = 1/x$. Расписав это векторно-матричное уравнение покомпонентно и заменив одно из уравнений на условие нормировки компонент вектора t, получаем систему линейных алгебраических уравнений для начальных приближений к $\{t_j\}$, $j = \overline{n, j_{\max}}$. Для $j = \overline{1, n-1}$ компоненты начальных векторов $\{t_j\}$ вновь примем равновероятными.

Начальные значения $\{z_j\}$ для $j > n$ следует считать как $1/x$, где x определяется согласно (7.5.15) с подстановкой v_A^2 из (7.10.10).

Начальные $\{z_j\}$ для первых ярусов диаграммы можно получить с помощью соображений, использованных при выводе (7.10.7). Проводя горизонтальные разрезы между точками j и $j-1$, убеждаемся, что

$$p_j^{(0)} = \frac{\lambda}{\bar{\mu}_j}[p_{j-1} + (1-f_1)p_{j-2} + \ldots + (1-f_1-\ldots f_{M-1})p_{j-M+1}].$$

Набор учитываемых вероятностей для каждого разреза ограничивается требованием неотрицательности индексов. Усредненные интенсивности обслуживания $\{\bar{\mu}_j\} = j/b_1$. Вероятность $p_0^{(0)}$ предполагается равной единице. Далее вычисляются начальные приближения к отношениям вероятностей

$$z_j^{(0)} = p_{j-1}^{(0)}/p_j^{(0)}, \qquad j = \overline{1, n}.$$

7.10.4. Вероятности состояний

Условие баланса заявок для n-канальной системы в данном случае имеет вид

$$\sum_{j=0}^{n-1}(n-j)p_j = n - \lambda\bar{f}b_1, \tag{7.10.12}$$

где $\{b_1\}$ — среднее время обслуживания заявки, а $\bar{f}$ — средний объем пачки. Входящие в (7.10.12) вероятности в нашей модели последовательно выражаются через p_0 и числа $\{z_j\}$: $p_1 = p_0/p_1$, $p_2 = p_1/z_2 = p_0/(z_1z_2)$ и т. д. Следовательно,

$$p_0\Big[n + \sum_{j=1}^{n-1}(n-j)/\prod_{i=1}^{j} z_i\Big] = n - \lambda\bar{f}b_1,$$

откуда и определяется p_0. Последующие вероятности вычисляются согласно $p_j = p_{j-1}/z_j, \quad j = 1, 2, \ldots$.

7.10.5. Результаты расчета

Расчет по описанной методике выполнялся для трехканальной системы с равномерным на [0,10] распределением обслуживания и равновероятным (от 1 до 6) объемом пачек заявок. Интенсивность потока пачек выбиралась для получения коэффициента загрузки $\rho = 0.8$. Опорное отношение вероятностей $z_\infty = 1/x$ составило 1.111. Для $j_{\max} = 39$ стабилизация $\{z_j\}$ (см. табл. 7.19) с точностью 10^{-9} потребовала 44 итераций.

Таблица 7.19. Расчетные отношения смежных вероятностей

j	z_j	j	z_j	j	z_j	j	z_j
0	1.0000	7	1.1719	14	1.0929	21	1.0957
1	2.0589	8	1.0837	15	1.0940	22	1.0958
2	1.1822	9	1.0776	16	1.0952	23	1.0958
3	1.0695	10	1.0824	17	1.0958	24	1.0958
4	0.9667	11	1.0918	18	1.0958	25	1.0958
5	0.9991	12	1.0972	19	1.0956	26	1.0958
6	1.0468	13	1.0976	20	1.0956	27	1.0959

Дальнейшие $\{z_j\}$ практически совпадают с $\{z_{27}\}$ и потому не приводятся. Как видно, они очень близки к опорному. То же можно сказать и о предельных векторах вероятностей условных микросостояний (табл. 7.20).

Таблица 7.20. Компоненты предельного вектора

Номер	Стартовая	Конечная
1	.01746 +.18314i	.01501 +.18323
2	.48576 +.25945i	.48498 +.26266
3	.48258 -.26134i	.48499 -.26264
4	.01419 -.18126i	.01502 -.18323

Теперь приведем итоговое распределение числа заявок в системе (табл. 7.21) в сопоставлении с результатами имитационного моделирования (10 млн. испытаний) при тех же исходных данных.

Таблица 7.21. Распределения количества заявок в системе

	Расчет	Имитация		Расчет	Имитация
0	1.3692e-1	1.3727e-1	20	1.4154e-2	1.4266e-2
1	6.6499e-2	6.5292e-2	21	1.2917e-2	1.3041e-2
2	5.6251e-2	5.8327e-2	22	1.1788e-2	1.1928e-2
3	5.2598e-2	5.2512e-2	23	1.0757e-2	1.0876e-2
4	5.4410e-2	5.4403e-2	24	9.8164e-3	9.8637e-3
5	5.4460e-2	5.4402e-2	25	8.9579e-3	9.0459e-3
6	5.2027e-2	5.1411e-2	26	8.1745e-3	8.2060e-3
7	4.4397e-2	4.4084e-2	27	7.4594e-3	7.4814e-3
8	4.0967e-2	4.0911e-2	28	6.8069e-3	6.7880e-3
9	3.8016e-2	3.7984e-2	29	6.2115e-3	6.1825e-3
10	3.5122e-2	3.5059e-2	30	5.6681e-3	5.6444e-3
11	3.2169e-2	3.2082e-2	31	5.1723e-3	5.1133e-3
12	2.9320e-2	2.9277e-2	32	4.7198e-3	4.6397e-3
13	2.6713e-2	2.6761e-2	33	4.3070e-3	4.2310e-3
14	2.4442e-2	2.4510e-2	34	3.9302e-3	3.8639e-3
15	2.2341e-2	2.2394e-2	35	3.5864e-3	3.5175e-3
16	2.0400e-2	2.0592e-2	36	3.2726e-3	3.2098e-3
17	1.8617e-2	1.8672e-2	37	2.9864e-3	2.9245e-3
18	1.6990e-2	1.7102e-2	38	2.7251e-3	2.6566e-3
19	1.5507e-2	1.5581e-2	39	2.4867e-3	2.4467e-3

Результаты согласуются очень хорошо. Отметим также, что сумма полученных вероятностей, дополненная суммой образующих геометрическую прогрессию неучтенных членов, с точностью 10^{-7} совпала с единицей, причем условие нормировки полных вероятностей в расчетной схеме не использовалось. Это является дополнительным подтверждением правильности расчетной схемы и ее программной реализации.

Имитационная модель допускала до 100 заявок в системе. Однако сходимость численного метода при увеличении $j_{\max}$ заметно ухудшается. Поэтому целесообразно выполнять его при умеренных $j_{\max}$, а требуемое число дополнительных вероятностей рассчитывать через $z = p_{j_{\max}-1}/p_{j_{\max}}$.

7.10.6. Распределение пачек и их задержек

Выше отмечалась необходимость расчета распределения числа «непочатых» пачек в очереди. Такая задача является обратной по отношению к естественному переходу от распределения числа пачек к распределению числа заявок и как таковая обладает вычислительной неустойчивостью. Это подтвердили попытки решения соответствующего уравнения в дискретных свертках несколькими разными методами. В связи с этим ниже предлагается хорошо зарекомендовавший себя полуэмпирический метод.

Анализ результатов имитации показал, что отношение смежных вероятностей искомого распределения близко к разумно ожидаемому $Z = z_\infty^{\bar{f}}$, где $\bar{f}$ — средний объем пачки. Полные пачки в очереди отсутствуют, если заявок в очереди нет или их количество не превышает среднего остатка пачки (скажем, $\bar{k}$ — половины ее среднего значения). Тогда можно считать, что вероятность этого события

$$q_0 = \sum_{i=0}^{n+\bar{k}} p_i,$$

а остальные вероятности образуют убывающую геометрическую прогрессию с начальным членом q_1 и знаменателем Z. Замыкающую схему недостающую вероятность находим из условия нормировки:

$$q_1 = (1 - 1/Z)(1 - q_0).$$

Имея распределение числа непочатых пачек в очереди (табл. 7.22), можно найти моменты распределения времени ожидания начала обслуживания пачки — см. табл. 7.23.

Приведенные результаты убедительно свидетельствуют о возможности практического использования предложенной расчетной схемы.

Таблица 7.22. Распределения количества пачек в очереди

j	Расчет	Имитация	j	Расчет	Имитация
0	4.7316e-1	4.5458e-1	15	1.6276e-3	1.4718e-3
1	1.4443e-1	1.5136e-1	16	1.1814e-3	1.0227e-3
2	1.0484e-1	1.1139e-1	17	8.5751e-4	7.1961e-4
3	7.6095e-2	8.0331e-2	18	6.2243e-4	5.1405e-4
4	5.5234e-2	5.7622e-2	19	4.5179e-4	3.7014e-4
5	4.0092e-2	4.1471e-2	20	3.2793e-4	2.5248e-4
6	2.9101e-2	2.9731e-2	21	2.3803e-4	1.6101e-4
7	2.1123e-2	2.0981e-2	22	1.7278e-4	1.2383e-4
8	1.5332e-2	1.4868e-2	23	1.2541e-4	9.8421e-5
9	1.1129e-2	1.0666e-2	24	9.1029e-5	6.8672e-5
10	8.0779e-3	7.6942e-3	25	6.6074e-5	5.0771e-5
11	5.8634e-3	5.5641e-3	26	4.7960e-5	3.3102e-5
12	4.2559e-3	3.9279e-3	27	3.4812e-5	1.9618e-5
13	3.0892e-3	2.8407e-3	28	2.5268e-5	1.3579e-5
14	2.2423e-3	2.0432e-3	29	1.8341e-5	6.8985e-6

Таблица 7.23. Моменты распределения ожидания пачек

Способ определения	w_1	w_2	w_3
Расчет	1.4013e+1	5.4106e+2	3.1337e+4
Имитация непосредственно	1.4079e+1	5.1615e+2	2.8040e+4
Имитация через MFACT	1.4067e+1	5.1369e+2	2.7659e+4

7.11. Задача с неоднородными каналами

Методы данной главы могут быть применены к задачам не только с неоднородными заявками, но и с неоднородными каналами [95].

7.11.1. Вероятности состояний

Обозначим

λ — параметр входящего потока,

k — количество типов каналов обслуживания,

μ_i — параметр показательного распределения времени обслуживания каналом i-го типа,

n_i — количество каналов i-го типа,

n — общее число каналов.

Зададим состояние системы кортежем $\{j; d_1, d_2, \ldots, d_k\}$, первый элемент j которого указывает общее число заявок в системе, а $\{d_i\}$ — количество занятых каналов i-го типа $(d_i \le n_i)$, $i = \overline{1,k}$. Без потери общности можно считать, что типы каналов упорядочены в порядке убывания предпочтительности их занятия. На рис. 7.9 показана диаграмма переходов по прибытию заявки между состояниями 6-канальной системы с $n_1 = 2$, $n_2 = 3$, $n_3 = 1$; индексы j вынесены в обозначение номера яруса.

Для управления процессом построения ключей j-го яруса полезно иметь верхнюю оценку их количества $K = \left(\begin{matrix} m+k-1 \\ m \end{matrix} \right)$, где $m = \min\{n_{\max}, j\}$, $n_{\max} = \max\limits_i \{n_i\}$. Сразу же оговорим, что общая схема алгоритма сохраняется независимо от правила выбора очередной заявкой свободного канала — это правило полностью учитывается структурой матриц $\{A_j\}$. Отметим, что при $j \ge n$ на каждом ярусе диаграммы имеется лишь одно микросостояние. Следовательно, при использовании метода разд. 7.5 итерации для расчета векторов условных вероятностей $\{t_j\}$ и отношений смежных вероятностей $\{x_j\}$ для этих значений j не нужны. Здесь

$$x_j = \lambda / M, \tag{7.11.1}$$

На рис. 7.10 показаны переходы в той же системе по завершению обслуживания (через $M = \sum\limits_{i=1}^{k} n_i \mu_i$ обозначена максимальная производительность системы).

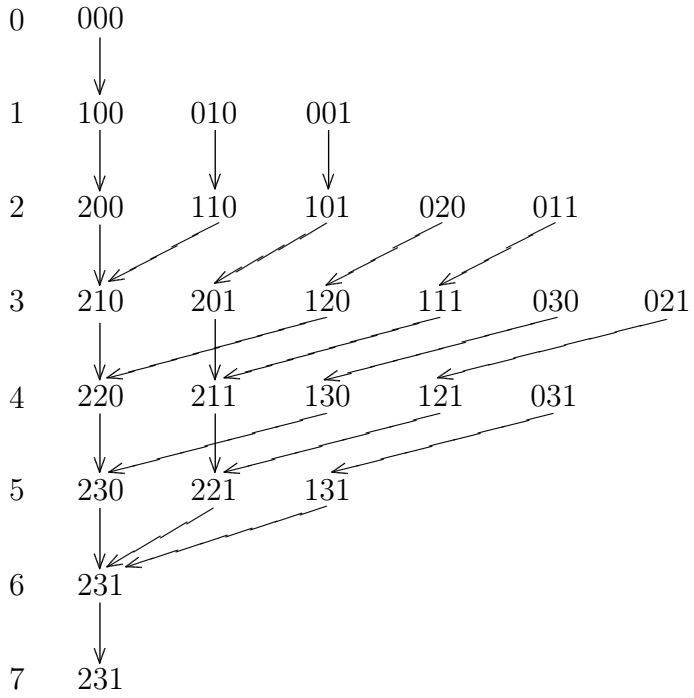

Рис. 7.9. Переходы в системе $M/\vec{M}_3/\vec{n}_3$ по прибытию заявки

После завершения итерационной части алгоритма исходная вероятность p_0 может быть определена из условия нормировки:

$$p_0 + \sum_{j=1}^{n-1} p_j + \sum_{j=n}^{\infty} p_j = 1. \tag{7.11.2}$$

Поскольку для $j > n$ $p_j = (\lambda/M)p_{j-1}$, последняя сумма оказывается геометрической прогрессией, и (7.11.2) сводится к

$$p_0 + p_0 \sum_{j=1}^{n-1} \Big(\prod_{i=0}^{j-1} x_i \Big) + \frac{p_0 \prod\limits_{i=0}^{n-1} x_i}{1 - \lambda/M} = 1,$$

откуда

$$p_0 = \Big[1 + \sum_{j=1}^{n-1} \Big(\prod_{i=0}^{j-1} x_i \Big) + \Big(\prod_{i=0}^{n-1} x_i \Big) \Big/ (1 - \lambda/M) \Big]^{-1}. \tag{7.11.3}$$

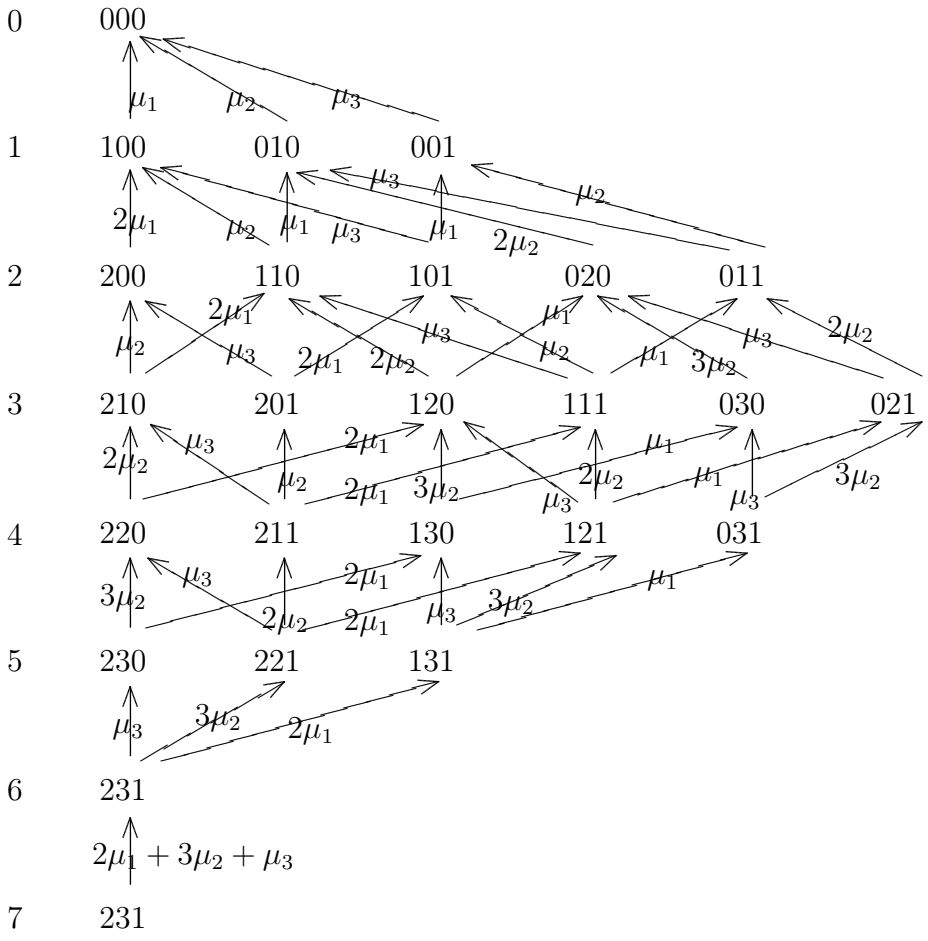

Рис. 7.10. Переходы в системе $M/\vec{M}_3/\vec{n}_3$ по завершению обслуживания

7.11.2. Варианты модели

Алгоритм легко модифицируется применительно к случаю ограниченного величиной R количества заявок в системе. При естественном условии $R > n$ все изменения сводятся к тому, что упоминавшаяся в связи с условием нормировки геометрическая прогрессия оказывается конечной, и ее сумма должна быть умножена на $1 - (\lambda/M)^{R-n}$.

Алгоритм может быть применен и для расчета *замкнутой* СМО, в которой параметр потока заявок зависит от состояния системы — обычно

предполагается $\lambda_j = \lambda(R-j)$. Это изменение прежде всего должно быть отражено в элементах матриц $\{A_j\}$ и $\{D_j\}$. Кроме того, (7.11.1) переходит в

$$x_j = \lambda_j / M. \tag{7.11.4}$$

Для замкнутой системы входящая в условие нормировки сумма «старших» вероятностей $\sum\limits_{j=n}^{R} p_j$ должна быть вычислена по общему правилу, так что вместо (7.11.3) получаем

$$p_0 = \Big[1 + \sum_{j=1}^{R} \Big(\prod_{i=0}^{j-1} x_i \Big) \Big]^{-1}. \tag{7.11.5}$$

7.11.3. Временные характеристики

При расчете моментов распределения времени пребывания заявки в данной системе приходится считаться с тем, что тип распределения времени обслуживания выбирается случайно — в зависимости от того, в какой из свободных каналов при недогруженной системе направляется вновь прибывшая заявка или какого типа канал раньше завершит обслуживание при полностью загруженной системе. Вероятности $\{\alpha_{j,i}\}$ выбора канала i-го типа при наличии в системе $j < n$ заявок подсчитываются суммированием вероятностей исходных микросостояний в соответствии с принятыми правилами предпочтения. Если, например, переходы по прибытию заявки описаны диаграммой рис. 7.9, то

$$\alpha_{3,1} = \sum_{i=3}^{6} p_{3,r}, \quad \alpha_{3,2} = p_{3,1} + p_{3,2}, \quad \alpha_{3,3} = 0.$$

Вероятности $\{q_i\}$ освобождения канала i-го типа в полностью занятой системе могут быть найдены по формуле $q_i = n_i \mu_i / M$.

Поскольку m-й момент распределения времени обслуживания в канале i-го типа $b_{i,m} = m\,!/\mu_i^m$, средневзвешенные моменты распределения времени обслуживания следует вычислять согласно

$$\begin{aligned} \bar{b}_m &= \sum_{j=0}^{n-1} p_j \sum_{i=1}^{k} \alpha_{j,i} m\,!/\mu_i^m + \Big(1 - \sum_{j=0}^{n-1} p_j\Big) \sum_{i=1}^{k} q_i m\,!/\mu_i^m \\ &= m\,! \sum_{i=1}^{k} \Big[q_i - \sum_{j=0}^{n-1} p_j (q_i - \alpha_{j,i}) \Big] / \mu_i^m. \end{aligned} \tag{7.11.6}$$

По этим моментам методами, обсуждаемыми в следующем разделе, можно получить аппроксимацию ПЛС $\delta(s)$ распределения интервалов между смежными обслуживаниями. Вновь прибывшая заявка, заставшая все каналы занятыми, будет ожидать завершения ближайшего текущего обслуживания и полного обслуживания стоящих в очереди. Обозначив через $\{q_k\}$ распределение длины очереди и $Q(z)$ — ее производящую функцию, имеем для ПЛС распределения времени ожидания выражение

$$\omega(s) = \frac{1-\delta(s)}{sd_1} \sum_{k=0}^{\infty} q_k \delta^k(s) = Q(\delta(s)).$$

Моменты распределения времени ожидания находим численным дифференцированием соответствующей таблицы.

Обобщение описанной задачи предлагается в работе [200]. Отмечается, что роль правила выбора сервера возрастает при уменьшении интенсивности трафика, увеличении числа каналов и уменьшении вариации интервалов между заявками.

7.12. Расчет выходящего потока

Задача о выходящем потоке является неотъемлемым элементом расчета *сетей обслуживания*, в которых потоки на входе узлов формируются из выходящих потоков других узлов. Вообще говоря, их нельзя считать потоками восстановления (рекуррентными), поскольку распределение интервалов между выходящими заявками зависит от количества ведущих обслуживание каналов. Мы будем считать выходной поток рекуррентным с соответственно усредненными характеристиками. В стационарном режиме выходящий поток узла сохраняет среднюю интенсивность (и средний интервал между заявками) входящего, но в общем случае меняет *вид* распределения интервалов между заявками.

Проблема строгого расчета выходящего потока была поставлена одним из отцов исследования операций Ф. Морзом в 1955 г. (студентом Морза был Д. Литтл) и частично — для многоканальных марковских систем — решена Берке [171]. Доказательство теоремы Берке см. также в [18, 190]. Для более общих случаев нужно найти способ оценки хотя бы второго момента этого распределения либо, что равноценно, дисперсии или второго коэффициента немарковости. Для последнего в [178, 202]

предложена эмпирическая формула, в наших обозначениях имеющая вид

$$\xi_2^D = \rho^2 \xi_2^B + (1-\rho^2)\xi_2^A. \tag{7.12.1}$$

Подозрение вызывает уже то, что в правой части не фигурирует число каналов n (можно предполагать, что с ростом n модуль ξ_2^D будет уменьшаться). К тому же, желательно оценить и $третий$ момент искомого распределения.

В [165] для квадрата коэффициента вариации выходящего потока приводятся три формулы:

$$v_D^2 = \rho^2(v_B^2+1) + (1-\rho)v_A^2 + \rho(1-\rho), \tag{7.12.2}$$

$$v_D^2 = 1 + \frac{\rho^2(v_B^2-1)}{\sqrt{n}} + (1-\rho)(v_A^2-1), \tag{7.12.3}$$

$$v_D^2 = -1 + \frac{\rho}{\mu}(v_B^2+1) + (1-\rho)(v_A^2+1+2\rho), \tag{7.12.4}$$

из которых вторая предложена Пюжолем, а третья — Геленбе. Все они являются аппроксимациями зависимостей, полученных на имитационных моделях.

Мы будем рассматривать проблему преобразования потока в узле обслуживания как серию $аналитических$ задач нарастающей сложности.

7.12.1. Марковские системы

В системе $M/M/n$, загруженной полностью, частное ПЛС интервала до очередного обслуживания при наличии в системе $j \geq n$ заявок есть

$$\delta_j(s) = n\mu/(n\mu+s), \qquad j = n, n+1, \ldots$$

При $j < n$ интенсивность обслуживания меняется с прибытием каждой очередной заявки. Будем интерпретировать s как параметр простейшего потока «катастроф». Тогда $\delta_j(s)$ можно считать вероятностью отсутствия катастроф за интервал от ухода заявки, оставившей в системе ровно j заявок, до следующего завершения обслуживания. Вероятности появления событий каждого из рассматриваемых простейших потоков (катастроф, завершений обслуживания и прибытия новых

заявок) пропорциональны их интенсивностям, т. е. s, $j\mu$ и λ соответственно. Теперь ясно, что

$$\delta_j(s) = \frac{j\mu}{j\mu+\lambda+s} + \frac{\lambda}{j\mu+\lambda+s}\delta_{j+1}(s), \qquad j = n-1, n-2, \ldots, 0.$$

Здесь первое слагаемое соответствует завершению обслуживания, а второе — прибытию новой заявки (отчего число заявок в системе увеличилось на единицу) и отсутствию катастроф до следующего завершения обслуживания при новом начальном условии.

Результирующее распределение интервалов между заявками выходящего потока задается ПЛС

$$D(s) = \sum_{j=0}^{n-1} \pi_j \delta_j(s) + \Big(1 - \sum_{j=0}^{n-1} \pi_j\Big)\delta_n(s),$$

где финальные вероятности $\{\pi_j\}$ наличия в системе сразу после завершения обслуживания ровно j заявок при $t \to \infty$ подсчитываются согласно

$$\pi_j = \mu_{j+1} p_{j+1} \Big/ \sum_{i=1}^{\infty} \mu_i p_i, \qquad j = 0, 1, \ldots \tag{7.12.5}$$

Здесь $\{p_j\}$ — стационарное распределение числа заявок в системе, а μ_j равно $j\mu$ при $j < n$ и $n\mu$ — в противном случае. Знаменатель формулы (7.12.5) сводится к

$$\sum_{i=1}^{\infty} \mu_i p_i = \mu\Big[\sum_{i=1}^{n-1} i p_i + n p_n/(1-\rho)\Big].$$

Моменты искомого распределения получаются многократным численным дифференцированием в нуле таблично заданной функции $D(s)$, $s = 0, h, 2h, \ldots,$ с последующей сменой знаков у нечетных производных.

Описанная методика была запрограммирована для $n = 1, 3, 5$, $\lambda = 1$ и коэффициента загрузки $\rho = 0.5, 0.9$, шага численного дифференцирования $h = 10^{-4}$, количества шагов $N = 6$. Средний интервал между выходящими заявками в полном согласии с законом сохранения заявок во всех случаях составил единицу.

В табл. 7.24 приведены отношения $\{k_i\}$ одноименных моментов распределений интервалов между заявками выходящего и входящего потоков. Отклонение результатов от единицы несущественно и может

быть полностью объяснено погрешностями численного дифференцирования. Таким образом, выходящий из системы $M/M/n$ поток обслуженных заявок является $простейшим$, что согласуется с аналитическим результатом Берке (см. [171, 246]). Это позволяет считать основную идею расчета правильной и распространить ее на более сложные немарковские системы, для которых теория расчета выходящего потока до сих пор отсутствовала.

Таблица 7.24. Сравнение «выходящих» моментов

n	ρ	Порядок моментов		
		1	2	3
1	0.5	1.00000	1.00000	1.00000
	0.9	1.00000	1.00000	1.00000
3	0.5	1.00000	1.00000	1.00027
	0.9	1.00000	1.00000	0.99999
5	0.5	1.00000	1.00000	0.99997
	0.9	1.00000	1.00000	1.00002

7.12.2. Система $M/H_k/n$

Систему $M/H_k/n$ можно интерпретировать как $\vec{M}_k/\vec{M}_k/n$ — с бесприоритетным обслуживанием k типов неоднородных заявок. Пусть $m_{j,i}^{(r)}$ — количество заявок r-го типа, которые обслуживаются при нахождении изображающей точки в i-м микросостоянии j-го яруса диаграммы переходов рис. 7.5–7.6. Тогда суммарная интенсивность завершения обслуживаний в этом микросостоянии

$$M_{j,i} = \sum_{r=1}^{k} \mu_r m_{j,i}^{(r)}, \qquad i = \overline{1, \sigma_j}.$$

Она равна сумме элементов i-й строки матрицы B_j интенсивностей переходов на вышележащий ярус.

Для микросостояний n-го и последующих ярусов ПЛС распределения времени ожидания ближайшего завершения обслуживания

$$\delta_{j,i}(s) = M_{n,i}/(M_{n,i} + s), \qquad i = \overline{1, \sigma_n}, \quad j = n, n+1, \ldots$$

Для вышележащих ярусов прибытие новой заявки, если оно произошло раньше завершения обслуживания, меняет «ключ» микросостояния и соответственно упомянутое распределение, так что

$$\delta_{j,i}(s) = \Bigl(M_{j,i} + \Lambda \sum_{l=1}^{\sigma_{j+1}} \tilde{a}_{i,l}^{(j)} \delta_{j+1,l}(s)\Bigr) \Big/ (M_{j,i} + \Lambda + s). \qquad (7.12.6)$$

Здесь Λ — суммарная интенсивность прибытия заявок, $\tilde{a}_{i,l}^{(j)}$ — вероятность перехода из i-го микросостояния j-го яруса в l-е микросостояние $(j+1)$-го в случае прибытия заявки. Вероятности $\{\tilde{a}_{i,l}^{(j)}\}$ легко выражаются через элементы $\{a_{i,l}^{(j)}\}$ матрицы A_j интенсивностей перехода на нижележащий ярус:

$$\tilde{a}_{i,l}^{(j)} = a_{i,l}^{(j)} / \Lambda,$$

так что формулу (7.12.6) можно переписать в виде

$$\begin{aligned} \delta_{j,i}(s) \;\; = \;\; & \Bigl(M_{j,i} + \sum_{l=1}^{\sigma_{j+1}} a_{i,l}^{(j)} \delta_{j+1,l}(s)\Bigr) \Big/ (M_{j,i} + \Lambda + s), \\ & i = \overline{1, \sigma_j}, \quad j = n-1, n-2, \ldots, 0. \end{aligned}$$

Для единственного микросостояния нулевого яруса можно предложить более простое выражение

$$\delta_0^{(1)}(s) = \frac{\Lambda}{\Lambda + s} \sum_{l=1}^{k} y_l \delta_{1,l}(s).$$

Итак, все $\{\delta_{j,i}(s)\}$ вновь можно определить рекуррентно.

Полученные частные ПЛС должны суммироваться с весами, равными вероятностям соответствующих микросостояний сразу после завершения обслуживания:

$$\pi_{j,i} = \sum_{l=1}^{\sigma_{j+1}} b_{l,i}^{(j+1)} \gamma_{j+1,l} \Big/ \sum_{j=0}^{\infty} \sum_{i=1}^{\sigma_j} \sum_{l=1}^{\sigma_{j+1}} b_{l,i}^{(j+1)} \gamma_{j+1,l},$$

где $\gamma_{j,l}$ — стационарная вероятность l-го микросостояния j-го яруса. Для каждого яруса вектор вероятностей микросостояний после завершения обслуживания

$$\pi_j = \gamma_{j+1} B_{j+1} \Big/ \sum_{i=1}^{\infty} \gamma_i B_i \mathbf{1}_{i-1}.$$

Здесь $\mathbf{1}_i = [1, 1, \ldots, 1]^T$ — вектор-столбец из σ_i единиц. В связи с проблемой вычисления знаменателя в последней формуле заметим, что согласно разд. 7.8

$$\sum_{i=n+1}^{\infty} \gamma_i B_i \mathbf{1}_{i-1} = \sum_{i=n+1}^{\infty} \gamma_{n+1} R^{i-n-1} B_{n+1} \mathbf{1}_n = \gamma_{n+1}(I-R)^{-1}(B_{n+1}\mathbf{1}_n),$$

где R — знаменатель матрично-геометрической прогрессии.

7.12.3. Система $H_k/H_k/n$

Пусть плотность распределения интервалов между смежными заявками рекуррентного потока

$$a(t) = \sum_{m=1}^{k} u_m \lambda_m e^{-\lambda_m t}.$$

Диаграмма состояний явится обобщением диаграммы рис. 7.5–7.6 и будет состоять из *слоев*, каждый из которых соответствует фиксированному числу j заявок в слое. Слои включают по k ярусов каждый. Для m-го яруса любого слоя интенсивность потока прибытия заявок равна λ_m, переход на r-й ярус $(j+1)$-го слоя осуществляется с вероятностью u_r, $r = \overline{1,k}$. Наборы микросостояний на всех ярусах слоя тождественны.

С учетом этих обстоятельств перепишем формулы, полученные выше для системы $M/H_k/n$, придавая ответным переменным нижний индекс слоя j и верхний индекс — яруса m:

$$\delta_{j,i}^{(m)}(s) = M_{n,i}/(M_{n,i}+s), \qquad j = n, n+1, \ldots$$

независимо от номера яруса m;

$$\begin{aligned}
\delta_{j,i}^{(m)}(s) &= \Big(M_{j,i} + \lambda_m \sum_{l=1}^{\sigma_{j+1}} \tilde{a}_{i,l}^{(j)} \sum_{r=1}^{k} u_r \delta_{j+1,l}^{(r)}(s)\Big) \Big/ (M_{j,i} + \lambda_m + s), \\
&\qquad\qquad j = n-1, n-2, \ldots, 1; \\
\delta_{0,1}^{(m)}(s) &= \frac{\lambda_m}{\lambda_m + s} \sum_{l=1}^{k} y_l \sum_{r=1}^{k} u_r \delta_{1,l}^{(r)}(s).
\end{aligned}$$

В формуле для расчета векторов вероятностей микросостояний яруса после завершения обслуживания

$$\pi_j^{(m)} = \gamma_{j+1}^{(m)} B_{j+1} \Big/ \sum_{i=1}^{\infty} \Big(\sum_{l=1}^{k} \gamma_i^{(l)}\Big) B_i \mathbf{1}_{i-1}$$

вероятности состояний с одинаковыми ключами можно суммировать в пределах слоя перед умножением на B_i.

7.12.4. Непосредственный расчет моментов

Обозначим $E(x)$ набор моментов показательного распределения с параметром x, имеющий компоненты

$$E_i(x) = i\,!/x^i, \qquad i = 1, 2, \ldots$$

Символом $*$ будем обозначать свертку распределений в моментах. Для наборов *моментов* распределений между выходящими заявками сохраним индексацию прежних обозначений, заменив лишь $\delta(s)$ на d. Тогда для модели $M/M/n$ при числе заявок в системе $j \geq n$

$$d_j = E(n\mu), \qquad j = n, n+1, \ldots$$

Перепишем формулу для $\delta_j(s)$ при $j < n$ в виде

$$\begin{aligned}\delta_j(s) &= \frac{j\mu+\lambda}{j\mu+\lambda+s}\Big(\frac{j\mu}{j\mu+\lambda}\cdot 1 + \frac{\lambda}{j\mu+\lambda}\,\delta_{j+1}(s)\Big)\\ &= \frac{j\mu}{j\mu+\lambda}\,\frac{j\mu+\lambda}{j\mu+\lambda+s} + \frac{\lambda}{j\mu+\lambda}\,\frac{j\mu+\lambda}{j\mu+\lambda+s}\delta_{j+1}(s).\end{aligned}$$

Заменяя преобразование Лапласа показательных плотностей их моментами и произведение ПЛС — сверткой, убеждаемся, что

$$d_j = \frac{j\mu}{j\mu+\lambda}E(j\mu+\lambda) + \frac{\lambda}{j\mu+\lambda}E(j\mu+\lambda) * d_{j+1}, \qquad j = n-1, n-2, \ldots, 0.$$

В частности,

$$d_0 = E(\lambda) * d_1.$$

Аналогичные рассуждения для системы $M/H_k/n$ приводят к рекуррентным выражениям для наборов моментов

$$\begin{aligned}d_{j,i} &= E(M_{n,i}), \qquad j = n, n+1, \ldots\\ d_{j,i} &= \frac{M_{j,i}}{M_{j,i}+\lambda}E(M_{j,i}+\lambda) + \frac{\lambda}{M_{j,i}+\lambda}E(M_{j,i}+\lambda) * \sum_{l=1}^{\sigma_{j+1}} a_{i,l}^{(j)} d_{j+1,l},\\ &\qquad\qquad j = n-1, n-2, \ldots, 1;\\ d_{0,1} &= \lambda^{-1}E(\lambda) * \sum_{l=1}^{\sigma_1} a\,(0)_{1,l}\, d_{1,l}.\end{aligned}$$

Соответствующие формулы для модели $H_k/H_k/n$ примут вид

$$d_{j,i}^{(m)} = E(M_{n,i}), \qquad j = n, n+1, \ldots$$

независимо от номера яруса m;

$$\begin{aligned} d_{j,i}^{(m)} &= \frac{M_{j,i}}{M_{j,i}+\lambda_m} E(M_{j,i}+\lambda_m) \\ &+ \frac{\lambda_m}{M_{j,i}+\lambda_m} E(M_{j,i}+\lambda_m) * \sum_{l=1}^{\sigma_{j+1}} \tilde{a}_{i,l}^{(j)} \sum_{r=1}^{k} u_r d_{j+1,l}^{(r)}, \\ & \qquad\qquad j = n-1, n-2, \ldots, 1; \\ d_{0,1}^{(m)} &= E(\lambda_m) * \sum_{l=1}^{k} y_l \sum_{r=1}^{k} u_r d_{1,l}^{(r)}. \end{aligned}$$

Вероятности микросостояний ярусов и слоев учитываются так же, как и при работе с преобразованиями Лапласа. Соответственно здесь при переходе от слоя к слою диаграммы для $j < n$ можно пользоваться средневзвешенными значениями моментов

$$\bar{d}_{j+1,l} = \sum_{r=1}^{k} u_r d_{j+1,l}^{(r)}.$$

7.12.5. «Дважды коксова» система

Обозначим для j-го яруса:

$M_{j,i}$ — суммарная интенсивность обслуживания в i-м микросостоянии каждой ветви;

$\hat{M}_{j,i}$ — интенсивность переходов *вверх* по завершению обслуживания;

$\vec{M}_{j,i}$ — интенсивность переходов *вправо* по завершению фазы обслуживания (сумма двух последних интенсивностей равна $M_{j,i}$);

$d_{j,i}^{(1)}, d_{j,i}^{(2)}$ — наборы моментов распределения времени до ближайшего обслуживания при соответствующем исходном состоянии (верхние индексы указывают левую и правую ветвь диаграммы соответственно);

σ_j — количество микросостояний в *одной* ветви диаграммы.

В этих обозначениях для $j \geq n$ и правой ветви диаграммы имеем

$$\begin{aligned} d^{(2)}_{n,i} &= \frac{\hat{M}_{n,i}}{M_{n,i}+\lambda_2} E(M_{n,i}+\lambda_2) \\ &+ E(M_{n,i}+\lambda_2) * \Big[\frac{\vec{M}_{n,i}}{M_{n,i}+\lambda_2} d^{(2)}_{n,i+1} + \frac{\lambda_2}{M_{n,i}+\lambda_2} d^{(1)}_{n,i} \Big], \\ & \qquad i = \sigma_n, \sigma_n - 1, \ldots, 1. \end{aligned}$$

Для тех же значений j и левой ветви диаграммы надо дополнительно учитывать возможные переходы в правую ветвь по завершению фазы прибытия заявки:

$$\begin{aligned} d^{(1)}_{n,i} &= \frac{\hat{M}_{n,i}}{M_{n,i}+\lambda_1} E(M_{n,i}+\lambda_1) \\ &+ E(M_{n,i}+\lambda_1) * \Big[\frac{\vec{M}_{n,i}}{M_{n,i}+\lambda_1} d^{(1)}_{n,i+1} + \frac{u\lambda_1}{M_{n,i}+\lambda_1} d^{(2)}_{n,i} + \frac{\bar{u}\lambda_1}{M_{n,i}+\lambda_1} d^{(1)}_{n,i} \Big], \\ & \qquad i = \sigma_n, \sigma_n - 1, \ldots, 1. \end{aligned}$$

После выполнения сверток в моментах замечаем, что для каждого i искомые $\{d_{n,i}\}$ входят в обе части приведенных уравнений, причем «с перекрестом». Поэтому для их определения приходится последовательно решать σ_n систем из 6 линейных алгебраических уравнений (по три момента в обеих ветвях на каждое микросостояние).

Для «ненасыщенных» ярусов $j = n-1, n-2, \ldots, 1$ в правой ветви диаграммы прибытие заявки приводит в одноименные состояния левой ветви нижележащего яруса. Итак, имеем рекуррентные формулы

$$\begin{aligned} d^{(2)}_{j,i} &= \frac{\hat{M}_{j,i}}{M_{j,i}+\lambda_2} E(M_{j,i}+\lambda_2) \\ &+ E(M_{j,i}+\lambda_2) * \Big[\frac{\vec{M}_{j,i}}{M_{j,i}+\lambda_2} d^{(2)}_{j,i+1} + \frac{\lambda_2}{M_{j,i}+\lambda_2} d^{(1)}_{j+1,i} \Big], \\ & \qquad i = \sigma_j, \sigma_j - 1, \ldots, 1. \end{aligned}$$

Для левой ветви зоны ненасыщенных состояний завершение фазы прибытия заявки может привести как в правую ветвь диаграммы, так и

в нижележащий ярус:

$$\begin{aligned} d_{j,i}^{(1)} &= \frac{\hat{M}_{j,i}}{M_{j,i}+\lambda_1} E(M_{j,i}+\lambda_1) \\ &+ E(M_{j,i}+\lambda_1) * \left[\frac{\vec{M}_{j,i}}{M_{j,i}+\lambda_1} d_{j,i+1}^{(1)} + \frac{\bar{u}\lambda_1}{M_{j,i}+\lambda_1} d_{j+1,i}^{(1)} + \frac{u\lambda_1}{M_{j,i}+\lambda_1} d_{j,i}^{(2)} \right], \\ & \qquad i = \sigma_j, \sigma_j - 1, \ldots, 1. \end{aligned}$$

Здесь первое слагаемое соответствует завершению обслуживания, а последующие — переходам по завершению первой фазы обслуживания и фазы прибытия заявки (окончательной или с переходом во вторую фазу ожидания прибытия) с рекуррентным учетом распределения времени до ближайшего обслуживания во вновь возникшем состоянии.

Наконец, в случае полного освобождения системы до очередного завершения обслуживания сначала придется дождаться прибытия хотя бы одной заявки. Здесь

$$\begin{aligned} d_{0,1}^{(2)} &= E(\lambda_2) * d_{1,1}^{(1)}, \\ d_{0,1}^{(1)} &= E(\lambda_1) * (\bar{u} d_{1,1}^{(1)} + u d_{0,1}^{(2)}). \end{aligned}$$

В «ненасыщенной» зоне искомые наборы моментов определяются последовательно через ранее вычисленные: для каждого i сначала вычисляются «правые», а затем «левые» моменты.

7.12.6. Численный эксперимент

Обсудим численные эксперименты над алгоритмом преобразования потоков в системе вида $H_2/H_2/n$ (гиперэкспонентой заменялись гамма-распределения с параметрами α и β соответственно). Прежде всего отметим, что *средний* интервал между обслуженными заявками во всех случаях оказался равным среднему интервалу a_1 между заявками входящего потока. Далее, табл. 7.25 подтверждает низкое качество приближения (7.12.1) для коэффициента немарковости выходящего потока.

Рис. 7.11 показывает влияние коэффициента загрузки ρ (ось абсцисс) и числа каналов (метки на кривых) на коэффициент немарковости ξ_2^D интервалов между заявками выходящего потока для системы $M/H_2/n$. При $n = 1$ и $\rho \to 1$, как и следовало ожидать, $\xi^D \to \xi^B$ (для верхнего пучка кривых $\xi_2^B = 3$, для нижнего -0.8).

Таблица 7.25. Сравнение расчета и аппроксимации ξ_2^D

β	α	Один канал				Два канала			
		$\rho = 0.6$		$\rho = 0.9$		$\rho = 0.6$		$\rho = 0.9$	
		Точно	Аппр.	Точно	Аппр.	Точно	Аппр.	Точно	Аппр.
0.5	0.3	1.540	1.853	1.112	1.253	1.223	1.853	0.838	1.253
	0.5	0.859	1.000	0.938	1.000	0.629	1.000	0.688	1.000
	1.0	0.360	0.360	0.810	0.810	0.204	0.360	0.576	0.810
	3.0	0.040	-0.067	0.727	0.683	-0.069	-0.067	0.502	0.683
	∞	-0.114	-0.280	0.687	0.620	-0.203	-0.280	0.465	0.620
3.0	0.3	1.007	1.253	-0.223	-0.097	1.017	1.253	-0.102	-0.097
	0.5	0.295	0.400	-0.403	-0.350	0.342	0.400	-0.268	-0.350
	1.0	-0.240	-0.240	-0.540	-0.540	-0.141	-0.240	-0.386	-0.540
	3.0	-0.598	-0.667	-0.637	-0.667	-0.438	-0.667	-0.461	-0.667
	∞	-0.777	-0.880	-0.692	-0.730	-0.572	-0.880	-0.499	-0.730

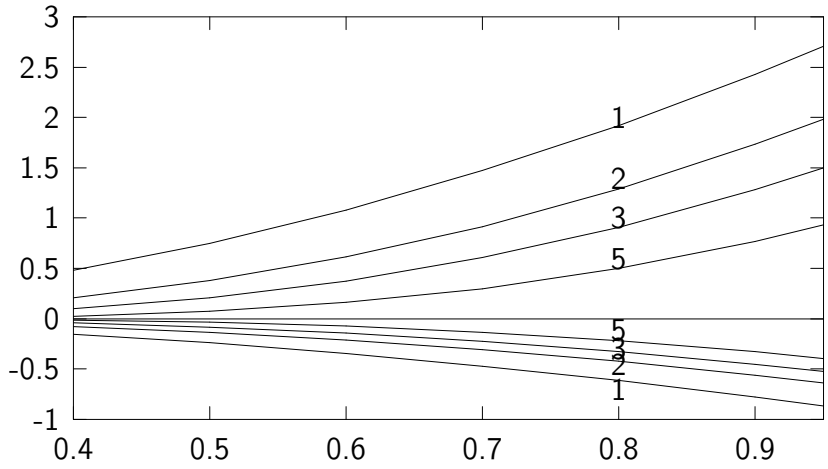

Рис. 7.11. Влияние загрузки на выходящий поток в $M/H_2/n$

При показательном распределении обслуживания ($\xi^B = 0$) и $\rho \to 1$ независимо от вида входящего потока к простейшему будут стремиться как выходящий поток каждого канала, так и их сумма, причем последняя тем быстрее, чем больше число каналов. Это предположение подтверждает рис. 7.12 (для верхнего пучка кривых $\xi_2^A = 3$, для нижнего -0.8).

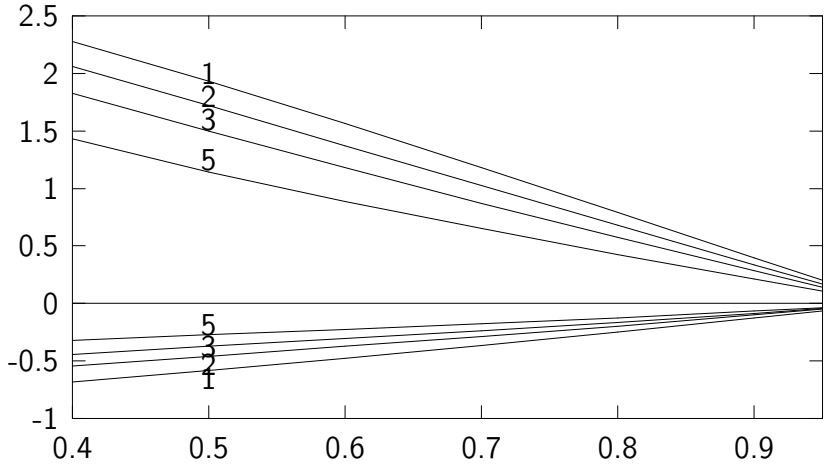

Рис. 7.12. Влияние загрузки на выходящий поток в $H_2/M/n$

Увеличение числа каналов обслуживания при сохранении коэффициента загрузки системы во всех случаях приближает выходящий поток к простейшему.

Важное значение имеет установленная в широком диапазоне условий практическая линейность зависимости коэффициентов немарковости выходящего потока от аналогичных коэффициентов входящего — см. рис. 7.13 для ξ_2^D и рис. 7.14 для ξ_3^D (число каналов и вариабельность обслуживания брались существенно различными).

В целом, исключая случаи большого числа каналов или одновременной близости обоих исходных распределений к показательным, поток обслуженных системой заявок заметно отличается от входящего, что делает задачу преобразования потока в узле необходимым элементом расчета сети обслуживания.

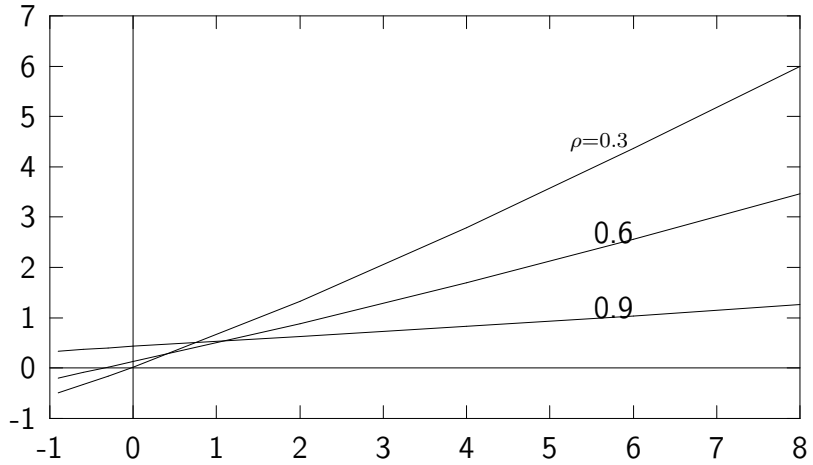

Рис. 7.13. Связь между ξ_2^D и ξ_2^A

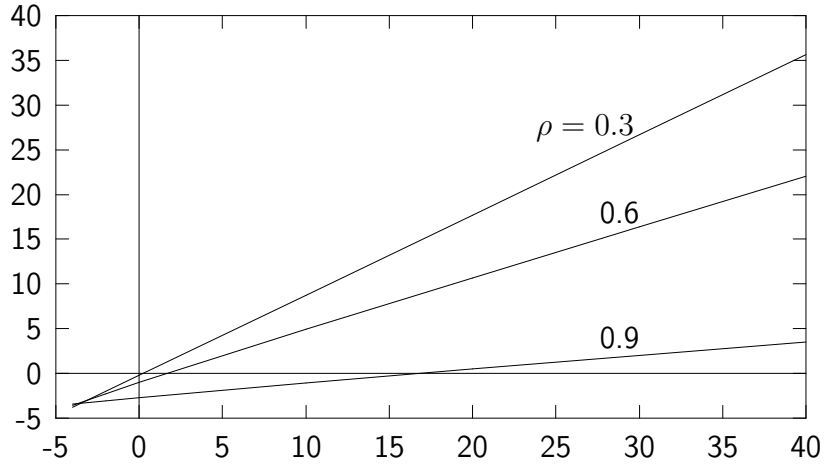

Рис. 7.14. Связь между ξ_3^D и ξ_3^A

7.12.7. Поток групповых заявок

Можно предполагать, что поток обслуженных заявок, прибывавших группами, будет иметь существенно меньший коэффициент вариации, чем входящий поток. Это обстоятельство смягчит «пачечный» эффект для узла-преемника. Методика расчета выходящего потока для узла вида $M^X/H_2/n$, принимающего пачки заявок, естественно обобщает форму-

лы (7.12.7) – (7.12.7):

$$d_{j,i} = E(M_{n,i}), \qquad j = n, n+1, \ldots \tag{7.12.7}$$

$$d_{j,i} = \frac{M_{j,i}}{M_{j,i}+\lambda} E(M_{j,i}+\lambda) + \frac{\lambda}{M_{j,i}+\lambda} E(M_{j,i}+\lambda) * \sum_{m=1}^{M} f_m \sum_{l=1}^{\sigma_{j+m}} A_{i,l}^{j,j+m} d_{j+m,l},$$
$$j = n-1, n-2, \ldots, 1; \tag{7.12.8}$$

$$d_{0,1} = E(\lambda) * \sum_{m=1}^{M} f_m \sum_{l=1}^{\sigma_m} A_{i,l}^{0,m} d_{m,l}. \tag{7.12.9}$$

Здесь $A_{i,l}^{j,j+m}$ есть элемент (i,l) матричного произведения $A_j A_{j+1} \ldots A_{j+m-1}$, задающего вероятности перехода по прибытию заявок с j-го на $(j+m)$-й ярус. Эти произведения целесообразно вычислить заранее — до начала рекуррентного счета. Программная реализация данной схемы требует специального рассмотрения переходов на ярусы ниже n-го.

Порядок взвешивания полученных частных моментов аналогичен рассмотренному для $M/H_k/n$.

Приведем результаты расчета выходящего потока в сопоставлении с имитационной моделью (10 млн. испытаний) при уже упоминавшихся исходных данных (табл. 7.26):

Таблица 7.26. Распределение интервалов между выходящими заявками

Способ получения	Моменты		
	1	2	3
Имитация	2.0842	1.1440e+1	1.6202e+2
Расчет	2.0833	1.1304e+1	1.6141e+2

Программа позволила сопоставить коэффициенты немарковости ξ_2 и вариации v входящего группового потока и выходящего в исследованной системе обслуживания (табл. 7.27). Выходящий поток оказывается заметно более близким к простейшему, чем входящий.

Таблица 7.27. Эффект сглаживания потока

Поток	Показатели	
	ξ_2	v
Входящий	2.898	1.974
Выходящий	0.604	1.267

Глава 8

Временные характеристики систем обслуживания

8.1. Общие соображения

Работа наиболее ответственных СМО оценивается главным образом по их оперативности, измеряемой различными временны́ми показателями (см. разд. 2.5). Ключом к решению этих проблем является расчет моментов $\{w_k\}$ распределения времени ожидания начала обслуживания при условии ненулевого ожидания. Далее они пересчитываются в безусловные умножением на вероятность ненулевого ожидания Π_0. Свернув результаты с моментами $\{b_k\}$ чистой длительности обслуживания (см. разд. 1.6), можно получить моменты $\{v_k\}$ распределения времени пребывания заявки в системе. По этим моментам методами разд. 1.10 строится ДФР.

В соответствии с изложенной схемой в данном разделе основное внимание уделяется расчету моментов распределения времени ожидания для СМО рассмотренных в предыдущих главах типов. Во всех случаях предполагается дисциплина очереди FCFS.

Для любых одноканальных систем, а также для многоканальных с регулярным обслуживанием, сохраняющих принцип FCFS для системы в целом, моменты распределения времени пребывания заявки в системе можно получать непосредственно.

8.2. Распределение числа заявок перед прибытием очередной заявки

Время ожидания вновь прибывшей заявки определяется состоянием системы непосредственно перед ее прибытием. Очевидно, распределение $\{\pi_j\}$ соответствующих вероятностей совпадает с распределением вероятностей состояний сразу после завершения обслуживания. Расчет последних уже обсуждался в связи с проблемой выходящего потока.

Для расчета моментов времени ожидания необходимо знать условное распределение длины очереди при ненулевом ожидании. Вероятность ненулевого ожидания

$$\Pi_0 = 1 - \sum_{i=0}^{n-1} \pi_j, \tag{8.2.1}$$

распределение числа заявок в очереди

$$\tilde{\pi}_j = \pi_{j+n}/\Pi_0, \qquad j = 0, 1, \ldots, \tag{8.2.2}$$

векторы условных вероятностей подсчитываются согласно

$$\begin{array}{rcll} \tilde{\alpha}_0 & = & \left(\sum\limits_{i=0}^{n} \alpha_j \right) \Big/ \Pi_0, & \\ \tilde{\alpha}_j & = & \alpha_{j+n}/\Pi_0, & j = 1, 2, \ldots \end{array} \tag{8.2.3}$$

Для реализации описываемых ниже методов существенно знать условное математическое ожидание длины очереди

$$\tilde{q} = \sum_{j=1}^{\infty} j\tilde{\pi}_j. \tag{8.2.4}$$

С его помощью по формуле Литтла определяем среднее значение условного времени ожидания

$$\tilde{w}_1 = a\tilde{q}, \tag{8.2.5}$$

где a — средний интервал между входящими заявками.

8.3. Простейший входящий поток

При простейшем входящем потоке моменты распределения времени ожидания (пребывания) выражаются через факториальные моменты распределения $\{\tilde{\pi}_j\}$ — соответственно $\{\pi_j\}$ — формулой (3.2.4).

8.4. Метод сверток

8.4.1. Показательное распределение обслуживания

При показательном распределении длительности обслуживания заявка, заставшая в очереди k заявок, будет ждать $(k+1)$-кратного завершения обслуживания, каждое из которых имеет показательно распределенную длительность с параметром $n\mu$. Соответственно ПЛС условного распределения времени ожидания

$$\tilde{\omega}(s) = \sum_{j=0}^{\infty} \tilde{\pi}_j \Big(\frac{n\mu}{n\mu + s} \Big)^{j+1}.$$

Фактически этот расчет, учитывающий и вид распределения $\{\tilde{\pi}_j\}$, был проделан в разд. 5.6 и 5.7.1. Используем этот подход как базу для более общего случая.

8.4.2. Обслуживание фазового типа

Введем следующие вспомогательные обозначения для режима полной занятости СМО:

$n_{i,j}$ — значение j-й позиции i-го ключа,

$M_i = \sum\limits_{j=1}^{k} \mu_j n_{i,j}$ — суммарная интенсивность обслуживания в i-м микросостоянии, k — разрядность ключа, $i = \overline{1, \sigma_n}$;

$M^{(c)}(s)$ — вектор-столбец с элементами $M_i/(M_i + s)$;

$M^{(d)}(s)$ — диагональная матрица с такими же элементами,

T — матрица вероятностей перехода по завершению обслуживания.

Тогда ПЛС условного распределения времени ожидания можно записать как

$$\tilde{\omega}(s) = \sum_{j=0}^{\infty} \tilde{\alpha}_j [M^{(d)}(s) T]^j M^{(c)}(s). \qquad (8.4.1)$$

Ограничив диапазон изменения j, можно рассчитать таблицу значений $\tilde{\omega}(s)$ для $s = 0, h, 2h, \ldots, nh$ и численным дифференцированием получить моменты распределения времени ожидания — см. формулу (1.6.3).

Расчет по этой схеме возможен также при ограниченной очереди и для замкнутых систем. Свертки могут быть выполнены непосредственно в моментах по аналогии с разд. 7.12.4.

Для разомкнутых систем с простейшим входящим потоком, анализируемых методом разд. 7.8, возможны дальнейшие упрощения. Полагая в (8.4.1) $\tilde{\alpha}_j = \tilde{\alpha}_0 R^j$, получаем

$$\tilde{\omega}(s) = \sum_{j=0}^{\infty} \tilde{\alpha}_0 [RM^{(d)}(s)T]^j M^{(c)}(s) = \tilde{\alpha}_0 [I - RM^{(d)}(s)T]^{-1} M^{(c)}(s). \tag{8.4.2}$$

Расчет моментов $\{\tilde{\omega}_i\}$ также выполняется посредством численного дифференцирования. Впрочем, в рассматриваемом частном случае лучше воспользоваться соотношениями (3.2.4), обобщающими формулу Литтла на высшие моменты.

8.4.3. Агрегированный вариант

Описанный выше метод опирается на детальный расчет системы (знание вероятностей микросостояний). Его реализация внутри стандартной процедуры расчета СМО с распределениями фазового типа, если временные характеристики не нужны, приводит к избыточным вычислениям. Вынесение же расчета в отдельную или главную процедуру требует существенного увеличения объема выдаваемой стандартной процедурой ответной информации.

Значительно меньший объем информации использует *агрегированный* вариант метода свертки, который опирается на представление ДФР распределения интервалов между обслуживаниями в постоянно занятой n-канальной системе формулой (2.2.3). Основная трудность состоит в расчете моментов искомого распределения. Возможен вариант такого расчета, опирающийся на распределение Вейбулла с поправочным многочленом (1.10.1). Проще, однако, воспользоваться гиперэкспоненциальной аппроксимацией:

1. Аппроксимировать $\bar{B}(t)$ согласно (1.8.5).
2. По формулам разд. 1.8.3 найти коэффициенты такой же аппроксимации случайной модификации этого распределения.
3. Применяя технику суммирования потоков из разд. 2.1.9, найти моменты распределения с ДФР $\bar{B}_n(t)$.

4. Согласно (1.7.3) получить моменты остаточного по отношению к нему $\bar{F}_n(t)$.

5. По этим моментам построить аппроксимации плотностей соответствующих распределений вида (1.9.1), обеспечивающие удобный расчет их преобразований Лапласа $\beta_n(s)$ и $\varphi_n(s)$.

Таким образом, основная расчетная формула агрегированного варианта метода сверток

$$\tilde{\omega}(s) = \sum_{j=0}^{\infty} \tilde{\pi}_j \varphi_n(s) \beta_n^j(s) = \varphi_n(s) \sum_{j=0}^{\infty} \tilde{\pi}_j \beta_n^j(s) = \varphi_n(s) Q(\beta_n(s)), \quad (8.4.3)$$

где $Q(\cdot)$ — производящая функция условного распределения длины очереди.

8.5. «Интегральный» метод

Основой этого метода является закон сохранения стационарной очереди в форме интегрального уравнения (3.2.1). ПЛС распределения интервалов между заявками просеянного потока «синих» заявок можно получить из (2.1.17) заменой z на $1 - z$:

$$\alpha_z(s) = \frac{(1-z)\alpha(s)}{1 - z\alpha(s)}. \quad (8.5.1)$$

Соответственно в формулу (3.2.1) нужно подставлять

$$\bar{A}_z(t) = L^{-1}\Big\{\frac{1}{s}\Big(1 - \frac{(1-z)\alpha(s)}{1-z\alpha(s)}\Big)\Big\} = L^{-1}\Big(\frac{1-\alpha(s)}{s(1-z\alpha(s))}\Big). \quad (8.5.2)$$

Приведем результаты этих выкладок для H_2- и E_2-потоков и аппроксимацию общего случая.

8.5.1. Система $H_2/G/n$

В этом случае $\alpha(s) = C_1\lambda_1/(\lambda_1 + s) + C_2\lambda_2/(\lambda_2 + s)$. Проделав соответствующие подстановки, можно показать, что в обозначениях

$$\begin{aligned} p &= [\lambda_1 + \lambda_2 - z(C_1\lambda_1 + C_2\lambda_2)]/2, \\ q &= (1-z)\lambda_1\lambda_2, \\ a &= p - \sqrt{p^2 - q}, \qquad b = p + \sqrt{p^2 - q}, \\ F_1 &= (1-z)(C_1\lambda_1 + C_2\lambda_2) - a \\ F_2 &= (1-z)(C_1\lambda_1 + C_2\lambda_2) - b \end{aligned} \tag{8.5.3}$$

общая формула (3.2.1) сводится к

$$\tilde{\Pi}(z) = \int_0^\infty \frac{1}{2\sqrt{p^2 - q}}(F_1 e^{-bt} - F_2 e^{-at})\tilde{w}(t)\,dt. \tag{8.5.4}$$

8.5.2. Система $E_2/G/n$

Как показано в разд. 1.8.3, распределение Эрланга второго порядка по отношению к H_2-аппроксимации является особым случаем. Здесь $\alpha(s) = [\lambda/(\lambda + s)]^2$, и аргументом правой части (8.5.2) становится выражение

$$\frac{\lambda^2(1-z)}{s[s^2 + 2\lambda + \lambda^2(1-z)]}.$$

Вычисляя обратное преобразование Лапласа, получаем конечный результат в форме

$$\tilde{\Pi}(z) = \int_0^\infty \frac{1}{2\sqrt{z}}\Big[(1+\sqrt{z})e^{-\lambda(1-\sqrt{z})t} - (1-\sqrt{z})e^{-\lambda(1+\sqrt{z})t}\Big]\tilde{w}(t)\,dt. \tag{8.5.5}$$

8.5.3. Численный алгоритм для двухфазных потоков

Сравнение формул (8.5.4) и (8.5.5) показывает, что в обоих случаях решение может быть представлено в форме

$$\tilde{\Pi}(z) = \int_0^\infty \Big[D_1(z)e^{-\lambda_1(z)t} - D_2(z)e^{-\lambda_2(z)t}\Big]\tilde{w}(t)\,dt. \tag{8.5.6}$$

Проведем параметризацию задачи, представив $\tilde{w}(t)$ в виде гамма-плотности с поправочным многочленом согласно (1.9.1), и будем решать уравнение (8.5.6) методом коллокации (совпадения в выбранных точках). Тогда (8.5.6) сводится к системе нелинейных уравнений

$$\sum_{i=0}^{N} g_i \frac{\Gamma(\alpha+i)}{\Gamma(\alpha)} \left\{ \frac{D_1(z)}{[\lambda_1(z)+\mu]^i} \left(\frac{\mu}{\lambda_1(z)+\mu} \right)^{\alpha} - \frac{D_2(z)}{[\lambda_2(z)+\mu]^i} \left(\frac{\mu}{\lambda_2(z)+\mu} \right)^{\alpha} \right\} = \tilde{\Pi}(z), \quad z = z_1, z_2, \ldots, z_{N+3} \tag{8.5.7}$$

относительно α, μ и $\{g_0, g_1, \ldots, g_N\}$.

Выделим из системы (8.5.7) ее нелинейную часть

$$D_1(z)\left(\frac{\mu}{\lambda_1(z)+\mu} \right)^{\alpha} - D_2(z)\left(\frac{\mu}{\lambda_2(z)+\mu} \right)^{\alpha} = \tilde{\Pi}(z), \tag{8.5.8}$$

решаемую численно (например, методом Ньютона) относительно параметров α и μ при двух фиксированных значениях z. Получив решение системы (8.5.8) с требуемой точностью, можно подставить α и μ в систему (8.5.7) и свести ее к линейной относительно $\{g_i\}$, $i = \overline{0, N}$.

Моменты плотности, представленной в форме (1.9.1), могут быть вычислены согласно

$$\tilde{w}_k = \sum_{i=0}^{N} \frac{g_i}{\mu^{k+i}} \frac{\Gamma(\alpha+k+i)}{\Gamma(\alpha)}, \qquad k = 0, 1, \ldots \tag{8.5.9}$$

8.5.4. Аппроксимация общего случая

Вычислим моменты распределения интервалов между «синими» заявками по формуле (2.1.19) с той же заменой z на $1-z$ и по этим моментам методами разд. 1.8.3 подберем гиперэкспоненциальную аппроксимацию

$$\bar{A}_z(t) = \sum_{i=1}^{k} u_i(z) e^{-\lambda_i(z) t}.$$

Вновь представляя $\tilde{w}(t)$ в виде (1.9.1), можно свести основное соотношение (3.2.1) к формуле

$$\begin{aligned}\tilde{\Pi}(z) &= \sum_{i=1}^{k} u_i(z)\Big(\frac{\mu}{\lambda_i(z)+\mu}\Big)^{\alpha}\sum_{j=0}^{N} g_j\frac{\Gamma(\alpha+j)}{(\lambda_i(z)+\mu)^j\Gamma(\alpha)} \\ &= \sum_{j=0}^{N} g_j\frac{\Gamma(\alpha+j)}{\Gamma(\alpha)}\sum_{i=1}^{k}\frac{u_i(z)}{(\lambda_i(z)+\mu)^j}\Big(\frac{\mu}{\lambda_i(z)+\mu}\Big)^{\alpha}.\end{aligned} \tag{8.5.10}$$

Выберем α и μ согласно (1.4.1) с тем расчетом, чтобы обеспечить выравнивание первых двух моментов гамма-распределением без поправочного многочлена. Необходимые значения моментов получаются: $\tilde{w}_1$ — по формуле Литтла, а $\tilde{w}_2$ — методом пересчета (см. разд. 8.6). Поскольку моменты $\tilde{w}_0 = 1$ и $\tilde{w}_1$ известны точно, имеет смысл включить в систему уравнений, линейных относительно $\{g_j\}$, уравнения вида (8.5.9) для $k = 0$ и $k = 1$. Остальные $N - 1$ уравнений берутся вида (8.5.10) для $z = z_1, z_2, \ldots, z_{N-1}; \quad z \in (0, 1)$. В качестве точек коллокации $\{z_i\}$ целесообразно выбирать чебышевские узлы, пересчитанные применительно к упомянутому интервалу.

Численные эксперименты показали, что регуляризованное описанным способом интегральное уравнение Фредгольма первого рода (3.2.1) при некоторых комбинациях параметров сохраняет характерную для задач этого класса вычислительную неустойчивость. Поэтому предпочтительно ставить задачу как проблему наименьших квадратов: найти

$$\min_{\{g_j\}}\sum_{m=1}^{M}\Big[\tilde{\Pi}(z_m) - \sum_{j=0}^{N} g_j F_j(z_m)\Big]^2$$

при двух ограничениях типа равенств (8.5.9). Здесь $M > N$, а $F_j(z)$ — множитель при g_j в формуле (8.5.10). Проблема решается методом неопределенных множителей Лагранжа.

Наконец, можно поставить задачу минимизации среднего квадрата отклонения на отрезке [0,1]. Сохраняя предположение об аппроксимации $\tilde{w}(t)$ гамма-плотностью с поправочным многочленом и прежний способ определения параметров α и μ, можно записать равенство (8.5.6) в виде

$$\tilde{\Pi}(z) = \sum_{j=0}^{N} g_j\varphi_j,$$

где $\{\varphi_j(z)\}$ — известные функции, определяемые способом аппроксимации ядра интегрального уравнения. При H_k–аппроксимации, как это

следует из (8.5.10),

$$\varphi_j(z) = \frac{\Gamma(\alpha+j)}{\Gamma(\alpha)} \sum_{i=1}^{k} \frac{u_i(z)}{(\lambda_i(z)+\mu)^j} \left(\frac{\mu}{\lambda_i(z)+\mu}\right)^{\alpha}.$$

При гиперэрланговской аппроксимации вида (1.8.13)

$$\varphi_j(z) = \left(\frac{\mu}{\lambda+\mu}\right)^{\alpha} \Big[\sum_{m=0}^{r-1} \frac{\lambda^m}{m!} \frac{\Gamma(\alpha+j+m)}{\Gamma(\alpha)(\lambda+\mu)^{j+m}} + y\,\frac{\lambda^r}{r!} \frac{\Gamma(\alpha+j+r)}{\Gamma(\alpha)(\lambda+\mu)^{j+r}}\Big].$$

Здесь λ — непрерывный параметр аппроксимации, r — полное число фаз, y — вероятность прохождения всех фаз (все названные параметры зависят от z). Выбор типа аппроксимации следует проводить по соотношению между моментами $\{d_i\}$ прореженного потока при текущем значении z. Если параметр $\beta = 1/(d_2/d_1^2 - 1)$ гамма-аппроксимации распределения интервалов между «синими» заявками находится в диапазоне $1 < \beta < 10$, предпочтительна гиперэрланговская аппроксимация, в остальных случаях — гиперэкспоненциальная.

Неизвестные $\{g_i\}$ будем определять из условия

$$\min_{\{g_j\}} \int_0^1 \Big(\sum_{j=0}^{N} g_j \varphi_j(z) - \tilde{\Pi}(z)\Big)^2 dz.$$

Дифференцируя интеграл по $\{g_i\}$ и приравнивая производные нулю, получаем условия оптимальности в виде системы линейных уравнений

$$\sum_{j=0}^{N} z_j \int_0^1 \varphi_j(z)\varphi_i(z)\,dz = \int_0^1 \tilde{\Pi}(z)\varphi_i(z)\,dz, \qquad i = \overline{0,N}.$$

Входящие в нее интегралы определяются численно (например, по квадратурной формуле Гаусса). Достоинством этого подхода является осмысленный выбор значений z, для которых выполняется расчет. Два уравнения этой системы, как и в предыдущем случае, заменяются на условия выравнивания начальных моментов.

Разумеется, сформулированные здесь соображения относительно выбора α и μ и регуляризации задачи относятся также к разд. 8.5.3.

8.6. Метод пересчета

Как показано в разд. 5.7.1, условное время ожидания заявки в системе $GI/M/n$ распределено показательно. Это обстоятельство позволяет считать, что *вид* распределения времени ожидания не зависит от распределения интервалов между смежными заявками и числа каналов и определяется только распределением времени обслуживания. Остается найти способ характеризации вида распределения и конструктивный метод его использования.

Будем определять интересующее нас распределение его первым моментом $\tilde{w}_1$ и набором введенных формулой (1.4.3) коэффициентов немарковости $\{\xi_i\}$. Тогда высшие моменты можно вычислить по формуле

$$\tilde{w}_i = \tilde{w}_1^i(\xi_i + i\,!), \qquad i = 2, 3, \ldots \tag{8.6.1}$$

В качестве базы для определения $\{\xi_i\}$ выберем систему $M/G/1$. Величину w_1 рассчитаем по формуле Полячека—Хинчина (3.4.1), а последующие моменты — рекуррентно из формулы (5.2.8). Разделив $\{w_k\}$ на $1 - p_0$, получим условные моменты $\{\tilde{w}_k\}$ распределения времени ненулевого ожидания. Далее нетрудно вычислить коэффициенты немарковости этого распределения, из которых мы приведем

$$\begin{aligned}
\xi_2 &= 2(1-\rho)\left(\frac{2}{3}\,\frac{b_1 b_3}{b_2^2} - 1\right), \\
\xi_3 &= 6(1-\rho)\left[\frac{b_1^2 b_4}{3b_2^3}(1-\rho) + \frac{4}{3}\,\frac{b_1 b_3}{b_2^2}\rho - \rho - 1\right], \\
\xi_4 &= 24(1-\rho)\left[\frac{2b_1^3 b_5}{15 b_2^4}(1-\rho)^2 + \left(\frac{2}{3}\frac{b_1^2 b_4}{b_2^3} + \frac{4}{9}\,\frac{b_1^2 b_3^2}{b_2^4}\right)\rho(1-\rho)\right. \\
&+ \left.\left(\frac{2b_1 b_3}{b_2^2} - 1\right)\rho^2 - \rho - 1\right].
\end{aligned} \tag{8.6.2}$$

Здесь для обеспечения возможности пересчета на многоканальные системы произведение λb_1 заменено на коэффициент загрузки $\rho = b_1/(na_1)$.

Отметим следующие важные обстоятельства:

- Непосредственная подстановка моментов показательного распределения в форме $b_k = k\,!\,b_1^k$ дает ожидавшиеся для модели $M/M/n$ значения $\xi_2 = \xi_3 = \xi_4 = 0$.

- В согласии с известными предельными результатами при $\rho \to 1$ все $\xi_k \to 0$, т. е. искомое распределение стремится к показательному независимо от вида распределения обслуживания.
- Коэффициенты при $\{\xi_k\}$ содержат множители $(1-\rho)^{k-1}$. Поэтому при $\rho \to 1$ влияние старших моментов убывает, начиная с высших.

Итак, метод пересчета сводится к следующим действиям:

1. По распределению числа заявок в исследуемой системе $GI/G/n$ найти ожидаемую длину очереди.
2. По формуле Литтла найти среднее время ожидания w_1.
3. Рассчитать среднее условное время ожидания $\tilde{w}_1 = w_1/(1-\sum_{j=0}^{n-1}\pi_j)$, где $\{\pi_j\}$ есть распределение числа заявок перед прибытием очередной заявки.
4. Согласно (8.6.2) рассчитать необходимое число коэффициентов немарковости распределения условного времени ожидания.
5. По (8.6.1) вычислить высшие моменты этого распределения.
6. Умножением их на $1-\sum_{j=0}^{n-1}\pi_j$ перейти к безусловным моментам.

Вычисление по этой схеме $\tilde{w}_2$ является наиболее простым путем к базовым значениям α и μ, необходимым для линеаризации систем уравнений интегрального метода.

8.7. Сравнение методов

Ниже сопоставлены полученные разными способами значения моментов w_2 и w_3 (первые моменты как правило совпадали и для экономии места опущены). Во всех случаях принималось $b_1 = 1$.

Табл. 8.1 содержит результаты тестирования при $\rho = 0.6$ интегрального метода (I) и пересчетного (R) на модели $GI/E_k/1$, для которой известен прямой метод (E) получения искомых зависимостей — см. разд. 5.8.2.

Таблица 8.1. Моменты распределения ожидания в системе $GI/E_k/1$

v_B^2	Способ расчета	$v_A^2=4$		$v_A^2=1/3$		$v_A^2=1/10$	
		w_2	w_3	w_2	w_3	w_2	w_3
1/3	E	3.85e+1	5.42e+2	6.26e-1	1.59e-0	2.06e-1	3.74e-1
	I	3.75e+1	5.03e+2	8.11e-1	2.04e-0	9.10e-1	1.63e-0
	R	3.75e+1	5.09e+2	6.24e-1	1.59e-0	2.01e-1	3.64e-1
1/4	E	3.69e+1	5.06e+2	4.61e-1	9.99e-1	1.21e-1	1.77e-1
	I	3.56e+1	4.60e+2	8.05e-1	1.75e-0	1.39e-0	2.00e-0
	R	3.56e+1	4.67e+2	4.60e-1	1.01e-0	1.19e-1	1.73e-1
1/10	E	3.41e+1	4.46e+2	2.30e-1	3.49e-1	2.87e-2	2.32e-2
	I	3.21e+1	3.84e+2	1.23e-0	1.91e-0	5.86e-0	4.77e-0
	R	3.23e+1	3.93e+2	2.33e-1	3.69e-1	2.84e-2	2.36e-2

В табл. 8.2 также для $\rho = 0.6$ приведены результаты расчета трехканальной системы с гиперэкспоненциальной аппроксимацией исходных распределений по трем моментам. Буквой С обозначен метод сверток. Распределение вероятностей микросостояний вычислялось методом матрично-геометрической прогрессии и являлось общей базой всех подходов.

Наконец, в табл. 8.3 показаны результаты расчета трех систем с помощью 11 процедур пакета МОСТ (см. заключительную главу) при коэффициенте загрузки $\rho = 0.7$, из которых две первых могут считаться эталонными. Прочерками отмечены случаи несовместимости процедуры и обсчитываемой модели.

Таблица 8.2. Моменты распределения ожидания в системе $H_2/H_2/3$

v_B^2	Способ расчета	$v_A^2=4$		$v_A^2=1/3$		$v_A^2=1/10$	
		w_2	w_3	w_2	w_3	w_2	w_3
1/3	C	3.29e-0	1.51e+1	3.52e-2	3.03e-3	7.14e-3	3.59e-3
	I	3.21e-0	1.41e+1	2.25e-0	2.03e-0	5.83e-0	3.39e-0
	R	3.20e-0	1.42e+1	3.58e-2	3.28e-3	7.70e-3	4.53e-3
1/4	C	3.17e-0	1.41e+1	2.62e-2	1.94e-2	3.22e-3	7.71e-4
	I	3.07e-0	1.30e+1	4.05e-0	3.22e-0	5.36e-0	3.25e-0
	R	3.07e-0	1.31e+1	2.69e-2	2.17e-2	7.27e-3	4.48e-3
1/10	C	2.95e-0	1.25e+1	1.37e-2	7.04e-3	—	—
	I	2.81e-0	1.09e+1	1.57e+1	9.44e-0	—	—
	R	2.81e-0	1.12e+1	1.44e-2	8.86e-3	1.01e-3	2.90e-4

Таблица 8.3. Сопоставление процедур расчета моментов ожидания

Процедура	$M/E_3/1$			$E_2/M/2$			$\Gamma_{1.5}/M/1$		
	w_1	w_2	w_3	w_1	w_2	w_3	w_1	w_2	w_3
GE1	1.56	6.57	41.31	—	—	—	1.84	10.49	89.53
GMN	—	—	—	0.63	1.65	6.43	1.84	10.49	89.53
MTIME	1.56	6.57	41.31	—	—	—	—	—	—
H2TIME	1.56	6.57	41.31	0.63	1.66	6.51	1.84	10.49	89.53
E2TIME	—	—	—	0.63	1.66	6.51	—	—	—
C2TIME	1.56	6.57	41.90	0.63	1.65	6.43	1.84	10.49	89.47
HETIME	1.56	7.09	44.14	0.63	1.79	6.99	1.84	11.40	97.31
GLTIME	1.56	6.57	41.69	0.63	1.65	6.43	1.84	10.49	89.43
XLTIME	1.56	6.58	40.93	0.63	1.79	6.97	1.84	11.01	93.92
RTIME	1.56	6.57	38.71	0.63	1.65	5.85	1.84	10.49	81.47
WFCFS	1.56	6.64	41.98	0.64	1.67	6.69	1.84	10.40	87.34

Анализ таблиц приводит к следующим выводам:

1. Большое число случаев хорошего согласия результатов свидетельствует о допустимости всех обсуждавшихся выше различных подходов и правильности (по крайней мере в принципе) реализующих их алгоритмов и программ.

2. Из процедур, применимых при рекуррентном входящем потоке, наилучшие результаты дают `H2TIME`, `GLTIME` и `C2TIME`. Несколько хуже других работает `HETIME`, основанная на гиперэрланговой аппроксимации.

3. Реализованный в `RTIME` метод пересчета во всех случаях обнаруживает удовлетворительное согласие с заслуживающими доверия эталонами и смежными элементами таблиц. Для его применения достаточно знать среднее время ожидания и вероятность ненулевого ожидания. В случае *ограниченной очереди* базы для пересчета нет.

Для *замкнутых систем* метод свертки альтернатив не имеет.

8.8. О приближенных оценках для многоканальных систем

Острая практическая потребность в численном определении показателей работы многоканальных СМО и отсутствие (до недавних пор) соответствующих методов вызвали появление многочисленных работ по приближенным оценкам этих показателей, прежде всего среднего времени ожидания начала обслуживания. Обширный обзор и частично — обоснование таких оценок приведены в монографии Д. Штойяна [149]. Как правило, оценки предлагались на базе немногих опорных моделей, поддававшихся численному анализу, с помощью полуэвристических рассуждений. К сожалению, эти оценки зачастую используются неправильно: не как *границы* для возможных значений среднего времени ожидания, а как его предполагаемое значение. Разница между ними (в особенности при умеренной загрузке) может быть весьма значительна и сильно зависит от исходных данных каждого конкретного случая.

Обширная подборка приближенных формул для среднего времени ожидания приводится в [165] на с. 231 и далее. При современном состоянии теории расчета систем обслуживания и наличии реализующих ее алгоритмы программных комплексов типа описанного в главе 13 работы в области приближенных оценок для классических типов СМО свою актуальность утратили.

В некоторых задачах (расчет сетей обслуживания, см. разд. 12.9, применение аппроксимаций для расчета среднего времени ожидания все же необходимо. В связи с этим была предпринята попытка уточнения формулы Моле хотя бы для ограниченного диапазона условий ($\rho \geq 0.5, n = \overline{1,4}$). Оказалось, что приемлемые результаты дает аппроксимация вида

$$w \approx \frac{\lambda b_2}{2n^2(1-\rho)} e^{-0.7(n-1)(1-\rho)}, \tag{8.8.1}$$

при $n = 1$ совпадающая, как и формула Моле, с формулой Полячека—Хинчина (3.4.1).

Формулу (8.8.1) можно переписать в более наглядной форме,

позволяющей одновременно получить среднюю длину очереди. Поскольку

$$\frac{\lambda b_2}{2n^2(1-\rho)} = \frac{\lambda b_1^2(1+v_B^2)}{2n^2(1-\rho)} = \frac{\rho b_1(1+v_B^2)}{2n(1-\rho)}$$
$$= \frac{\rho\cdot\rho n a_1(1+v_B^2)}{2n(1-\rho)} = \frac{\rho^2}{1-\rho}\,\frac{1+v_B^2}{2}a_1,$$

можно записать

$$w \approx \frac{\rho^2}{1-\rho}\,\frac{1+v_B^2}{2}e^{-0.7(n-1)(1-\rho)}a_1, \qquad (8.8.2)$$

и на основании формулы Литтла произведение трех первых сомножителей (включая экспоненту) дает оценку средней длины очереди.

8.9. Неоднородный входящий поток

Рассмотрим систему бесприоритетного обслуживания с неоднородным входящим потоком $\overrightarrow{GI_m}/\overrightarrow{G_m}/n$. Аппроксимируем эту систему однородной моделью вида $A/B/n$, где распределение A интервалов между входящими заявками строится по схеме суммирования потоков (разд. 2.1.9), а распределение B — по средневзвешенным моментам частных распределений обслуживания:

$$\bar{b}_r = \Lambda^{-1}\sum_{i=1}^{m}\lambda_i b_{i,r}, \qquad r = 1,2,\ldots \qquad (8.9.1)$$

Здесь $\{\lambda_i = 1/a_i\}$ — интенсивности частных потоков, а $\Lambda = \sum_{i=1}^{m}\lambda_i$ — интенсивность суммарного.

Распределение *общей* очереди и времени ожидания в ней определяется только вышеупомянутыми суммарными характеристиками (для системы $\vec{M}/\vec{G}/1$ известно формальное доказательство этого факта). Эти распределения могут быть найдены уже обсуждавшимися методами после подбора соответствующих аппроксимаций для усредненной модели.

Моменты распределения времени пребывания в системе заявки конкретного i-го типа вычисляются сверткой в моментах *общего* распределения времени ожидания и распределения длительности обслуживания заявки i-го типа на основе символического разложения

$$v_i^r = (w + b_i)^r, \qquad r = 1,2,\ldots \qquad (8.9.2)$$

8.10. Случайный выбор на обслуживание

Подавляющее большинство работ по теории очередей посвящено системам с выборкой из очереди по принципу «первый пришел — первый обслужен». Однако значительный интерес вызывает и другой принцип — случайный выбор из очереди (SIRO). Он особенно типичен для связных приложений и ситуаций с разъездными бригадами обслуживания (ремонтными, аварийными, скорой помощи, групп быстрого реагирования и т. п.), где очередность обработки заявок не определяется моментами их поступления. Различные подходы к этой задаче предлагаются или воспроизводятся в [66, 87, 132, 176, 195, 207, 227, 235, 248]. В работах [87, 132, 227] обсуждаются чисто марковские системы $M/M/n$. В частности, в них выводятся формулы для коэффициентов увеличения высших моментов времени ожидания в сравнении с дисциплиной FCFS:

$$\begin{aligned} R_2 &= (1-\rho/2)^{-1}, \\ R_3 &= (4+2\rho)/(2-\rho)^2. \end{aligned}$$

В статье Кингмана [207] выполнен анализ системы $M/G/1$, к сожалению, не доведенный до расчетных зависимостей. Ряд конечных результатов для этой же системы приводится в справочнике Дж. Мартина [66], но без вывода и без указания источников заимствования. В нем дано выражение для дисперсии времени ожидания

$$D_w = \frac{2\lambda b_3}{3(1-\rho)(2-\rho)} + \frac{\lambda^2 b_2^2(2+\rho)}{4(1-\rho)^2(2-\rho)} + D_B. \qquad (8.10.1)$$

Приведены также выражения для среднеквадратического отклонения времени ожидания при показательном и регулярном обслуживании:

$$\begin{aligned} \sigma_M &= \frac{b_1}{1-\rho}\sqrt{\frac{2\rho^2+2-\rho}{2-\rho}}, \\ \sigma_D &= \frac{b_1}{1-\rho}\sqrt{\frac{8\rho-2\rho^2+3\rho^3}{12(2-\rho)}}. \end{aligned}$$

В работе Розенлунда [248] применительно к модели $GI/M/n$ получены преобразования Лапласа—Стилтьеса для распределения времени ожидания и рекуррентный алгоритм вычисления моментов. Воспроизведем фрагмент его итоговой таблицы:

Таблица 8.4. Поправки к FIFO-моментам

Порядок моментов	$D/M/n$			$E_4/M/n$			$M/M/n$		
	0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9
2	1.2550	1.5005	1.8125	1.2884	1.5164	1.8148	1.3333	1.5385	1.8182
3	1.9782	3.0602	4.7698	2.0776	3.1149	4.7793	2.2222	3.1953	4.7934
4	3.6889	7.6951	16.238	3.9570	7.8927	16.286	4.3704	8.1896	16.356

Наконец, в статье Картера и Купера [176] рассматриваются $GI/M/n$ и $M/G/n$, подтверждены известные результаты для $M/M/n$ и $M/D/1$, рассчитана система $M/E_k/1$.

Перечисленные результаты выводятся различными для разных моделей методами, с трудно прослеживаемой аргументацией и с привлечением эвристических приемов. Желательна их проверка на надежной имитационной модели. В табл. 8.5 приведены результаты имитации системы $M/M/3$ с дисциплинами FCFS и SIRO при 500 тыс. испытаний.

Таблица 8.5. Расчет и имитация марковской системы

ρ	Моменты	FCFS (расчет)	FCFS (имитация)	RANDOM (имитация)	Отношение RANDOM/FCFS
0.5	w_1	0.2368	0.2346	0.2346	1.0000
	w_2	0.4737	0.4628	0.5362	1.1585
	w_3	1.4211	1.3572	2.2109	1.6290
0.7	w_1	1.1488	1.1494	1.1494	1.0000
	w_2	5.3611	5.2108	7.1150	1.3654
	w_3	37.528	34.162	86.600	2.5350
0.9	w_1	7.3535	7.6507	7.3628	0.9624
	w_2	132.36	140.17	226.70	1.6173
	w_3	3573.8	3687.5	14980	4.0624

Средний интервал между заявками предполагался единичным, интенсивность обслуживания назначалась из условия обеспечения заданного коэффициента загрузки ρ. Результаты аналитического расчета хорошо согласуются с имитационной моделью FCFS — с учетом общеизвестного нарастания погрешностей при увеличении порядка моментов и коэффициента загрузки системы. Переделка модели под случайный выбор из очереди затронула лишь четыре оператора программы, и в ней было невозможно ошибиться. Правильность SIRO-модели

подтверждается и практическим совпадением первых моментов (средних времен ожидания). $Отношения$ моментов времени ожидания вычислены по результатам имитационного моделирования. Они заметно отличаются от приведенных в табл. 8.4, что вынуждает поставить вопрос о новых методах расчета систем обслуживания с SIRO-дисциплиной.

8.10.1. Вложенная цепь Маркова для простейшего входящего потока

Рассмотрим режим полной занятости системы $GI/G/n/R$ и обозначим:

$J = R - n$ — максимальную длину очереди;

$B_n(t)$ — распределение интервалов между последовательными завершениями обслуживания;

$B_n^*(t)$ — распределение интервала от прибытия «меченой» заявки до ближайшего завершения обслуживания;

λ — интенсивность входящего потока.

Вычислим вероятности прибытия ровно j новых заявок за время до ближайшего обслуживания

$$q_j^* = \int_0^\infty \frac{(\lambda t)^j}{j\,!} e^{-\lambda t} dB_n^*(t), \quad j = 0, 1, \ldots$$

и — с заменой $B_n^*(t)$ на $B_n(t)$ — аналогичный набор $\{q_j\}$ вероятностей прибытия ровно j заявок за время между смежными обслуживаниями. Введем также вектор $P = \{p_0, p_1, \ldots p_J\}$ распределения вероятностей длины очереди в момент прибытия меченой заявки (стационарное распределение). Построим вложенную цепь Маркова изменения числа заявок в очереди до выбора на обслуживание меченой заявки по моментам непосредственно перед очередным выбором на обслуживание.

Элементы матрицы R^* переходов за время ожидания меченой заявкой ближайшего обслуживания должны вычисляться согласно

$$r_{i,j}^* = q_{j-i}^*, \quad i = \overline{1, J}, \quad j = \overline{i, J-1}.$$

В случае продолжения процесса за интервал между смежными обслуживаниями систему покидает ровно одна заявка. Следовательно, элементы матрицы «продолжающихся» переходов должны вычисляться согласно

$$\begin{aligned} r_{1,j} &= q_j, \quad j = \overline{1, J-1}, \\ r_{i,j} &= q_{j-i+1} \quad i = \overline{2, J}, \quad j = \overline{i-1, J-1}. \end{aligned}$$

Вероятности $\{r^*_{i,J}\}$ для всех строк i определяются из условия нормировки суммы строки к единице. Все прочие переходы имеют нулевые вероятности.

Вероятности выбора меченой заявки и тем самым попадания процесса в *поглощающее* состояние задаются вектором-столбцом T с компонентами $\{1/i\}$. Соответственно вероятности его продолжения задаются вектором C, компоненты которого дополняют соответствующие компоненты T до единицы.

Введем ПЛС распределений временны́х интервалов:

$\beta^*(s)$ — от прибытия меченой заявки до первого обслуживания;

$\beta(s)$ — между смежными завершениями обслуживания;

$\omega_k(s)$ — ожидания меченой заявки, завершаемого через k шагов марковской цепи, $k = 1, 2, \ldots$.

Нетрудно видеть, что

$$\begin{aligned} \omega_1(s) &= \beta^*(s) P R^* T, \\ \omega_2(s) &= \beta^*(s) P R^* [C\beta(s) R] T, \\ \omega_3(s) &= \beta^*(s) P R^* [C\beta(s) R]^2 T, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \omega_k(s) &= \beta^*(s) P R^* [C\beta(s) R]^{k-1} T, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}$$

Суммируя эти выражения по всем k, находим

$$\begin{aligned} \omega(s) &= \beta^*(s) P R^* \Big(\sum_{k=1}^{\infty} [C\beta(s) R]^{k-1} \Big) T \\ &= \beta^*(s) P R^* [I - CR\beta(s)]^{-1} T. \end{aligned}$$

Искомые *моменты* распределения могут быть получены численным дифференцированием $\omega(s)$ (точнее, построенного по значениям в окрестности $s = 0$ интерполяционного многочлена) с последующей сменой знака у нечетных производных.

8.10.2. Тестирование схемы на модели $M/M/n$

Описанную расчетную схему целесообразно тестировать применительно к простейшей ситуации — с показательными распределениями. При этом нет проблем с потоком заявок: все остаточные распределения остаются показательными с исходным параметром λ, а связанные с процессом обслуживания распределения

$$B_n(t) = B_n^*(t) = 1 - e^{-n\mu}$$

и соответственно

$$\beta(s) = \beta^*(s) = \frac{n\mu}{n\mu + s}.$$

При $\lambda = 1$, коэффициенте загрузки $\rho = 0.7$ и шаге $h = 10^{-3}\mu$ построения таблицы $\omega(s)$ для интерполирования по Стирлингу были получены значения моментов времени ожидания

$$w_1 = 1.1479,\ w_2 = 7.3481,\ w_3 = 91.338.$$

Соотнесенные с аналогичными результатами при дисциплине FCFS, они дают коэффициенты роста

$$k_1 = 1.0000,\ k_2 = 1.3706,\ k_3 = 2.4339.$$

Хорошее их согласие с полученными на имитационной модели позволяет считать основной алгоритм правильным и применить его к более сложным случаям: с немарковским распределением обслуживания и произвольным рекуррентным потоком заявок.

8.10.3. Произвольное распределение длительности обслуживания

В данном случае получение распределения интервалов до очередного обслуживания становится самостоятельной проблемой. Дополнительная функция распределения (ДФР) интервала от прибытия меченой

заявки до ближайшего завершения обслуживания может быть подсчитана согласно

$$\bar{B}_n^*(t) = [\bar{B}^*(t)]^n. \tag{8.10.2}$$

Для работы с этой формулой вычислим моменты $B_n^*(t)$ через моменты $B^*(t)$, аппроксимируя ДФР последнего распределением Вейбулла

$$\bar{B}^*(t) = \exp(-t^k/W) \tag{8.10.3}$$

с моментами

$$b_m^* = W^{m/k}\Gamma(1+m/k), \qquad m = 1, 2, \ldots \tag{8.10.4}$$

Подставляя (8.10.3) в формулу (8.10.2), убеждаемся, что интересующее нас распределение $\bar{B}_n^*(t)$ вновь описывается формулой (8.10.3) с заменой W на W/n. Соответственно его моменты следует вычислять как

$$b_{m,n}^* = (W/n)^{1/m}\Gamma(1+m/k), \qquad m = 1, 2, \ldots \tag{8.10.5}$$

Вторым удобным вариантом аппроксимации ДФР является гиперэкспоненциальная H_2. Если

$$\bar{B}(t) = \sum_{i=1}^{2} y_i e^{-\mu_i t},$$

то остаточное распределение описывается аналогичной функцией с заменой параметров $\{y_i, \mu_i\}$ на $\{z_i, \nu_i\}$. Соответственно

$$\bar{B}_n^*(t) = \left(\sum_{i=1}^{2} z_i e^{-\nu_i t}\right)^n = \sum_{j=0}^{n} \binom{n}{j} z_1^j z_2^{n-j} e^{-[j\nu_1 + (n-j)\nu_2]t}.$$

Моменты этого распределения

$$b_{m,n}^* = \int_0^\infty t^{m-1}\bar{B}_n^*(t)\,dt = m! \sum_{j=0}^{n} \binom{n}{j} \frac{z_1^j z_2^{n-j}}{[j\nu_1 + (n-j)\nu_2]^m}, \qquad m = 1, 2, \ldots$$

Для основного режима марковской цепи приходится строить распределение интервалов между смежными завершениями обслуживания в n-канальной системе. Здесь (см. Саати, [132])

$$\bar{B}_n(t) = [\bar{B}^*(t)]^{n-1}\bar{B}(t).$$

В данной ситуации аппроксимация распределениями Вейбулла не имеет смысла, а гиперэкспоненциальная аппроксимация приводит к

$$\begin{aligned}\bar{B}_n(t) &= \Big(\sum_{i=1}^{2} z_i e^{-\nu_i t}\Big)^{n-1}\Big(\sum_{i=1}^{2} y_i e^{-\mu_i t}\Big)\\ &= \sum_{j=0}^{n-1}\binom{n-1}{j} z_1^j z_2^{n-1-j} e^{-[j\nu_1+(n-1-j)\nu_2]t}\Big(\sum_{i=1}^{2} y_i e^{-\mu_i t}\Big)\\ &= \sum_{i=1}^{2} y_i \sum_{j=0}^{n-1}\binom{n-1}{j} z_1^j z_2^{n-1-j} e^{-[j\nu_1+(n-1-j)\nu_2+\mu_i]t}.\end{aligned}$$

Его моменты можно считать по формулам

$$b_{m,n} = m!\sum_{i=1}^{2} y_i \sum_{j=0}^{n-1}\binom{n-1}{j}\frac{z_1^j z_2^{n-1-j}}{[j\nu_1+(n-1-j)\nu_2+\mu_i]^m}, \qquad m=1,2,\ldots \tag{8.10.6}$$

Далее по найденным моментам подбирается аппроксимация, обеспечивающая удобное вычисление $\{q_j\}$. Так, для гамма-плотности

$$b(t) = \frac{m(mt)^{\alpha-1}}{\Gamma(\alpha)} e^{-mt}$$

параметры выражаются через среднее b_1 и дисперсию D согласно

$$\alpha = b_1^2/D, \qquad m = \alpha/b_1.$$

Теперь

$$\begin{aligned}q_0 &= \Big(\frac{m}{\lambda+m}\Big)^{\alpha},\\ q_j &= q_{j-1}\frac{\lambda}{\lambda+m}\frac{\alpha+j-1}{j}, \qquad j=1,2,\ldots\end{aligned}$$

Применение гамма-плотности с поправочным многочленом, который строится на основе теории обобщенных многочленов Лагерра (см. разд. 1.6.4 и 2.1.5) позволяет выравнивать *произвольное* число моментов. При этом сохраняется возможность рекуррентного вычисления $\{q_j\}$, но по более сложному алгоритму.

8.10.4. Рекуррентный поток

При входящем потоке общего вида приходится учитывать следующие дополнительные обстоятельства:

1. Для начального шага марковской цепи коэфициенты $\{q_j\}$ считаются как вероятности прибытия ровно j заявок рекуррентного потока за модифицированный интервал между смежными обслуживаниями (отсчет времени начинается с прибытием меченой заявки).

2. На последующих шагах расчет начинается с завершения очередного обслуживания, которое приходится на случайную точку интервала между смежными заявками. Соответственно поток заявок представляется как рекуррентный $c\ запаздыванием$, а интервалы между обслуживаниями не модифицируются.

Вопрос о расчете вероятностей $\{q_j(t)\}$ за фиксированный интервал длины t для указанных видов потоков рассматривался в главе 2. Полученные там выражения здесь должны интегрироваться с учетом вышеуказанных распределений длительности шагов марковской цепи. При этом могут возникнуть дополнительные трудности. В частности, при гамма-плотности интервалов между обслуживаниями и аппроксимации интервалов между заявками дополнительной функцией распределения Вейбулла попытка разложения последней в степенной ряд привела к

$$\begin{aligned} q_0 &= \int_0^\infty e^{-t^k/T}\frac{\mu(\mu t)^{\beta-1}}{\Gamma(\beta)}e^{-\mu t}\,dt = \frac{\mu^\beta}{\Gamma(\beta)}\int_0^\infty\sum_{i=0}^\infty\frac{1}{i\,!}(-t^k/T)^i t^{\beta-1}e^{-\mu t}\,dt \\ &= \frac{\mu^\beta}{\Gamma(\beta)}\sum_{i=0}^\infty\frac{(-1)^i}{i\,!T^i}\int_0^\infty t^{ki+\beta-1}e^{-\mu t}\,dt = \sum_{i=0}^\infty\frac{(-1)^i}{i\,!(\mu^k T)^i}\frac{\Gamma(ki+\beta)}{\Gamma(\beta)}. \end{aligned}$$

При эрланговском потоке 3-го порядка с единичным средним и обслуживании по закону Эрланга 2-го порядка со средним $b_1 = 3$ оказалось, что произведение $\mu^k T \approx 0.4$, и полученный степенной ряд расходился. Поэтому интегрирование выполнялось численно для исходных распределений — сначала на полуинтервале (0,1]. Этот интервал разбивался на два: $(0, c]$ и $[c, 1]$, $c_0 = 0.5$. Интегрирование на левом участке выполнялось по квадратурной формуле Гаусса, а на правом — по Ромбергу, и результаты суммировались. Далее половина левого участка присоединялась к правому, интегрирование по Гауссу выполнялось

заново, а по Ромбергу — только по добавленной половинке. Пересчет на интервале (0,1] продолжался, пока уточнение не становилось пренебрежимо малым. Далее вычислялись по Ромбергу и добавлялись к сумме (пока давали значимые приращения) интегралы на смежных интервалах единичной длины.

8.11. О нестационарных задачах

Нестационарные задачи ТМО решаются со значительно бóльшим трудом, чем стационарные. Явный вид таких решений известен лишь в простейших случаях — для немногих марковские моделей [18, 51, 190, 221]. В приложении к [153] дается вывод нестационарного распределения числа требований в системе M/G/1 с обслуживанием в порядке поступления.

Марковские (марковизированные) системы с *конечным* числом состояний в принципе могут быть просчитаны численно — решением соответствующих систем обыкновенных дифференциальных уравнений с постоянными коэффициентами при заданных начальных условиях (такие уравнения можно получить, применив закон сохранения вероятностей из разд. 3.3 в дифференциальной форме). Эти и матричные методы решения подобных задач в связи с большой размерностью последних мало перспективны.

Достаточно актуален расчет систем обслуживания с циклически (суточный цикл) меняющейся нагрузкой. В подобных случаях приходится делить цикл на интервалы с относительно стабильными потоками заявок и обсчитывать интервалы отдельно.

Из нестационарных задач ТМО наибольший практический интерес вызывают проблемы временной перегрузки системы и последующего рассасывания накопившейся очереди. Причинами такой перегрузки могут быть, например, суточные флуктуации потока заявок или отключение части каналов обслуживания для ремонтных либо профилактических работ. Проблемы рассасывания очередей возникают и в связи с организацией рабочего цикла, если поток заявок постоянен (аварии, вызовы медицинской помощи и т. п.), а обслуживание прерывается на обед, в ночное время и/или на выходные дни. Излагаемый ниже с некоторыми упрощениями (подробнее см., например, [33, 234]) аппарат *диффузионной аппроксимации* существенно использует предположение о

постоянной занятости системы на изучаемом отрезке нестационарности. Подчеркнем, что термин «нестационарный» здесь относится к режиму изменения очереди, но не к процессам прибытия и обслуживания заявок порознь.

Известно [16, с. 91], что распределение количества событий потока восстановления со средним интервалом a за время t асимптотически нормально со средним и дисперсией

$$M[N] = t/a, \qquad D[N] = tD_A/a^3. \tag{8.11.1}$$

соответственно. В частности, эти соотношения (в других обозначениях) следуют из теории семиинвариантов — см. разд. 2.1.10. Аналогичные формулы можно установить и для процесса обслуживания. Тогда изменение Δ числа заявок в системе за достаточно большое время t распределено по закону, близкому к нормальному, с параметрами

$$\begin{aligned} M_\Delta &= t(1/a - 1/b), \\ D_\Delta &= t(D_A/a^3 + D_B/b^3). \end{aligned} \tag{8.11.2}$$

Следовательно,

$$P\{\Delta \le N\} = \Phi_0\Big(\frac{\Delta - M_\Delta}{\sqrt{D_\Delta}}\Big) + 0.5,$$

где $\Phi_0(z) = \dfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{z} e^{-x^2/2}\, dx.$

Найдем теперь распределение времени рассасывания накопившейся очереди длины N. *Уменьшение* очереди за время t имеет математическое ожидание

$$M_\Delta^- = t(1/b - 1/a)$$

(при условии $M_\Delta^- < N$) и дисперсию, вновь вычисляемую согласно второй из формул (8.11.2). Пусть $X(t)$ — случайное уменьшение очереди за время t, T_N — случайное время уменьшения очереди на N единиц. Тогда

$$P\{T_N < t\} = P\{X(t) > N\} = 0.5 - \Phi_0\Big(\frac{N - M_\Delta^-}{\sqrt{D_\Delta}}\Big).$$

При решении этих задач для многоканальных систем b и D_B определяются по методике суммирования потоков, описанной в разд. 2.1.9, применительно к процессам обслуживания (с учетом задействованного числа каналов).

Диффузионная аппроксимация динамики очереди рассматривается в [234] применительно к системе с большим числом каналов.

На основе теории семиинвариантов можно построить алгоритмы, учитывающие большее число моментов исходных распределений.

Глава 9

Приоритетные режимы

Приоритетными называются СМО, в которых одни заявки неоднородного потока имеют преимущественное право обслуживания перед другими. Основные варианты приоритетных СМО перечислены в разд. 2.3. К ним можно добавить часто применяемую в организации связи систему с чередованием приоритетов: вызовы других приоритетов обслуживаются только после исчерпания текущей очереди; затем начинается обслуживание непустой очереди с наивысшим приоритетом. Среди известных приоритетных дисциплин эта влечет за собой наименьшее число переключений прибора. Ряд приоритетных задач специального вида из области передачи данных в сетях связи рассмотрен в [58, 59].

Приоритет может назначаться в зависимости от момента прибытия заявки в систему и/или требуемого времени обслуживания; в этом случае имеет место непрерывная шкала приоритетов. При назначении фиксированных (внешних) приоритетов отдельным составляющим неоднородного потока нумерация последних выполняется в порядке убывания приоритетов.

Мы ограничимся рассмотрением одноканальных СМО, в которых выполнены следующие условия:

1) все заявки остаются в системе до завершения их обработки;

2) канал занят, если в системе имеется хотя бы одна заявка.

В случае абсолютного приоритета начатое обслуживание отмеченной заявки может многократно прерываться, так что определение времени ожидания начала обслуживания W практической ценности не имеет.

Поэтому в качестве основной результирующей характеристики мы примем распределение времени V пребывания в системе заявки каждого приоритетного класса и его моменты.

9.1. Элементарная теория

Для *консервативных* дисциплин, не создающих в системе дополнительной работы в случае прерывания (относительный приоритет и абсолютный приоритет с дообслуживанием прерванной заявки), расчет среднего времени пребывания в системе заявки j-го приоритета может быть выполнен элементарными средствами.

9.1.1. Относительный приоритет

До начала обслуживания прибывшей в систему с относительным приоритетом меченой j-заявки должно быть завершено начатое обслуживание (независимо от типа заявки), обслужены все ранее прибывшие заявки приоритетов $i \le j$ и вновь поступающие заявки приоритетов $i < j$. На основании закона сохранения объема работы (см. разд. 3.4) для средних длительностей названных событий можно записать уравнение баланса

$$w_j = \sum_{i=1}^{k} \lambda_i b_{i,2}/2 + \sum_{i=1}^{j} w_i \lambda_i b_{i,1} + w_j \sum_{i=1}^{j-1} \lambda_i b_{i,1}. \qquad (9.1.1)$$

Суммы в правой части (9.1.1) представляют ожидаемое время завершения: первая — начатого обслуживания (независимо от j), вторая — обслуживания ранее прибывших заявок данного и более высоких приоритетов, а третья — обслуживания заявок более высокого приоритета, прибывших *после* меченой. Среднее количество заявок каждого типа в очереди подсчитывается по формуле Литтла и умножается на среднюю длительность их обслуживания.

Из (9.1.1) получаем рекуррентную формулу

$$w_j = \Big(\sum_{i=1}^{k} \lambda_i b_{i,2}/2 + \sum_{i=1}^{j-1} \rho_i w_i\Big) \Big/ (1 - R_j), \qquad j = \overline{2,k},$$

где $R_j = \sum\limits_{i=1}^{j} \rho_i$ — суммарный коэффициент загрузки системы заявками до j-го приоритета включительно. Очевидно начальное выражение

$$w_1 = \sum_{i=1}^{k} \lambda_i b_{i,2}/[2(1-R_1)].$$

Выполняя последовательные подстановки, можно показать, что

$$w_j = \sum_{i=1}^{k} \lambda_i b_{i,2} \Big/ [2(1-R_{j-1})(1-R_j)]. \tag{9.1.2}$$

Справедливость этой формулы при $j=1$ непосредственно следует из (9.1.1), а для бóльших j доказывается по индукции.

Формула (9.1.2) естественно [183] обобщается на *непрерывный* случай, когда приоритет предоставляется по возрастанию длительностей обслуживания (дисциплина SPT — Shortest Processing Time):

$$w_t = \frac{\lambda b_2}{2[1-\lambda \int\limits_0^t x\, dB(x)]^2}. \tag{9.1.3}$$

При случайных длительностях обслуживания эта дисциплина не реализуема.

При перегрузке системы формула (9.1.2) нуждается в некоторой коррекции. Пусть j^* — индекс типа заявок, при котором наступает перегрузка. Тогда в стационарном режиме будет обслуживаться лишь часть потока заявок этого типа интенсивности $\hat{\lambda}_{j^*} = (1-R_{j^*-1})/b_{j^*,1}$, и вместо (9.1.2) применяется формула

$$w_j = \Big(\sum_{i=1}^{j^*-1} \lambda_i b_{i,2} + \hat{\lambda}_{j*} b_{j^*,2} \Big) \Big/ [2(1-R_{j-1})(1-R_j)], \qquad j = \overline{1, j^*-1}.$$

Эта оговорка относится и к рассмотренному ниже случаю смешанного приоритета.

Среднее время пребывания j-заявки в системе

$$v_j = w_j + b_{j,1}. \tag{9.1.4}$$

9.1.2. Абсолютный приоритет

При абсолютном приоритете наличие заявок типов $i > j$ не влияет на обслуживание j-заявок. Следовательно, среднее время *ожидания начала обслуживания* j*-заявки* можно получить по формуле вида (9.1.2), ограничив суммирование в числителе индексом j:

$$w'_j = \sum_{i=1}^{j} \lambda_i b_{i,2}/[2(1-R_{j-1})(1-R_j)]. \tag{9.1.5}$$

Для получения среднего времени *пребывания* j-заявки в системе к w'_j в данном случае нужно прибавить среднее время чистого обслуживания, которое при прерывании с дообслуживанием не зависит от числа прерываний и равно $b_{j,1}$, и среднюю длительность прерываний начатого обслуживания w''_j. *Продолжительность каждого прерывания равна длине периода непрерывной занятости системы заявками классов выше* j*-го*.

Вывод средней длины периода занятости $\bar{T}_B$ мы для простоты проведем применительно к однородному случаю. Пусть в систему $M/G/1$ поступает простейший поток заявок с параметром λ. Тогда среднее число заявок, обслуженных за период занятости, составит $\lambda\bar{T}_B$. Каждая из них в среднем потребует b единиц времени, а открывающая период непрерывной занятости головная заявка — b^*. Таким образом, $\bar{T}_B = b\lambda\bar{T}_B + b^*$, откуда

$$\bar{T}_B = b^*/(1-\lambda b).$$

В этой формуле числитель имеет смысл средней длительности обслуживания головной заявки периода занятости, а вычитаемое в знаменателе — коэффициент загрузки системы заявками, формирующими период занятости. Если головная заявка специфики обслуживания не имеет, то

$$\bar{T}_B = b/(1-\lambda b). \tag{9.1.6}$$

Теперь можно записать среднюю длительность одного прерывания j-заявки в виде

$$\bar{T}_B^{(j)} = \sum_{i=1}^{j-1} \frac{\lambda_i}{\Lambda_{j-1}} \frac{b_{i,1}}{1-R_{j-1}} = \frac{R_{j-1}}{\Lambda_{j-1}(1-R_{j-1})}.$$

Ожидаемое количество прерываний j-заявки составит $\Lambda_{j-1}b_{j,1}$. Таким образом, дополнительная задержка из-за прерываний начатого обслуживания j-заявки в среднем равна

$$w''_j = \Lambda_{j-1}b_{j,1}\bar{T}_B^{(j)} = \frac{R_{j-1}}{1-R_{j-1}}b_{j,1},$$

а суммарное время ожидания

$$w_j = \frac{R_{j-1}}{1-R_{j-1}}b_{j,1} + \frac{\sum\limits_{i=1}^{j}\lambda_i b_{i,2}}{2(1-R_{j-1})(1-R_j)}. \qquad (9.1.7)$$

В литературе часто предлагается учитывать ограниченную надежность обслуживающего устройства, рассматривая поток отказов как заявки высшего приоритета с прерыванием, а процесс ремонта — как их обслуживание. Этот прием допустим лишь при условии, что дополнительными отказами за время ремонта можно пренебречь.

9.1.3. Смешанный приоритет

Если в схеме абсолютных приоритетов имеется k типов заявок, то наивысшая кратность прерывания составляет $k-1$. В частности, для работы вычислительной системы необходимо иметь столько же областей сохранения (запоминания состояния) прерванной заявки. При больших k имеет смысл группировать типы заявок в *классы* с общим абсолютным приоритетом и относительными приоритетами внутри классов. Такой подход снижает суммарную частоту прерываний и количество потребных областей сохранения (оставляя по одной на каждый класс, кроме первого) и в то же время обеспечивает значительную дифференциацию оперативности обслуживания. Рассмотрим расчет среднего времени ожидания для заявки j-го типа. Обозначим j^+ наибольший индекс заявок, прерывающих j-ю, а j^- — нижнюю границу (т. е. наибольший индекс) класса, содержащего j-заявки. Эти обозначения иллюстрируются рис. 9.1. Заявка j-го типа будет ждать завершения начатого обслуживания, если она не имеет права его прерывания.

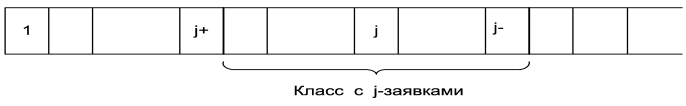

Рис. 9.1. Схема классов

Следовательно,

$$w'_j = \sum_{i=1}^{j^-} \lambda_i b_{i,2} \Big/ [2(1-R_{j-1})(1-R_j)]. \tag{9.1.8}$$

Период недоступности канала для j-заявки может быть начат любой заявкой, имеющей право ее прерывания. Средняя длительность обслуживания головной заявки такого периода

$$\bar{B}_{j^+} = \sum_{i=1}^{j^+} (\lambda_i/\Lambda_{j+}) b_{i,1} = R_{j+}/\Lambda_{j+}.$$

Образуют этот период занятости все заявки, имеющие приоритет перед j-й — с правом прерывания или даже без оного. В первом случае средняя длительность одного прерывания

$$T_j = \bar{B}_{j+}/(1-R_{j+}).$$

Во втором случае к прерыванию присоединяются и «старшие сестры» меченой j-заявки по классу, и

$$T_j = \bar{B}_{j+}/(1-R_{j-1}).$$

Ожидаемое количество прерываний составит $\Lambda_{j+} b_{j,1}$. Итак, суммарная ожидаемая длительность прерываний j-заявки

$$w''_j = \Lambda_{j+} b_{j,1} T_j. \tag{9.1.9}$$

Рассмотренная схема охватывает как частные случаи абсолютный приоритет (один тип заявок в каждом классе), относительный (единственный класс) и стандартный вариант смешанного (все типы заявок с абсолютным приоритетом образуют отдельные классы, заявки с относительным приоритетом объединяются в самый младший класс).

Заслуживает внимания еще один вариант смешанного приоритета — *по разности рангов*, не связанный с жесткой фиксацией классов. Здесь заявка типа i может прервать j-ю, если $j - i \geq m$, а в случае $1 \leq j - i \leq m-1$ пользуется только преимущественным правом выбора из очереди. В данном варианте среднее время ожидания начала обслуживания вновь подсчитывается по формуле (9.1.2), но во всех формулах

этого раздела индекс j^+ заменяется на $j-m$, а j^- — на $j+m-1$ (разумеется, с учетом границ всего множества типов заявок). Потери на прерывания учитываются только для заявок с индексом приоритета $j>m$.

При пороге $m=1$ схема вырождается в обычную систему чисто абсолютных приоритетов, при $m=k$ — в систему относительных приоритетов.

9.1.4. Замкнутые системы

В приложениях часто встречаются приоритетные системы замкнутого типа — с конечными объемами источников $\{K_j\}$, $j=\overline{1,k}$. Рассмотрим технику их расчета на уровне средних значений временны́х характеристик.

Прежде всего отметим, что полный цикл оборота j-заявки составляет

$$T_j = t_j + w_j + b_{j,1},$$

где t_j — средняя задержка в источнике и w_j — среднее время ожидания обслуживания. Соответственно средняя интенсивность их потока

$$\lambda_j = K_j/(t_j + w_j + b_{j,1}) \tag{9.1.10}$$

зависит от искомой средней длительности ожидания, что определяет итерационный характер алгоритма:

1. Задаться начальными значениями $w_j=0$.
2. Для всех j вычислить $\{\lambda_j\}$ согласно (9.1.10).
3. Для всех j, применяя формулы предыдущих разделов этой главы в зависимости от типа приоритета, рассчитать средние времена ожидания $\{w_j\}$ и согласно (9.1.10) — новые интенсивности $\{\lambda'_j\}$.
4. Если $\max\limits_j\{|\lambda'_j/\lambda_j-1|>\varepsilon\}$, заменить $\{\lambda_j\}$ на $\{\lambda'_j\}$. Перейти к этапу 3.
5. Конец алгоритма.

Возможен «зейделевский» вариант этого процесса с немедленной корректировкой λ_j до перехода к $(j+1)$-му типу.

9.1.5. Динамический приоритет

Важное прикладное значение имеет задача о $динамических$ приоритетах. В этом случае диспетчерский приоритет каждой заявки растет как известная функция времени ее ожидания, что исключает непомерно большие задержки заявок низкоприоритетных классов. В социальной сфере эта форма приоритета представляется наиболее целесообразной.

Следуя [50], рассмотрим простейший вариант этой задачи — расчет средних времен ожидания без прерывания начатого обслуживания при линейном росте диспетчерских приоритетов с угловыми коэффициентами $\{\beta_j\}$, $\beta_1 > \beta_2 > \ldots > \beta_k$. Среднее время ожидания j-заявки будет суммой четырех средних времен:

$S_0 = \sum_{i=1}^{k} \lambda_i b_{i,2}/2$ — завершения начатого обслуживания;

$S_1 = \sum_{i=1}^{j} \lambda_i w_i b_{i,1} = \sum_{i=1}^{j} \rho_i w_i$ — ожидания обслуживания ранее пришедших заявок приоритетов $i = \overline{1,j}$ (их j-заявка обгонять не может);

S_2 — ожидания обслуживания заявок приоритетов $i = \overline{1,j-1}$, которые прибыли после меченой j-заявки и успевают ее обогнать;

S_3 — ожидания обслуживания ранее пришедших заявок типов $i = \overline{j+1,k}$, которые j-заявка обгонять может, но не успевает.

Для определения S_2 и S_3 рассмотрим графики рис. 9.2 возрастания диспетчерских приоритетов, где началу координат соответствует момент прихода в систему меченой j-заявки.

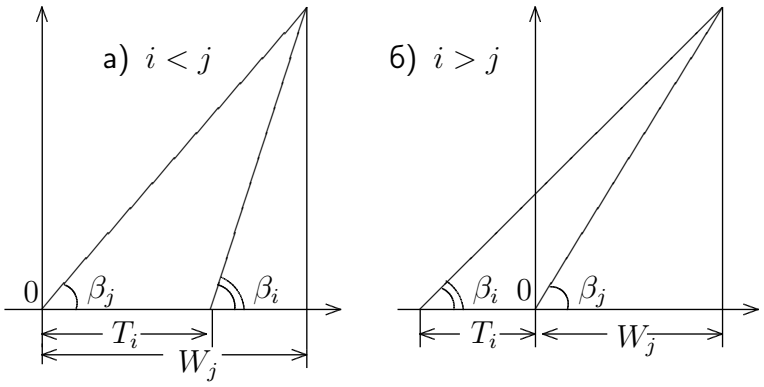

Рис. 9.2. Рост диспетчерских приоритетов

Из показанного в варианте а) предельного условия равенства динамических приоритетов в момент W_j выбора меченой j-заявки на обслуживание находим $(W_j - T_i)\beta_i = W_j\beta_j$, откуда следует выражение для критического момента поступления i-заявки $T_i = W_j(1 - \beta_j/\beta_i)$. Заявка типа i, пришедшая не более чем на T_i позже меченой, успевает ее обогнать. Математическое ожидание числа i-заявок, обогнавших меченую, составит $\lambda_i w_j(1-\beta_j/\beta_i)$; для получения их трудоемкости найденное число заявок должно быть умножено на $b_{i,1}$. Значит,

$$S_2 = \sum_{i=1}^{j-1} w_j(1 - \beta_j/\beta_i)\lambda_i b_{i,1} = w_j \sum_{i=1}^{j-1}(1 - \beta_j/\beta_i)\rho_i.$$

В свою очередь, меченая заявка сама может обгонять ранее пришедшие типов $i > j$, которые имеют по отношению к ней опережение меньше T_i — см. вариант б). Очевидно, $\beta_i(T_i + W_j) = \beta_j W_j$, откуда $T_i = W_j(\beta_j/\beta_i - 1)$. Ожидаемая загрузка системы от прибывших за это время низкоприоритетных заявок равна

$$w_j \sum_{i=j+1}^{k} \lambda_i b_{i,1}(\beta_j/\beta_i - 1) = w_j \sum_{i=j+1}^{k} \rho_i(\beta_j/\beta_i - 1).$$

Общая ожидаемая загрузка по заявкам этих типов составит $\sum\limits_{i=j+1}^{k} w_i\rho_i$. Таким образом, дополнительная задержка j-заявок из-за менее приоритетных

$$S_3 = \sum_{i=j+1}^{k} w_i\rho_i - w_j \sum_{i=j+1}^{k} \rho_i(\beta_j/\beta_i - 1).$$

Объединяя найденные составляющие в условие баланса объема работы, имеем

$$w_j = S_0 + \sum_{i=1}^{j} w_i\rho_i + w_j \sum_{i=1}^{j-1} \rho_i(1 - \beta_j/\beta_i) + \sum_{i=j+1}^{k} w_i\rho_i - w_j \sum_{i=j+1}^{k} \rho_i(\beta_j/\beta_i - 1).$$

Но согласно (3.4.3) $\sum\limits_{i=1}^{k} w_i\rho_i = S_0R/(1 - R)$, где $R = R_k$ — суммарный

коэффициент загрузки. Далее,

$$\begin{aligned} & \sum_{i=1}^{j-1} \rho_i(1-\beta_j/\beta_i) - \sum_{i=j+1}^{k} \rho_i(\,\beta_j/\beta_i - 1) \\ = \;& \sum_{i\neq j} \rho_i - \beta_j\Big(\sum_{i\neq j} \rho_i/\beta_i\Big) = R - \rho_j - \beta_j\Big(\sum_{i=1}^{k} \rho_i/\beta_i - \rho_j/\beta_j\Big) \\ = \;& R - \beta_j \sum_{i=1}^{k} \rho_i/\beta_i. \end{aligned}$$

Теперь условие баланса можно переписать в форме

$$w_j = S_0/(1-R) + w_j\Big(R - \beta_j \sum_{i=1}^{k} \rho_i/\beta_i\Big).$$

Окончательная формула

$$w_j = \frac{S_0/(1-R)}{1 - R + \beta_j \sum\limits_{i=1}^{k} \rho_i/\beta_i} \tag{9.1.11}$$

оказывается неожиданно простой. Она исправляет ошибки (различные), допущенные в [50, с. 158] и [82, с. 103], и, в отличие от названных источников и [5], не требует рекуррентного счета. Отметим в заключение, что коэффициенты $\{\beta_j\}$ входят в эту формулу лишь через их отношение, и один из них можно выбрать произвольно (например, положить $\beta_1 = 1$).

Несложно получить обобщение описанной задачи на случай динамических приоритетов *со стартовыми вкладами* при убывании последних по индексам типов заявок — см. разд. 9.8.7.

В [5] обсуждается также система с абсолютным динамическим приоритетом. На наш взгляд, состоятельность этой модели сомнительна хотя бы потому, что она требует наращивания приоритета заявки, уже проходящей обслуживание, и *прогнозирования* момента прерывания.

9.2. Эффект и назначение приоритетов

9.2.1. Эффект приоритетов

В соответствии с формулой (3.4.3) введение любой системы приоритетов приводит к сокращению среднего времени ожидания заявок

одних типов *за счет его увеличения для других*. Для количественной иллюстрации влияния типа дисциплины на среднее время ожидания рассмотрим совокупность 10 одинаковых потоков с показательно распределенной длительностью обслуживания.

На рис. 9.3 представлены средние времена ожидания при бесприоритетном обслуживании, относительном приоритете, абсолютном приоритете с дообслуживанием и комбинированном варианте с группировкой в классы 1–3, 4–7 и 8–10 типов. Номера типов отложены по оси абсцисс. Видно, что улучшение обслуживания заявок высшего приоритета достигается за счет низшего, причем этот эффект в случае абсолютного приоритета проявляется сильнее. Относительный выигрыш заметно больше относительного проигрыша.

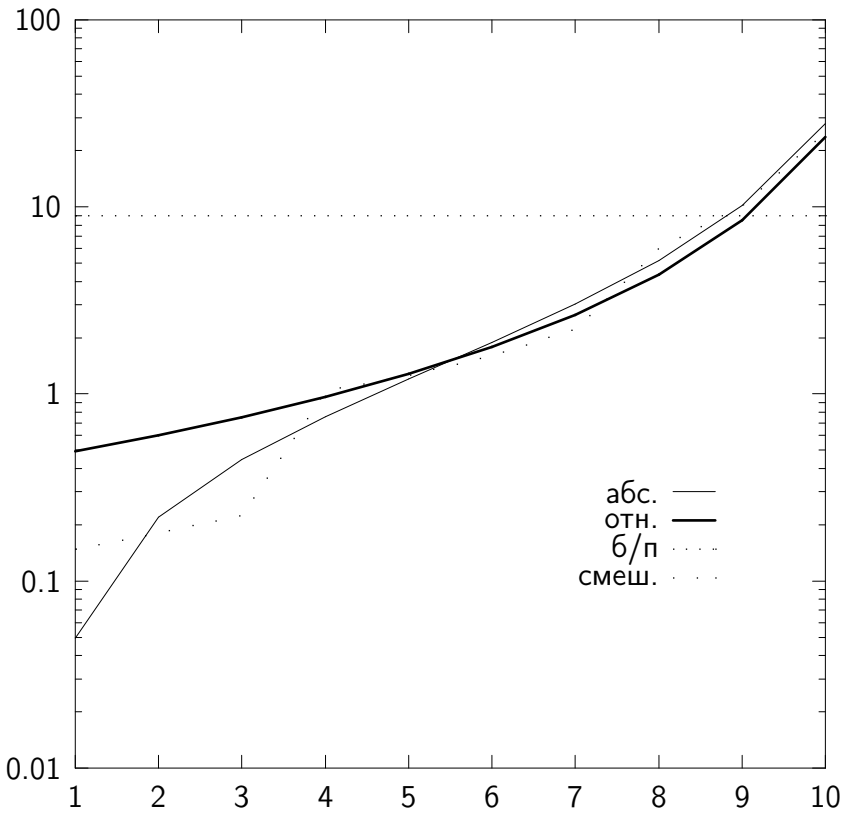

Рис. 9.3. Средние времена ожидания и дисциплины обслуживания

На рис. 9.4 качественно показано влияние увеличение интенсивности суммарного потока Λ на средние времена ожидания при неизменных $\{b_{i,1}\}$ и долях $\{\lambda_i/\Lambda\}$ заявок каждого вида с относительным приоритетом в суммарном потоке. Абсциссы $\Lambda_1, \Lambda_2, \Lambda_3$ вертикальных асимптот суть точки, где обращаются в единицу кумулянтные коэффициенты загрузки R_1, R_2 и R_3 соответственно. При значениях $\Lambda > \Lambda_3$ заявки 3-го типа обречены на бесконечное в среднем ожидание, тогда как остальные могут продолжать обслуживаться с приемлемым качеством.

При $\Lambda_2 < \Lambda < \Lambda_1$ конечное среднее время ожидания будут иметь только заявки 1-го типа.

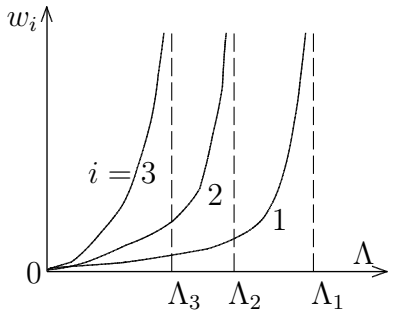

Рис. 9.4. Влияние интенсивности потока на ожидание

Таким образом, введение приоритетного обслуживания может рассматриваться как мера частичной защиты системы от временной перегрузки путем откладывания до лучших времен обслуживания менее приоритетных заявок. Среднее время последующего рассасывания скопившейся очереди можно определить по методике разд. 8.11. При малой загрузке введение приоритетов смысла не имеет.

9.2.2. Оптимальное назначение приоритетов

С точки зрения пользователя естественно считать, что «штрафы» в единицу времени пропорциональны средней длине очереди заявок каждого вида. Для простоты и наглядности рассмотрим случай двух видов заявок [20, с. 17]. Введем целевую функцию

$$U = \sum_{i=1}^{2} c_i q_i.$$

Нужно найти условия, при которых перестановка номеров типов (и, соответственно, инверсия их относительных приоритетов) приводит к невозрастанию целевой функции:

$$c_1 q_1 + c_2 q_2 \le c_1 \bar{q}_1 + c_2 \bar{q}_2. \tag{9.2.1}$$

Здесь надчеркиваниями помечены средние длины очередей заявок соответствующих типов при перестановке их приоритетов. В силу закона сохранения объема работы при любом варианте приоритета

$$\sum_i \rho_i w_i = \sum_i \lambda_i b_{i,1} w_i = \sum_i q_i b_{i,1} = \text{const}.$$

Поэтому разделим левую часть неравенства (9.2.1) на $(q_1 b_{1,1} + q_2 b_{2,1})$, а правую — на $(\bar{q}_1 b_{1,1} + \bar{q}_2 b_{2,1})$:

$$\frac{c_1 q_1 + c_2 q_2}{q_1 b_{1,1} + q_2 b_{2,1}} \le \frac{c_1 \bar{q}_1 + c_2 \bar{q}_2}{\bar{q}_1 b_{1,1} + \bar{q}_2 b_{2,1}}.$$

После элементарных преобразований находим

$$b_{2,1} c_1 (\bar{q}_1 q_2 - q_1 \bar{q}_2) \ge b_{1,1} c_2 (\bar{q}_1 q_2 - q_1 \bar{q}_2).$$

Поскольку выражение в скобках заведомо положительно (в уменьшаемом собраны «длинные» очереди, а в вычитаемом — «короткие»), после сокращения на него получаем

$$c_1 / b_{1,1} \ge c_2 / b_{2,1}.$$

Заметим, что это условие не зависит от интенсивностей потоков. В [143, с. 223] оно выводится применительно к $произвольному$ числу типов заявок, но менее очевидным способом.

При экспоненциальном обслуживании оно справедливо и в классе абсолютных приоритетов. Последний результат может быть обобщен. Действительно, поскольку целевая функция

$$U = \sum_{i=1}^{k} c_i \lambda_i w_i$$

зависит только от первых двух моментов распределений $\{B_i(t)\}$, то и ограничения должны определяться этими парами моментов. Последо-

вательность абсолютных приоритетов будет оптимальна, если дополнительно выполнены неравенства

$$\begin{cases} \dfrac{2b_{i,1}c_i}{b_{i,2}} > \dfrac{c_j}{b_{j,1}} & \text{при } i < j; \\ \dfrac{2b_{i,1}c_i}{b_{i,2}} < \dfrac{c_j}{b_{j,1}} & \text{при } i > j. \end{cases}$$

Для показательных распределений они выполняются автоматически. Такая последовательность абсолютных приоритетов дает меньшее значение целевой функции, чем относительные.

Безотносительно к выполнению вышеприведенных условий назначение приоритетов по этой схеме считается хорошей эвристикой и в общем случае.

Рассмотренные выше основы элементарной теории приоритетных систем позволяют выполнить чрезвычайно поучительные расчеты. В частности, нас будет интересовать взвешенная коэффициентами загрузки сумма средних времена ожидания заявок разных приоритетов (заявленный ранее *инвариант объема работы*).

Зададим доли $\{\alpha_i\}$ заявок каждого типа в суммарном потоке интенсивности Λ и средние длительности обслуживания $\{b_{i,1}\}$. Общий коэффициент загрузки

$$\rho = \sum_i \Lambda \alpha_i b_{i,1},$$

откуда следует

$$\Lambda = \rho \Big/ \sum_i \alpha_i b_{i,1}.$$

Длительности обслуживания будем считать подчиненными гамма-распределению с общим коэффициентом формы β. Соответственно второй момент

$$b_2 = \frac{\beta(\beta+1)}{\mu^2} = b_1^2(1+1/\beta).$$

Ниже приводится таблица результатов расчета.

Таблица 9.1. Проверка сохранения объема работы

ρ	$\beta = 3$			$\beta = 1$			$\beta = 0.5$		
	PR	NP	FCFS	PR	NP	FCFS	PR	NP	FCFS
.50	1.268	1.319	1.319	0.879	0.879	0.879	0.619	0.586	0.586
.55	1.711	1.773	1.773	1.182	1.182	1.182	0.829	0.788	0.788
.60	2.299	2.373	2.373	1.582	1.582	1.582	1.104	1.055	1.055
.65	3.095	3.183	3.183	2.122	2.122	2.122	1.473	1.415	1.415
.70	4.204	4.307	4.307	2.872	2.872	2.872	1.983	1.914	1.914
.75	5.813	5.933	5.933	3.956	3.956	3.956	2.717	2.637	2.637
.80	8.300	8.439	8.439	5.626	5.626	5.626	3.843	3.751	3.751
.85	12.543	12.702	12.702	8.468	8.468	8.468	5.751	5.645	5.645
.90	21.180	21.360	21.360	14.240	14.240	14.240	9.614	9.494	9.494
.95	47.396	47.600	47.600	31.733	31.733	31.733	21.291	21.155	21.155

Проверка инварианта $\sum_i \rho_i w_i$ показала, что строгое равенство этой суммы ее аналогу для дисциплины FCFS имеет место только для относительного приоритета и, кроме того, для всех сопоставляемых дисциплин — *при показательно распределенных длительностях обслуживания* ($\beta = 1$). Для абсолютного приоритета сумма занижается при бóльших коэффициентах вариации и увеличивается — при меньших. В исследованном диапазоне значений разбаланс измерялся десятыми долями процента. Тем не менее, этот результат опровергает общепринятое в мировой литературе утверждение о постоянстве упомянутой суммы для *всех* консервативных дисциплин без оговорок относительно вариации распределений (единственное известное автору исключение — [165]).

Малость обнаруженного разбаланса позволяет пользоваться упомянутым утверждением для практических расчетов как хорошим приближением.

9.3. Распределение периода занятости

Более полная теория приоритетных СМО, которую мы изложим в основном по книге [55] и работам школы МГУ [51, 68, 86], существенно использует *распределение* периода непрерывной занятости системы.

9.3.1. Функциональное уравнение

Пусть $\beta(s)$ — ПЛС распределения длительности обслуживания заявки, $\pi(s)$ — то же для распределения периода непрерывной занятости. Интерпретируем s как параметр простейшего потока «катастроф» [204]. Тогда $\beta(s)$ и $\pi(s)$ можно истолковать как вероятности отсутствия катастроф за случайное время обслуживания и период занятости соответственно.

Назовем заявку «плохой», если в течение открываемого ею периода занятости происходит катастрофа. Простейший поток плохих заявок будет иметь интенсивность $\lambda(1-\pi(s))$, а суммарный поток неблагоприятных событий — $(s+\lambda(1-\pi(s)))$. Таким образом,

$$\pi(s)=\beta(s+\lambda(1-\pi(s))). \tag{9.3.1}$$

Для решения функционального уравнения (9.3.1) перепишем (9.3.1) в виде

$$y=\beta(s+\lambda-\lambda y) \tag{9.3.2}$$

и будем рассматривать y как переменную, зависящую от s. Правая часть (9.3.2) может быть вычислена при любом y, $0\le y\le 1$, по формуле вида (1.9.3). Следовательно, уравнение (9.3.2) определяет итерационный процесс, сходящийся при $y<1+s/\lambda$ и $\lambda b_1<1$ к его решению $y=y(s)$, и можно построить таблицу значений $y(s)$. Дальнейший ход решения соответствует методике разд. 1.6.

Стандартным методом разложения ПЛС по степеням аргумента можно показать, что моменты распределения периода непрерывной занятости

$$\begin{array}{rcl}\pi_1 &=& b_1/(1-\rho),\\ \pi_2 &=& b_2/(1-\rho)^3,\\ \pi_3 &=& b_3/(1-\rho)^4+3\lambda b_2^2/(1-\rho)^5.\end{array} \tag{9.3.3}$$

Первая из этих формул уже была выведена из элементарных соображений — см. (9.1.6). Еще два момента можно получить по рекуррентным соотношениям

$$\begin{aligned}\pi_4 &= \{b_4z^4+6b_3\lambda\pi_2z^2+b_2[3(\lambda\pi_2)^2+4\lambda\pi_3z]\}/(1-\rho),\\ \pi_5 &= \{b_5z^5+10b_4\lambda\pi_2z^3+b_3[15(\lambda\pi_2)^2z+10\lambda\pi_3z^2]\\ &+ b_2(5\lambda\pi_4z+10\lambda^2\pi_2\pi_3)\}/(1-\rho).\end{aligned}$$

Здесь $z = 1 + \lambda\pi_1$.

Квадрат коэффициента вариации для периода занятости

$$v_{\text{п}}^2 = \pi_2/\pi_1^2 - 1 = (\rho + v_B^2)/(1-\rho) \gg v_B^2. \tag{9.3.4}$$

Отметим, что в классе консервативных дисциплин распределение длительности периода занятости инвариантно к дисциплинам обслуживания.

9.3.2. Инверсионное обслуживание

Рассмотрим СМО с инверсионным (обратным) порядком обслуживания и прерываниями текущего обслуживания (дисциплина LCFS.PR). Вновь прибывшая (меченая) заявка может считаться открывающей период занятости, который окончится завершением ее обслуживания. Предварительно должны быть обслужены все заявки, прибывшие $после$ меченой. Таким образом, при этой дисциплине распределение длительности пребывания заявки в системе совпадает с распределением периода непрерывной занятости.

9.3.3. Период занятости с разогревом

Поступление заявки в свободную систему может потребовать некоторых дополнительных операций («разогрева»). Пусть их длительность подчинена распределению $F(t)$ с моментами $\{f_k\}$ и имеет ПЛС $\varphi(s)$. Применив к ПЛС длительности цикла с разогревом $\gamma(s)$ «катастрофическую» интерпретацию, получаем

$$\gamma(s) = \varphi(s + \lambda - \lambda\pi(s)). \tag{9.3.5}$$

Стандартными методами выводим моменты $\{g_i\}$ ПНЗ с разогревом:

$$\begin{aligned}
g_1 &= f_1 z, \\
g_2 &= f_1\lambda\pi_2 + f_2 z^2, \\
g_3 &= f_1\lambda\pi_3 + 3f_2\lambda\pi_2 z + f_3 z^3, \\
g_4 &= f_1\lambda\pi_4 + f_2[3(\lambda\pi_2)^2 + 4\lambda\pi_3 z] + 6f_3\lambda\pi_2 z^2 + f_4 z^4, \\
g_5 &= f_1\lambda\pi_5 + f_2(5\lambda\pi_4 z + 10\lambda^2\pi_2\pi_3) \\
&+ f_3[15(\lambda\pi_2)^2 z + 10\lambda\pi_3 z^2] + 10f_4\lambda\pi_2 z^3 + f_5 z^5.
\end{aligned} \tag{9.3.6}$$

9.4. Абсолютный приоритет

Дисциплину с абсолютным приоритетом (с прерыванием) обычно рассматривают в трех вариантах, различающихся судьбой прерванной заявки [55, 86]:

1) дообслуживание с места прерывания (`preemptive resume` — PR);

2) обслуживание заново с новой случайной длительностью в соответствии с тем же законом распределения (`preemptive repeat with resampling` — RS);

3) обслуживание заново с прежней случайной длительностью обслуживания (`preemptive repeat without resampling` — RW).

При необходимости повторения прерванного обслуживания вариант RS имеет место, когда случайность вносится главным образом обслуживающим устройством. Если же длительность обслуживания определяется индивидуальностью заявки, то ситуация описывается схемой RW.

При рассмотрении обслуживания j-заявок в системе с абсолютным приоритетом заявки классов $\overline{j+1,k}$ можно игнорировать. Для $j=1$ распределение времени пребывания можно определить по формуле Полячека—Хинчина (5.2.6).

Обозначим для заявки j-го типа через λ_j интенсивность входящего потока, положим $\Lambda_j = \sum\limits_{i=1}^{j} \lambda_i$ и определим ПЛС распределений:

$\beta_j(s)$ — потребной длительности не прерываемого обслуживания,

$\omega_j(s)$ — времени ожидания $\mathit{начала}$ обслуживания,

$h_j(s)$ — активного времени (с момента начала обслуживания до ухода заявки из системы),

$\nu_j(s)$ — полного времени пребывания заявки в системе,

$\pi_j(s)$ — непрерывной занятости системы заявками j-го типа и более приоритетными.

Очевидно, что для всех рассматриваемых дисциплин

$$\nu_j(s) = \omega_j(s) h_j(s). \tag{9.4.1}$$

9.4.1. Периоды непрерывной занятости

Период непрерывной занятости (ПНЗ) системы заявками типов $i = \overline{1, j-1}$ по отношению к j-заявкам выступает как время прерывания. Время прерывания и активное время зависят друг от друга рекуррентно: активное время заявок типа $j-1$ определяет время прерывания j-заявок и, следовательно, их активное время, и т. д. Заявки первого типа не прерываются, и ПНЗ для них определяется из функционального уравнения

$$\pi_1(s) = \beta_1(s + \lambda_1 - \lambda_1\pi_1(s)).$$

Соответственно моменты этого распределения можно вычислить согласно (9.3.3).

Сложнее обстоит дело с последующими типами заявок. Обозначим $\pi_{j,i}(s)$ ПЛС распределения цикла занятости системы заявками приоритета j и выше при условии, что цикл начался обслуживанием i-заявки, $i = \overline{1, j}$. Тогда

$$\begin{aligned} \pi_{j,j}(s) &= h_j(s + \lambda_j - \lambda_j\pi_{j,j}(s)), \\ \pi_{j,i}(s) &= \pi_{j-1,i}(s + \lambda_j - \lambda_j\pi_{j,j}(s)), \qquad i = \overline{1, j-1}, \\ \pi_j(s) &= \Lambda_j^{-1} \sum_{i=1}^{j} \lambda_i \pi_{j,i}(s). \end{aligned} \tag{9.4.2}$$

В первом из этих уравнений $\pi_{j,j}(s)$ определяется как ПЛС цикла занятости (формула (9.3.1)). Последующие $j-1$ уравнений задают $\{\pi_{j,i}(s)\}$ рекуррентно (катастрофы не будет, если за время занятости системы заявками приоритета до $j-1$ включительно не случится ни катастрофы, ни j-цикла, связанного с катастрофой). Заключительное уравнение усредняет $\{\pi_{j,i}(s)\}$ с весами $\{\lambda_i/\Lambda_j\}$.

Легко видеть, что определяемые $\pi_{j,j}(s)$ моменты снова вычисляются согласно (9.3.3) с заменой распределения обслуживания на распределение активного времени. Моменты последующих распределений находим как моменты распределения ПНЗ j-заявками с разогревом согласно (9.3.6), причем роль разогрева играет ПНЗ заявками до $(j-1)$-го приоритета включительно. Усреднение также можно выполнить на уровне моментов.

В разд. 9.4.2 будет получена также общая для всех дисциплин с прерыванием формула вычисления $\omega_j(s)$. В последующих пунктах этого раздела для каждой дисциплины выводятся упомянутые выше рекуррентные зависимости между $h_j(s)$ и $\pi_j(s)$, используемые при получении

$\omega_j(s)$ и $\nu_j(s)$, причем распределения активного времени определяются как ПНЗ с разогревом. В заключительном пункте приведены выражения для среднего времени пребывания заявки при дисциплинах RS и RW, дополняющие результаты первого раздела главы.

9.4.2. Распределение времени ожидания

В момент прибытия очередной заявки система может находиться в одном из следующих состояний:

а) «0» — канал свободен для непосредственного приема j-заявки (в частности, обслуживает заявки младших приоритетов);

б) «1» — канал в периоде занятости, начатом прибытием в свободную систему более приоритетных заявок;

в) «2» — канал в периоде занятости, начатом прибытием в свободную систему j-заявки.

Прежде всего определим вероятности этих состояний через средние времена $\{\tau_i\}$, $i = \overline{0,2}$, пребывания в них. Очевидно, $\tau_0 = 1/\Lambda_j$. В остальных случаях мы имеем цикл занятости j-заявками, причем в случае 1 — с разогревом, определяемым периодом занятости СМО заявками высших приоритетов. Согласно разд. 9.3,

$$\begin{aligned} \tau_1 &= \pi_{j-1,1}/(1-\lambda_j h_{j,1}), \\ \tau_2 &= h_{j,1}/(1-\lambda_j h_{j,1}), \end{aligned}$$

а средний интервал между началами свободных периодов

$$\tau = \tau_0 + (\Lambda_{j-1}\tau_1 + \lambda_j\tau_2)/\Lambda_j.$$

Вероятности состояний

$$x_i = \tau_i/\tau, \qquad i = \overline{0,2}.$$

В случае 0 плотность распределения времени ожидания представляет собой δ-функцию, ПЛС от которой равно 1. В случаях 1 и 2 искомая плотность — свертка остатка периода разогрева (см. формулу (1.7.2)) и

условного распределения времени ожидания, подсчитываемого по формуле Полячека—Хинчина при $p_0 = 1 - \lambda_j h_{j,1}$. Соответственно

$$\begin{aligned} \tilde{\omega}_1(s) &= \frac{p_0(1-\pi_{j-1}(s))}{\pi_{j-1,1}(s-\lambda_j+\lambda_j h_j(s))}, \\ \tilde{\omega}_2(s) &= \frac{p_0(1-h_j(s))}{h_{j,1}(s-\lambda_j+\lambda_j h_j(s))}. \end{aligned}$$

Результирующее ПЛС распределения времени ожидания

$$\omega_j(s) = x_0 + x_1\tilde{\omega}_1(s) + x_2\tilde{\omega}_2(s) = c\frac{s+\Lambda_{j-1}(1-\pi_{j-1}(s))}{s-\lambda_{j-1}(1-h_j(s))}, \qquad (9.4.3)$$

где

$$c = (1-\lambda_j h_{j,1})/(1+\Lambda_{j-1}\pi_{j-1,1}).$$

Стандартная техника разложений ПЛС дает формулы для рекуррентного нахождения моментов $\{w_i\}$ распределения ожидания j-заявки (индекс типа заявок для простоты опущен):

$$w_i = \left[c\frac{\Lambda\pi_{i+1}}{i+1} + \lambda_j \sum_{m=0}^{i-1} \frac{i\,!}{m\,!\,(i+1-m)\,!} w_m h_{i+1-m}\right] \Big/ (1-\lambda h_1), i = 1, 2, \ldots \qquad (9.4.4)$$

при начальном w_0=1.

9.4.3. PR-прерывания

За активное время j-заявки не случится «катастроф», если за чистое время ее обслуживания (при этой дисциплине нет потерь на повторение) не произойдет ни одной катастрофы ни непосредственно, ни в период прерываний. Поток катастроф на прерываниях имеет параметр $\Lambda_{j-1}(1-\pi_{j-1}(s))$. Следовательно, для $j \geq 2$

$$h_j(s) = \beta_j(s + \Lambda_{j-1}(1-\pi_{j-1}(s))). \qquad (9.4.5)$$

9.4.4. RS-прерывания

Определим как нежелательные события катастрофы и прерывания. За активный период не произойдет катастроф, если

а) за время обслуживания j-заявки не произошло нежелательных событий — вероятность $\beta_j(s+\Lambda_{j-1})$);

б) такое событие произошло (с вероятностью $1-\beta_j(s+\Lambda_{j-1})$), но

- это было прерывание, а не катастрофа (вероятность $\frac{\Lambda_{j-1}}{\Lambda_{j-1}+s}$),
- за время прерывания катастроф не произошло (вероятность $\pi_{j-1}(s)$),
- в последующей попытке выполнить обслуживание заново за активное время катастроф не было (вероятность $h_j(s)$).

Таким образом,

$$h_j(s)=\beta_j(s+\Lambda_{j-1})+[1-\beta_j(s+\Lambda_{j-1})]\frac{\Lambda_{j-1}}{\Lambda_{j-1}+s}\pi_{j-1}(s)h_j(s),$$

откуда

$$h_j(s)=\frac{\beta_j(s+\Lambda_{j-1})}{1-\frac{\Lambda_{j-1}}{\Lambda_{j-1}+s}[1-\beta_j(s+\Lambda_{j-1})]\pi_{j-1}(s)}. \qquad (9.4.6)$$

Многократное дифференцирование преобразованной формы уравнения (9.4.6) с последующей подстановкой $s=0$ позволило получить формулы для моментов распределения активного времени j-заявки (индекс типа заявки для простоты обозначений опущен):

$$\begin{aligned}
h_1 &= (1/\Lambda+\pi_1)(1/\beta(\Lambda)-1),\\
h_2 &= \frac{1}{\Lambda\beta(\Lambda)}\{2h_1+2\beta'(\Lambda)[1+\Lambda(\pi_1+h_1)]+\Lambda[1-\beta(\Lambda)](\pi_2+2\pi_1h_1)\},\\
h_3 &= \frac{1}{\Lambda\beta(\Lambda)}\{3h_2-3\beta''(\Lambda)[1+\Lambda(\pi_1+h_1)]+3\Lambda\beta'(\Lambda)(\pi_2+2\pi_1h_1+h_2)\\
&+ \Lambda[1-\beta(\Lambda)](\pi_3+3\pi_2h_1+3\pi_1h_2)\},\\
h_4 &= \frac{1}{\Lambda\beta(\Lambda)}\{4h_3+4\beta'''(\Lambda)[1+\Lambda(\pi_1+h_1)]\\
&- 6\Lambda\beta''(\Lambda)(\pi_2+2\pi_1h_1+h_2)+4\Lambda\beta'(\Lambda)(\pi_3+3\pi_2h_1+3\pi_1h_2+h_3)\\
&+ \Lambda[1-\beta(\Lambda)](\pi_4+4\pi_3h_1+6\pi_2h_2+4\pi_1h_3)\}.
\end{aligned} \qquad (9.4.7)$$

В этих формулах легко усмотреть подвыражения, задающие высшие моменты свертки распределений ПНЗ и активного времени; их следует вычислять последовательными (по мере получения «активных» моментов) обращениями к процедуре свертки в моментах. Входящие в (9.4.7)

производные

$$\beta^{(k)}(s) = (-1)^k \int_0^\infty t^k e^{-st}\, dB(t)$$

при $s = \Lambda$ можно преобразовать к виду

$$\beta^{(k)}(\Lambda) = \frac{k\,!}{(-\Lambda)^k} \int_0^\infty \frac{(\Lambda t)^k}{k\,!} e^{-\Lambda t}\, dB(t) \tag{9.4.8}$$

и вычислить методами, обсуждавшимися в разд. 2.1.8. В частности, для вырожденного распределения со средним b

$$\beta^{(k)}(\Lambda) = (-1)^k b^k e^{-\Lambda b}.$$

9.4.5. RW-прерывания

В отличие от варианта RS, здесь новые попытки обслуживания предпринимаются при одной и той же реализации длительности обслуживания t. Вероятность того, что очередная попытка окажется неудачной, но «катастрофы» не произойдет, есть произведение следующих вероятностей:

$1 - e^{-(s+\Lambda_{j-1})t}$ — произошло неблагоприятное событие;

$\Lambda_{j-1}/(\Lambda_{j-1} + s)$ — им оказалось прерывание;

$\pi_{j-1}(s)$ — за время прерывания катастрофы не будет.

Вероятность отсутствия катастроф и успешного завершения попытки равна $e^{-(s+\Lambda_{j-1})t}$.

Полная вероятность отсутствия катастрофы при распределении длительности обслуживания $B_j(t)$ и любом числе неудачных попыток

$$\begin{aligned} h_j(s) &= \int_0^\infty \sum_{n=0}^\infty \left\{ [1 - e^{-(s+\Lambda_{j-1})t}] \frac{\Lambda_{j-1}}{\Lambda_{j-1}+s} \pi_{j-1}(s) \right\}^n e^{-(s+\Lambda_{j-1})t}\, dB_j(t) \\ &= \int_0^\infty \frac{e^{-(s+\Lambda_{j-1})t}}{1 - \frac{\Lambda_{j-1}}{\Lambda_{j-1}+s}[1 - e^{-(s+\Lambda_{j-1})t}]\pi_{j-1}(s)}\, dB_j(t). \end{aligned} \tag{9.4.9}$$

Эффективный расчет этого интеграла возможен при подходящей аппроксимации $B_j(t)$. Применение гамма-плотности с поправочным

многочленом (1.9.1) требовало вычисления узлов и весов квадратурной формулы Чебышева—Лагерра в зависимости от среднего и дисперсии распределения $B_j(t)$, было чрезвычайно трудоемко и не позволяло оценить достигнутую точность интегрирования. При гиперэкспоненциальной аппроксимации, введя вспомогательные величины $q = s + \Lambda_{j-1}$ и $r = \Lambda_{j-1}\pi_{j-1}(s)$, можно переписать (9.4.9) в виде

$$h_j(s) = \sum_{i=1}^{m} y_i\mu_i \int_0^{\infty} \frac{e^{-qx}e^{-\mu_i x}}{1 - \frac{r}{q}[1 - e^{-qx}]}\, dx = \frac{q}{r}\sum_{i=1}^{m} y_i\mu_i \int_0^{\infty} \frac{e^{-qx}e^{-\mu_i x}}{z + e^{-qx}}\, dx,$$

где $z = q/r - 1$. После подстановки $u = e^{-qx}$ интеграл сводится к

$$h_j(s) = \frac{1}{r}\sum_{i=1}^{m} y_i\mu_i \int_0^{1} \frac{u^{\mu_i/q}}{z + u}\, du. \tag{9.4.10}$$

Интегрирование по этой формуле с применением составных квадратурных формул Гаусса и последовательным делением интервалов пополам снизило трудоемкость счета почти на два порядка.

Дальнейшая рационализация связана с малостью параметра z. Имея в виду это обстоятельство, запишем

$$\begin{aligned} I &= \int_0^{1} \frac{u^{\mu_i/q}}{z + u}\, du = \int_0^{1} u^{\mu_i/q}\Big(\frac{1}{u} + \frac{1}{z + u} - \frac{1}{u}\Big)\, du \\ &= \int_0^{1} u^{\mu_i/q - 1}\, du - z\int_0^{1} \frac{u^{\mu_i/q - 1}}{z + u}\, du. \end{aligned} \tag{9.4.11}$$

Первый интеграл равен отношению q/μ_i, а второй определяется численно, причем допустимая погрешность его вычисления по крайней мере на два порядка больше исходной.

Расчет $\pi_j(s)$ вновь производится согласно (9.4.2).

Определим загрузку системы заявками до j-го приоритета включительно. Пусть требуемая продолжительность обслуживания j-заявки равна t. Тогда условная средняя продолжительность неудачной попытки

$$\bar{x}_1(t) = \int_0^{t} \frac{x\Lambda_{j-1}e^{-\Lambda_{j-1}x}}{1 - e^{-\Lambda_{j-1}t}}\, dx = \frac{1 - (1 + \Lambda_{j-1}t)e^{-\Lambda_{j-1}t}}{\Lambda_{j-1}(1 - e^{-\Lambda_{j-1}t})}.$$

Ожидаемое число неудачных попыток равно $(1-\alpha)/\alpha$, где $\alpha = e^{-\Lambda_{j-1}t}$. Следовательно, общие потери

$$\tau_1(t) = \frac{e^{\Lambda_{j-1}t}}{\Lambda_{j-1}}[1-(1+\Lambda_{j-1}t)e^{-\Lambda_{j-1}t}] = (e^{\Lambda_{j-1}t}-1-\Lambda_{j-1}t)/\Lambda_{j-1}.$$

Добавление к ним длительности успешной попытки t и последующее интегрирование с учетом распределения t дают

$$\bar{\tau}_j = \frac{1}{\Lambda_{j-1}}\int_0^\infty (e^{\Lambda_{j-1}t}-1)\,dB_j(t) = [\beta_j(-\Lambda_{j-1})-1]/\Lambda_{j-1}. \qquad (9.4.12)$$

Итак, при дисциплине RW кумулянтная загрузка

$$R_j = \lambda_1 b_{1,1} + \frac{\lambda_2}{\Lambda_1}(\beta_2(-\Lambda_1)-1) + \ldots + \frac{\lambda_j}{\Lambda_{j-1}}(\beta_j(-\Lambda_{j-1})-1). \qquad (9.4.13)$$

9.4.6. Первые моменты распределения пребывания

Простейший расчет варианта PR уже приводился в разд. 9.1.2. Для двух других вариантов рассуждения оказываются значительно более сложными. Мы приведем только конечные результаты согласно [55]. Положим

$$x_j = \lambda_j \bar{\tau}_j^2. \qquad (9.4.14)$$

При дисциплине RS

$$\bar{\tau}_j^2 = \frac{2}{[\Lambda_{j-1}\beta_j(\Lambda_{j-1})]^2}[1-\beta_j(\Lambda_{j-1})-\Lambda_{j-1}E(T_j e^{-\Lambda_{j-1}T_j})], \qquad (9.4.15)$$

а среднее время пребывания j-заявки в системе

$$v_{j,1} = \frac{\bar{\tau}_j}{1-R_{j-1}} + \frac{\sum_{i=1}^{j}\{x_i(1-R_{i-1})+2R_{i-1}\lambda_i\bar{\tau}_i^2\}}{2(1-R_{j-1})(1-R_j)}. \qquad (9.4.16)$$

Для дисциплины RW

$$\bar{\tau}_j^2 = \frac{2}{\Lambda_{j-1}^2}[\beta_j(-2\Lambda_{j-1})-\beta_j(-\Lambda_{j-1})-\Lambda_{j-1}E(T_j e^{\Lambda_{j-1}T_j})], \qquad (9.4.17)$$

$$v_{j,1} = \frac{\bar{\tau}_j}{1-R_{j-1}} + \frac{\sum\limits_{i=1}^{j}\{x_i(1-R_{i-1}) + 2R_{i-1}\lambda_i E[(e^{\Lambda_{i-1}T_i}-1)^2]/\Lambda_{i-1}^2\}}{2(1-R_{j-1})(1-R_j)}. \tag{9.4.18}$$

В этих формулах T_i — случайная продолжительность не прерываемого обслуживания заявки i-го приоритета, а E — оператор математического ожидания.

Можно показать, что при использовании плотности распределения T_i вида (1.9.1)

$$E(T_i e^{-\Lambda T_i}) = b_{i,1}\Big(\frac{\mu}{\Lambda+\mu}\Big)^{\alpha+1}\sum_{i=0}^{N}\frac{g_i}{(\Lambda+\mu)^i}\frac{\Gamma(\alpha+1+i)}{\Gamma(\alpha+1)}, \tag{9.4.19}$$

т. е. множитель при $b_{i,1}$ подсчитывается согласно (1.9.3) с заменой α на $\alpha+1$.

9.5. Относительный приоритет

Для сравнения предпочтительности обслуживания заявок классов i и j важно лишь отношение порядка между i и j (иначе говоря, роль играет только знак разности $i-j$). Это позволяет для любого фиксированного j рассматривать три укрупненных класса приоритетов:

$a = \{1,2,\ldots,j-1\}$ — с высшим приоритетом,

j — с данным приоритетом,

$e = \{j+1, j+2, \ldots, k\}$ — с низшим приоритетом.

Для $j=1$ $a=\oslash$, а для $j=k$ (самый младший приоритет) $e=\oslash$. Параметры потоков по укрупненным классам должны суммироваться:

$$\Lambda_a = \Lambda_{j-1} = \sum_{i=1}^{j-1}\lambda_i, \qquad \Lambda_e = \sum_{i=j+1}^{k}\lambda_i,$$

а характеристики обслуживания (моменты) — усредняться: начальные моменты m-го порядка чистого времени обслуживания

$$\bar{b}_{a,m} = \Lambda_a^{-1}\sum_{i=1}^{j-1}\lambda_i b_{i,m}, \qquad \bar{b}_{e,m} = \Lambda_e^{-1}\sum_{i=j+1}^{k}\lambda_i b_{i,m}.$$

Усредненные распределения времени обслуживания (и преобразования Лапласа $\bar{\beta}_a(s)$ и $\bar{\beta}_e(s)$ от них) строятся по усредненным моментам методами, описанными в главе 1. При относительном приоритете загрузка системы заявками до j-го типа включительно

$$R_j = \sum_{i=1}^{j} \lambda_i b_{i,1}.$$

Для ожидающей j-заявки длительность обслуживания стоящей перед ней заявки того же типа равна чистой продолжительности ее обслуживания T_j плюс время, необходимое для обслуживания всех поступивших за время T_j a-заявок. Этот суммарный интервал мы назовем временем блокировки и обозначим T_{ja}. Время блокировки фактически эквивалентно периоду занятости с разогревом, в котором разогрев представляет собой обслуживание j-заявки, а в периоде занятости происходит обслуживание a-заявок. Тогда на основании (9.3.5) ПЛС плотности распределения времени блокировки

$$\gamma_{ja}(s) = \beta_j(s + \Lambda_a - \Lambda_a \pi_a(s)), \tag{9.5.1}$$

где $\pi_a(s)$ представляет собой ПЛС распределения периода занятости системы a-заявками и находится из уравнения типа (9.4.10) с заменой a на $j-1$. Первый момент времени блокировки согласно (9.3.6) равен

$$\bar{\tau}_{ja} = b_{j,1}/(1-\rho_a). \tag{9.5.2}$$

Соответственно создаваемая j-заявками приведенная загрузка

$$\rho_{ja} = \lambda_j \bar{\tau}_{j,a} = \rho_j/(1-\rho_a).$$

Заявка типа j, поступающая в свободную систему, немедленно принимается на обслуживание. В противном случае она приходит во время цикла с разогревом, который начинается с обслуживания a-, j- или e-заявки и заканчивается освобождением системы от a- и ранее пришедших j-заявок. Используя индекс $i \in \{a, j, e\}$ для указания типа цикла, аналогично (9.4.6) имеем условное ПЛС распределения длительности ожидания

$$\tilde{\omega}_i(s) = \frac{(1-\rho_{ja})[1-\eta_i(s)]}{\bar{T}_i[\lambda_j \gamma_{ja}(s) - \lambda_j + s]}. \tag{9.5.3}$$

Для a-цикла разогрев представляет собой ПНЗ a-заявками. Следовательно,

$$\eta_a(s) = \pi_a(s)$$

и на основании (9.3.3)

$$\bar{T}_a = \bar{b}_{a,1}/(1-\rho_a).$$

При j-цикле разогрев есть время блокировки для j-заявок, так что

$$\eta_j(s) = \gamma_{ja}(s), \qquad \bar{T}_j = b_{j,1}/(1-\rho_a).$$

В $e-$ цикле роль разогрева играет обслуживание e-заявки, а затем следует период занятости a-заявками. В этом случае

$$\eta_e(s) = \bar{\beta}_e(s+\Lambda_a - \Lambda_a\pi_a(s)), \qquad \bar{T}_e = \bar{b}_{e,1}/(1-\rho_a).$$

Для перехода к безусловному распределению вычислим вероятности соответствующих условий. Очевидно, вероятность свободного состояния

$$x_0 = 1-\rho,$$

где ρ — суммарный коэффициент загрузки. Средние интервалы между последовательными возвращениями системы в интересующие нас состояния

$$l_a = [(\Lambda_a(1-\rho)]^{-1}, \qquad l_j = [(\Lambda_j(1-\rho)]^{-1}.$$

Поскольку обслуживание каждой заявки класса e начинает e-цикл, имеем

$$l_e = \Lambda_e^{-1}.$$

Средние длительности соответствующих циклов можно получить по формуле

$$m_i = \bar{T}_i/(1-\rho_{ja}).$$

Выполняя соответствующие подстановки, получаем стационарные вероятности типов циклов по правилу $x_i = m_i/\lambda_i$:

$$\begin{aligned} x_a &= \rho_a(1-\rho)/(1-\rho_a-\rho_j), \\ x_j &= \rho_j(1-\rho)/(1-\rho_a-\rho_j), \\ x_e &= \rho_e/(1-\rho_a-\rho_j). \end{aligned}$$

С их учетом ПЛС безусловного распределения ожидания j-заявок

$$\begin{aligned}\omega_j(s) &= x_0 + \sum_{i\in\{a,j,e\}} x_i\tilde{\omega}_i(s) \\ &= \frac{(1-\rho)[s+\Lambda_a-\Lambda_a\pi_a(s)]+\Lambda_e[1-\bar{\beta}_e(s+\Lambda_a-\Lambda_a\pi_a(s))]}{\lambda_j\beta_j(s+\Lambda_a-\Lambda_a\pi_a(s))-\lambda_j+s}.\end{aligned}$$

Возвращаясь к обозначениям разд. 9.4 и полагая

$$\mu_j(s) = s + \Lambda_{j-1}(1-\pi_{j-1}(s)), \tag{9.5.4}$$

для ПЛС распределения времени ожидания j заявки в системе имеем формулу

$$\omega_j(s) = \frac{(1-\rho)\mu_j(s)+\Lambda_e[1-\bar{\beta}_e(\mu_j(s))]}{s-\lambda_j(1-\beta_j(\mu_j(s)))}. \tag{9.5.5}$$

В заключение этого раздела упомянем задачу с «инерционными приоритетами» (в терминологии [86, гл. 6] — с чередованием приоритетов): здесь система обслуживает заявки определенного вида до полного освобождения от них, после чего переходит к обслуживанию непустой очереди с наивысшим приоритетом.

9.6. Смешанный приоритет

При обслуживании в схеме классов и PR-прерываниях ПЛС распределения времени ожидания начала обслуживания j-заявки также определяется формулой (9.5.5), но с заменой общего коэффициента загрузки ρ на R_{j-}. При определении времени *пребывания* j-заявки следует учитывать длительность ее прерываний заявками старших приоритетных классов. Распределение активного времени j-заявки имеет ПЛС

$$\eta_j(s) = \beta_j(s+\Lambda_{j+}(1-\pi_{j-1}(s))).$$

9.7. Численные эксперименты

Для тестирования разработанных процедур и демонстрации влияния приоритетов на показатели обслуживания была выбрана задача с тремя потоками одинаковой интенсивности $\lambda = 0.4$ и средней трудоемкостью заявки $b_1 = 0.55$, что обеспечивает докритическую загрузку даже

для дисциплин с повторением прерванного обслуживания. Коэффициент загрузки без потерь на обслуживание заново составил 0.66.

На рис. 9.5 и 9.6 приведены графики ДФР (по типам заявок) для системы $\overrightarrow{M_3}/\overrightarrow{M_3}/1$ с абсолютным приоритетом и дообслуживанием (дисциплина PR) и с относительным приоритетом (NP).

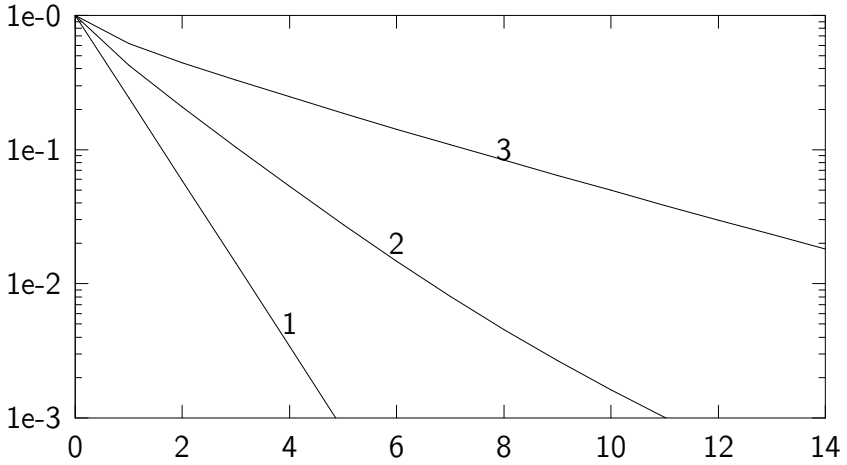

Рис. 9.5. ДФР пребывания в системе $\vec{M}_3/\vec{M}_3/1$ (PR)

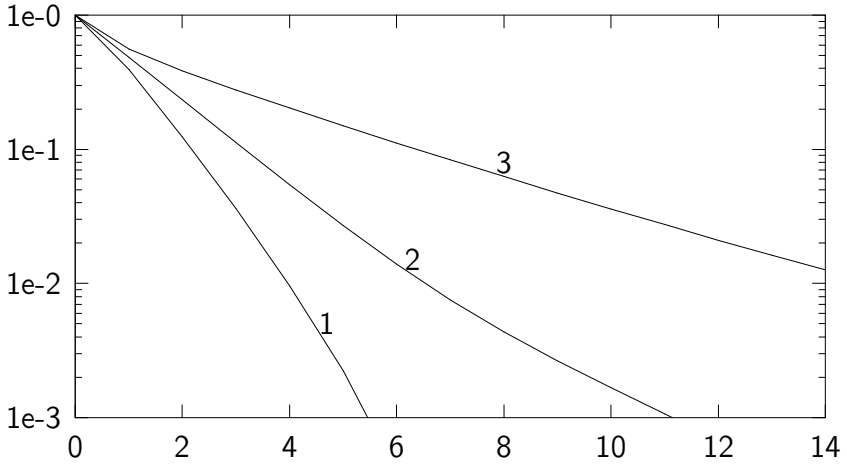

Рис. 9.6. ДФР пребывания в системе $\vec{M}_3/\vec{M}_3/1$ (NP)

Бросается в глаза большая разница в оперативности обслуживния заявок разных приоритетных классов. Переход к относительному приоритету

заметно ухудшает обслуживание заявок высших приоритетов и несколько улучшает его для низкоприоритетных заявок.

Рис. 9.7 и 9.8 иллюстрируют влияние дисциплин обслуживания и типа распределений чистого времени обслуживания на ДФР времени пребывания в системе заявок третьего типа, наиболее чувствительных к изменению входных данных. Как и следовало ожидать, уменьшение вариации длительности обслуживания перераспределяет время пребывания в пользу более коротких интервалов.

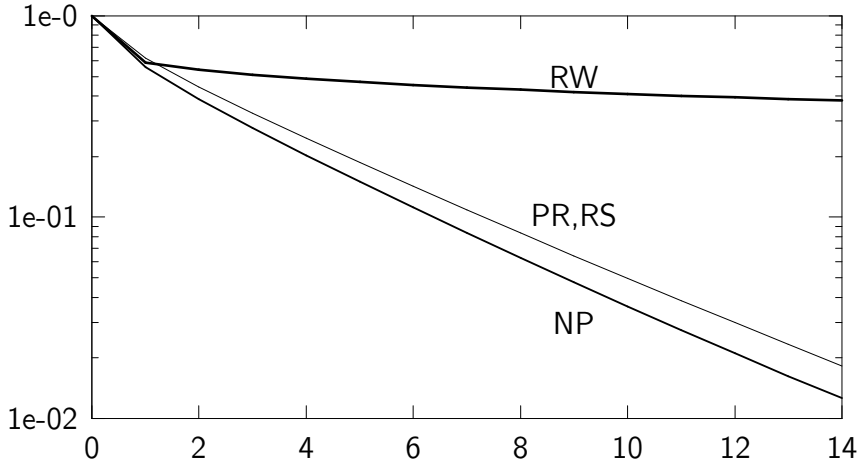

Рис. 9.7. Влияние дисциплины обслуживания — система $\vec{M}_3/\vec{M}_3/1$

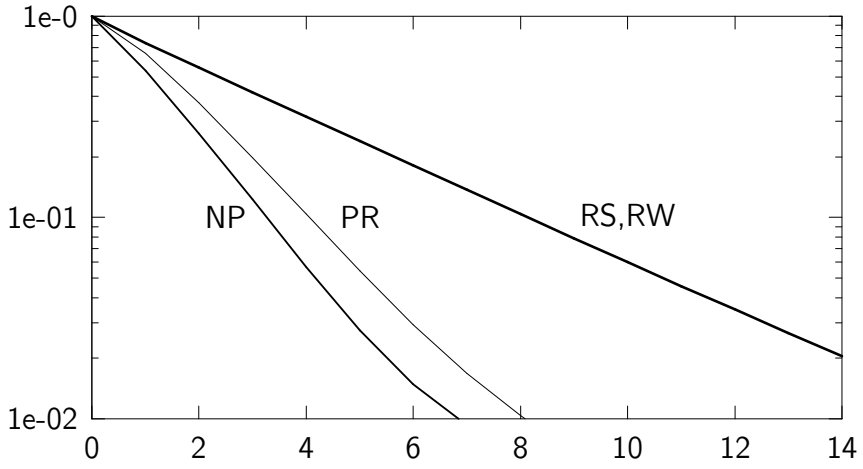

Рис. 9.8. Влияние дисциплины обслуживания — система $\vec{M}_3/\vec{D}_3/1$

Дисциплина RW (повторное обслуживание прерванной заявки с прежней длительностью) заметно затягивает «хвост» распределения времени пребывания.

Расчеты подтвердили предсказываемое теорией совпадение результатов по дисциплинам PR и RS для $\overrightarrow{M_3}/\overrightarrow{M_3}/1$ и по дисциплинам RS и RW — для $\overrightarrow{M_3}/\overrightarrow{D_3}/1$. Первые моменты, полученные по формулам разделов 9.1 и 9.4.6, практически совпали с найденными через ПЛС [100].

9.8. Многоканальные приоритетные системы

Системы обслуживания любого рода (технические и организационные) и назначения (производственные процессы, передача данных, здравоохранение, бытовое обслуживание, охрана общественного порядка, военное дело) время от времени нуждаются в профилактическом осмотре, проведении регламентных и ремонтных работ и, наконец, в отдыхе персонала. Необходимость *непрерывного* выполнения ответственных функций, хотя бы и с уменьшенной производительностью, определяет применение *многоканальных* систем обслуживания. Такие системы, особенно в период функционирования в неполном составе, эксплуатируются в режиме, близком к насыщению. Поэтому для достижения приемлемой оперативности обслуживания хотя бы по наиболее важным заявкам приходится вводить *приоритеты*.

Теория многоканальных систем обслуживания с приоритетами до настоящего времени практически не разработана. В обсуждающих ее немногих статьях делаются настолько стеснительные оговорки (как правило, одинаковое и к тому же экспоненциальное распределение времени обслуживания для всех типов заявок — см., например, [20]), что результаты оказываются бесполезными. Такие задачи обычно решаются лишь в простейшем (экспоненциальном) случае — см. обзор в [86, с. 400], причем средние длительности обслуживания предполагались различными буквально в единичных работах [65, 141, 194, 223]. Решение (методом производящих функций) оказывалось весьма громоздким.

В статье А. Д. Хомоненко [145] предлагается фазовая аппроксимация системы с абсолютным приоритетом и двумя классами заявок с экспоненциальным обслуживанием, рассчитываемая итерационным методом Такахаси—Таками [93, 94, 265]. Отказ от замены показатель-

ными законами распределений длительности обслуживания существенно усложняет диаграммы переходов. Еще сложнее анализ многоканальных систем в случае относительных приоритетов. Здесь редуцированная схема требует учета $трех$ потоков — данного (j-го) и двух объединенных потоков высших и низших приоритетов соответственно. Заметим, что группирование типов заявок требует перехода к $средневзвешенному$ распределению обслуживания, которое при различных экспоненциальных составляющих аппроксимировать с приемлемой погрешностью показательным законом нельзя. Практическую ценность могут иметь только алгоритмы, обобщенные в указанных направлениях. Добавим к этому, что все попытки создания реально применимых методик, учитывающих находящиеся в каналах и в очередях количества заявок каждого вида, заведомо обречены на неудачу в связи с непомерным разрастанием размерности пространства состояний.

В работе автора [123] намечена свободная от этого ограничения схема решения задачи для средних значений ожидания (пребывания) в системе заявок каждого вида и демонстрируются удовлетворительные результаты ее применения. Однако некоторые элементы этой схемы построены на эвристиках, полученных по результатам имитационного моделирования и не гарантирующих достаточную точность. Объективная сложность вероятностной ситуации оставляет мало надежды на строгое решение основных частей проблемы (распределения длительностей периодов недоступности системы для заявок данного приоритета и кратностей их прерывания).

Ниже предлагаются способы расчета средних времен $\{w_j\}$ ожидания и пребывания в многоканальной системе с относительным и абсолютным приоритетами при $произвольных$ распределениях длительностей чистого обслуживания, свободные от упомянутых ограничений. Методика основана на применении инвариантов отношения [20] и верифицируется с помощью имитационных моделей.

9.8.1. Инварианты теории очередей

Под инвариантами теории очередей в [20] понимаются соотношения, определяемые не распределениями в целом, но некоторым количеством их начальных моментов. К примеру, инвариантами являются формула Полячека—Хинчина для среднего времени ожидания в системе

$M/G/1$

$$w = \frac{\lambda b_2}{2(1-\lambda b_1)},$$

где λ — интенсивность входящего потока, а $\{b_i\}$ — моменты распределения длительности обслуживания, а также связывающие средние длину очереди и число заявок в системе со средними временами ожидания и пребывания в системе формулы Литтла. С другой стороны, параметр ω, определяющий геометрическое распределение числа заявок перед прибытием очередной заявки в систему $GI/M/1$

$$\pi_k = (1-\omega)\omega^k,$$

задается уравнением

$$\omega = \int_0^\infty e^{-\mu(1-\omega)t}\, dA(t), \tag{9.8.1}$$

где $A(t)$ — распределение интервалов между смежными заявками и μ — параметр экспоненциального распределения обслуживания. Следовательно, ω зависит от распределения $A(t)$ *в целом*, а не от конечного числа его моментов, и уравнение (9.8.1) инвариантом в указанном выше смысле не является.

Формулы разд. 9.1 теории одноканальных приоритетных систем обслуживания также могут рассматриваться как инварианты.

9.8.2. Инварианты отношения

К сожалению, перечень инвариантов для многоканальных систем обслуживания весьма ограничен. В частности, в [20] отмечено, что точная формула для среднего времени ожидания в системе $M/G/n$ не была получена в течение 40 лет, а с тех пор прошло еще 35. Еще хуже обстоит дело для многоканальных систем с приоритетами. В таких случаях для приближенного решения задачи можно применить *инварианты отношения* [20], основанные на символических пропорциях для искомых средних показателей. В интересующем нас случае это будет

$$\frac{M/G/n}{M/G/1} \approx \frac{\overrightarrow{M_k}/\overrightarrow{G_k}/n}{\overrightarrow{M_k}/\overrightarrow{G_k}/1}. \tag{9.8.2}$$

Заметим, что этот инвариант в [20] не обсуждается, поскольку к моменту выхода упомянутой книги ее авторы не располагали методами расчета

немарковских многоканальных систем, исключая метод Кроммелина для системы $M/D/n$ (обслуживание с постоянной длительностью). Инварианты отношения не слабее прямых инвариантов, обсуждавшихся в предыдущем разделе — в том смысле, что если верны прямые инварианты, то верны и соответствующие инварианты отношения.

Из (9.8.2) выводим базовое соотношение

$$\overrightarrow{M_k}/\overrightarrow{G_k}/n \approx \overrightarrow{M_k}/\overrightarrow{G_k}/1 \cdot \frac{M/G/n}{M/G/1}. \qquad (9.8.3)$$

9.8.3. Реализация инвариантов отношения

Прежде всего отметим, что символическое равенство (9.8.2) является приближенным, и точность построенного на его основе алгоритма должна подтверждаться при $n = 1$ — согласием с результатами счета по формулам разд. 9.1, а для $n > 1$ — с результатами имитационного моделирования.

Ошибка в использовании инвариантов отношения будет тем меньше, чем ближе к интересующей нас модели $\overrightarrow{M_k}/\overrightarrow{G_k}/n$ будут условия обсчета моделей $M/G/1$ и $M/G/n$. Для дисциплины *с прерываниями* наличие заявок с индексами $i > j$ никак не скажется на обслуживании заявок j-го типа. Поэтому упомянутые модели должны обсчитываться при суммарном потоке интенсивности $\Lambda_j = \sum_{i=1}^{j} \lambda_i$ и средневзвешенными моментами распределения обслуживания $b_{j,m} = \Lambda_j^{-1} \sum_{i=1}^{j} \lambda_i b_{i,m}, \quad m = 1, 2, \ldots$. Среднее время пребывания в одноканальной системе считается для уменьшенных в n раз интенсивностей потоков по формулам разд. 9.1.

Практически расчеты удобно выполнять применительно к модели $M/H_2/n$ с гиперэкспоненциальным 2-го порядка распределением обслуживания, параметры которого подбираются по трем средневзвешенным моментам. Расчет стационарных вероятностей состояний выполняется на основе итерационного метода Такахаси—Таками или метода матрично-геометрической прогрессии (см. главу 7).

Через стационарные вероятности состояний можно вычислить среднее число заявок (при его расчете можно ввести поправку на неограниченность очереди, дополнив реально вычисленные вероятности бесконечной убывающей геометрической прогрессией). Наконец, среднее время пребывания в системе определяется по формуле Литтла.

При *обслуживании без прерывания* расчетная схема оказывается несколько сложнее. Здесь для системы $M/G/n$ указанным выше способом по интенсивности входящего потока Λ_k и средневзвешенным моментам обслуживания $\bar{b}_{k,m} = \Lambda_j^{-1} \sum_{i=1}^{k} \lambda_i b_{i,m}$ определяются стационарные вероятности и через них — средняя длина очереди. Далее по формуле Литтла определяется общее для всех типов среднее время ожидания и добавлением к нему средней длительности обслуживания заявки каждого типа — среднее время пребывания их в системе. Расчет среднего времени пребывания заявок каждого типа в одноканальной системе выполняется аналогично — с учетом того, что среднее время ожидания определяется по суммарному потоку и средневзвешенным моментам обслуживания через формулу Полячека—Хинчина.

9.8.4. Имитационный эталон

Для тестирования предложенных алгоритмов были разработаны на современном Фортране программы соответствующих имитационных моделей. В модели системы с прерываниями на каждую заявку заводился паспорт, включающий в себя тип заявки, момент ее входа в систему, случайную трудоемкость, кратность прерываний, начало последнего прерывания данной заявки и суммарную длительность ее прерываний. Этот набор сведений был ориентирован на перспективную разработку аналитического метода расчета приоритетных систем. Содержание паспортов заносилось в каналы (при наличии свободных) или в очереди — раздельно по типам заявок по возрастанию моментов прибытия. После каждого события каналы переупорядочивались — с тем, чтобы кандидат на прерывание всегда оказывался последним.

Логика модели без прерываний оказалась значительно проще и в дополнительных комментариях не нуждается.

9.8.5. Численные результаты

При проведении численных экспериментов базовые интенсивности входящих простейших потоков принимались равными {0.222, 0.333, 0.445}, а средние длительности обслуживания — {0.45, 0.90, 1.35}. Для получения заданного суммарного коэффициента загрузки (0.9) интенсивности потоков умножались на этот коэффициент. Длительности обслуживания предполагались подчиненными гамма-распределению с парамет-

рами формы $\alpha = 3$ и $\alpha = 0.25$, что соответствовало коэффициентам вариации 0.577 и 2.0. Высшие моменты $\{b_i\}$ распределений обслуживания вычислялись через первые согласно

$$b_i = b_{i-1} \cdot b_1 \cdot [1 + (i-1)/\alpha], \qquad i = 2, 3.$$

В процессе расчета по этим моментам подбиралась H_2-аппроксимация (для первого α ее параметры оказывались комплексными, а для второго — вещественными).

В таблицах 9.2 и 9.3 приведены результаты расчета среднего времени пребывания заявок типов 1–3 в системе $\overrightarrow{M_k}/\overrightarrow{G_k}/n$ при относительном и абсолютном приоритетах в сопоставлении с результатами имитационных экспериментов. Прогон моделей осуществлялся до обработки 300 тыс. заявок первого типа.

Таблица 9.2. Относительный приоритет

Тип обслуж.	Метод	$n=1$			$n=3$			$n=5$		
		1	2	3	1	2	3	1	2	3
E_3	Имит.	1.20	2.06	11.99	0.72	1.33	4.52	0.61	1.16	3.08
	Инвар.	1.19	2.06	11.93	0.68	1.26	4.63	0.58	1.10	3.20
H_2	Имит.	3.20	5.18	37.52	1.09	2.10	12.54	0.76	1.52	7.67
	Инвар.	3.23	5.25	41.03	1.28	2.19	13.12	0.89	1.58	7.60

Таблица 9.3. Абсолютный приоритет

Тип обслуж.	Метод	$n=1$			$n=3$			$n=5$		
		1	2	3	1	2	3	1	2	3
E_3	Имит.	0.48	1.31	12.74	0.45	0.95	4.86	0.45	0.91	3.30
	Инвар.	0.48	1.31	12.69	0.45	0.83	4.85	0.45	0.80	3.41
H_2	Имит.	0.56	2.19	35.20	0.45	1.02	11.65	0.45	0.92	7.16
	Инвар.	0.56	2.20	41.79	0.45	0.89	13.33	0.45	0.81	7.79

Из этих таблиц можно сделать следующие выводы:

1) Расхождение результатов «инвариантного» счета и имитационного моделирования не превышает 15 %, что свидетельствует о корректности как расчетных схем обоих сравниваемых подходов, так и реализующих их компьютерных программ.

2) Упомянутая корректность дополнительно подтверждается: качественным соответствием результатов разумным ожиданиям — уменьшением по числу каналов среднего времени пребывания и разброса его по типам заявок, а также заметным увеличением этого разброса при переходе от относительных к абсолютным приоритетам.

3) Расхождение результатов допустимо в очень широком диапазоне коэффициентов вариации распределений длительности обслуживания (в рассмотренных примерах — от 0.577 до 2.0) и не обнаруживает тенденцию к росту при увеличении числа каналов.

Основная доля трудоемкости метода приходится на расчет модели $M/H_2/n$, повторяемый для абсолютного приоритета по числу типов заявок, а для относительного приоритета — однократно.

Разработанный метод опирается на максимальную приближенность опорных вариантов к обсчитываемой ситуации и дифференцированный учет влияния вида приоритетов. Так, для относительного приоритета по методу инвариантов вычислялось среднее время ожидания, к которому затем прибавлялось известное априорно среднее время обслуживания. Поэтому результаты счета для относительного приоритета согласуются с имитационным моделированием заметно лучше, чем для абсолютного. Таким образом, инварианты отношения следует применять для оценки только той части результирующей величины, которая не может быть вычислена точно.

Метод после очевидных модификаций позволяет распространить на многоканальный случай и дисциплину смешанного приоритета, а также использовать обсуждаемые в главе 7 модели с *немарковскими* входящими потоками.

Предложенная схема использования инвариантов отношения позволяет с достаточной для практических целей точностью рассчитывать средние времена пребывания заявок в *многоканальной* системе с *произвольными* распределениями длительности обслуживания. Она избавляет от необходимости имитационного моделирования, трудоемкого как на этапе программирования, так и в процессе счета (имитационная модель использовалась только для верификации расчетных методик). Попутно отметим, что стандартные системы моделирования типа GPSS вообще не позволяют моделировать прерывания многоканальных систем, а в случае одноканальных исключают вложенные прерывания.

9.8.6. Замкнутые системы

Рассмотрим технику приближенного расчета замкнутых приоритетных систем — с конечными объемами источников $\{K_j\}$, $j = \overline{1,k}$. Прежде всего отметим, что полный цикл оборота j-заявки составляет

$$T_j = t_j + w_j + b_{j,1},$$

где t_j — средняя задержка в источнике и w_j — среднее время ожидания обслуживания. Соответственно средняя интенсивность их потока

$$\lambda_j = K_j/(t_j + w_j + b_{j,1}), \tag{9.8.4}$$

она максимальна при нулевых $\{w_j\}$. Умножая обе части этого равенства на $b_{j,1}$ и суммируя по j, получаем *достаточное* условие корректности исходных данных в форме

$$\sum_{j=1}^{k} K_j b_{j,1}/(t_j + w_j + b_{j,1}) < n$$

(тогда гарантируется отсутствие перегрузки «разомкнутой» системы).

Поскольку согласно (9.8.4) интенсивности входящих потоков зависят от искомых средних длительностей ожидания, последние приходится определять методом итераций:

1. Задаться начальными значениями $w_j = 0$.
2. Для всех j вычислить $\{\lambda_j\}$ согласно (9.8.4).
3. Для всех j, применяя формулы предыдущих разделов в зависимости от типа приоритета, рассчитать средние времена ожидания $\{w_j\}$ и согласно (9.8.4) — новые интенсивности $\{\lambda'_j\}$.
4. Если $\max\limits_j\{|\lambda'_j/\lambda_j - 1| > \varepsilon\}$, заменить $\{\lambda_j\}$ на $\{\lambda'_j\}$. Перейти к этапу 3.
5. Конец алгоритма.

Итерации по этой схеме оказались неустойчивыми с явными признаками «слишком энергичной» пошаговой коррекции. Поэтому интенсивности потоков во всех итерациях, начиная со второй, определялись через полусумму средних времен ожидания j-заявки на двух последних

шагах. Стабилизация 5 значащих цифр наступала примерно за 10 шагов практически независимо от исходных данных.

Расчет выполнялся для трех типов заявок при средних длительностях обслуживания {0.45, 0.90, 1.35} и задержках в источнике {50, 30, 10}. Численность популяций при $n = 3$ и $n = 1$, удовлетворяющих условию (9.8.4), составила {4, 13, 21} и {1, 4, 7} соответственно. При обслуживании с коэффициентом вариации 2 были получены средние времена ожидания {0.526, 0.647, 3.254} и {2.056, 2.328, 6.517}, а при детерминированном обслуживании с теми же средними — {0.251, 0.290, 1.506} и {0.510, 0.581, 2.773}.

9.8.7. Системы с динамическим приоритетом

Рассмотрим методику расчета многоканальных систем с динамическим приоритетом по времени ожидания и дополнительным обобщением — стартовыми вкладами в приоритеты.

Компоненты задержки. Введем стартовые значения диспетчерских приоритетов $a_1 > a_2 > \ldots > a_k$ и угловые коэффициенты их последующего линейного роста $\{\beta_j\}$, $\beta_1 > \beta_2 > \ldots > \beta_k$. Среднее время ожидания j-заявки будет суммой четырех средних времен:

S_0 — завершения ближайшего начатого обслуживания;

$S_1 = \sum\limits_{i=1}^{j} \lambda_i w_i b_{i,1} = \sum\limits_{i=1}^{j} \rho_i w_i$ — ожидания обслуживания ранее пришедших заявок приоритетов $i = \overline{1,j}$ (их j-заявка обгонять не может);

S_2 — ожидания обслуживания заявок приоритетов $i = \overline{1,j-1}$, которые прибыли после меченой j-заявки и успевают ее обогнать;

S_3 — ожидания обслуживания ранее пришедших заявок типов $i = \overline{j+1,k}$, которые j-заявка обгонять может, но не успевает.

Для определения S_2 и S_3 рассмотрим графики рис. 9.9 возрастания диспетчерских приоритетов, где началу координат соответствует момент прихода в систему меченой j-заявки.

Из показанного в варианте а) предельного условия равенства динамических приоритетов в момент W_j выбора меченой j-заявки на обслуживание находим $(W_j - T_i)\beta_i + a_i = W_j\beta_j + a_j$, откуда следует выражение для критического момента поступления i-заявки $T_i =$

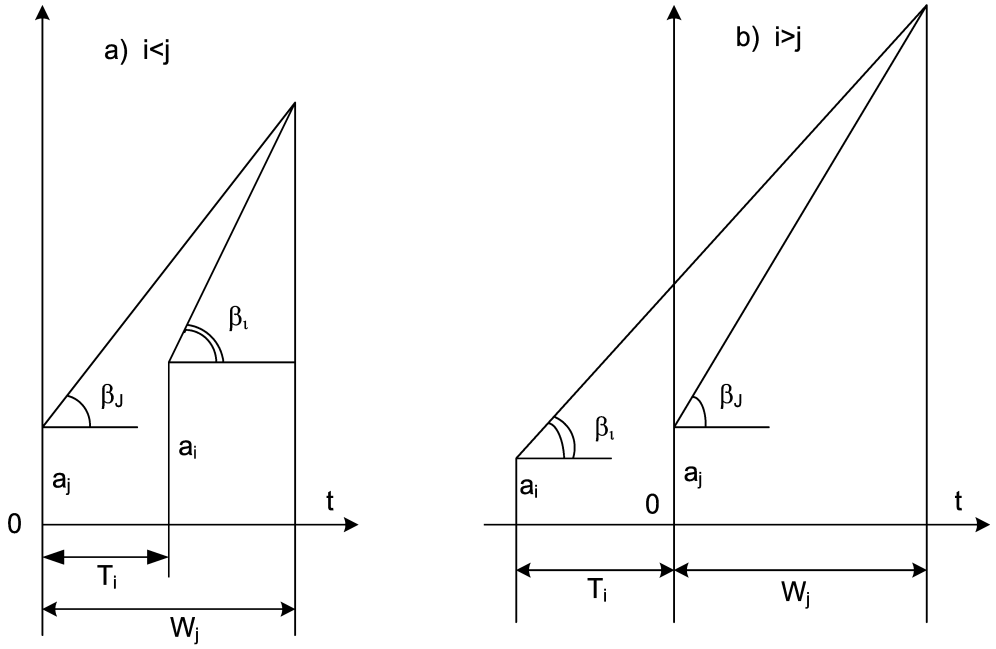

Рис. 9.9. Рост динамических приоритетов со стартовыми вкладами

$(a_i - a_j)/\beta_i + W_j(1 - \beta_j/\beta_i)$. Заявка типа i, пришедшая не более чем на T_i позже меченой, успевает ее обогнать. Математическое ожидание числа i-заявок, обогнавших меченую, составит $\lambda_i T_i$; для получения времени их обработки найденное число заявок должно быть умножено на $b_{i,1}/n$. Значит,

$$S_2 = \sum_{i=1}^{j-1} \frac{\rho_i}{\beta_i}(a_i - a_j) + w_j \sum_{i=1}^{j-1}(1 - \beta_j/\beta_i)\rho_i.$$

В свою очередь, меченая заявка сама может обгонять ранее пришедшие типов $i > j$, которые имеют по отношению к ней опережение меньше T_i — см. вариант б). Очевидно,

$$\beta_i(T_i + W_j) + a_i = \beta_j W_j + a_j,$$

откуда $T_i = (a_j - a_i)/\beta_i + W_j(\beta_j/\beta_i - 1)$. Вычитая из приходящейся на заявки этих типов доли инварианта Клейнрока загрузку системы не опережающими заявками, получаем дополнительную задержку j-заявок

из-за менее приоритетных

$$S_3 = \sum_{i=j+1}^{k} w_i\rho_i - \sum_{i=j+1}^{k} \frac{\rho_i}{\beta_i}(a_j - a_i) - w_j \sum_{i=j+1}^{k} \rho_i(\beta_j/\beta_i - 1).$$

Объединяя найденные составляющие в условие баланса объема работы, имеем

$$\begin{aligned} w_j = S_0 + \sum_{i=1}^{j} w_i\rho_i \quad &+ \sum_{i=1}^{j-1} \frac{\rho_i}{\beta_i}(a_i - a_j) \quad + w_j \sum_{i=1}^{j-1} \rho_i(1 - \beta_j/\beta_i) \\ &+ \sum_{i=j+1}^{k} w_i\rho_i + \sum_{i=j+1}^{k} \frac{\rho_i}{\beta_i}(a_i - a_j) + w_j \sum_{i=j+1}^{k} \rho_i(1 - \beta_j/\beta_i) \\ = S_0 + \sum_{i=1}^{k} w_i\rho_i \quad &+ \sum_{i=1}^{k} \frac{\rho_i}{\beta_i}(a_i - a_j) \quad + w_j \sum_{i=1}^{k} \rho_i(1 - \beta_j/\beta_i). \end{aligned}$$

Окончательно имеем равенство

$$w_j = S_0 + \sum_{i=1}^{k} w_i\rho_i + \sum_{i=1}^{k} \frac{\rho_i}{\beta_i}a_i - a_j \sum_{i=1}^{k} \frac{\rho_i}{\beta_i} + w_j\Big(R - \beta_j \sum_{i=1}^{k} \frac{\rho_i}{\beta_i}\Big).$$

Конечная формула

$$w_j = \frac{S_0 + \sum\limits_{i=1}^{k} \rho_i w_i + \sum\limits_{i=1}^{k} \frac{\rho_i}{\beta_i}a_i - a_j \sum\limits_{i=1}^{k} \frac{\rho_i}{\beta_i}}{1 - R + \beta_j \sum\limits_{i=1}^{k} \rho_i/\beta_i}, \qquad j = 1, 2, \ldots k. \qquad (9.8.5)$$

Осталось найти эффективные способы вычисления S_0 и $\sum\limits_{i=1}^{k} \rho_i w_i$.

Инвариант Клейнрока. В работе Л. Клейнрока [50] утверждается, что интересующая нас сумма $\sum_{i=1}^{k} \rho_i w_i$ инвариантна в классе *консервативных* дисциплин обслуживания, реализация которых не задает системе дополнительную работу. Автор с помощью надежно отлаженных программ для одноканальных моделей убедился в справедливости этого инварианта для относительных приоритетов и очень малой погрешности — для приоритета с прерыванием и дообслуживанием (погрешность была нулевой для показательных распределений времени обслуживания, а для прочих не превышала десятых долей процента). Имитационное моделирование подтвердило эти тезисы и для n-канальных систем. На этом основании можно считать, что интересующая нас сумма

$$\sum_{i=1}^{k} \rho_i w_i = RW, \tag{9.8.6}$$

где R — суммарный коэффициент загрузки, а W — среднее время ожидания, которое можно подсчитать для средневзвешенных трудоемкостей заявок с помощью известного алгоритма анализа модели $M/H_2/n$.

Среднее время до ближайшего обслуживания. Теперь попытаемся найти S_0. Вычислим средневзвешенные моменты распределения длительности обслуживания

$$\bar{b}_j = \sum_{i=1}^{k} \lambda_i b_{i,j}/\Lambda, \quad j = \overline{1,3}.$$

Здесь $\{b_{i,j}\}$ суть j-е моменты распределения длительности обслуживания i-заявки, а Λ — суммарная интенсивность потока заявок. По этим моментам найдем средневзвешенные моменты остаточной длительности обслуживания

$$\tilde{b}_j = \bar{b}_{j+1}/((j+1)\bar{b}_1), \quad j = 1, 2,$$

и аппроксимируем распределение остаточной длительности законом Вейбулла с дополнительной функцией распределения (ДФР)

$$\bar{F}(t) = \exp(-t^\gamma/T).$$

Моменты этого распределения

$$f_j = T^{j/\gamma}\Gamma(1 + j/\gamma), \quad j = 1, 2, \ldots$$

При обслуживании без прерывания среднее время ожидания вновь прибывшей заявкой ближайшего завершения обслуживания

$$\bar{F}_n(t) = \bar{F}_1^n(t). \tag{9.8.7}$$

Здесь индекс указывает число каналов. Интересующий нас результат

$$W = \int_0^\infty \bar{F}_n(t)\, dt.$$

Применяя формулу (9.8.7) к ДФР Вейбулла, убеждаемся, что она приводит к распределению того же типа с пересчитанным параметром

$T_n = T/n$ и прежней γ. Соответственно интересующее нас среднее время ожидания завершения ближайшего обслуживания в n-канальной системе

$$W(n) = (T/n)^{1/\gamma}\Gamma(1 + 1/\gamma).$$

Легко убедиться в том, что при переходе к n каналам уменьшение среднего времени ожидания

$$W(1)/W(n) = n^{1/\gamma}.$$

Ожидание завершения текущего обслуживания имеет место с вероятностью $1 - \sum_{i=0}^{n-1} p_i$, где $\{p_i\}$ суть вероятности наличия в системе i заявок, определяемые расчетом вышеупомянутой модели $M/H_2/n$. Итак,

$$S_0 = \frac{\tilde{b}_1}{n^{1/\gamma}}\Big(1 - \sum_{i=0}^{n-1} p_i\Big).$$

Сопоставление результатов. Предложенные подходы были запрограммированы на Фортране 90 для потока заявок трех типов. Обслуживание предполагалось регулярным с длительностями 2, 3 и 6; скорости роста приоритетов — 1, 0.7 и 0.5; доли заявок по типам — 0.0968, 0.2581, 0.6451. Суммарная интенсивность потока задавалась по формуле $0.155 * n$, что обеспечивало коэффициент загрузки 0.75. В качестве эталонных использовались результаты имитационного моделирования многоканальных систем при 200 тыс. наблюдений по заявкам высшего приоритета.

Таблица 9.4. Варианты расчета систем с динамическим приоритетом

Число каналов	Метод	Тип заявок		
		1	2	3
1	Аналитика	4.8677	6.5301	8.4552
	Аналитика с setup	4.6720	6.4385	8.4841
	Имитация с setup	5.1441	6.6139	8.4937
2	Аналитика	2.2660	3.0399	3.9360
	Аналитика с setup	2.0695	2.9461	3.9635
	Имитация с setup	2.2991	2.9087	3.7144
3	Аналитика	1.4084	1.8894	2.4464
	Аналитика с setup	1.2117	1.7963	2.4734
	Имитация с setup	1.3671	1.7442	2.2213

Прежде всего отметим качественное совпадение результатов с разумным прогнозом: увеличение числа каналов приводит к резкому уменьшению среднего времени ожидания. Введение стартовых (setup-) вкладов в динамические приоритеты, согласованных по порядку со скоростью их роста, приводит к дополнительному перераспределению времен ожидания в пользу заявок первого типа. Результаты аналитической setup-методики и соответствующей имитационной модели достаточно близки. Наблюденные расхождения вполне объясняются погрешностями выравнивания моментов распределений Вейбулла, гиперэкспоненциальной аппроксимацией при обсчете модели $M/H_2/n$, статистическими погрешностями имитации и неидеальностью датчиков случайных чисел.

Таким образом, предложенная методика расчета моделей с setup-вкладами в динамические приоритеты представляется достаточно надежно верифицированной и пригодной для практического применения.

Глава 10

Квантованное обслуживание

10.1. Общие положения

Одной из ведущих тенденций в современной информатике является *коллективное* использование имеющихся вычислительных ресурсов. В сравнении с монопольным доступом такой подход дает следующие преимущества:

1. Клиент системы платит лишь за фактически использованные ресурсы.

2. В связи с относительным удешевлением единицы продукции на более мощных установках эти ресурсы обходятся дешевле (согласно закону Гроша стоимость вычислительной установки растет пропорционально *квадратному корню* из ее производительности).

3. Улучшаются показатели обслуживания за счет масштабного эффекта — см. формулу (4.2.5).

4. Уменьшается доля необходимых страховых ресурсов мощностей.

5. На мощных установках поднимается потолок пиковых потребностей и появляется возможность оказывать качественно новые виды услуг.

Идея коллективного использования ресурсов по фактической потребности при сохранении психологических преимуществ индивидуального

доступа нашла свое отражение в операционных системах мощных стационарных ЭВМ (mainframes).

Ключевым элементом технологии коллективного использования процессора является $квантованное$ обслуживание, при котором каждая из находящихся в системе активных задач поочередно получает «квант» времени процессора. В зависимости от порядка его предоставления рассматриваются различные дисциплины обслуживания:

- циклические (RR — Round Robin, в вольном переводе — «карусель»);
- многоуровневые с N очередями разного приоритета ($\mathrm{FB_N}$ — Foreground-Background, т. е. с передним и задним планом);
- уравнительное разделение процессора (EPS — Egalitarian Processor Sharing).

Важнейшим показателем эффективности математической эксплуатации вычислительной системы коллективного доступа является количество задач, решенных на ней в единицу времени. Этот показатель будет максимален, если предоставить статический приоритет коротким заявкам. Однако в типичных условиях случайной трудоемкости эта стратегия физически не реализуема, поскольку длины заявок априорно не известны. Напротив, перечисленные выше дисциплины $автоматически$ обеспечивают приоритет коротким заявкам и к тому же по фактической, а не по ожидаемой длительности. Естественно, что делается это $за\ счет$ длинных заявок и $ценой$ дополнительных системных издержек на прерывания и обмены. Поэтому при анализе упомянутых дисциплин показатели обслуживания определяются в зависимости от длины заявки t. Основным показателем является среднее время v_t ее пребывания в системе или среднее время v_k пребывания в системе заявки, завершающей обслуживание на k-м кванте.

10.2. Циклическое обслуживание

В простейшем RR-варианте режима квантования времени каждая из находящихся в системе задач поочередно получает квант времени фиксированной длительности q. Если за это время задача не успевает решиться, то она возвращается в конец очереди. В противном случае

обслуженная заявка покидает систему, и на ее место процессор выбирает следующую. Переход к новой заявке всегда связан с затратами времени процессора τ=const. Отметим, что той же схемой описывается взаимодействие мультиплексного канала системной ЭЦВМ с устройствами ввода-вывода.

При анализе квантованных режимов обычно интересуются средним временем v_t пребывания в системе задачи с указанной длительностью счета t. Такая задача требует прохождения

$$r_t = [t/q] + 1 \tag{10.2.1}$$

циклов обслуживания, в каждый из которых, помимо собственно обслуживания, войдут системные потери τ и среднее время w ожидания в очереди к процессору. Следовательно,

$$v_t = t + r_t(\tau + w), \tag{10.2.2}$$

и ключом к получению результата является вычисление w.

10.2.1. Показательно распределенная трудоемкость

Расчет RR-системы удается провести сравнительно просто только при показательно распределенной продолжительности счета с плотностью вида $b(t) = \mu e^{-\mu t}$. Тогда длительность незавершенного обслуживания подчиняется тому же распределению независимо от номера цикла, и вероятность возврата на дообслуживание сохраняет постоянное значение $\alpha = e^{-\mu q}$. Работа такой системы описывается схемой рис. 10.1.

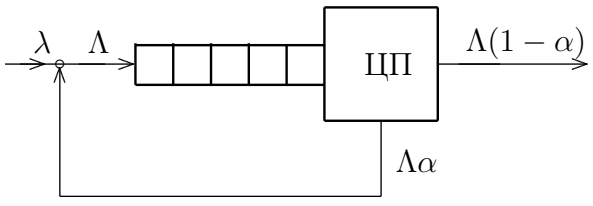

Рис. 10.1. Схема циклического обслуживания

Из рисунка ясно, что параметр потока заявок на входе в процессор $\Lambda = \lambda + \Lambda\alpha$, откуда

$$\Lambda = \lambda/(1 - \alpha). \tag{10.2.3}$$

Расчет моментов k-го порядка длительности неполного кванта можно свести к вычислению функции распределения закона Эрланга порядка $k+1$:

$$f_k = \int_0^q t^k \mu e^{-\mu t}\, dt = \frac{k\,!}{\mu^k}\Big(1 - e^{-\mu q} \sum_{i=0}^{k} \frac{(\mu q)^i}{i\,!}\Big).$$

Моменты распределения фактической длительности кванта

$$\varphi_k = f_k + \alpha q^k, \qquad k = 1, 2$$

(вероятность неполного кванта уже учтена в расчете $\{f_k\}$). В любом случае заход заявки на обслуживание связан с системными потерями τ, так что моменты распределения использованного времени

$$\begin{aligned} b_1 &= \tau + \varphi_1, \\ b_2 &= \tau^2 + 2\tau\varphi_1 + \varphi_2. \end{aligned}$$

Теперь можно рассчитать среднее время w ожидания заявки при очередном заходе на обслуживание по формуле Полячека—Хинчина (3.4.1), подставляя интенсивность входящего потока Λ из (10.2.3). Далее применяется формула (10.2.2). Отметим, что длительности ожидания заявки в последовательных турах не являются независимыми: если заявке пришлось долго ждать при очередном заходе, то ситуация скорее всего повторится и в следующем. Это обстоятельство, однако, никак не сказывается на расчете *первых* моментов.

На рис. 10.2 качественно показана зависимость $v_t(t)$ — ступенчатая ломаная со скачками в точках, кратных длительности кванта. Для сравнения штриховой линией показана аналогичная зависимость в обычной системе $M/G/1$. В данном случае RR-дисциплина оказывается предпочтительнее для заявок, укладывающихся в два кванта. Уменьшение длительности квантов заставляет систему более строго обеспечивать приоритет коротких заявок, но увеличивает сумму системных потерь.

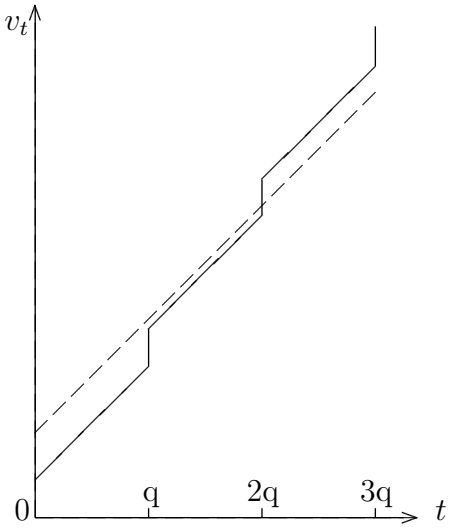

Рис. 10.2. Зависимость времени пребывания от трудоемкости

Минимальная величина кванта может быть найдена из условия $\Lambda b_1 = 1$, или

$$\frac{\lambda}{1-\alpha}\left\{\frac{1}{\mu}[1-\alpha(1+\mu q)]+\alpha q\right\}=1.$$

Его можно свести к $\alpha = 1 - \lambda\tau/(1-\lambda/\mu)$. Поскольку вероятность возврата $\alpha = e^{-\mu q}$, получаем пороговое значение кванта

$$q_{\min} = \frac{1}{\mu}\ln\frac{\mu-\lambda}{\mu-\lambda(1+\mu\tau)}.$$

При $\tau = 0$ ограничений по существу нет $(q_{\min} = 0)$. По опыту эксплуатации ЕС ЭВМ системные потери τ измеряются десятками миллисекунд (время страничного обмена с дисковой памятью).

10.2.2. Произвольное распределение трудоемкости

Пусть $B(t)$ — распределение потребной чистой длительности обслуживания. Тогда при постоянной длине q рабочего кванта вероятность недостаточности k квантов $\bar{F}_k = \bar{B}(kq)$. Обслуживание заявки завершается на k-м кванте с вероятностью $F_k - F_{k-1} = \bar{F}_{k-1} - \bar{F}_k$. Условная вероятность возврата на дообслуживание после исчерпания

i-го кванта равна $\bar{F}_i/\bar{F}_{i-1}$, а ее среднее значение за k туров составит $\left(\sum_{i=1}^{k} \bar{F}_i/\bar{F}_{i-1}\right)\Big/k$. Таким образом, средняя по распределению трудоемкости вероятность возврата

$$\alpha = \sum_{k=1}^{\infty} \frac{\bar{F}_{k-1} - \bar{F}_k}{k} \sum_{i=1}^{k} \bar{F}_i/\bar{F}_{i-1}. \qquad (10.2.4)$$

Легко убедиться, что при $\bar{B}(t) = e^{-\mu t}$

$$\alpha = \sum_{k=1}^{\infty} \frac{\bar{F}_{k-1} - \bar{F}_k}{k} \sum_{i=1}^{k} e^{-\mu q} = e^{-\mu q} \sum_{k=1}^{\infty} (\bar{F}_{k-1} - \bar{F}_k) = e^{-\mu q},$$

т. е. совпадает с ожидаемым результатом. Проверка формулы (10.2.4) для эрланговского распределения обслуживания 4-го порядка с $\mu = 4$ и $q = 0.7$ дала $\alpha =$0.526 (на имитационной модели при 10 тыс. испытаний — 0.519). Таким образом, формулу (10.2.4) можно считать вполне приемлемой. Заметим, что члены суммы по k, если квант не слишком мал, убывают очень быстро.

Для квантов фиксированной длительности q алгоритм расчета RR-системы выглядит следующим образом:

1. Вычислить в цикле по номерам захода $k = \overline{1, K}$, где предельный номер K выбирается из условия пренебрежимой малости разности $\bar{F}_{K-1} - \bar{F}_K$, следующие характеристики:

- вероятность возврата на дообслуживание $\bar{F}_k/\bar{F}_{k-1}$;
- частичные моменты суммарной трудоемкости обслуживания

$$b_{k,m}^{(p)} = \int_{Q_{k-1}}^{Q_k} t^m b(t)\, dt, \quad m = \overline{1, M}, \quad \text{где } Q_k = kq;$$

- «остаточные» моменты длительности неполного кванта

$$b_{k,m}^{(o)} = \int_{Q_{k-1}}^{Q_k} (t - Q_{k-1})^m b(t)\, dt = \sum_{i=0}^{m} \binom{m}{i} (-Q_{k-1})^i \int_{Q_{k-1}}^{Q_k} t^{m-i} b(t)\, dt$$
$$= \sum_{i=0}^{m} \binom{m}{i} (-Q_{k-1})^i b_{k,m-i}^{(p)};$$

- произвести накопления внутренних и внешних сумм в формуле (10.2.4) и аналогичной ей $b_m^{(o)} = \sum\limits_{k=1}^{\infty} \frac{\bar{F}_{k-1}-\bar{F}_k}{k} \sum\limits_{i=1}^{k} b_{i,m}^{(o)}$ для подсчета средневзвешенной длины неполного кванта.

2. Найти моменты распределения средневзвешенной длины кванта

$$b_m^{(W)} = \alpha q^m + (1-\alpha) b_m^{(o)}.$$

3. Свернуть их с системными потерями: $B^{(L)} = B^{(W)} * T$, где T — моменты распределения системных потерь (большими буквами обозначены *наборы* соответствующих моментов).

4. Вычислить коэффициент загрузки системы $R = \Lambda b_1^{(L)}$, где Λ вновь считается согласно (10.2.3).

5. При $R \geq 1$ система перегружена. Перейти к этапу 8.

6. Вычислить среднее время ожидания в цикле по формуле Полячека—Хинчина. В текущих обозначениях $w = \Lambda b_2^{(L)}/2(1-R)$.

7. Для $k = \overline{1,K}$ вычислить среднее время пребывания k-заявки в системе

$$v_k = (k-1)(w+\tau_1+q) + w + \tau_1 + b_1^{(o)} = k(w+\tau_1+q) + b_1^{(o)} - q.$$

8. Конец алгоритма.

Алгоритм допускает случайные потери на переключение, имеющие место, например, при страничном обмене или запросе на ввод-вывод. Безусловная средняя длительность ожидания при дисциплине RR меньше аналогичной характеристики при FCFS в классе УФИ-распределений и больше — в классе ВФИ-распределений.

Описанная расчетная схема опирается на предположение о независимости времен ожидания в последовательных турах, невыполнение которого должно сказаться на высших моментах. Имитационное моделирование показало, однако, что тривиальное обобщение расчетной схемы на высшие моменты посредством замены суммирования математических ожиданий на свертки в моментах дает вполне удовлетворительную точность для второго и третьего моментов, исключая высокие кратности захода $(k > 5)$. Более строгая, но громоздкая теория для зависимых туров предложена С. Ф. Яшковым в [153].

Выделение квантов фиксированной длительности — это наиболее простое, но не всегда лучшее решение. Прерывание решения в произвольной точке может вести к избыточным системным затратам (если

производится перед самым концом задания или в моменты максимального объема промежуточных данных) и как правило требует запоминания большого объема промежуточных результатов. Если разрешать прерывание задания только в указанных программистом точках, то эти недостатки будут сняты, но длина кванта станет случайной величиной. Тогда вероятность недостаточности k квантов нужно считать по формуле

$$\bar{F}_k = \int_0^\infty \bar{B}(t)\, dQ_k(t), \tag{10.2.5}$$

где $Q_k(t) = Q^{k*}(t)$ — k-кратная свертка распределения длительности кванта.

10.2.3. Продление квантов

Если в момент исчерпания кванта очередь оказывается пустой, имеет смысл немедленно предоставить очередной квант прерванной заявке. Системные потери при этом можно считать практически нулевыми. Указанный эффект может заметно повлиять на результаты расчета слабо загруженных систем. Рассмотрим способ его учета при постоянных τ и q. Пусть p_0 — вероятность отсутствия заявок в очереди. Тогда средняя длительность расширенного кванта

$$B_1 = (1-p_0)\tau + \alpha q + (1-\alpha)q/2 = (1-p_0)\tau + (\alpha+1)q/2. \tag{10.2.6}$$

Но $p_0 = 1 - \Lambda B_1$, где Λ вычисляется согласно (10.2.3). Подставляя в это равенство выражение для B_1, находим

$$p_0 = 1 - \frac{\lambda q(1+\alpha)}{2(1-\alpha-\lambda\tau)}.$$

Таким образом, коэффициент загрузки в нашем случае

$$\rho = 1 - p_0 = \frac{\lambda q(1+\alpha)}{2(1-\alpha-\lambda\tau)}.$$

Подстановка этих результатов в (10.2.6) дает для B_1 явное выражение

$$B_1 = \frac{q}{2}\,\frac{(1+\alpha)(1-\alpha)}{1-\alpha-\lambda\tau}.$$

Для дальнейшего нам понадобятся моменты распределения длительности расширенного кванта

$$q_i^* = \alpha q^i + (1-\alpha)\frac{q^{i+1}}{q(i+1)} = \frac{q^i}{i+1}(i\alpha+1). \tag{10.2.7}$$

Теперь можно вычислить

$$\begin{aligned} B_2 &= p_0\frac{q^2}{3}(2\alpha+1) + (1-p_0)[\frac{q^2}{3}(2\alpha+1) + \tau^2 + 2\tau\frac{q}{2} + 2\tau\frac{q}{2}(\alpha+1)] \\ &= \frac{q^2}{3}(2\alpha+1) + \rho\tau[\tau + q(\alpha+1)]. \end{aligned} \tag{10.2.8}$$

Дальнейший ход решения совпадает с описанным в разд. 10.2.1.

10.3. Разделение процессора

В случае пренебрежимо малых системных потерь имеет смысл рассмотреть предельное поведение системы при $q \to 0$. Здесь центральный процессор в каждый момент времени оказывается *разделенным* между всеми находящимися в системе заявками, и на основании закона сохранения объема работы $v_t = t + \lambda b v_t$, откуда следует

$$v_t = t/(1-\rho).$$

Таким образом, при данной дисциплине (EPS — Egalitarian Processor Sharing) среднее время нахождения в системе каждой заявки пропорционально ее длине — точнее, требуемому времени обслуживания при использовании всей мощности процессора. Эта дисциплина обеспечивает «естественный» режим функционирования вычислительной установки, поскольку пользователи лишаются основанных на нелинейных эффектах стимулов к агрегации коротких заявок или разбивке длинных.

Распределение числа заявок в одноканальных системах с дисциплинами EPS и LCFS.PR инвариантно к виду распределения (высшим моментам) распределения длительности обслуживания [18, 153] и имеет геометрическую форму:

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

По смыслу дисциплины EPS все результаты ее анализа переносятся на n-канальный случай соответствующим пересчетом производительности. Установлено [153], что дисциплина EPS хорошо аппроксимирует более сложную для анализа дисциплину RR даже при довольно больших размерах кванта. Эта гипотеза существенно упрощает расчеты.

10.4. Многоуровневое обслуживание

Затраты τ времени работы процессора на каждое переключение увеличивают общую загрузку системы циклического обслуживания и особенно сильно сказываются на прохождении «длинных», много раз прерываемых задач. Многоуровневая система типа показанной на рис. 10.3 [5, 50, 82, 92, 153, 183] дает наибольшие преимущества коротким заявкам и к тому же уменьшает частоту переключений. Эта дисциплина в [183] называется `Shortest Elapsed Time` (SET) и актуальна для однопроцессорных суперЭВМ конвейерного типа. В такой системе заявка, не завершившая обслуживание на i-м уровне в течение кванта q_i, переходит на следующий уровень и получает на нем квант $q_{i+1} > q_i$. Нижний уровень N предоставляет обслуживание до его завершения. Обслуживание заявок из разных очередей производится по схеме относительного приоритета. Переход к каждой новой заявке требует времени процессора τ независимо от уровня обслуживания.

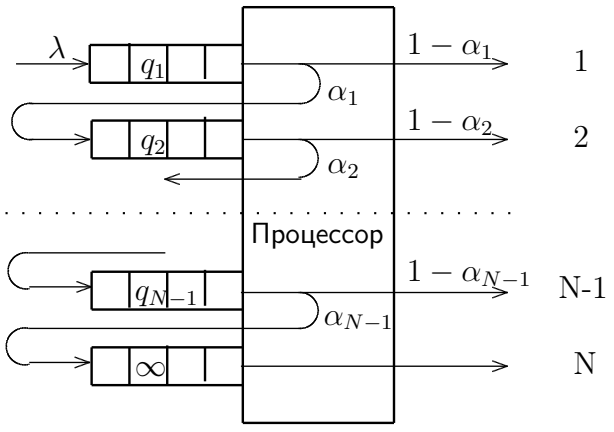

Рис. 10.3. Многоуровневая система с квантованием времени

В системах подобного типа используются расценки, дифференцированные по уровням обслуживания. Стоимость обслуживания заявки трудоемкости t, покидающей систему с уровня k, может подсчитываться, например, по формуле

$$\tilde{L}_k = \sum_{i=1}^{k} c_i w_i + \sum_{i=1}^{k-1} d_i(q_i + \tau) + d_k(t - \sum_{i=1}^{k-1} q_i + \tau), \qquad (10.4.1)$$

где c_i и d_i — стоимости пребывания заявки в единицу времени соответственно в очереди и на обслуживании с приоритетом i. Учитывая, что фактическая длительность обслуживания не зависит от дисциплины очереди, имеет смысл перейти от (10.4.1) к функции

$$L_k = \sum_{i=1}^{k} c_i w_i + d_k \tau.$$

Тогда суммарные затраты

$$L = \sum_{k=1}^{N} \lambda_k L_k \;\; = \;\; d\tau \sum_{i=1}^{N} i\lambda_i + \sum_{i=1}^{N} c_i w_i \sum_{k=i}^{N} \lambda_k = \sum_{i=1}^{N} (c_i w_i \sum_{k=i}^{N} \lambda_k + i\lambda_i \tau).$$

Во всех вышеперечисленных источниках поставленная задача решается при существенных ограничениях. В [5, 82] общая длительность обслуживания предполагается показательно распределенной; в [82] то же допущение делается относительно длительности кванта; в [5, 183] кванты считаются постоянными и не зависящими от номера уровня, а в [50] — бесконечно малыми. Описываемые методы как правило неконструктивны: к примеру, в [50] предлагается решать функциональное уравнение в преобразованиях Лапласа, а в [153] — бесконечную систему линейных уравнений, причем среднее время ожидания на j-м уровне зависит от таковых на последующих уровнях (это явная ошибка). Нигде не учитываются играющие принципиальную роль потери на переключения и не приводится верификация расчетных методик.

Опишем свободный от указанных недостатков метод расчета среднего времени v_k пребывания в системе заявки, обслуживание которой завершается на k-м уровне, $k = \overline{1, N}$. Для всех уровней определим:

$Q_k = \sum_{i=1}^{k} q_{i,1}$ — интегральный квант;

$L_k = \sum\limits_{i=1}^{k} \tau_{i,1}$ — средние интегральные потери до k-го уровня включительно;

$\alpha_k = \bar{B}(Q_k)$ — вероятность перехода заявки на следующий уровень ($\alpha_0 = 1$);

$b_{k,m}^{(p)} = \int\limits_{Q_{k-1}}^{Q_k} t^m b(t)\,dt$ — частичный момент трудоемкости;

$b_{k,m}^{(o)} = \int\limits_{Q_{k-1}}^{Q_k} (t - Q_{k-1})^m b(t)\,dt$ — момент длительности неполного кванта;

$b_{k,m}^{(W)} = \alpha_k q_{k,m} + b_{k,m}^{(o)}$ — «взвешенный» момент длительности k-обслуживания, учитывающий вероятную потребность в полном кванте;

$b_k^{(L)} = b_k^{(W)} * \tau_k$ — «нагрузочные моменты», получаемые сверткой взвешенных моментов и системных потерь (распределение последних зависит от номера уровня, если информация по соответствующим заявкам размещается в запоминающих устройствах разных типов);

$\lambda_k = \lambda\alpha_{k-1}$ — параметр потока на k-м уровне;

w_k — среднее время ожидания обслуживания на k-м уровне;

W_k — среднее время ожидания в очередях до k-го уровня включительно.

Интересующий нас конечный результат

$$v_k = W_k + Q_{k-1} + L_k + b_{k,1}^{(o)}.$$

Ключевым элементом этой технологии является расчет средних времен $\{w_k\}$ ожидания меченой заявкой очередного кванта на k-м уровне для $k \geq 2$. Выделим моменты:

«X» — прибытия меченой заявки в систему,

«Y» — начала ее обслуживания на $(k-1)$-м уровне,

«Z» — ее перехода на уровень k

и назовем А-заявкой обслуживаемую в момент «X» (эта заявка с вероятностью ρ_i выбирается из i-й очереди).

Обслуживание меченой заявки на $(k-1)$-м уровне может начаться лишь при отсутствии заявок на всех вышележащих уровнях. К этому моменту («Y») все заявки, которые находились в системе в момент «X», окажутся уже на уровне k[1], но в связи с наличием более приоритетной меченой обслуживаться пока не будут. Они создадут для меченой среднюю задержку T_1. Им будет предшествовать А-заявка (задержка T_2).

Все заявки, пришедшие после меченой, окажутся позади нее на уровне $(k-1)$, и после момента Z меченая будет ждать прохождения ими этого уровня (задержка T_3). Наконец, за время обслуживания меченой на $(k-1)$-м уровне в систему придут новые заявки, и меченая будет ждать прохождения ими уровней $i = \overline{1, k-1}$ — задержка T_4.

Запишем формулы подсчета средних значений упомянутых задержек:

$$\begin{aligned}
T_1 &= r_{k,k}\Big[\bar{B}(Q_{k-1})\sum_{i=1}^{k-1}\lambda_i w_i + \lambda_k w_k\Big];\\
T_2 &= r_{k,k}\bar{B}(Q_{k-1})\sum_{i=1}^{k-1}\rho_i;\\
T_3 &= \lambda r_{k-1,k-1}\bar{B}(Q_{k-2})\Big[\sum_{i=1}^{k-2}(w_i+\tau_i+q_i)+w_{k-1}\Big];\\
T_4 &= \lambda(\tau_{k-1}+q_{k-1})r_{1,k-1}.
\end{aligned}$$

После приравнивания w_k их суммы и элементарных преобразований получаем для всех $k \geq 2$ формулу

$$w_k = F_k/D_k,$$

в которой

$$\begin{aligned}
F_k &= \bar{B}(Q_{k-1})r_{k,k}\Big[\lambda\sum_{i=1}^{k-1}\bar{B}(Q_{i-1})w_i + \sum_{i=1}^{k-1}\rho_i\Big]\\
&+ \lambda\Big\{(\tau_{k-1}+q_{k-1})r_{1,k-1} + \bar{B}(Q_{k-2})\Big[\lambda\sum_{i=1}^{k-2}(w_i+\tau_i+q_i)+w_{k-1}\Big]r_{k-1,k-1}\Big\},\\
D_k &= 1-\lambda\Big[r_{1,k-1}+\bar{B}(Q_{k-1})r_{k,k}\Big].
\end{aligned}$$

[1] Здесь и далее имеется в виду та часть заявок, которая добралась до соответствующего уровня.

Для первого уровня расчет выполняется как для обычных систем с относительным приоритетом.

Трудоемкости $r_{i,j}$ обработки заявки на уровнях $\overline{i,N}$ включительно с учетом отсева полностью обслуженных вычисляются рекуррентно:

$$\begin{aligned} r_{N,N} &= \tau_N + \frac{1}{\bar{B}(Q_{N-1})}\int\limits_{Q_{N-1}}^{\infty}(t-Q_{N-1})\,dB(t) \\ &= \tau_N + \frac{1}{\bar{B}(Q_{N-1})}\Big[b_1 - \int\limits_{0}^{Q_{N-1}} t\,dB(t) - Q_{N-1}\bar{B}(Q_{N-1})\Big], \\ r_{i,N} &= \tau_i + \frac{1}{\bar{B}(Q_{i-1})}\Big\{\int\limits_{Q_{i-1}}^{Q_i}(t-Q_{i-1})\,dB(t) + \bar{B}(Q_i)[q_i + r_{i+1,N}]\Big\}, \\ & \quad i = N-1,\ N-2,\ldots,1. \end{aligned}$$

Для ярусов $j = \overline{1,N-1}$ аналогичные формулы имеют вид

$$\begin{aligned} r_{j,j} &= \tau_j + \frac{1}{\bar{B}(Q_{j-1})}\Big[\int\limits_{Q_{j-1}}^{Q_j}(t-Q_{i-1})\,dB(t) + \bar{B}(Q_j)q_j\Big], \\ r_{i,j} &= \tau_i + \frac{1}{\bar{B}(Q_{i-1})}\Big\{\int\limits_{Q_{i-1}}^{Q_i}(t-Q_{i-1})\,dB(t) + \bar{B}(Q_i)[q_i + r_{i+1,j}]\Big\}, \\ & \quad i = j-1,\ j-2,\ldots,1. \end{aligned}$$

Для замыкания методики осталось обсудить технику расчета частичных моментов. Эти моменты, если соответствующие интегралы явно не берутся, удобно считать численным интегрированием по Симпсону с последовательным делением шага пополам и вычислением только дополнительных ординат. Алгол-программа реализующей этот метод процедуры `SIMFAST` приведена в приложении 3.2 [92]. Возможно также (после аппроксимации $b(t)$ гамма-плотностью с поправочным многочленом) использование известных степенных рядов и рекуррентных соотношений для неполной гамма-функции [152]. К несложным формулам приводит и H_2-аппроксимация распределения времени обслуживания.

Для последнего уровня $Q_N = \infty$, и частичные моменты получаются вычитанием из полных ранее полученных:

$$b_{N,m}^{(p)} = b_m^{(p)} - \sum_{i=1}^{N-1} b_{i,m}^{(p)}.$$

Описанная методика была запрограммирована и реализована применительно к пятиуровневой системе с начальным квантом $\tau_1 = 0.2$, удвоением кванта на уровнях до 4-го включительно, постоянных потерях на переключение $\tau = 0.1$ и длительностью обслуживания по закону Эрланга 2-го порядка с единичным средним. Результаты (И — имитация, Р — расчет) представлены в табл. 10.1.

Таблица 10.1. Средние времена ожидания по уровням

Уровни	$\lambda = 0.5$		$\lambda = 0.6$	
	И	Р	И	Р
1	0.222	0.222	0.275	0.276
2	0.289	0.287	0.429	0.427
3	1.382	1.328	2.506	2.411
4	3.669	3.384	8.541	7.922
5	4.757	4.623	11.459	11.489

Многоуровневая система несколько ухудшает оперативность обслуживания коротких заявок, которые могут застать ЦП занятым выдачей длинного кванта. Поэтому имеет смысл наиболее длинным заявкам выдавать требуемый квант по частям, в интервалах между которыми возможно обслуживание коротких заявок. Соответственно последний уровень на рис. 10.3 разбивается на несколько подуровней. Такая схема может быть применена для анализа квантованного приоритетного обслуживания, реализуемого на фоне вычислений большой трудоемкости. При малых «дробных» квантах она фактически предоставляет коротким заявкам приоритет с прерыванием, хотя и несколько задержанным.

Укажем еще несколько обобщений многоуровневой схемы:

- прямой доступ заявок из внешнего источника на промежуточные уровни;
- обслуживание части верхних уровней с абсолютным приоритетом (в этом случае к системе в целом применяется алгоритм расчета модели со смешанными приоритетами);
- выдача «дробных квантов» случайной длительности с объединением подуровней в одну RR-систему (этот вариант хорошо моделиру-

ет длительный счет в фоновом разделе, прерываемый в моменты промежуточных выдач);

- распределение недообслуженных заявок с определенными вероятностями между *несколькими* нижележащими уровнями.

Комбинация двух последних вариантов особенно удачно описывает реальные процессы обработки неоднородных заявок (на первом уровне вновь прибывшие заявки классифицируются по системным очередям).

В любом варианте среднее время пребывания t-заявки складывается из средних времен ожидания на каждом проходимом ею уровне, системных потерь и чистого времени обслуживания.

10.5. Кольцевая система очередей

В системах коллективного доступа (в особенности при передаче данных) практикуется обслуживание одним устройством кольцевой системы очередей (рис. 10.4).

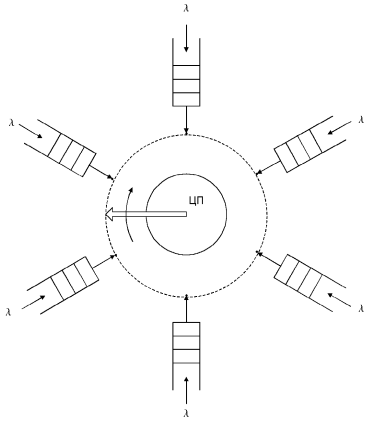

Рис. 10.4. Кольцо очередей

Эта схема реализуется в трех вариантах обслуживания очередей:

ALL — до исчерпания заявок в ней (включая вновь пришедшие);

CUM — до исчерпания заявок, накопленных к началу ее обслуживания;

SOL — по одной заявке из каждой непустой очереди.

Расчет такой системы опирается на анализ одноканальной системы с «прогулками» обслуживающего устройства — отключением его от данной очереди на время обхода кольца. В связи с громоздкостью полной теории вопроса мы приведем лишь выражения из [20] для среднего времени ожидания заявки в симметричной системе (с одинаковыми статистическими характеристиками потоков заявок в каждой из N очередей):

$$\begin{aligned} w_{\text{all}} &= \frac{\tau_2}{\tau_1} + \frac{(N-1)\tau_1 + N\lambda b_2}{2(1-\rho)}, \\ w_{\text{cum}} &= \frac{1}{2}\Big[\frac{N\lambda b_2}{1-\rho} + \frac{\tau_2}{\tau_1} + \Big(\frac{N+\rho}{1-\rho} - 1\Big)\tau_1\Big], \\ w_{\text{sol}} &= \frac{N\tau_1(1+\lambda b_1) + \lambda b_2}{2[1 - N\lambda(b_1+\tau_1)]}. \end{aligned}$$

Здесь $\{\tau_i\}, \quad i = \overline{1,2}$ — моменты распределения длительности перехода к смежной очереди; λ — интенсивность потока, входящего в каждую очередь; $\rho = N\lambda b_1$. Последняя из этих формул (см. [173]) предполагает длительность переключения постоянной.

Более общая схема опроса очередей позволяет рассчитывать системы с альтернирующими приоритетами, когда наивысший приоритет отдается заявкам текущей очереди. При этом уменьшаются потери на переналадку обслуживающего устройства.

Глава 11

Сети обслуживания

11.1. Проблема сетей обслуживания

Анализ реальных процессов обслуживания часто показывает, что заявка получает «полное удовлетворение» лишь после прохождения нескольких видов обслуживания в специализированных узлах. В таких случаях принято говорить о $сетях$ обслуживания. Создание теории сетей обслуживания стимулировалось как возвратом к старым задачам на новом (более детальном) уровне, так и принципиально новыми приложениями, среди которых прежде всего нужно отметить вычислительные сети и сети передачи данных с использованием вычислительных устройств. Одной из первых сетевых моделей ЭВМ стала схема с «центральным вычислителем», представляющая работу вычислительной системы коллективного доступа (рис. 11.1). Подробное описание современной ЦВМ как сети массового обслуживания приведено в [3]. В этой статье, в частности, отмечалось, что в ЦВМ всегда имеются узлы, участвующие во всех этапах решения. Процесс вычислений может начинаться до окончания процесса ввода, а вывода — до окончания вычислений. В процессе ввода-вывода для одной задачи можно выполнять счет для другой (мультипрограммирование).

Большое значение имеют модели ВС как программно-аппаратных комплексов, учитывающие алгоритмы операционной системы [5, 161]. В частности, реентерабельные модули могут рассматриваться как одноканальные СМО с очередями и приоритетным обслуживанием. Наконец, важнейшим объектом применения теории сетей массового обслуживания (СеМО) являются цифровые системы передачи данных (СПД) и вычис-

лительные сети локального, регионального и глобального масштаба, а в последние годы — и сети интегрального обслуживания.

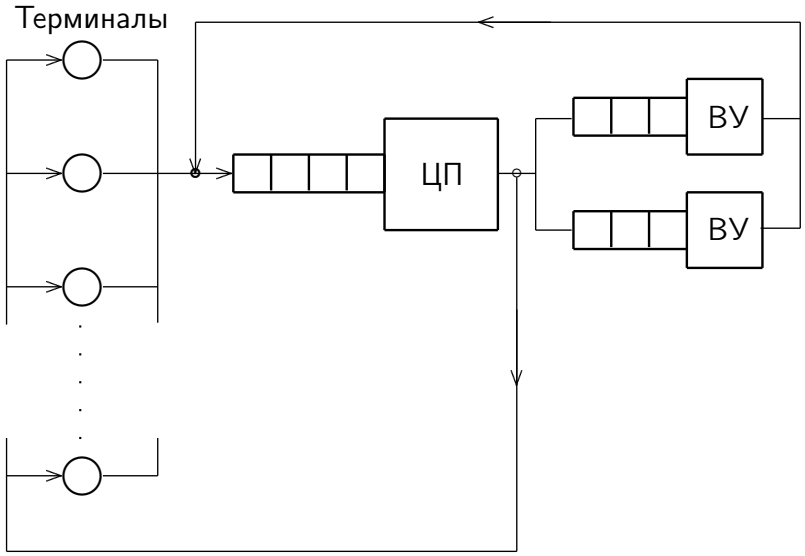

Рис. 11.1. Сетевая модель вычислительной системы

Модели сетей распространяются на биологию (нейросети), лечение болезней, процессы полимеризации.

Центральной проблемой теории сетей обслуживания, определяющей ее специфику по отношению к классической ТМО, является учет *взаимодействия* узлов сети. Здесь мы снова отметим принципиальное отличие сети (тандема очередей) от эрлангова обслуживания в одном канале — в последнем случае новая заявка не начинает обслуживаться, пока предыдущая не прошла последнюю фазу.

При расчете сетей приходится иметь дело с комбинаторным ростом количества состояний сети в целом. Поэтому при разработке численных методов расчета СеМО особое внимание уделяется преодолению «проклятия размерности».

Обстоятельные обзоры литературы по сетям обслуживания приведены в [9, 18, 28, 36, 41, 42, 165]. Специалисты полагают [165, с. 571], что сетевые модели, особенно подпадающие под условия обсуждаемой ниже теоремы ВСМР, легки для понимания и компактно описывают задачу. Автор придерживается *противоположного* мнения и поэтому сосредоточил основное внимание на новых алгоритмах приближенного анализа сетей.

11.2. Определения и допущения

Сеть обслуживания состоит из $рабочих\ узлов$, занумерованных от 1 до M, $источника$ (узел «0») и $стока$ (узел «M+1»). Новая заявка рождается в источнике; с попаданием заявки в сток фиксируется окончание ее пребывания в сети[1]. Если суммарная интенсивность входящего потока не зависит от количества находящихся в сети заявок, сеть считается $открытой$ (open), в противном случае — $замкнутой$ (closed).

Наиболее важный частный случай замкнутых сетей — это сети с постоянной популяцией (числом заявок) K. К сожалению, многие уважаемые авторы [18, 164, 165, 168, 221, 233, 276] утверждают, что в замкнутых сетях «заявки не уходят из сети и не приходят в нее», что лишает смысла процесс обслуживания. Фактически же вместо каждой попавшей в сток заявки в источнике мгновенно генерируется (и передается в один из рабочих узлов) новая. В связи с этим мы будем во всех сетях $выделять\ источник\ и\ сток$.

Замкнутые сети хорошо моделируют вычислительные системы коллективного пользования (популяцию K можно рассматривать как степень мультипрограммирования) и локальные вычислительные сети, а разомкнутые — СПД. При обслуживании сетью неоднородных заявок она может быть замкнутой по одним типам и разомкнутой по другим. Такая сеть считается $смешанной$.

Для каждого j-го узла сети задаются моменты распределения чистой длительности обслуживания $\{b_{j,l}\}, \quad l = \overline{1,L}$, число каналов n_j и дисциплина обслуживания. Наименование последней рассматривается как $тип\ узла$.

Маршрут заявки в сети, вообще говоря, случаен и определяется неразложимой матрицей передач $R = \{r_{i,j}\}, \quad i,j = \overline{0,M+1}$, образованной вероятностями перехода из i-го в j-й узел. Эти вероятности считаются не зависящими от маршрута, уже пройденного заявкой. Предполагаемая неразложимость (неприводимость) сети означает невозможность ее разделения на не связанные допустимыми переходами компоненты.

При наличии заявок нескольких типов $q = \overline{1,Q}$ они разбиваются на классы «замкнутых» Q_c и «открытых» Q_o, а всем перечисленным

[1] Некоторые авторы [165] отождествляют сток с источником. На наш взгляд, это мешает адекватному восприятию процесса обслуживания в сети.

выше величинам приписывается дополнительный (верхний) индекс типа заявки q. В наиболее общих случаях рассматриваются возможные смены типа заявок с q на s одновременно с переходом в следующий узел, управляемые вероятностями $\{r_{i,j}^{(q,s)}\}$. Совокупность типов заявок, между которыми происходят (в том числе транзитивно) взаимные превращения, называется *цепью*. В каждой замкнутой цепи суммарная популяция постоянна. Мы будем считать типы заявок неизменными.

Как было показано в предыдущих главах, основные показатели работы СМО можно рассчитать, получив распределение вероятностей ее состояний. Разумеется, предварительно должна быть проведена марковизация процесса. Состояние *сети* обслуживания в общем случае приходится характеризовать *совместным* распределением вероятностей состояний, учитывающим взаимную зависимость происходящих в узлах процессов. Совместные распределения крайне неудобны в работе, в связи с чем при анализе сетей делаются *допущения*, сводящие упомянутую зависимость к минимуму:

- вероятности $\{r_{i,j}\}$ перехода заявки из узла i в узел j не зависят от предыстории ее (в частности, от кратности прохождения узла) и от состояния узла-преемника j;
- распределения длительности обслуживания в узлах определяются только типом заявки и номером узла и не зависят от ее предыстории и обслуживания других заявок.

Обоснованность перечисленных допущений о независимости процессов (по крайней мере как разумных аппроксимаций) в общем не вызывает серьезных сомнений, исключая случай петель в графе переходов. Здесь эта зависимость становится явной, а поток в узле, охваченном петлей непосредственной обратной связи, перестает быть рекуррентным. По указанной причине ряд авторов [212, 275] рекомендует исключать такие петли, соответственно пересчитывая исходные данные. В заключительном разделе этой главы рассмотрены теоретические основы такого пересчета.

Режим работы сети мы будем предполагать стационарным. Очевидный критерий возможности такого режима — это докритическая загрузка всех узлов сети, что обеспечивает конечные средние длины очередей. В указанной проверке должны учитываться только заявки

«открытых» типов, поскольку популяции «замкнутых» типов фиксированы и интенсивности потоков по ним саморегулируются.

В качестве показателей работы сетей обслуживания имеет смысл рассматривать:

- «маргинальные» распределения числа заявок в узлах, т. е. усредненные по распределениям заявок в дополняющих подсетях (указанные распределения нужны, в частности, для обоснования требований к емкости накопителей заявок);
- коэффициенты загрузки узлов (необходимы для оптимизации сети);
- распределение времени пребывания заявки в сети и его моменты (дифференцированно по видам заявок);
- производительность сети (интенсивность потока обслуженных заявок) — только по замкнутым типам (по разомкнутым в стационарном режиме она равна интенсивности входящего потока).

11.3. Условия строгой декомпозиции сети

Выполнение условий разд. 11.2 позволяет разделить задачи расчета узлов и сети в целом. Теория сетей обслуживания по мере своего развития постоянно расширяла класс задач, для которых можно строго обосновать представление вероятностей состояний сети в виде произведения вероятностей состояний узлов:

$$P(\mathbf{k}) = \frac{1}{G_M}\Lambda(\mathbf{k}) \prod_{j=1}^{M} z_j(k_j). \qquad (11.3.1)$$

Здесь $\mathbf{k} = \{k_1, k_2, \ldots, k_M\}$, G_M — нормирующая константа. Такие (мультипликативные, они же сепарабельные) сети мы будем для краткости называть П-сетями. В англоязычной литературе для них используется сокращение PFQN — Product Form Queueing Networks.

Состояние узлов сети в неоднородном случае определяется числом находящихся в них заявок каждого типа $\{k_j^{(q)}\}$, $q = \overline{1, Q}$, а в однородном случае — просто числом заявок.

Множители $\{z_j(k_j)\}$ зависят от интенсивностей обслуживания в соответствующих узлах и относительных интенсивностей $\{e_j\}$ входящих в узлы потоков. Последние могут быть получены из системы уравнений баланса потоков в форме

$$\begin{array}{lcll} e_j - \sum\limits_{i=1}^{M} e_i r_{i,j} & = & 0, & j = \overline{1, M-1}, \\ e_M & = & 1 & \end{array} \qquad (11.3.2)$$

(выделенным узлом с единичным потоком здесь произвольно выбран последний). Очень важно, что относительные потоки через узлы не зависят от популяций, дисциплины и распределений обслуживания, т. е. инвариантны в широком диапазоне условий. Относительные потоки используются во многих схемах расчета сетей. Заметим, что каждая из функций $z_j(k_j)$ при автономной нормировке ее значений на множестве $k_j = \overline{0, K}$ в общем случае не дает маргинального распределения вероятностей состояний j-го узла [199].

Согласно агрегированной форме теоремы BCMP, названной по первым буквам фамилий авторов [163], узлы в общем случае смешанной П-сети должны принадлежать к одному из следующих типов:

Тип 1. Показательно распределенное обслуживание с дисциплиной FCFS и интенсивностью обслуживания $\mu(k)$, общей для всех классов заявок и зависящей от суммарного числа k заявок в узле (в частности, пропорциональной числу занятых каналов). Типовой множитель

$$z(\mathbf{k}) = \frac{k\,!}{\mu^k} \prod_{q=1}^{Q} \frac{e_q^{k^{(q)}}}{k^{(q)}}$$

(здесь и далее индексы, задающие номер узла, для упрощения обозначений опущены).

Тип 2. Коксово обслуживание (см. разд. 1.8.4) в одноканальном узле с дисциплиной EPS. Типовой множитель

$$z(\mathbf{k}) = k\,! \prod_{q=1}^{Q} \frac{1}{k^{(q)}\,!} \left(\frac{e_q}{\mu^{(q)}}\right)^{k^{(q)}}.$$

Тип 3. Многолинейный узел вида $M/G/\infty$ с вышеупомянутой аппроксимацией распределения длительности обслуживания. Такие

узлы обозначаются IS (immediate service). Типовой множитель для них

$$z(\mathbf{k}) = \prod_{q=1}^{Q} \frac{1}{k^{(q)}\,!} \left(\frac{e_q}{\mu^{(q)}}\right)^{k^{(q)}}.$$

Тип 4. Одноканальный узел с той же аппроксимацией распределения длительности обслуживания и дисциплиной LCFS.PR. Вид множителя тот же, что для второго типа.

Требования к распределениям могут быть ослаблены. Фактически *именно коксово* распределение не обязательно: важна возможность сведéния распределений к фазовым. В [168, с. 114] утверждается, что распределения должны иметь рациональную ПЛС. М. Ньютс полагает [233, с. 238–239], что достаточна дифференцируемость функций распределения.

В результирующие формулы BCMP входят только средние соответствующих распределений. Во всех перечисленных случаях параметр $\mu^{(q)}$ — это *средняя* интенсивность обслуживания заявок типа q. Итоговое распределение (11.3.1) зависит только от средних характеристик обслуживания. Множитель

$$\Lambda^*(\mathbf{k}) = \prod_{i=0}^{M\cdot|\mathbf{k}|-1} \Lambda(i),$$

если входящий внешний поток определяется суммарной популяцией $|\mathbf{k}| = k^{(1)} + k^{(2)} + \ldots + k^{(Q)}$. Если поток состоит из Q независимых, то в аналогичной ситуации

$$\Lambda^*(\mathbf{k}) = \prod_{q=1}^{Q} \prod_{i=0}^{M\cdot|\mathbf{k}|-1} \Lambda^{(q)}(i).$$

При постоянной популяции $\Lambda^*(\mathbf{k}) = 1$.

Позже (см. [165, с. 307]) класс П-сетей был расширен на сети, где вероятность попасть в очередной узел зависит от числа заявок в нем.

Доказательство теоремы BCMP основывается на прямой проверке уравнений баланса переходов между детально (с учетом фаз процессов) определенными состояниями сети — см. [36, 163]. При этом существенно используется *попарный* (локальный) баланс переходов — равенство интенсивностей встречных переходов для каждой пары непосредственно сообщающихся состояний. Из попарного баланса всегда следует

глобальный (но не наоборот!). На рис. 11.2 буквой G помечен разрез для глобального баланса и буквой L — для локального.

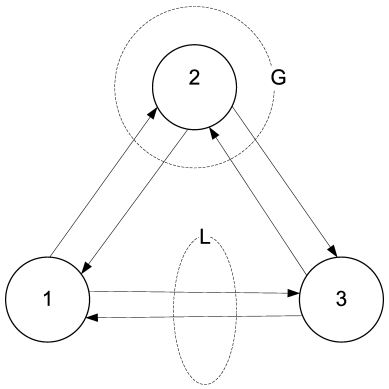

Рис. 11.2. Локальный и глобальный балансы переходов

Попарный баланс является следствием *обратимости* процессов обслуживания [206], критерий которого есть сохранение показательного распределения интервалов между входящими заявками и в выходящем потоке. Именно этим свойством обладают перечисленные выше четыре типа узлов. Все они согласно [197] гарантируют пуассоновский выход при простейшем входящем потоке.

Способ *априорной* проверки условий попарного баланса для произвольно выбранных подмножеств состояний пока не известен [230].

Прямой расчет нормализующей константы чрезвычайно трудоемок даже в однородном случае, когда он требует вычисления суммы $\binom{M+K-1}{M-1}$ слагаемых. Представление о числе слагаемых дает помещенное ниже извлечение из табл.2.1 [15].

Таблица 11.1. Число состояний сети

Число узлов	Объем популяции K				
	5	10	20	50	100
5	126	$1 \cdot 10^3$	$1 \cdot 10^4$	$3 \cdot 10^5$	$5 \cdot 10^6$
10	$2 \cdot 10^3$	$9 \cdot 10^4$	$1 \cdot 10^7$	$1 \cdot 10^{10}$	$4 \cdot 10^{12}$
20	$4 \cdot 10^4$	$2 \cdot 10^7$	$7 \cdot 10^{10}$	$4 \cdot 10^{16}$	$5 \cdot 10^{21}$
50	$3 \cdot 10^6$	$6 \cdot 10^{10}$	$1 \cdot 10^{17}$	$5 \cdot 10^{28}$	$7 \cdot 10^{39}$
100	$9 \cdot 10^7$	$4 \cdot 10^{13}$	$2 \cdot 10^{22}$	$1 \cdot 10^{40}$	$4 \cdot 10^{58}$

Для более экономного расчета П-сетей были разработаны рекуррентные методы, из которых наиболее известны:

- метод сверток (Бузен, [174]);
- метод анализа средних (Рейзер и Лавенберг, [244, 245]);
- метод локального баланса,
- метод соединенных вычислений (Чэнди и Сойер, [179]).

Ниже будут даны описания двух первых в их простейшей (однородной) форме и сводная характеристика рекуррентных методов. Более подробное изложение и многочисленные обобщения их, а также дополнительная литература приводятся в [9, 18, 15, 36, 42, 165, 178, 179, 224, 246]. Во введении к гл. 5 [28] указаны некоторые расширения применимости теоремы BCMP. В [165, с. 307] обуждается случай, где вероятность попасть в очередной узел зависит от числа заявок в нем. В главе 12 [164] описывается *инструментальный* анализ систем и сетей средствами математических пакетов `MATLab` и `Mathematica`. Приведен псевдокод алгоритма свертки и MVA на языке C.

11.4. Метод средних значений

Среднее время пребывания заявки в i-м (одноканальном) узле

$$V_i(K) = b_i[1 + \nu_i(K)],$$

где ν_i — среднее число ранее пришедших заявок, которое застает в i-м узле вновь пришедшая заявка, и b_i — средняя длительность обслуживания в i-м узле. Можно показать, что $\nu_i(K)$ равно стационарному среднему числу заявок в узле $L_i(K-1)$ для сети с популяцией, на единицу меньшей («теорема прибытия»), так что

$$V_i(K) = b_i[1 + L_i(K-1)]. \tag{11.4.1}$$

Определим интенсивности $\{e_i\}$ потоков через узлы сети по отношению к произвольному узлу i^* из системы уравнений баланса типа (11.3.1). Эти $\{e_i\}$ можно трактовать как средние кратности посещения заявкой узлов сети между смежными возвращениями в узел i^*. Общая

длительность такого цикла составит $T = \sum_{i=1}^{M} e_i V_i(K)$. При популяции K интенсивность потока заявок через выделенный узел

$$\lambda^* = K \Big/ \sum_{i=1}^{M} e_i V_i(K). \tag{11.4.2}$$

Потоки в остальных узлах

$$\lambda_i = \lambda^* e_i, \qquad i = \overline{1, M}. \tag{11.4.3}$$

Применяя формулу Литтла, находим среднее число заявок в узлах

$$L_i(K) = \lambda_i V_i(k). \tag{11.4.4}$$

Общая схема алгоритма:

1. Решить уравнения баланса.
2. Для $i = \overline{1, M}$ положить $L_i(0) = 0$.
3. Принять $K = 1$.
4. Положить T=0.
5. Для всех i:
   - вычислить V_i согласно (11.4.1),
   - провести накопление суммы $T := T + e_i V_i(K)$.
6. Если K равно расчетной популяции, принять T в качестве результата (среднее время пребывания заявки в сети). Перейти к этапу 10.
7. Вычислить согласно (11.4.2) $\lambda^* = K/T$.
8. Для всех i:
   - получить λ_i согласно (11.4.3),
   - рассчитать $L_i(K)$ согласно (11.4.4).
9. Увеличить K на единицу и перейти к этапу 4.
10. Конец алгоритма.

11.5. Метод сверток

Метод сверток — единственный, который допускает расчет всех сетей, удовлетворяющих условиям теоремы BCMP [251].

11.5.1. Вычисление нормализующей константы

Определим

$$G_m(k) = \sum_{\mathbf{k}\in D(k,m)} \prod_{i=1}^{m} z_i(k_i). \tag{11.5.1}$$

Здесь $D(k,m)$ — множество допустимых векторов $\mathbf{k}$, удовлетворяющих условию $\sum\limits_{i=1}^{m} k_i = k$. Искомая константа $G(K) \stackrel{\text{def}}{=} G_M(K)$. Перепишем (11.5.1) в форме, фиксирующей количество заявок в узле m — последнем из рассмотренных:

$$\begin{aligned} G_m(k) &= \sum_{j=0}^{k}\Bigg[\sum_{\substack{\mathbf{k}\in D(k,m)\\ k_m=j}} \prod_{i=1}^{m} z_i(k_i)\Bigg] \\ &= \sum_{j=0}^{k} z_m(j)\Bigg[\sum_{\mathbf{k}\in D(k-j,m-1)} \prod_{i=1}^{m-1} z_i(k_i)\Bigg] \\ &= \sum_{j=0}^{k} z_m(j) G_{m-1}(k-j). \end{aligned} \tag{11.5.2}$$

Сформируем $(k+1)$-мерные векторы

$$\begin{aligned} \mathbf{z}_m &= \{z_m(0), z_m(1), \ldots, z_m(K)\}, \qquad m = \overline{1,M}, \\ \mathbf{g}_0 &= \{1, 0, \ldots, 0\}. \end{aligned}$$

Тогда вектор $\mathbf{g}_m = \{G_m(k)\}$ в соответствии с (11.5.2) можно получить как свертку векторов: $\mathbf{g}_m = \mathbf{z}_m * \mathbf{g}_{m-1}$. Схема этих вычислений наглядно представлена в таблице на рис. 11.3, заполняемой *по столбцам*.

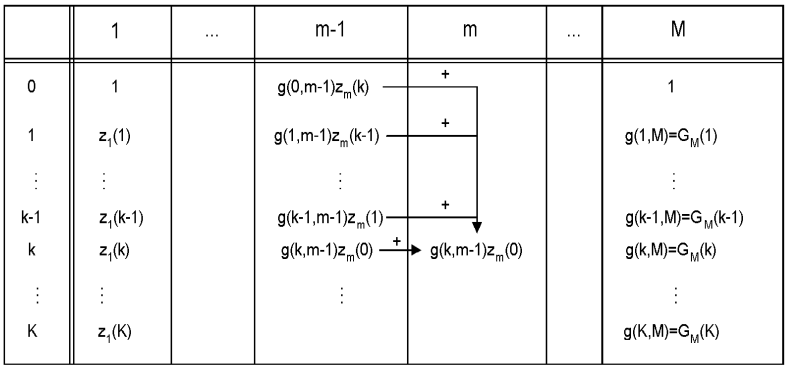

Рис. 11.3. Реализация алгоритма свертки

Запишем маргинальное распределение числа заявок в последнем (с номером M) узле замкнутой сети:

$$\begin{aligned} P_M(k,K) &= G_M^{-1}(K) \sum_{\substack{\mathbf{k}\in D(K,M)\\ k_M=k}} \prod_{i=1}^{M} z_i(k_i) \\ &= G_M^{-1}(K)\cdot z_M(k) \sum_{\mathbf{k}\in D(K-k,M-1)} \prod_{i=1}^{M-1} z_i(k_i). \end{aligned}$$

Последняя сумма по смыслу представляет собой $G_{M-1}(K-k)$. Следовательно,

$$P_M(k,K) = z_M(k)G_{M-1}(K-k)/G_M(K). \qquad (11.5.3)$$

11.5.2. Потоки через узлы

По определению, $z_M(k) = z_M(k-1)e_M/\mu_M(k)$. Подставив это выражение и вычисляемые согласно (11.5.3) маргинальные вероятности в условие баланса

$$\lambda_M(K) = \sum_{k=1}^{K} P_M(k,K)\mu_M(k)$$

для ожидаемого потока через M-й узел, получим

$$\lambda_M(K) = e_M G_M^{-1}(K) \sum_{k=1}^{K} z_M(k-1)G_{M-1}(K-k).$$

Последняя сумма представляет собой свертку $G_M(K-1)$. Итак, поток через M-й узел

$$\lambda_M(K) = e_M G_M(K-1)/G_M(K). \qquad (11.5.4)$$

Поочередно ставя все узлы в конец списка, мы можем использовать (11.5.3) и (11.5.4) для расчета любого узла сети.

11.6. Сравнение рекуррентных методов

Принципиальным недостатком BCMP-методов является необходимость рекуррентного счета по объему популяции K: чтобы получить результат для требуемого $\hat{K}$, нужно предварительно выполнить расчет для всех $K = 1, \hat{K}-1$ включительно. Это определяет их значительную трудоемкость. В табл. 11.2 приведены заимствованные из [165] сравнительные характеристики методов.

Приблизительно так же методы этой группы и родственные им оцениваются в [179, 253]. По *общности* наиболее универсальным является метод сверток CONV, применимый для любых П-сетей. Метод соединенных вычислений CCNC допустим только для одноканальных и IS-узлов. Методы анализа средних MVA и локального баланса LBANC допускают зависимость интенсивности обслуживания только от суммарной длины очереди. MVA неприменим для смешанных сетей.

Таблица 11.2. Требования к памяти и количество операций

Метод	Память	Трудоемкость
CONV	$O(\prod_{r=1}^{R}(K_r+1))$	$O(2\cdot R(N-1)\prod_{r=1}^{R}(K_r+1))$
MVA	$O(N\cdot\prod_{r=1}^{R}(K_r+1))$	$O(2\cdot R(N-1)\prod_{r=1}^{R}(K_r+1))$
RECAL	$O\Big(\frac{K^{N+1}+1}{(N-1)!}\Big)$	$O\Big(\frac{4N-1}{(N+1)!}R(N+1)\Big)$
FES	$>O(3\prod_{r=1}^{R}(K_r+1))$	$O(2\cdot R(N-1)\prod_{r=1}^{R}(K_r+1))$

Для RECAL популяция K одинакова по всем типам.

Таблица 11.3. Сравнение методов

Метод	Преимущества	Недостатки
MVA	Средние могут быть получены без нормализующей констаты (NC), если не нужны вероятности состояний.	Нужно много памяти при большом числе узлов и типов заявок. Проблемы переполнения, потери значимости при расчете маргинальных вероятностей.
CONV	Меньше памяти, чем MVA.	Необходим расчет NC. Проблемы переполнения, потери значимости при расчете NC.
RECAL	Меньше памяти и трудоемкости для задач большого размера, чем MVA и CONV.	Больше памяти и трудоемкости для малого числа узлов (типов), чем MVA и CONV.
FES	Снижаются требования к памяти и быстродействию при исследовании чувствительности к изменению параметров. отдельных узлов. Базис для расчета несепарабельных сетей.	Многократное применение CONV или MVA. Требуется преобразование сети.

По *трудоемкости* в [179] эти методы квалифицируются как приблизительно равноценные, хотя таблицы из [253] свидетельствуют о разбросе в 2–4 раза (наилучшим оказался LBANC, наихудшим — MVA). Разница существенно зависит от типа узлов сети, а также от перечня вычисляемых характеристик.

Потребность в *памяти* как правило минимальна у CCNC, а в неоднородных задачах — у CONV.

Устойчивость вычислений лучше всего обеспечивается методом MVA. Остальные методы предполагают вычисление нормализующей константы $G(K)$ и потому нуждаются в крайне нетипичном при работе на ЭВМ с плавающей точкой нетривиальном масштабировании. При реализации MVA и LBANC, кроме того, в процессе расчета распределения числа заявок имеет место накопление погрешностей. Эта же проблема возникает и в CONV при вычислении функций $\{g_i(k)\}$.

11.7. Немультипликативные сети — общий подход

Уже первое тридцатилетие развития методов строгой декомпозиции сетей обслуживания (а сейчас завершается второе) практически исчерпало их возможности. Здесь уместно повторно процитировать П. Швейцера: «мы дошли до конца дороги с точными моделями... Мы затратили слишком много времени на такие модели, как П-сети и матрично-геометрические решения» [219, с.418]. «Классические» методы расчета сетей имеют слишком быстрый рост трудоемкости при увеличении размерности задачи (в особенности по объему популяции и числу типов заявок).

Но дело не только в связанных с ними вычислительных трудностях. Не существует нетривиальных сетей, для которых были бы верны пуассоновская гипотеза и типовые рассуждения. Ряд особенностей реальных задач вообще исключает возможность П-решения. К таким особенностям прежде всего относятся узлы типа FCFS с немарковским обслуживанием и с различными распределениями обслуживания для заявок разных видов. Напомним в связи с этим о показанной в главах 5–10 сильной зависимости характеристик работы СМО от вида распределений и дисциплины обслуживания, возможность учета которой перекрывает ошибки от «силовой» декомпозиции сети. В рамках классических подходов к сетям невозможно также учесть:

- неэкспоненциальное обслуживание в многоканальных узлах,
- рекуррентные входящие потоки,
- приоритетное обслуживание,
- одновременное удержание нескольких ресурсов,
- порождение и ликвидацию параллельных подзадач,
- блокировку узлов из-за ограниченной емкости очередей на входе их преемников,
- зависимость маршрутизации от длин очередей в последующих узлах, и т. д.

Особенно огорчительно то, что они вынуждают ограничиться *средними* характеристиками, недостаточными для большинства приложений.

В практике проектирования вычислительных систем и сетей численные методы применяются в основном для быстрой оценки и отбора перспективных вариантов структуры. Практика требует разработки разумных аппроксимаций, применимых в перечисленных выше «неудобных» задачах и обладающих приемлемой точностью, умеренной трудоемкостью и хорошей *робастностью*. Необходимость применения пусть менее строгих, но более реалистических методов осознается все большим числом исследователей. Г.П. Башарин и А. Л. Толмачев в [8, с. 56] признают, что в достаточно разветвленной сети зависимость длительности обслуживания заявки в отдельных узлах не имеет существенного значения. В. М. Вишневский [28, с. 179] в качестве возможных направлений работы называет распространение требования мультипликативности распределений вероятностей состояний узлов за пределы условий теоремы BCMP; постулирование независимости процессов там, где она в строгом смысле отсутствует; применение эрлангова и гиперэкспоненциального распределений и, наконец, потокоэквивалентную декомпозицию сети. Он отмечает (там же, с. 184), что «декомпозиция является одним из наиболее перспективных методов, позволяющих, с одной стороны, повысить эффективность анализа СеМО в вычислительном отношении, а с другой — получить достаточно точные оценки характеристик моделей, которые принципиально не могут быть рассчитаны точными методами».

При этом, учитывая ограниченную точность реальных исходных данных, к точности оценок следует предъявлять разумные требования: в той же книге [28, с. 203], а также в [234] указывается относительная погрешность 10 %. Еще бóльшие допуски разрешает Г. П. Башарин с соавторами [10, с. 8]: «Во многих случаях теория СеМО обеспечивает подходящий компромисс между требованиями к точности модели, сложностью получения численных результатов и возможностью их интерпретации. Согласно данным многих исследователей, аккуратные аналитические модели ИВС приводят к оценкам, погрешность которых достигает 10 % для среднего быстродействия, для коэффициентов использования отдельных устройств и для интенсивностей внутренних потоков в замкнутых моделях и 25-30 % — для времени отклика и количества заявок в узлах». Последняя цифра фигурирует и в [184] — следовательно, она является мнением *международного* сообщества исследователей.

Относительную погрешность для высших моментов случайных величин общепринято считать пропорциональной порядку момента. Согласно [56, с. 611], дисперсия статистической оценки момента k-го порядка

$$D[f_k] = \frac{f_{2k} - f_k^2}{n}.$$

Из этого следует, что для оценок первого момента допустима погрешность в 20 %, второго — 40 %, а третьего — 60 %.

Как резерв для повышения точности рассматривается учет высших моментов распределений и корреляции процессов в смежных узлах, а также автокорреляции. Средством детального анализа и прогнозирования рабочих характеристик отобранного варианта остается *имитационное моделирование*, обзор основных концепций и технологии которого приведен, в частности, в работах автора [104, 117].

Умеренная трудоемкость приближенных методов может быть достигнута:

- декомпозицией сети на определенном этапе расчета на отдельные подсети, в предельном случае — узлы;
- экстраполяцией возможности П-решений за пределы условий теоремы BCMP и позднейших расширений ее;
- отказом от рекуррентных по численности популяций расчетных схем и замены их итерациями с разумно достаточными требованиями по точности;
- заменой математических моделей узлов на возможно более простые с учетом известных предельных тенденций.

В качестве последних мы отметим возможности:

- представления узлов с большим числом каналов (в особенности при малых популяциях) IS-моделями, т. е. схемой $M/G/\infty$;
- замены произвольных распределений обслуживания в тех же условиях показательными с сохранением среднего;
- замены узлов с циклическим обслуживанием моделью типа EPS (учитывая реальное соотношение между длиной кванта и системными потерями);

- замены дисциплины FCFS на случайный выбор заявок из очереди, не требующий запоминания позиций заявок в ней;

- допущения о том, что все потоки в сети — простейшие (при достаточно равномерном ветвлении маршрутов);

- замены замкнутых моделей узлов на разомкнутые при больших популяциях (в неоднородном случае — суммарной численности популяций), обслуживаемых данным узлом;

- замены произвольных распределений показательными для получения нижних/верхних оценок в классах ВФИ/УФИ-распределений соответственно;

- предположение о постоянной занятости наиболее загруженного узла (узкого места), формирующего рекуррентный поток обслуженных заявок.

Робастность метода состоит в устойчивости результатов по отношению к вариациям исходных данных. Она может быть обеспечена:

- предварительной ранжировкой влияющих на определяемые показатели факторов и соблюдением принципа разумной достаточности;

- подбором допущений парами противоположно действующих;

- корректностью математической постановки задачи и при необходимости — ее регуляризацией;

- физической наглядностью основных этапов расчета (о методе свертки, например, этого не скажешь);

- опорой на надежно измеряемые исходные данные (с этой концепцией связано одно из новых направлений в расчете СеМО — операционный анализ, идейно близкий методу средних);

- заданием исходных данных по возможности интегральными характеристиками (например, распределений — конечным набором моментов);

- максимальным использованием в расчетной схеме разного рода стабилизирующих инвариантов (законов сохранения и следствий из них: формулы Литтла, сохранения потоков заявок, формулы Полячека—Хинчина для среднего времени ожидания; инвариантных к закону обслуживания распределений числа заявок в IS-узле и средней длительности пребывания в EPS-узле и т. д.);

- применением иных специфических для рассматриваемых моделей контрольных соотношений;

- обязательным включением в расчетную схему этапов агрегации подмоделей: нормировка по подмножествам состояний и баланс переходов между *группами* состояний — см. итерационный метод в главе 7, проверка сохранения популяции для сети в целом (разд. 11.9), и т. п.

Все приближенные методы расчета СеМО опираются на ту или иную форму *декомпозиции* сети. Декомпозиционная аппроксимация является одним из наиболее перспективных методов, позволяющих повысить вычислительную эффективность анализа СеМО и получить достаточно точные оценки характеристик моделей, которые принципиально не могут быть получены точными методами [28, с. 184].

Декомпозиция предоставляет возможности:

- сведения большой задачи к серии задач меньшей размерности;

- однократного анализа типовой подструктуры общей модели;

- аналитического исследования простых подмоделей;

- упрощения параметрических исследований (разделяются фиксированные и изменяемые подмодели);

- применения к подмоделям различных временных масштабов;

- применения к разным подмоделям различных методов исследования, наиболее подходящих в конкретных условиях.

Проиллюстрируем эффективность декомпозиции на примере линейных алгебраических уравнений. Известно, что трудоемкость решения системы из n уравнений пропорциональна n^3. Если ее удалось разбить на k

подзадач, то суммарная трудоемкость составит $k*(n/k)^3$, т. е. снизится в k^2 раз[2].

Стандартный подход к декомпозиции в теории математического моделирования [187] связан с разделением высокочастотной и низкочастотной динамики. При исследовании первой медленно меняющиеся переменные рассматриваются как параметры; при исследовании второй высокочастотные процессы сглаживаются и представляются их ожидаемыми характеристиками. С этих позиций можно исследовать, например, взаимодействие центрального процессора с электромеханическими внешними устройствами. Проводимая при указанном подходе агрегация состояний редко совпадает с физической структурой моделируемой системы, да и существенной разницы в частоте переходов обычно не наблюдается.

В общем случае декомпозиция проводится на отдельные подсети (см., например, кластерный подход [261]); однако проще всего решать задачу следующим образом:

1. Решением системы типа (11.3.2) или Q таких систем в неоднородном случае определить средние потоки через узлы с точностью до постоянного множителя.

2. Из сети последовательно выделять по одному узлы $j = \overline{1, M}$ с присущими им числовыми характеристиками обслуживания, заменяя дополняющие подсети кусочно-постоянным пуассоновским источником заявок одного из следующих видов (в порядке усложнения расчета интенсивности):

   – постоянной,

   – линейно убывающей по числу заявок в узле,

   – нелинейно убывающей.

   Выделенный узел рассчитать как изолированную систему.

3. В случае ограниченной популяции по результатам второго этапа провести пропорциональную коррекцию потоков, после чего повторить этап — до выполнения условия сходимости итераций.

[2]Реальный выигрыш будет меньше вследствие неизбежных накладных затрат на декомпозицию и последующую агрегацию результатов.

4. После стабилизации потоков рассчитать каждый узел как изолированную СМО с присущей ей дисциплиной обслуживания.

5. Определить агрегированные характеристики распределения времени пребывания заявок для сети в целом.

Наличие этапов агрегации 1, 3 и 4, обеспечение баланса потоков (этап 1) и постоянства популяций (этап 3), независимость этапа балансировки от дисциплины обслуживания, а также применение итераций, исключающих характерное для рекуррентных методов длительное накопление погрешностей, обеспечивают $робастность$ методов указаного типа. Достоинства, варианты и многочисленные примеры итерационного подхода к анализу моделей СеМО приводятся в [224, 260, 274].

Подход, аналогичный описанному методу декомпозиции, в теории электрических цепей составляет содержание известной теоремы Нортона, откуда он и был заимствован для теории СеМО [165, 177]. Покажем, как его использовать в наиболее сложном случае нелинейной зависимости потока в m-м узле $\lambda_m(k, K)$ от числа k заявок в этом узле. Заменим дополнение к выделенному узлу сети эквивалентной экспоненциальной подсетью и вычислим для нее согласно алгоритму свертки (разд. 11.5) нормализующие константы $\{g(k,i)\}$, $k = \overline{1, K}$, где

$$i = \begin{cases} M, & \text{если } i = M, \\ M-1 & \text{при } i \neq M. \end{cases}$$

Через эти константы согласно (11.5.4) легко определяется искомая зависимость

$$\lambda_m(k) = e_m g(k-1, i)/g(k, i).$$

Погрешность такого расчета порождается двумя причинами:

- вынужденной экспоненциальной аппроксимацией узлов дополнения, не отвечающих условиям теоремы BCMP;

- неучетом возможной зависимости интенсивности входящего в выделенный узел потока от $фаз$ проходящих в нем процессов.

Методы, основанные на теореме Нортона, принципиально рекурсивны по объему популяции, поэтому их трудоемкость весьма значительна. Дальнейшим их развитием является итерационное уточнение параметров экспоненциальной сети по результатам расчета узлов (метод Мари —

см. первоисточник [218] и [165, с. 452–459], где даны подробный разбор, схема алгоритма и числовой пример). Этот метод применим лишь для одноканальных узлов с распределениями обслуживания фазового типа и обладает рекордно большой трудоемкостью.

11.8. Разомкнутая сеть

11.8.1. Баланс и преобразование потоков

Для разомкнутой сети все потоки между узлами считаются рекуррентными, т. е. имеющими одинаковые и независимые распределения интервалов между смежными заявками. Расчет начинается с определения интенсивностей входящих в узлы потоков из уравнений баланса заявок

$$\lambda_i = \Lambda r_{0,i} + \sum_{j=1}^{M} \lambda_j r_{j,i}, \qquad i = \overline{1, M}. \tag{11.8.1}$$

Здесь Λ — суммарная интенсивность потока от внешних источников. Попутно отметим, что в некоторых работах [202] допускаются множители $\{\gamma_j\}$ коррекции числа заявок в узлах:

$$\lambda_i = \Lambda r_{0,i} + \sum_{j=1}^{M} \lambda_j \gamma_j r_{j,i}.$$

Далее для всех узлов должно быть проверено условие отсутствия перегрузки

$$\lambda_i b_{i,1}/n_i < 1, \tag{11.8.2}$$

обеспечивающее существование в сети стационарного режима. Положительный результат проверки позволяет продолжить вычисления в зависимости от алгоритмических и вычислительных возможностей, которыми располагает исследователь.

Если входной поток пуассонов и распределение обслуживания — показательное, то на основании теоремы Берке все потоки в сети — также пуассоновы. Это их свойство часто можно постулировать безотносительно к распределениям обслуживания (с учетом случайного прореживания и суммирования выходящих потоков и наличия многоканальных узлов обслуживания). Тогда можно непосредственно приступить к расчету маргинальных распределений числа заявок в узлах с заданными

для них распределениями длительности обслуживания, числом каналов и полученными из системы (11.8.1) интенсивностями $\{\lambda_i\}$ входящих потоков.

Отказ от упомянутой гипотезы предполагает детальный учет:

- свойств внешних потоков,
- расщепления выходящих из узлов потоков вследствие случайного (с вероятностями $\{r_{i,j}\}$) либо регулярного распределения обслуженных заявок по узлам-преемникам,
- суммирования в преемнике потоков от нескольких источников,
- преобразования входящих потоков в узлах обслуживания.

11.8.2. Общая схема расчета разомкнутой сети

Алгоритм расчета разомкнутой сети с учетом преобразования потоков в общем случае является итерационным и включает в себя следующие шаги:

1. Задание начальных приближений для потоков заявок, выходящих из узлов сети. Указанные потоки принимаются простейшими с интенсивностями, определяемыми решением системы уравнений баланса потоков (11.8.1), а коэффициенты немарковости интервалов между смежными заявками $\xi_i = 0,\ i = \overline{1, M}$.
2. Для узлов $i = \overline{1, M}$:
   - расчет прореженных потоков, поступающих с выхода узлов $j \neq i$ на вход i-го, согласно разд. 2.1.7;
   - суммирование этих потоков по методике разд. 2.1.9;
   - расчет новых коэффициентов немарковости ξ_i' на входе i-го узла, определение модуля уточнения $\Delta_i = |\xi_i' - \xi_i|$ и обновление ξ_i;
   - расчет узла как СМО соответствующего типа;
   - расчет потока, выходящего из i-го узла, по методике разд. 7.12.
3. Если $\max\limits_i \Delta_i > \varepsilon$, повторение шага 2.

4. Расчет моментов распределения времени пребывания заявки в узлах при каждом посещении согласно рекомендациям главы 8.

5. Расчет моментов распределения времени пребывания заявки в сети в целом (разд. 11.14) и при необходимости — построение по ним ДФР (разд. 1.10).

6. Конец алгоритма.

Итерационное повторение шагов 2–3 необходимо лишь для сетей с циклическими маршрутами.

Объем вычислений при реализации этого алгоритма можно заметно сократить, если воспользоваться отмеченной в разд. 7.12 линейностью зависимости коэффициентов немарковости выходящего из узла потока от аналогичных коэффициентов входящего. Тогда достаточно:

- рассчитать каждый узел для двух типов входящих потоков (с двумя наборами коэффициентов немарковости) при заданной средней интенсивности;

- по результатам этих вычислений определить параметры упомянутых линейных зависимостей для каждого узла;

- методом итераций или прямым решением соответствующей системы линейных уравнений (после линеаризации операций суммирования потоков) определить установившиеся коэффициенты немарковости входящих в узлы потоков;

- с их помощью по формуле (8.6.1) найти моменты интервалов между заявками на входе узлов и рассчитать эти узлы.

Этот подход требует $3M$-кратного обращения к процедурам расчета узлов. Дальнейшего выигрыша в объеме вычислений (порядка 30%) и повышения устойчивости работы программы можно добиться, если вторым типом всех потоков считать простейший и соответственно пользоваться менее трудоемкими процедурами расчета узлов.

11.8.3. Расчет сети с неоднородными потоками

Алгоритм разд. 11.8.2 естественно обобщается на случай Q типов неоднородных заявок с дифференцированными характеристиками

бесприоритетного обслуживания каждого типа. Обобщенный алгоритм состоит из следующих шагов:

1. Решение системы уравнений баланса (11.8.1) по каждому q-типу заявок относительно средних интенсивностей потоков $\{\lambda_j^{(q)}\}$ в узлах $j = \overline{1, M}$; заметим, что матрица передач R, вообще говоря, будет зависеть от q.

2. Проверка отсутствия перегрузки:

$$\max_j \Big\{ \sum_q \lambda_j^{(q)} b_{j,1}^{(q)} / n_j \Big\} < 1$$

   (при его нарушении необходима коррекция исходных данных).

3. Расчет удельного веса заявок каждого типа во всех узлах и средневзвешенных моментов распределений длительности обслуживания по каждому узлу.

4. Расчет начальных приближений для суммарных потоков (все — простейшие) и задание их коэффициентов немарковости $\xi_j = 0$.

5. Расчет узлов и потоков до стабилизации значений $\{\xi_j\}$ по базисному алгоритму разд. 11.8.2. Выходящий поток j-го узла по q-му типу заявок получается случайным или регулярным прореживанием с учетом относительного веса заявок этого типа в данном узле.

6. Расчет моментов распределения времени ожидания, общих для всех типов заявок, по каждому из узлов.

7. Для всех типов заявок:

   – получение для всех узлов моментов распределения времени пребывания при одном посещении посредством свертки моментов распределений времени ожидания и чистой длительности обслуживания заявок этого типа;

   – расчет моментов распределения времени пребывания заявки этого типа в сети;

   – построение ДФР времени пребывания.

8. Конец алгоритма.

Этот же подход применяется в случае приоритетного обслуживания в узлах. Здесь расчет моментов распределения времени пребывания в узле проводится методами главы 9, а входящие и выходящие потоки вынужденно принимаются простейшими.

11.9. Замкнутые сети

11.9.1. Постановка задачи

Для замкнутой однородной сети условия баланса заявок можно записать в виде

$$\begin{aligned} \sum_{i=1}^{M} \lambda_i (r_{i,j} + r_{i,M+1} r_{0,j}) - \lambda_j &= 0, \qquad j = \overline{1, M-1}, \\ \sum_{i=1}^{M} \lambda_i (b_{i,1} + w_i) &= K. \end{aligned} \tag{11.9.1}$$

В уравнениях первой группы второе слагаемое в скобках описывает переход заявки из i-го узла в сток и появление в источнике новой заявки, направляемой в j-й узел. Эти уравнения определяют интенсивности потоков с точностью до постоянной C: $\lambda_i = Ce_i$.

Последнее уравнение (11.9.1) играет роль нормировки. Здесь w_i — среднее время ожидания в i-м узле начала обслуживания. Это уравнение есть применение формулы Литтла для среднего количества заявок в узле. Среднее время w_i ожидания в узле i сложным образом зависит от интенсивности входящего потока, числа каналов и моментов распределения обслуживания. Фактически задача распадается на линейную часть (расчет $\{e_i\}$) и нелинейную — последнее уравнение, которое может быть решено только методом итераций.

Система уравнений (11.9.1) была независимо предложена Д. Куватсосом в [226] и автором этой книги в [101] еще в середине 1980-х гг. Данное уравнение требует равенства числу K *среднего* числа заявок в системе, тогда как реально равенство должно выполняться в любой момент времени. Таким образом, реальное ограничение в этой расчетной схеме ослабляется и реализуется лишь приближенно.

Для решения системы (11.9.1) необходимы предварительные оценки средних времен ожидания $\{w_i\}$ в узлах, получить которые невозможно, не зная интенсивностей потоков, определяемых ... из той же системы.

Поскольку произведение $Ce_ib_{i,1}$ равно среднему числу занятых в i-м узле каналов, логично заменить систему (11.9.1) на

$$\begin{aligned} x_i &= \sum_{j=1}^{M} x_j(r_{j,i} + r_{j,M+1}r_{0,i}), \quad i = \overline{1, M-1}, \\ K &= C\sum_{i=1}^{M} x_i[w_i(Cx_i) + b_{i,1}], \end{aligned} \tag{11.9.2}$$

где N — суммарное количество каналов обслуживания в узлах сети. Начальное значение C выбирается из условия докритической загрузки узлов

$$\max_i \frac{Cx_ib_{i,1}}{n_i} < 1,$$

или

$$C < 1 \Big/ \max_i \frac{x_ib_{i,1}}{n_i}. \tag{11.9.3}$$

Можно, например, принять C_0 равной произведению правой части (11.9.3) на $1 - 1/K$.

11.9.2. Стратегии «сетевых» итераций

Теперь рассмотрим организацию перевычисления коэффициента C. В ходе итераций это можно делать согласно

$$\begin{aligned} C &= K \Big/ \sum_{i=1}^{M} x_i[w_i(C) + b_{i,1}\Big] \\ &= K \Big/ \Big[\sum_{i=1}^{M} x_iw_i(C) + \sum_{i=1}^{M} x_ib_{i,1}\Big]. \end{aligned} \tag{11.9.4}$$

Заметим, что вторая сумма в пересчете по ходу «сетевых» итераций не нуждается.

Опыт вычислений согласно (11.9.4) обнаружил медленную сходимость, а в ряде случаев — и расходимость «сетевых» итераций. Поэтому в дальнейшем простые итерации были заменены на итерации по Вегстейну [118, с. 114–115], которые для убывающей по C правой части (11.9.4) обеспечивают очень быструю сходимость.

В качестве альтернативы этой схеме рассматривалось также решение уравнения

$$C\sum_{i=1}^{M} x_i[w_i(C) + C\sum_{i=1}^{M} x_i b_{i,1} - K = 0 \qquad (11.9.5)$$

методом секущих. Здесь возникала проблема построения начальной «вилки» — нахождения интервала значений C со сменой знака левой части (11.9.5). Она решалась пошаговым изменением C до смены знака у левой части уравнения (11.9.5) — с одновременной корректировкой стартовой границы. При использовании этого метода возникает проблема построения начальной «вилки» — нахождения интервала значений C со сменой знака левой части исходного уравнения.

11.9.3. Инварианты отношения

Рассмотренные выше подходы опирались на расчет среднего времени ожидания в системе $M/H_2/n$ по схеме Такахаси—Таками или методом МГП. Поскольку нас интересует только среднее время ожидания, для одноканальных «разомкнутых» узлов можно воспользоваться формулой Полячека—Хинчина. Однако для сетей большой размерности с многоканальными узлами расход времени мог бы оказаться недопустимо велик. Предлагавшиеся рядом авторов приближенные формулы ([149, 217] и многие другие источники) ненадежны и недостаточно точны. Кроме того, зачастую они дают не приближенные значения искомых величин, а их верхние оценки.

В связи с этим была проверена возможность использования для оценки средних времен ожидания в многоканальных системах предложенных И. М. Духовным *инвариантов отношения*, которые уже обсуждались в главе 9. Из табл. 11.4 следует, что при коэффициенте загрузки 0.7 и более интересующие нас отношения с достаточной для практики точностью не зависят от вида распределения обслуживания.

Таблица 11.4. Отношения $M/G/n$ к $M/G/1$

Модель	ρ	Число каналов n			
		2	3	4	5
$M/D/n$	0.5	3.5319e-1	1.7393e-1	9.8720e-2	6.0704e-2
	0.7	4.2367e-1	2.4599e-1	1.6314e-1	1.1656e-1
	0.9	4.7747e-1	3.0671e-1	2.2275e-1	1.7313e-1
$M/E_3/n$	0.5	3.4423e-1	1.6665e-1	9.3362e-2	5.6798e-2
	0.7	4.1838e-1	2.4083e-1	1.5863e-1	1.1269e-1
	0.9	4.7581e-1	3.0490e-1	2.2101e-1	1.7150e-1
$M/M/n$	0.5	3.3333e-1	1.5789e-1	8.6957e-2	5.2149e-2
	0.7	4.1176e-1	2.3445e-1	1.5309e-1	1.0795e-1
	0.9	4.7368e-1	3.0262e-1	2.1882e-1	1.6944e-1
$M/H_2/n$	0.5	3.1610e-1	1.4390e-1	7.6693e-2	4.4697e-2
	0.7	4.0065e-1	2.2373e-1	1.4380e-1	9.9972e-2
	0.9	4.6996e-1	2.9859e-1	2.1459e-1	1.6427e-1

11.9.4. «Замкнутая» модель узла

Можно надеяться, что при большом числе узлов и больших популяциях «разомкнутые» модели узлов (с постоянными интенсивностями потоков) окажутся достаточно адекватными ситуации в замкнутой сети. Однако в общем случае интенсивность входящего в узел потока будет убывающей функцией от количества k заявок, уже находящихся в данном узле. Как простейший вариант примем эту зависимость линейной:

$$\lambda_k = \lambda^*(K-k), \quad k = \overline{0,K}.$$

В этом случае условие баланса пришедших и обслуженных заявок имеет вид $Cx = \lambda^*(K-\bar{k})$, где $\bar{k}$ — среднее число заявок, находящихся в узле. Отсюда следует, что удельная интенсивность потока

$$\lambda^* = Cx/(K-\bar{k}). \tag{11.9.6}$$

Для расчета начальных $\{\bar{k}_i\}$ к ожидаемому количеству $Cx_ib_{i,1}$ заявок, находящихся непосредственно на обслуживании, добавляется остаток популяции $K - C\sum_{i=1}^{M} x_ib_{i,1}$, распределяемый между узлами сети обратно пропорционально разностям $n_i - Cx_ib_{i,1}$.

После обсчета всех узлов сети выполняется критически важная перенормировка $\{\bar{k}_i\}$ к расчетной популяции K.

Итерации по этой схеме часто расходились — видимо, из-за неадекватности постулируемой линейной зависимости. Поэтому в дальнейшем использовалась разомкнутая модель узлов, степень реалистичности которой возрастает при увеличении количества узлов сети и связей между ними.

11.9.5. Имитационная модель

Для окончательного тестирования расчетной схемы была необходима эталонная имитационная модель. Поскольку в замкнутой сети первичными событиями являются только завершения обслуживания, при моделировании на Фортране 90 была использована закольцованная цепь будущих событий, число элементов которой, естественно, равнялось общему числу каналов обслуживания N. В элементе указывался момент времени, номер узла сети и номер канала в узле. Цепь упорядочивалась по моментам наступления очередных событий; при отработке головного элемента соответственно генерировался и вставлялся в должное место цепи момент завершения нового обслуживания.

Описанная технология строится на нетривиальной логике работы с закольцованной цепью событий, но резко сокращает объем требуемой памяти (в частности, делает его независимым от общего числа проходящих через сеть заявок) и сводит к минимуму трудоемкость поиска ближайшего события.

В целях тестирования вышеописанной технологии был разработан альтернативный вариант модели с «иерархической» структурой ЦБС: на нижнем уровне прогнозировались наиболее ранние моменты завершения обслуживания в каждом из каналов, на верхнем фиксировались минимальные из них. При этом на нижнем уровне обновлялись данные только для тех узлов, которые затрагивались очередным переходом.

Результаты моделирования по обоим вариантам полностью совпали.

11.9.6. Численные результаты

Предлагаемые методы тестировались на сети из четырех рабочих узлов, представленной на рис. 11.4. Вероятности межузловых переходов проставлены на соответствующих дугах. Количество каналов по узлам задается набором {2,1,1,3}, средние времена обслуживания — {2,3,3,5}.

Тип распределения времени обслуживания был назначен общим: распределение Эрланга третьего порядка. Соответственно определялись высшие моменты. Результаты расчета сведены в табл. 11.5.

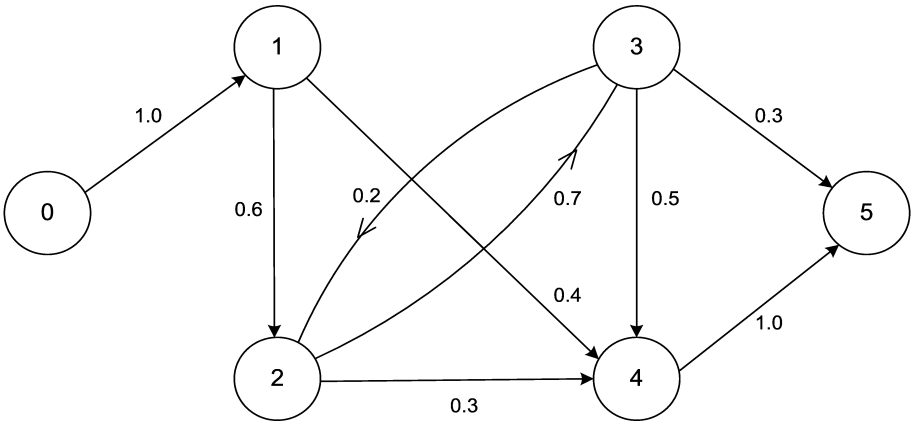

Рис. 11.4. Схема замкнутой сети обслуживания

Таблица 11.5. Средние времена пребывания в сети

Популяция K	Вариант	Распределения обслуживания		
		E_3	M	H_2
8	1	17.373	18.663	22.113
	2	19.806	21.061	25.296
	3	19.818	21.040	25.181
16	1	33.481	33.825	37.469
	2	35.721	36.635	41.103
	3	35.570	36.632	41.073
24	1	50.223	50.300	53.059
	2	52.417	52.868	57.260
	3	52.046	52.992	57.252
32	1	66.902	66.894	68.965
	2	69.137	69.464	73.632
	3	68.672	69.554	73.628
40	1	83.564	83.687	84.440
	2	85.868	86.129	90.128
	3	85.349	86.193	90.122

Приведем расшифровку вариантов:

1 — имитация (500 тыс. испытаний);

2 — «разомкнутые» (M/H/n) модели узлов, итерации по Вегстейну;

3 — «разомкнутые» узлы с расчетом среднего времени ожидания по инвариантам отношения и уравниванием популяции по методу хорд.

Из опыта отладки и сопоставления результатов вытекают следующие выводы:

1. Тенденции изменения результатов счета в зависимости от входных данных соответствуют качественным ожиданиям: среднее время пребывания в сети возрастает при увеличении коэффициентов вариации распределения длительности обслуживания и росте популяции.

2. Расхождение результатов имитационного моделирования (вариант 1) и аналитико-численных методик не превышает 10 % и имеет тенденцию к уменьшению с ростом популяции. Таким образом, использование «разомкнутых» моделей узлов в расчете замкнутых сетей представляется вполне оправданным. При увеличении размеров сети и увеличении количества межузловых связей можно ожидать дальнейшего уменьшения погрешностей.

3. Практическое совпадение результатов в вариантах 2 и 3 подтверждает допустимость использования для расчета средних длительностей пребывания заявок в сети инвариантов отношения. Поскольку вариант 3 не требует трудоемкого пересчета (сотен итераций) характеристик узлов на каждом шаге коррекции интенсивности потоков, его использование в указанных целях явно предпочтительней.

4. При необходимости получения ДФР распределения времени пребывания заявки в замкнутой сети можно применить тот же подход, что и в случае открытых сетей. В этом случае потребуются высшие моменты распределения времени пребывания в узлах и соответственно — счет по варианту 2.

11.10. Неоднородные заявки

Если в сети циркулируют заявки нескольких типов с бесприоритетным обслуживанием, допущение о возможности представления узлов разомкнутыми СМО становится более обоснованным. При этом среднее

время ожидания в узле не зависит от типа заявки и считается по формуле (8.8.1) со средневзвешенными моментами. Расчет замкнутой сети с неоднородными заявками состоит из следующих этапов:

1. Решить системы вида (11.9.1) с дополнительным индексом q типа заявок для всех $q \in Q_c$.

2. Полагая для начала коэффициент пропорциональности общим по всем типам заявок, из условия

$$C\Big(\sum_{q\in Q_c} e_i^{(q)} b_{i,1}^{(q)}\Big)\Big/ n_i < 1, \qquad i = \overline{1,M},$$

найти граничное значение

$$C_{\max} = (1 - 1/K_\Sigma)\min_i\Big\{n_i \Big/ \sum_{q\in Q_c} e_i^{(q)} b_{i,1}^{(q)}\Big\}. \tag{11.10.1}$$

Здесь $K_\Sigma = \sum\limits_{q\in Q_c} K^{(q)}$.

3. Принять исходное значение $C = C_{\max}$.

4. Начать решение системы нелинейных уравнений

$$\sum_{i=1}^{M} C^{(q)} e_i^{(q)} (b_{i,1}^{(q)} + w_i) - K^{(q)} = 0, \qquad \forall q \in Q_c, \tag{11.10.2}$$

в которой w_i зависит от всех $\{C^{(q)}\}$ и $\{e_i^{(q)}\}$, методом скорейшего спуска (мы использовали его «сглаженный» вариант, уменьшающий характерную для скорейшего спуска колебательность направлений).

5. При достаточно малых невязках продолжить это решение методом Ньютона.

6. Конец алгоритма.

В ходе решения (11.10.2) необходима коррекция получаемых решений $\{C^{(q)}\}$ при обнаружении перегрузки хотя бы одной из представляющих узлы сетей разомкнутых СМО. Заметим также, что каждый тип неоднородных заявок в общем случае имеет свою матрицу передач, возможно, исключающую некоторые узлы. Учет последнего обстоятельства был одной из основных проблем в программной реализации алгоритмов расчета неоднородных сетей.

Предложенный подход учитывает взаимное влияние потоков (через общие очереди в узлах) и характерную для замкнутых сетей

отрицательную обратную связь между числом заявок в узле и интенсивностью входящего потока заявок.

Точность алгоритма может быть повышена, если для вычисления $\{w_i\}$ вместо аппроксимаций (8.8.1) пользоваться расчетом узлов как систем вида $M/H_k/n$ по средневзвешенным характеристикам обслуживания с последующим применением формулы Литтла. Необходимые для вычисления матрицы Якоби производные $\{\partial w_i/\partial C^{(q)}\}$ можно по-прежнему получать на основе формулы (8.8.1). Разумеется, в этом случае следует снизить требования к получаемым невязкам.

11.11. Смешанные сети

В смешанной сети циркулируют заявки как из ограниченных популяций $\{K^q\}$, $q \in Q_c$, так и из неограниченных, заявки от которых поступают из внешнего источника с интенсивностями $\{\Lambda^{(q)}\}$, $q \in Q_o$. Интенсивности потоков «открытых» заявок могут быть определены из систем уравнений

$$\lambda_j^{(q)} - \sum_{i=1}^{M} \lambda_i^{(q)} r_{i,j}^{(q)} = \Lambda^{(q)} r_{0,j}^{(q)}, \qquad j = \overline{1, M}$$

для всех $q \in Q_o$ и далее не пересчитываются. Дальнейший ход решения соответствует описанному в разд. 11.10 со следующими отличиями:

1. В формуле (11.10.1) из числа каналов n_i вычитается $\sum\limits_{q \in Q_o} \lambda_i^{(q)} b_{i,1}^{(q)}$. Тем самым вводится поправка на загрузку узлов «открытыми» заявками.

2. При вычислении величин $\{w_i\}$ средних времен ожидания в узлах учитываются и «открытые» заявки.

Таким образом, в общем случае процесс декомпозиции сводится к решению $|Q_c| + |Q_o|$ систем линейных алгебраических уравнений и системы из $|Q_c|$ нелинейных уравнений. При уплотнении маршрутной информации (составлении для каждого типа заявок q списка посещаемых узлов, что и было выполнено практически) метод легко реализуется для задач большой размерности с разреженными матрицами передач.

При расчете по этой методике смешанных сетей определение средневзвешенных характеристик обслуживания проводится с учетом заявок

всех типов, а средняя интенсивность «замкнутого» потока на входе каждого узла корректируется с учетом загрузки, создаваемой «открытыми» заявками. Такой подход был применен, в частности, при разработке описанного в [169] пакета программ анализа сетей обслуживания.

Рассмотрим зависимость входящего потока от общего числа заявок в смешанном узле (индекс узла для упрощения обозначений опустим). Обозначим $p = \Lambda^c/(\Lambda^c + \Lambda^o)$ вероятность принадлежности заявки к одному из замкнутых типов и положим $q = 1 - p$. Тогда при наличии в системе ровно k заявок, $k < K^c$, ожидаемое количество «замкнутых» составит $\bar{k}^c(k) = kp$. При $k \geq K^c$

$$\bar{k}^c(k) = \sum_{i=0}^{K^c} i \left(\begin{array}{c} k \\ i \end{array} \right) p_i q_{k-i}.$$

Соответственно $\bar{k}^c(k)$ для каждого k в рамках выбранной модели определяется интенсивность замкнутого потока и к ней добавляется Λ^o.

Заметим, что заявки «замкнутых типов» не влияют на существование стационарных решений в смешанных сетях (но влияют на сами эти решения).

В табл. 11.6 показаны результаты расчета смешанной сети из четырех узлов с одним типом «разомкнутых» заявок и двумя типами замкнутых с популяциями 14 и 20 соответственно.

Таблица 11.6. Средние времена пребывания в смешанной сети

Метод расчета	Типы заявок		
	1	2	3
Имитационная модель (5000 испытаний)	59.06	38.39	113.50
Разомкнутая с применением (8.8.1)	66.21	44.15	113.23
То же с расчетом узлов по схеме $M/H_2/n$	63.00	44.11	114.00

Попытки использовать «замкнутые» и «смешанные» модели *узлов* не обеспечили повышения точности — по-видимому, вследствие неадекватности постулированных зависимостей $\{\lambda_m(k)\}$ реальным. Более того, в ряде случаев отмечались ухудшение сходимости и даже расходимость итераций при балансировке популяций, вызванные несоответствием аппроксимированной матрицы Якоби реальному процессу получения невязок.

Близость результатов, полученных по двум последним вариантам, при несоизмеримой трудоемкости (в описанном примере — 0.47 и 12.2 с процессорного времени ЕС-1061 соответственно) позволяют рекомендовать следующую стратегию:

1. При достаточности средних характеристик работы сети ограничиться исходным вариантом методики, полностью опирающимся на формулу (8.8.1).

2. При необходимости вычисления высших моментов распределения времени пребывания заявки в сети по методике разд. 11.14, требующей соответствующих моментов времени пребывания в узлах, пользоваться исходной методикой для получения интенсивностей потоков через узлы и моделью $M/H_k/n$ — для вычисления окончательных характеристик задержек в узлах.

Реализация комбинированной стратегии позволила в том же примере получить высшие моменты за 0.97 с процессорного времени счета вместо 12.2 с.

Заметное различие между аналитическими и имитационными результатами объясняется прежде всего ограниченностью популяций. Это подтверждается сравнением табл. 11.6 с табл. 11.7, в которой приведены результаты счета для популяций {20,30}.

Таблица 11.7. Средние времена пребывания при увеличенных популяциях

Метод расчета	Типы заявок		
	1	2	3
Имитационная модель	117.70	67.49	166.80
Комбинированный	107.30	65.24	171.70

Зависимость относительной погрешности результатов от объема популяции имеет явно выраженный максимум при умеренных объемах. При малых популяциях очереди практически не возникают, а при больших аппроксимация узлов сети разомкнутыми СМО оказывается достаточно точной.

11.12. Технология расчета смешанной сети

Данный раздел имеет целью показать читателю реальную сложность процесса расчета смешанной сети и вооружить его базой для самостоятельного программирования задач подобного вида.

11.12.1. Системы нелинейных уравнений и задачи минимизации

Для решения системы нелинейных уравнений вида

$$f_i(x) = 0, \qquad i = \overline{1,n}, \tag{11.12.1}$$

где x — вектор из n элементов, можно использовать быстро сходящийся метод Ньютона, в котором очередное приближение

$$x_k = x_{k-1} - W(x_{k-1})f(x_{k-1}). \tag{11.12.2}$$

Здесь $f(x)$ — вектор значений левых частей (невязок), а $W(x)$ — матрица Якоби производных вектор-функции по векторному аргументу:

$$W(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}.$$

Метод Ньютона имеет малую область сходимости и практически применим только в комбинации с методами *спуска*, позволяющими выйти в достаточно близкую окрестность решения. В этих методах минимизируется сумма квадратов невязок

$$U(x) = \sum_{i=1}^{n} f_i^2(x) = (f(x), f(x)) \tag{11.12.3}$$

(скалярное произведение векторной функции на себя же). Направление спуска связывается с понятием *градиента* $U(x)$ — вектора с компонентами

$$\frac{\partial U}{\partial x_j} = 2\sum_{i=1}^{n} f_i(x)\frac{\partial f_i(x)}{\partial x_j}.$$

В векторно-матричной форме градиент записывается как

$$g(x) \equiv \nabla U(x) = 2W^T(x)f(x).$$

Простейший вариант спуска связан с движением по антиградиенту и при поверхностях уровня, заметно отличающихся от гиперсферических, приводит к замедляющим спуск осцилляциям направлений. Этот недостаток устранен при использовании взаимно сопряженных (относительно матрицы Якоби) направлений спуска. В методе Полака—Рибьера ([73, с. 100–104]) первый шаг делается по антиградиенту U, а последующие направления строятся как линейная комбинация текущего антиградиента и предыдущего направления движения:

$$d_k = -g_k + \alpha_k d_{k-1}.$$

«Вес» предыдущего направления

$$\begin{aligned} \alpha_k &= g_k^T(g_k - g_{k-1})/(g_{k-1}^T g_{k-1}) \\ &= (g_k^T g_k - g_k^T g_{k-1})/(g_{k-1}^T g_{k-1}). \end{aligned}$$

По выбранному направлению делается шаг величины

$$\beta_k = (g_k^T g_k - g_k^T g_{k-1})/(g_k^T g_k).$$

Первый шаг в каждом цикле из n шагов, где n — число неизвестных, делается по антиградиенту. Величина шага может корректироваться с учетом ограничений на компоненты вектора x — например, требований неотрицательности.

11.12.2. Невязки и производные

Невязка популяции по i-му типу вычисляется согласно

$$L_i = C_i \sum_m l_{i,m}(b_{i,m,1} + w_m) - K_i,$$

причем среднее время ожидания заявок w_m зависит от всего набора $\{C_i\}$. Значит,

$$\frac{\partial L_i}{\partial C_j} = \begin{cases} C_i \sum\limits_m l_{i,m} \dfrac{\partial w_m}{\partial C_j}, & \text{если } j \neq i; \\ C_i \sum\limits_m l_{i,m} \dfrac{\partial w_m}{\partial C_i} + \sum\limits_m l_{i,m}(b_{i,m,1} + w_m) & \text{для } j = i. \end{cases}$$

Среднее время ожидания в m-м узле мы запишем в виде

$$w_m = \frac{A_m}{N_m} e^{-B_m N_m},$$

где

$$\begin{aligned} A_m &= \frac{1}{2n_m^2}\Big(S_{m,2}^{(o)} + \sum_{q\in Q_c} C_q l_{q,m} b_{q,m,2}\Big), \\ N_m &= 1 - \frac{1}{n_m}\Big(S_{m,1}^{(o)} + \sum_{q\in Q_c} C_q l_{q,m} b_{q,m,1}\Big), \\ B_m &= 0.7(n_m - 1). \end{aligned}$$

Опуская индекс у C, по которой берется производная, имеем

$$\begin{aligned} \frac{\partial w_m}{\partial C} &= \frac{A'_m N_m - A_m N'_m}{N_m^2} e^{-B_m N_m} - \frac{A_m}{N_m} e^{-B_m N_m} \cdot B_m N'_m \\ &= \frac{e^{-B_m N_m}}{N_m}\Big(A'_m - A_m \frac{N'_m}{N_m} - A_m B_m N'_m\Big) \\ &= \frac{e^{-B_m N_m}}{N_m}[A'_m - A_m N'_m (1/N_m + B_m)] \\ &= w_m\Big[\frac{A'_m}{A_m} - N'_m\Big(\frac{1}{N_m} + B_m\Big)\Big]. \end{aligned}$$

Производные от входящих в эту формулу подвыражений

$$\frac{\partial A_m}{\partial C_j} = \frac{1}{2n_m^2} l_{j,m} b_{j,m,2}, \qquad \frac{\partial N_m}{\partial C_j} = -l_{j,m} b_{j,m,1}/n_m.$$

Эти зависимости используются в процедуре расчета смешанной сети `MIXNETO` — см. главу 12.

11.13. Расчет сети с блокировками

В некоторых задачах (например, расчета сетей передачи данных) емкость накопителей в узлах предполагается конечной. При этом сообщение, передаваемое из узла m в узел j и заставшее у адресата предельное число заявок l_j, *блокируется* в узле-отправителе и прекращает его

работу до завершения адресатом очередного обслуживания. Сети с блокировками часто встречаются при анализе гибких автоматизированных производств.

Приближенно эффект блокировок учитывает следующий итерационный алгоритм:

1. Принять начальные задержки $\tau_m^{(q)} = 0, \quad m = \overline{1, M},\ q = \overline{1, Q}$.

2. Пересчитать средние времена обслуживания в узлах согласно

   $$\tilde{b}_{m,1}^{(q)} = b_{m,1}^{(q)} + \tau_m^{(q)} .$$

3. Рассчитать высшие моменты соответствующих распределений, постулируя сохранение коэффициентов немарковости исходных.

4. С помощью комбинированного алгоритма разд. 11.10 получить для каждого узла m:

   а) интенсивности потоков $\lambda_m^{(q)}$,

   б) моменты распределения ожидания $w_{m,*}$,

   в) средние времена обслуживания с учетом блокировок

   $$\bar{b}_m = \sum_q \lambda_m^{(q)} \tilde{b}_{m,1}^{(q)} \Big/ \sum_q \lambda_m^{(q)} ,$$

   г) вероятность блокировки данным узлом $\pi_m = \sum\limits_{i=l_m}^{\infty} p_{m,i}$, где $\{p_{m,i}\}$ — распределение числа заявок в узле m.

   Применение именно комбинированного алгоритма диктуется необходимостью получать вероятности блокировок. Эффект блокировки рекомендуется учитывать, если ее вероятность превышает 0.1.

4. Для всех узлов и типов заявок подсчитать ожидаемые задержки

   $$\tau_m^{(q)} = \sum_{j=1}^{M} r_{m,j}^{(q)} \pi_j \bar{b}_{j,1} / n_j .$$

5. Проверить условие сходимости процесса (например, максимальное изменение задержек меньше δ). При его невыполнении перейти к этапу 2.

6. Выполнить расчет итоговых характеристик распределения времени пребывания заявки в сети.

7. Конец алгоритма.

11.14. Многоресурсные задачи

Процесс обслуживания в сети может лимитироваться несколькими последовательно выделяемыми ресурсами. На рис. 11.5 показан вариант схемы вычислительной системы коллективного пользования, в котором явно отражена конкуренция заявок за место в оперативной памяти. Заявка удерживает выделенный ей раздел при прохождении нескольких циклов использования ЦП и дисковой памяти. Завершение обслуживания (∇) снимает блокировку $(\triangle)$ на доступ следующей заявки. Подобным образом можно организовать использование и большего числа типов ресурсов. Процессы выделения—освобождения обычно оказываются вложенными наподобие циклов в программах, что снимает известную из теории операционных систем проблему взаимной блокировки задач.

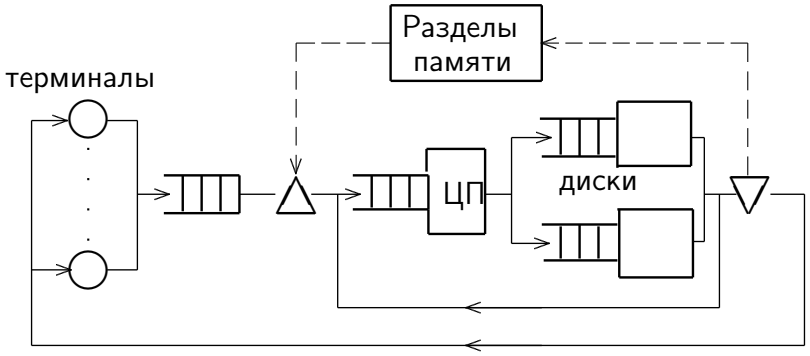

Рис. 11.5. Одновременное удержание нескольких ресурсов

Опишем метод анализа этой модели, предложенный Бардом [161]. Обозначим для q-го типа заявок

$t^{(q)}$ — среднее время задержки («раздумья») на терминале,

$h^{(q)}$ — среднее время удержания памяти,

$K^{(q)}$ — популяция q-заявок.

Кроме того, положим

w — среднее время ожидания памяти (общее для всех типов при дисциплине FCFS),

S — количество разделов памяти, равное числу активных терминалов.

Целью расчета является определение среднего времени реакции системы

$$T^{(q)} = w + h^{(q)}, \qquad q = \overline{1, Q} \tag{11.14.1}$$

(от выдачи запроса до получения ответа) по всем типам заявок.

Прежде всего отметим, что доля q-заявок, находящихся в активной фазе, составляет $h^{(q)}/(t^{(q)} + h^{(q)} + w)$. Соответственно ожидаемое число таких заявок

$$L^{(q)} = \frac{K^{(q)} h^{(q)}}{t^{(q)} + h^{(q)} + w}. \tag{11.14.2}$$

В типичной ситуации постоянно занятой памяти

$$\sum_{q=1}^{Q} \frac{K^{(q)} h^{(q)}}{t^{(q)} + h^{(q)} + w} = S. \tag{11.14.3}$$

Алгоритм расчета является итерационным:

1. Задать начальные приближения для $\{h^{(q)}\}$ как суммы ожидаемых длительностей обслуживания ЦП и дисковой памятью. Положить $w = 0$.

2. Решить (например, методом Ньютона) уравнение (11.14.3) относительно w.

3. По всем типам заявок $q = \overline{1, Q}$ рассчитать средние интенсивности потоков

   $\lambda^{(q)} = K^{(q)}/(t^{(q)} + h^{(q)} + w)$.

4. Для подсети, охваченной на рисунке штриховыми стрелками, найти новые средние времена $\{h^{(q)}\}$ пребывания q-заявки в ней с помощью алгоритмов разд. 11.9.

5. Рассчитать по формуле (11.14.1) новые средние времена реакции $\{T_1^{(q)}\}$.

6. Если $\max\limits_q(T_1^{(q)}/T^{(q)} - 1) > \varepsilon$, заменить $\{T^{(q)}\}$ на $\{T_1^{(q)}\}$. Перейти к этапу 2.

7. Конец алгоритма.

Более точный алгоритм, основанный на теореме Нортона, описан в статье Сойера [250].

11.15. Распределение времени пребывания заявки в сети

11.15.1. Решение в средних

Согласно [143, 164], формула Литтла справедлива не только для отдельных систем обслуживания, но и для *сетей в целом*. Это подтвердили и численные эксперименты автора над имитационными моделями. Соответственно среднее время пребывания заявки в *разомкнутой* сети

$$v = \sum_{i=1}^{M} \bar{k}_i \Big/ \Lambda \tag{11.15.1}$$

(среднее число заявок в сети делится на суммарную интенсивность внешнего потока).

В *замкнутой* сети покидающая ее обслуженная заявка немедленно замещается новой, так что здесь

$$v = K \Big/ \sum_{i=1}^{M} \lambda_i r_{i,M+1}. \tag{11.15.2}$$

В знаменателе этой формулы второй множитель отбирает заявки, покидающие сеть после прохождения i-го узла. Весь знаменатель — это суммарная интенсивность потока обслуженных заявок, по закону сохранения заявок равная интенсивности входящего потока.

Для замкнутой сети упомянутый знаменатель является одной из важных ответных величин.

11.15.2. О зависимости времен пребывания в узлах

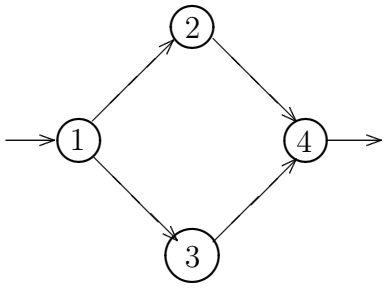

Рис. 11.6. Обгон в сети

В разд. 11.2 были сделаны оговорки, обеспечивающие независимость процессов изменения числа заявок в узлах и как следствие — возможность строгой декомпозиции сети. Однако независимость состояний узлов не означает независимости времен пребывания в них. Рассмотрим фрагмент сети, представленный на рис. 11.6. Если $\mu_2 \gg \mu_3$, то «длинная» заявка, идущая в узел 4 через узел 3, может получить дополнительную задержку из-за тех заявок, которые прибыли в узел 1 после нее, но прошли в узел 4 через «быстрый» узел 2. Строгая независимость времен пребывания гарантируется только в линейной цепочке (тандеме) узлов, из которых первым может быть система $M/M/n$, последним — $M/G/n$, а промежуточными — $M/M/1$ [172]. Отмеченное обстоятельство не сказывается на определении средних и, следовательно, на применимости формул разд. 11.15.1, но влияет на *высшие* моменты. Можно надеяться, однако, что в хорошо спроектированных сетях, близких к сбалансированным, случаи обгона будут сравнительно редки и описываемый ниже метод даст приемлемые результаты и для высших моментов, что подтверждается имитационным моделированием.

11.15.3. Преобразование Лапласа и высшие моменты

Для наиболее ответственных применений знать среднее время пребывания заявки в сети недостаточно: в подобных случаях обычно встает вопрос и о высших моментах и/или построении функции распределения. Ниже рассматривается *приближенный* способ решения этой задачи, игнорирующий отмечавшиеся ранее многообразные зависимости и

корреляции. Автор полагает, что такой подход предпочтительнее тупого отстаивания «строгих» методов, не дающих ответа на поставленные практикой вопросы и неприменимых к реальным ситуациям.

Выделим из матрицы передач:

- вектор-строку $P = \{r_{0,1}, r_{0,2}, \ldots, r_{0,M}\}$ вероятностей перехода из источника в конкретные рабочие узлы;
- вектор-столбец $T = \{r_{1,M+1}, r_{2,M+1}, \ldots, r_{M,M+1}\}^T$ вероятностей перехода из рабочих узлов в сток;
- матрицу $Q = \{r_{i,j}\}, \quad i,j = \overline{1,M}$, вероятностей переходов между рабочими узлами.

Кроме того, определим диагональную матрицу $N(s)$ преобразований Лапласа $\{\nu_i(s)\}$ распределений времени пребывания в рабочих узлах сети и сформируем матрицу

$$\Gamma(s) = N(s)Q. \tag{11.15.3}$$

Легко видеть, что ПЛС распределения длительности одношаговых переходов в сток

$$\gamma_1(s) = PN(s)T.$$

Для двухшаговых переходов

$$\gamma_2(s) = P\Gamma(s)N(s)T$$

и вообще для k-шаговых

$$\gamma_k(s) = P\Gamma^{k-1}(s)N(s)T.$$

Полное ПЛС распределения времени пребывания заявки в сети

$$\gamma(s) = \sum_{k=1}^{\infty} \gamma_k(s) = P\Big(\sum_{k=0}^{\infty} \Gamma^k(s)\Big)N(s)T = P(I - \Gamma(s))^{-1}N(s)T.$$

Существенным элементом этой технологии является вычисление преобразований Лапласа по моментам распределения (разд. 1.9.2).

Моменты $\{g_i\}$ распределения времени пребывания в сети можно получить численным дифференцированием $\gamma(s)$ в нуле. С. В. Кокорин

вывел аналитические матричные выражения для первых трех производных, подстановка в которые нуля должна давать искомые моменты. В частности, среднее время ожидания может быть подсчитано по формуле

$$v = P(I - Q)^{-1}V[Q(I - Q)^{-1} + I]T. \qquad (11.15.4)$$

Здесь V — диагональная матрица средних длительностей пребывания заявки в узлах при однократном посещении их. Результаты применения (11.15.4) согласуются (проверено автором!) с полученными согласно (11.15.1) и численным дифференцированием. Эта технология снимает нетривиальную проблему выбора шага построения таблиц для численного дифференцирования (при слишком малом шаге резко уменьшается количество верных цифр в разностях высокого порядка, при большом снижается точность вычисления младших производных).

Пользователей сети обычно интересует вопрос о вероятности пребывания заявки в сети долее заданного времени, причем само это время указывается со значительным элементом произвола. Практически в таких случаях нужно строить по найденным моментам ДФР времени пребывания заявки в сети для значений аргумента, перекрывающих диапазон возможных изменений директивного срока. Таблица ДФР может быть построена на базе ДФР Вейбулла или интегрированием гамма-плотности с поправочным многочленом. Напомним, что не следует ожидать хорошей относительной точности для значений ДФР, меньших 0.01.

В заключение отметим, что очевидная модификация этой методики позволяет считать суммарный доход, полученный заявкой в процессе ее миграции по сети, или понесенные ею убытки.

11.15.4. Численные эксперименты

Для численного эксперимента был взят тот же пример замкнутой сети, что и в разд. 11.9.1. Сравнительные результаты аналитического расчета (A) и имитационного моделирования (M) по данным прохождения через сеть 5000 заявок после выхода из нее всех 14 первоначальных, распределенных по узлам случайным образом с вероятностями $\{r_{0,j}\}$, приведены в табл. 11.8.

Таблица 11.8. Моменты распределения пребывания в замкнутой сети

Распределение обслуживания	Способ расчета	Порядок моментов			
		1	2	3	4
Показательное	A	3.63e1	2.32e3	2.17e5	2.70e7
	M	3.72e1	2.44e3	2.32e5	2.87e7
Эрланговское 3-го порядка	A	3.64e1	2.32e3	2.17e5	2.70e7
	M	3.59e1	2.26e3	2.12e5	2.72e7

Процессорное время счета на ЕС-1033 по аналитической методике для показательного и эрланговского распределений составило соответственно 5.5 с и 5 мин. Столь большая относительная разница объясняется тем, что в первом случае счет велся по элементарным формулам, а во втором — по трудоемкому итерационному алгоритму разд. 7.5. Имитационное моделирование потребовало соответственно 10 и 15 мин. процессорного времени.

Разомкнутая сеть из четырех узлов рассчитывалась при той же матрице передач с заменой нулевой строки упомянутой матрицы на [0, 0.2, 0.3, 0.4, 0.1, 0], количестве каналов в узлах $n_1 = 1$, $n_2 = 2$, $n_3 = 3$, $n_4 = 3$ и прежними средними длительностями обслуживания. Интенсивность внешнего потока была принята $\Lambda = 1/3$; при этом максимальный из коэффициентов загрузки узлов составил ρ_3=0.632 (метод дает хорошие результаты при $\rho_{\max} \leq 0.8$). Результаты аналитического расчета и имитационного моделирования (2000 заявок) при простейших потоках заявок представлены в табл. 11.9.

Таблица 11.9. Моменты распределения пребывания в разомкнутой сети

Распределение обслуживания	Способ расчета	Порядок моментов			
		1	2	3	4
Показательное	A	1.75e1	5.80e2	2.84e4	1.85e6
	M	1.81e1	6.51e2	3.72e4	2.96e6
Вырожденное	A	1.56e1	4.02e2	1.52e4	7.68e5
	M	1.56e1	4.12e2	1.60e4	8.27e5

Отметим, что первые моменты, подсчитанные по аналитической методике, совпали с результатами применения формул (11.14.1) и (11.14.2) по крайней мере на 6 знаков.

Представляет интерес сопоставление результатов аналитического расчета и имитационного моделирования $открытых$ сетей обслуживания. Для примера была выбрана сеть рис. 11.7 с пятью рабочими узлами и числом каналов обслуживания в узлах {1,2,1,2,1} соответственно.

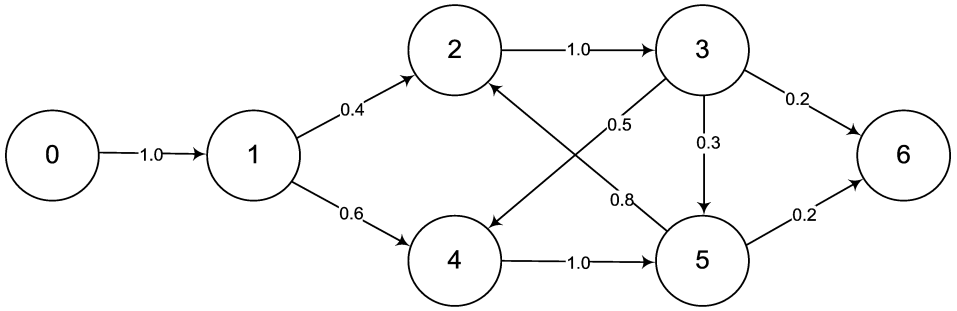

Рис. 11.7. Пример сети обслуживания

Вероятности переходов между узлами сети проставлены на стрелках. Длительности обслуживания в каждом из узлов предполагались распределенными по закону Эрланга 3-го порядка с интенсивностями {1,2,3,4,5}. При интенсивности внешнего (простейшего) входящего потока $\Lambda = 0.3$ узлы сети имели коэффициенты загрузки 0.900, 0.550, 0.733, 0.205 и 0.460.

Расчет выполнялся тремя способами: с помощью имитационной модели и двух численных процедур `OPNETMH` и `OPNETHH`. Первая из них предполагала все внутрисетевые потоки простейшими, а вторая пересчитывала выходящие потоки с учетом преобразования в узлах, их прореживания и суммирования на входе узлов-преемников согласно матрице передач. Имитировалось обслуживание в сети 2 млн заявок. Полученные временны́е характеристики приведены в табл. 11.10.

Таблица 11.10. Моменты распределения времени пребывания в сети

Способ расчета	Моменты		
	1	2	3
OPNETMH	3.80e1	2.09e3	1.51e5
OPNETHH	3.53e1	2.20e3	1.30e5
Имитация	3.52e1	1.86e3	1.35e5

Согласие результатов обеих численных процедур, в особенности второй из них, с имитацией следует считать очень хорошим — если учесть, что из-за неидеальной работы генераторов псевдослучайных чисел имитационная модель не может рассматриваться как совершенный эталон.

Расчетные *распределения* вероятностей (см. табл. 11.11) согласуются со статистическим намного хуже.

Таблица 11.11. Распределение $\{p_k\}$ числа заявок в сети

k	Способ расчета			k	Способ расчета		
	OPNETMH	OPNETHH	Имитация		OPNETMH	OPNETHH	Имитация
0	1.68e-2	7.80e-2	9.00e-3	15	3.34e-2	1.94e-2	2.93e-2
1	3.18e-2	8.34e-2	2.64e-2	16	3.00e-2	1.81e-2	2.51e-2
2	4.35e-2	7.89e-2	4.56e-2	17	2.69e-2	1.69e-2	2.15e-2
3	5.17e-2	7.16e-2	6.15e-2	18	2.39e-2	1.59e-2	1.85e-2
4	5.70e-2	6.38e-2	7.17e-2	19	2.12e-2	1.50e-2	1.62e-2
5	5.98e-2	5.63e-2	7.59e-2	20	1.88e-2	1.42e-2	1.37e-2
6	6.05e-2	4.95e-2	7.60e-2	21	1.66e-2	1.34e-2	1.17e-2
7	5.98e-2	4.35e-2	7.30e-2	22	1.46e-2	1.27e-2	1.01e-2
8	5.78e-2	3.84e-2	6.78e-2	23	1.28e-2	1.20e-2	8.78e-3
9	5.51e-2	3.41e-2	6.22e-2	24	1.12e-2	1.13e-2	7.54e-3
10	5.18e-2	3.05e-2	5.59e-2	25	9.82e-3	1.07e-2	6.55e-3
11	4.82e-2	2.74e-2	4.98e-2	26	8.58e-3	1.01e-2	5.59e-3
12	4.44e-2	2.49e-2	4.43e-2	27	7.47e-3	9.47e-3	4.69e-3
13	4.07e-2	2.27e-2	3.87e-2	28	6.51e-3	8.89e-3	4.06e-3
14	3.70e-2	2.09e-2	3.34e-2	29	5.66e-3	8.33e-3	3.52e-3

Естественно проверить также возможность аналитического расчета распределения числа заявок в сети через распределение числа заявок простейшего потока, прибывающих за время пребывания заявки в ней. На рис. 11.8 расчетные результаты, полученные по трем моментам модельного распределения времени пребывания заявки в сети (`alag`), сопоставлены с модельным же распределением числа заявок (`imit`). Согласие оставляет желать много лучшего. Тому есть несколько причин. Прежде всего это нарушение для сети в целом[3] принципа FCFS: заявка,

[3]Исключая тандемы из одноканальных систем.

пришедшая позже, может обогнать ранее пришедшую. Кроме того, играют роль и допущения, сделанные при декомпозиции сети. Однако *среднее число заявок в сети* будет иметь ту же относительную погрешность, что и среднее время пребывания.

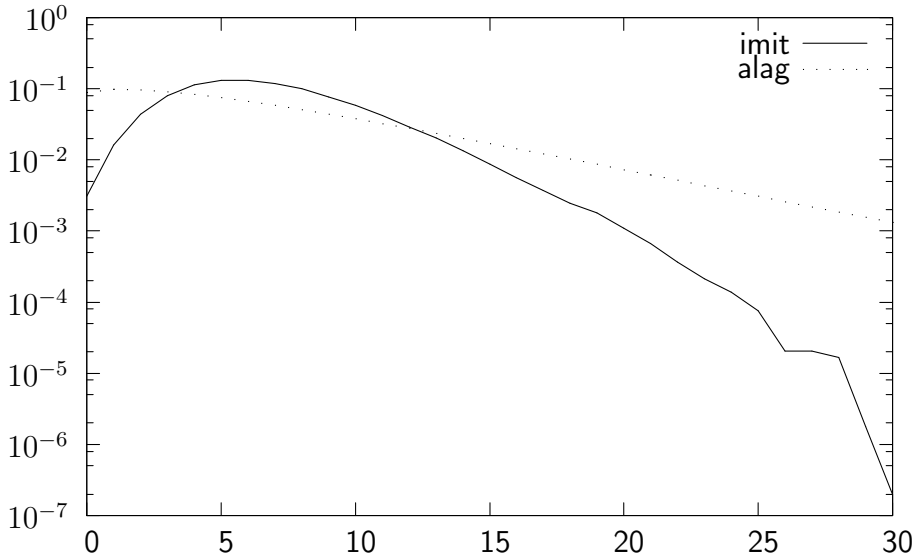

Рис. 11.8. Сопоставление расчетного и модельного распределения $\{p_k\}$ числа заявок в сети

11.16. Дальнейшие обобщения

11.16.1. Анализ вычислительных систем

В этом разделе мы лишь перечислим обсуждавшиеся в литературе применения теории очередей к анализу вычислительных процессов, попутно отмечая вероятностную специфику возникающих задач.

В [165] рассматриваются многопроцессорная вычислительная система, сеть клиент — сервер, система связи, ядро UNIX, гибкая производственная система, мультипроцессор с различными $cash$-стратегиями. Анализируется, в частности, одновременное использование нескольких ресурсов (например, оперативной памяти и процессора). При этом вводятся дополнительные пассивные узлы, содержащие фишки (tokens)

и очереди. Только получив их, прибывшая заявка направляется на обслуживание. Затем она поступает в пассивный узел, где освобождает фишки. В этом подходе просматриваются четкие аналогии с сетями Петри и G-сетями.

Содержательным объектом рассмотрения является иерархия памятей — электронные и механические устройства. Средневзвешенная задержка при обращении к памяти рассчитывается в зависимости от ее емкостей на каждом уровне и от потребностей (вероятности перехода на следующий уровень). Соответственно расчетная схема предполагает последовательное обращение к нижележащим уровням — с очередями или без них.

На [165, с. 501] рассмотрены подсистемы ввода-вывода. Операция обращения к дисковой памяти делится на фазы: установка радиального смещения и доворот диска для выбора сектора. Затем производится запись или считывание. Учитывается, что на первых двух фазах заняты одновременно диск и канал (диск может быть занят другим каналом). Можно приближенно учесть связанную с этим дополнительную задержку. Диск моделируется как система M/G/1. Общая идея решения подобных задач предполагает введение первичных и вторичных узлов, которые могут быть заняты с перекрытием или без него. Соответственно создаются две модели и ведется их итерационный пересчет.

При проектировании вычислительных систем оптимизируются число процессоров, их быстродействие, количество и быстродействие блоков памяти, топология. На Soft-уровне решаются проблемы распределения процессов по ресурсам, выбор дисциплины обслуживания и структуры приоритетов, а также проблемы поддержания надежности, размещения информации и разграничения доступа. Поучителен, в частности, рассмотренный в [229] пример работы с распределенной базой данных (при записи необходимо дополнительно поддерживать идентичность копий, размещенных в разных узлах сети).

11.16.2. Проблема «узких мест»

Сильно загруженные узлы замкнутых и смешанных сетей по отношению к заявкам «замкнутых» типов в принципе не могут быть представлены моделями разомкнутых СМО. В подобных ситуациях имеет смысл выделить наиболее загруженный узел и удостовериться в том, что его загрузка близка к единице (например, превышает 0.9). Такой

узел как правило один (в моделях вычислительных систем им обычно является процессор) и должен рассматриваться как «узкое место» сети (bottleneck), ограничивающее ее пропускную способность. Характеристики этого узла далее будем помечать индексом «B».

Расчет сетей с узкими местами базируется на следующих допущениях:

- «узкое место» постоянно занято;
- свободный ресурс $\tilde{n}_B = n_B(1 - \rho_B^o)$, оставшийся после открытых заявок, распределяется между типами $q \in Q^c$ замкнутых заявок , посещающими данный узел, пропорционально произведениям $\{x_B^{(q)} b_B^{(q)}\}$ (во всех формулах данного пункта мы используем только первые моменты распределения длительности обслуживания и потому порядок моментов не указываем).

Укрупненная схема такого расчета представляется следующей:

1. Для всех открытых типов заявок $q \in Q^o$ и узлов $i = \overline{1, M}$ рассчитать базовые интенсивности потоков $\{x_i^{(q)}\}$ и остаточные ресурсы узлов $\{\tilde{n}_i\}$.

2. Для всех замкнутых типов $q \in Q^c$ и узлов $i = \overline{1, M}$ рассчитать базовые интенсивности потоков $\{x_i^{(q)}\}$.

3. Рассчитать предельный коэффициент пропорциональности

$$C = \min_i \left\{ \tilde{n}_i \Big/ \sum_{q \in Q^c} x_i^{(q)} b_i^{(q)} \right\},$$

обозначить через B индекс, соответствующий минимуму.

4. Зафиксировать интенсивности потоков

$$\lambda_i^{(q)} = C x_i^{(q)}, \qquad i = \overline{1, M}, \quad \forall q \in Q^c.$$

Для всех узлов, посещаемых заявками этих («узких») типов, присоединить создаваемую ими нагрузку к «открытой».

5. Решая задачу расчета смешанной сети для всех замкнутых типов $q \in Q^c \backslash Q_B^c$, добиться балансировки числа заявок указанных типов в сети и зафиксировать средние времена $\{w_i\}$ ожидания в узлах. Заметим, что узел «B» при этом автоматически исключается из сети.

6. Рассчитать для всех «узких» типов $q \in Q^c$ ожидаемое число заявок в узле $\bar{k}_i^{(q)} = \lambda_i^q(w_i + b_i^{(q)})$, $i = \overline{1,M},\ i \neq B$, и суммарное число их в узлах, кроме «В»:

$$\bar{K}^{(q)} = \sum_{i \neq B} \bar{k}_i^{(q)}.$$

7. Если среди разностей $\{\bar{K}^{(q)} - K^{(q)}\}$ есть положительные, перейти к этапу 10.

8. Рассчитать среднее время ожидания в узле «В» согласно

$$w_B = \Big[\sum_q (\bar{K}^{(q)} - K^{(q)}) - \tilde{n}_B\Big] \Big/ \sum_q \lambda_B^{(q)}.$$

9. Если $w_B > \max\limits_{i \neq B} w_i$, перейти к этапу 11.

10. Считать, что узкого места в сети нет. Продолжить ее расчет стандартным для смешанной сети методом.

11. Рассчитать глобальные характеристики сети.

12. Конец алгоритма.

В связи с этим алгоритмом остаются нерешенными следующие вопросы:

1) Рассогласованность оценок для среднего времени ожидания w_B, получаемых отдельно для всех $q \in Q_B^c$:

$$\delta = \max_q \{[(\bar{K}^{(q)} - K^{(q)})/\lambda_B^{(q)} - b_B^{(q)}]/w_B - 1\}.$$

Здесь необходим численный эксперимент.

2) Расчет высших моментов распределения времени ожидания заявки в узле «В». Возможный вариант — представление этого узла замкнутой СМО с линейно убывающей интенсивностью входящего потока.

Описанный алгоритм может быть применен и при наличии нескольких узких мест с последовательным сокращением количества узлов и типов «замкнутых» заявок. В этом случае этапы 3–10 целесообразно оформить в виде *рекурсивной* процедуры.

11.16.3. Интервальная оценка параметров

Для оценки работы сети обслуживания могут потребоваться граничные значения времени ее реакции или иных показателей. В таких случаях для получения худших показателей следует взять комбинацию верхних оценок потоков и нижних — производительности, т. е. $\{\lambda^+, \mu^-\}$, а для лучших — $\{\lambda^-, \mu^+\}$.

11.16.4. Индивидуальности и корреляции

Серьезным недостатком всех известных методов расчета сетей является независимость выбора реализации длительности обслуживания в каждом из узлов сети. Между тем, для прикладных задач достаточно типичны ситуации, когда заявка оказывается «трудной» или «легкой» для всех узлов сети или вообще имеет индивидуальную судьбу. Указаное обстоятельство можно учесть в рамках алгоритма разд. 11.15.3, если предварительно разбить все заявки на классы с малым разбросом характеристик обслуживания внутри каждого класса.

Описанные выше методы излагались применительно к обслуживанию по схеме FCFS. Тот же подход можно реализовать и в случае приоритетного обслуживания в узлах, причем тип и назначение приоритетов можно менять от узла к узлу. При этом среднее время ожидания q-заявки в узле сети будет зависеть от ее приоритета в данном узле и должно определяться по соответствующей формуле разд. 9.1.

При расчете распределения времени пребывания заявок в сети с приоритетным обслуживанием следует после вычисления потоков в узлах методами главы 9 получить моменты распределения времени пребывания в них заявок различной приоритетности, а затем применить к каждому из классов заявок алгоритм разд. 11.15.3 для сети. Кстати отметим, что в системах передачи данных приоритетными являются квитанции о приеме сообщения. Эффективно присоединение квитанций к пакетам данных, следующим в обратном направлении. Именно так делается в известных сетях ARPA, TYMNET [203].

Не вызывает принципиальных затруднений обобщение расчетных методик на случай «превращающихся» заявок (см. разд. 11.2).

Точность расчета сетей обслуживания повышается при учете *корреляции* длительности ожидания в последовательных (по маршруту движения заявки) узлах [17, 201]. В работе Гаррисона [201] сеть анализируется

по $парам$ узлов. В ней показано, что упомянутая корреляция отрицательна и, следовательно, стандартные оценки дисперсии времени пребывания заявки в сети завышены.

При обслуживании в режиме уравнительного разделения процессора (EPS) среднее время пребывания заявки в узле согласно разд. 10.3 выражается формулой

$$v_1 = b_1/(1-\rho). \tag{11.16.1}$$

Соответственно ожидаемое число q-заявок, находящихся в узле с дисциплиной обслуживания типа EPS, составляет

$$\bar{k}^{(q)} = \lambda^{(q)} b_1^{(q)}/(1-\rho) \tag{11.16.2}$$

(индекс узла для простоты опущен). Напомним, что по смыслу дисциплины EPS количество каналов в таком узле учитывается соответствующим пересчетом производительности. Формула (11.16.2) используется для составления уравнения баланса q-заявок в сети с узлами типа EPS. Детальный расчет подобного узла затруднен тем, что для системы с упомянутой дисциплиной не решена задача расчета распределения времени пребывания заявки в ней. Известно лишь [153], что распределение числа заявок

$$p_k = (1-\rho)\rho^k, \qquad k = 0, 1, \ldots \tag{11.16.3}$$

независимо от вида распределения времени обслуживания. Совпадение формул (11.16.1) и (11.16.3) с соответствующими результатами разд. 4.2 дает основание считать время пребывания в обсуждаемой системе распределенным показательно с параметром $1/b_1 - \lambda$.

В заключение напомним, что дисциплина EPS считается хорошей аппроксимацией для весьма важных при анализе вычислительных систем режимов квантованного обслуживания.

11.16.5. Моделирование процессов с подзадачами

Важное прикладное значение, в особенности при анализе вычислительных процессов в многопроцессорных суперЭВМ и агрегатного ремонта с фазами разборки и сборки, имеют процессы типа Split–Join (с разветвлением первичных заданий на подзадачи и последующим объединением результатов). Задача расчета сети обслуживания со сборкой–разборкой решается при следующих допущениях относительно выполняющих эти функции подсетей:

1) процессы в различных подсетях независимы друг от друга;

2) не расщепляемые заявки в них не обрабатываются;

3) новый агрегат принимается в подсеть после сборки предыдущего.

Процесс обработки заявки в подсети состоит из трех фаз: разборка агрегата на блоки, их параллельный ремонт, сборка и настройка агрегата. Общая длительность пребывания агрегата в подсети определяется сверткой распределений длительностей этих фаз. Свертку удобно выполнять в моментах по многократно обсуждавшейся технологии.

В численном анализе фазы $ремонта$ существенно используется гиперэкспоненциальная аппроксимация распределения длительности обслуживания с ДФР вида (1.8.5). Поскольку общая длительность фазы ремонта определяется максимальной продолжительностью ремонта блока, ее ДФР в случае двух блоков

$$\bar{F}(t) = \bar{F}_1(t)\bar{F}_2(t) = \sum_{i=1}^{k} y_i e^{-\mu_i t} \sum_{j=1}^{k} u_j e^{-\lambda_j t},$$

а соответствующие моменты можно вычислить согласно (2.1.24). Снова построив по этим моментам гиперэкспоненту с k составляющими, мы можем учесть обслуживание третьего блока, и т. д. — аналогично процессу суммирования потоков. Выполнив вышеупомянутые свертки, получаем моменты распределения времени восстановления агрегата в подсети и заменяем ее одним эквивалентным узлом. Проделав подобные замены для всех подсетей, можно анализировать первичную сеть уже описанным методом.

Проблема синхронизации в иерархической вычислительной системе обсуждается в статье [203].

11.16.6. G-сети

Растущий размах применения сетей обслуживания (в первую очередь сетей обработки данных) открыл и новое поле деятельности для злоумышленников разного рода — хакеров, жуликов, диверсантов и т. п. Наиболее характерный пример — это проблема компьютерных вирусов.

Представляются весьма актуальными попытки создания математического аппарата, позволяющего оценивать ущерб от негативных

информационных воздействий. В частности, набирает популярность появившаяся в 1990-х гг. теория сетей Геленбе (G-сети), первоначально ориентированная на описание работы нейронных сетей (с активаторами и ингибиторами нейрона). Здесь рассматриваются два типа заявок. Заявки первого типа — обычные («положительные», «целевые»). «Отрицательная» заявка, пришедшая в узел сети, вынуждает имевшуюся в нем положительную покинуть узел. Дальнейшие обобщения состояли в удалении пачки случайного объема или всех заявок, отказе узла с потерей заявок, мгновенном перемещении одиночной заявки в другой узел, удалении случайного объема заданной работы (в том числе частично выполненной), необходимости «спецобслуживания» отрицательных заявок. К примеру, в модели передачи данных в качестве отрицательной заявки выступает сбой передачи. При этом сеть должна повторить передачу «с нуля». При моделировании G-сетью развития вирусной атаки деструктивный вирус должен быть предварительно «обслужен» (выполнен как программа).

Маршрутизация заявки может управляться внешним событием — *сигналом*. Сигнал, поступивший в узел i, переадресует заявку из его очереди в узел j с вероятностью $q_{i,j}$ или вызывает потерю заявки либо группы заявок с вероятностью $D_i = 1 - \sum_j q_{i,j}$. Сигнал может быть экзогенным или порождаться в процессе перемещения заявки по сети. При отработке сигнала возможно введение случайной задержки во времени.

Обстоятельный обзор соответствующих работ (60 источников) приведен в статье Арталехо [158]. Эти работы классифицированы по областям применения: вирусы в компьютерных сетях; удаление транзактов в базах данных; сети интегрального обслуживания; управление запасами; миграция людей и животных; городское хозяйство; синхронизация параллельных вычислений; задача коммивояжера и т. д.

В [158], а также в монографии В. М. Вишневского [28] обсуждается множество «математических» вариантов моделей G-сетей. Отмечается значительное *многообразие вариантов действия* отрицательных заявок: на обычную заявку в канале и/или в очереди, на канал или узел сети, а также по судьбе самой отрицательной заявки после воздействия: исчезновение, преобразование в положительную, миграция по сети, размножение и т. п. Решение задач в подавляющем большинстве случаев ограничивается записью уравнений баланса для интенсивностей потоков. Продолжаются попытки строить решения этих очень сложных задач на алгоритмах, связанных с теоремой BCMP — бессмысленные уже по-

тому, что условия последней с требованиями к практическим задачам не пересекаются. Не обсуждается вопрос о целевых функциях при расчете G-сетей. Обходятся проблемы расчета узлов сети (в особенности многоканальных с немарковским обслуживанием), принципиально разных временны́х масштабов процесса обслуживания заявки и ликвидации отказа канала или последствий вирусной атаки.

Для снятия ограничений BCMP мы вновь будем использовать *потокоэквивалентную декомпозицию* открытой однородной сети. Предполагается, что в сеть поступают «положительные» заявки, каждая из которых в процессе своего обслуживания (т. е. находясь непосредственно в канале) может быть «убита» пришедшей из внешнего источника «отрицательной» заявкой. Убитая заявка заново генерируется в источнике и продвигается по сети — до следующего убиения или благополучного завершения обработки. Решается задача расчета моментов распределения времени пребывания в сети положительной заявки — *от первого появления до попадания в сток*. Результаты расчета сопоставляются с имитационным моделированием.

Нумерация вершин. Ключевым элементом описанной ниже расчетной схемы является расчет длительности неполного маршрута в зависимости от места гибели положительной заявки, что невозможно при повторном прохождении узлов. В связи с этим приходится делать допущение об отсутствии в сети циклических маршрутов (исключая повторную генерацию убитой заявки). Тогда можно перенумеровать узлы на основе отношений предшествования — например, с помощью известного алгоритма Форда. Перенумерация позволяет искать предшественников узла j только среди узлов с номерами $i = \overline{1, j-1}$. Более того, на уровне программной реализации имеет смысл просматривать этот список в обратном порядке и прекращать просмотр при первом же нарушении непосредственного предшествования.

Далее мы всюду предполагаем, что упомянутая перенумерация проведена.

Расчет узлов. Прибытие отрицательной заявки «выбивает» из канала положительную, уменьшая число заявок в системе на единицу, и в этом смысле эквивалентно завершению обслуживания. Однако этот эффект не растет по числу занятых каналов и, следовательно, не может быть учтен пересчетом их быстродействия. С другой стороны, его нельзя отразить и через уменьшение интенсивности прибытия положительных заявок λ^+, поскольку прибытие отрицательной заявки в свободную

систему не оказывает на нее никакого действия. Следовательно, для обсчета узлов с положительными и отрицательными заявками необходимо разрабатывать *специальные* методы.

Узлы сети в общем случае могут быть многоканальными с произвольным распределением длительности обслуживания. Соответственно для их аппроксимации приходится воспользоваться моделью $M/H_2/n$, в которой обслуживание каждой заявки с вероятностью y_i подчинено показательному распределению с параметром μ_i, $i = 1, 2$. Эту ситуацию можно интерпретировать как обслуживание потока заявок двух типов с различными экспоненциальными законами. Такая модель обсуждалась в главе 7 и обсчитывалась на основе модифицированного варианта метода Такахаси—Таками. Важным достоинством H_2-аппроксимации является применимость при любых коэффициентах вариации (при малых — с комплексными параметрами).

В нашем случае диаграмма переходов по прибытию положительных заявок, сохраняет прежний вид — см. главу 7. *Уходы заявок* для модели $M/H_2/3$ показаны на рис. 11.9. Цифры слева (номера ярусов) обозначают общее количество заявок в системе, а кодовые числовые комбинации дополнительно идентифицируют *микросостояния*, определяя количество заявок первого и второго типа, проходящих обслуживание. Этим количествам пропорциональны интенсивности как обслуживания, так и «убиения» заявок соответствующих типов. Величины y_1 и y_2 в нижней части диаграммы суть вероятности выбора из очереди на обслуживание заявок первого и второго типов и рассматриваются как множители при интенсивностях потоков, проставленных у корня стрелки.

Согласно этой диаграмме строятся матрицы $\{B_j\}$ интенсивностей перехода на вышележащий ярус. Матрицы $\{A_j\}$ переходов вниз сохраняют свою структуру; ко всем элементам диагональных матриц $\{D_j\}$ интенсивностей ухода из микросостояний j-го яруса добавляется интенсивность λ^- прибытия отрицательных заявок.

Схема расчета узла в основном совпадает с описанной в главе 7. Начальные приближения к $\{t_j\}$ считаются распределенными биномиально с нормированными к единице вероятностями $\{y_1/\mu_1, y_2/\mu_2\}$. Итерации продолжаются до стабилизации с заданной точностью отношений вероятностей состояний смежных ярусов $x_j = p_{j+1}/p_j, \quad j = \overline{0, N-1}$.

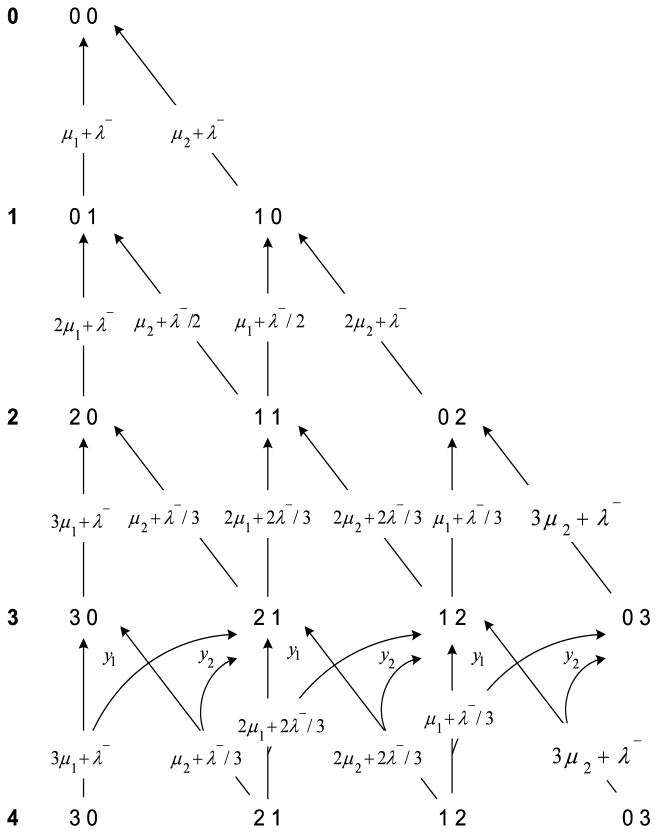

Рис. 11.9. Переходы с уменьшением числа заявок

Далее принимается $p_0 = 1$, а последующие вероятности считаются согласно $p_j = p_{j-1}x_{j-1}, \quad j = \overline{0, N}$, с одновременным накоплением суммы. Вероятности для $j > N$ представляются геометрической прогрессией со знаменателем x_{N-1}, так что к сумме добавляется $p_N x_{N-1}/(1 - x_{N-1})$. Затем все вычисленные вероятности делятся на упомянутую сумму.

Перейдем к расчету условной вероятности q «благополучного» завершения обслуживания в узле. Запишем условие баланса заявок на входе и выходе узла

$$\lambda^+ = \lambda^-(1 - p_0) + q\lambda^+, \tag{11.16.4}$$

(первое слагаемое в правой части (11.16.4) — это интенсивность прошедших в систему отрицательных заявок, каждая из которых «убивает»

положительную. Отсюда следует, что

$$q = 1 - \frac{\lambda^-}{\lambda^+}(1 - p_0). \qquad (11.16.5)$$

Расчет характеристик времени пребывания заявки в узле проводится уже обсуждавшимися методами. Моменты распределения времени пребывания в узле обслуженной и прерванной заявок получаются сверткой в моментах распределений общего времени ожидания и полной либо остаточной длительностей обслуживания.

Баланс межузловых потоков. Обозначим Λ^+ суммарную интенсивность внешнего потока и зададим матрицу вероятностей межузловых переходов $R = \{r_{i,j}\}$, $i, j = \overline{0, M+1}$. Здесь узел «0» — источник заявок, «M+1» — сток (при попадании в него фиксируется окончание пребывания заявки в сети), а прочие узлы считаются рабочими и производят фактическое обслуживание. Заявка, получившая полное обслуживание, переходит в один из очередных рабочих узлов или покидает сеть; «убитая» заявка мгновенно вновь появляется в источнике и адресуется в один из рабочих узлов.

Указанные соображения приводят к системе линейных алгебраических уравнений относительно интенсивностей $\{\lambda_i^+\}$ входящих в узлы потоков:

$$\lambda_i^+ = \Lambda^+ r_{0,i} + \sum_{j=1}^{M} \lambda_j^+ [q_j r_{j,i} + (1 - q_j) r_{0,i}], \qquad i = \overline{1, M}. \qquad (11.16.6)$$

Второе слагаемое в квадратных скобках соответствует доле повторно генерируемых заявок, направляемых из источника в i-й узел. Отметим, что вероятности $\{q_j\}$ зависят от интенсивностей входящих в узлы положительных и отрицательных потоков, что определяет необходимость решения системы (11.16.6) внутри итерационного цикла. На первом шаге для всех j разумно принять

$$q_j = \int_0^\infty e^{-\lambda_j^- t}\, dB_j(t) \qquad (11.16.7)$$

(вероятность неприбытия отрицательных заявок за время обслуживания положительной), а далее уточнять их согласно (11.16.5).

Конечные результаты решения системы (11.16.6) следует проверить на загрузку: для всех узлов должно выполняться условие $n/b + \lambda^- > \lambda^+$

(максимальная интенсивность ухода заявок должна превышать интенсивность их прибытия). Таким образом, необходимо выполнение условия

$$n/b > \lambda^+ - \lambda^-.$$

При его нарушении требуется корректировка исходных данных. Это могут быть изменение маршрутизации заявок, увеличение количества каналов в критических узлах или повышение быстродействия каналов — с учетом экономических и иных соображений.

По окончательному решению системы (11.16.6) можно определить приведенную интенсивность потока из источника, состоящую из внешнего потока и потоков заявок, обслуживание которых было прервано в различных узлах:

$$\Lambda_\Sigma^+ = \Lambda^+ + \sum_{j=1}^{M} \lambda_j (1 - q_j). \qquad (11.16.8)$$

Отметим также, что при расчете распределения времени ожидания необходимо учитывать возможное продвижение j-очереди как по обслуживанию (с вероятностью q_j), так и вследствие «убиения» заявки (с дополнительной вероятностью).

Продвижение по сети. Здесь при организации вычисления сумм учитываются соображения, изложенные в связи с нумерацией вершин. Первый узел имеет только одного предшественника (источник), единичную вероятность предшественника и нулевую задержку в нем.

После завершения итераций для каждого узла вычисляются:

а) Моменты $\{w_j\}$ распределения времени ожидания в очереди; моменты $\{v_j\}$ пребывания в узле обслуженной заявки и $\{\hat{v}_j\}$ — прерванной.

б) Кумулянтная вероятность Q'_j добраться до j-го узла. Она равна отношению интенсивности входящего в узел потока к общей интенсивности потока из источника:

$$Q'_j = \lambda_j / \Lambda_\Sigma^+.$$

в) Вероятность успешного прохождения j-го узла

$$Q_j = q_j Q'_j.$$

г) Вероятность убиения заявки именно в j-м узле

$$\hat{Q}_j = (1 - q_j) Q'_j.$$

д) Кумулянтные моменты распределения времени пребывания в сети j-убитой заявки

$$\hat{V}_{j,:} = \left(\sum_{i=1}^{j-1} r_{i,j} V_{i,:} \right) * \hat{v}_{j,:}$$

и аналогичные моменты для j-прошедшей

$$V_{j,:} = \left(\sum_{i=1}^{j-1} r_{i,j} V_{i,:} \right) * v_{j,:}$$

(здесь $*$ означает оператор свертки в моментах).

Теперь можно рассчитать наборы моментов:
средней длительности жизни в цикле для убиенной заявки

$$A = \sum_{j=1}^{M} \hat{Q}_j \hat{V}_j,$$

моментов времени пребывания в сети при успешном (полном) ее прохождении

$$B = \sum_{j=1}^{M} r_{j,M+1} V_{j,:}$$

и, наконец, вероятность «счастливого» прохождения сети

$$x = \sum_{j=1}^{M} r_{j,M+1} Q_j.$$

Искомое распределение полного времени ожидания ответа сети мы сначала получим в терминах преобразований Лапласа—Стилтьеса (ПЛС), построенных по моментам соответствующих распределений. Пусть

$\alpha(s)$ — ПЛС распределения длительности неполного прохода,

$\beta(s)$ — то же для успешного прохода.

Тогда результирующая ПЛС

$$\gamma(s) = \sum_{k=1}^{\infty} \Big[(1-x)\alpha(s) \Big]^{k-1} x\beta(s) = \frac{x\beta(s)}{1-(1-x)\alpha(s)}.$$

Умножим обе части этого равенства на знаменатель правой и разложим все ПЛС по степеням параметра. В коэффициенты этого разложения

будут входить моменты известных и искомого распределений. Приравнивая множители при одинаковых степенях s в левой и правой частях равенства, получаем рекуррентное выражение для моментов времени реакции сети

$$g_k = b_k + \frac{1-x}{x}\sum_{i=0}^{k-1}\frac{k!}{i!(k-i)!}g_i a_{k-i}, \quad k = 1, 2, \ldots$$

при начальных

$$g_0 = 1, \qquad g_1 = b_1 + \frac{1-x}{x}\,a_1.$$

По найденным моментами можно построить аппроксимацию дополнительной функции распределения времени пребывания заявки в сети до ее «полного удовлетворения».

Численные эксперименты. Центральным элементом предлагаемой методики является расчет узла сети с отрицательными заявками как изолированной системы обслуживания — точнее, определение вероятности q «неубиения» положительной заявки. Для контроля была выбрана модель $M/E_3/2$ с интенсивностью потока положительных заявок $\lambda^+ = 1.6$, отрицательных — $\lambda^- = 0.6$ и средней длительностью обслуживания $b = 1$. Имитационное моделирование (500 тыс. испытаний) дало $q = 0.8941$, а расчет итерационным методом — $q = 0.8940$. Кроме того, было проверено используемое при расчете сети разумное начальное приближение

$$q_0 = \int_0^\infty e^{-\lambda^- t}\,dB(t) \tag{11.16.9}$$

(вероятность неприбытия ни одной отрицательной заявки за время обслуживания положительной). Оказалось, что при тех же исходных данных $q = 0.8240$.

Далее рассчитывалась сеть с шестью рабочими узлами, матрицей переходов

$$R = \begin{bmatrix}
0.300 & 0.500 & 0.200 & 0.000 & 0.000 & 0.000 & 0.000 \\
0.000 & 0.300 & 0.000 & 0.400 & 0.300 & 0.000 & 0.000 \\
0.000 & 0.000 & 0.700 & 0.000 & 0.300 & 0.000 & 0.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.400 & 0.600 & 0.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.600 & 0.000 & 0.400 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.500 & 0.500 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000
\end{bmatrix}$$

(нулевой столбец опущен), числом каналов в рабочих узлах $\{1, 3, 2, 1, 3, 2\}$ и единичными средними длительностями обслуживания (предполагалось эрлангово обслуживание 3-го порядка). Суммарная интенсивность внешнего потока положительных заявок принималась $\Lambda^+ = 2.50$. Отрицательные заявки суммарной интенсивности $\Lambda^- = 0.6$ распределялись по рабочим узлам с вероятностями {0.1, 0.2, 0.3, 0.2, 0.1, 0.1}.

В табл. 11.12 иллюстрируется процесс сходимости решения уравнений баланса межузловых потоков при требовании $\varepsilon = 10^{-7}$ к максимальному по узлам уточнению последовательных приближений к вероятностям $\{q_j\}$.

Таблица 11.12. Ход балансировки межузловых потоков

Итерации	1	2	3	4	5	6
Δq	7.286e-2	4.347e-3	2.320e-4	1.124e-5	5.066e-7	2.176e-8

В табл. 11.13 представлены окончательные результаты расчета в сопоставлении с результатами имитационного моделирования той же сети (500 тыс. испытаний).

Таблица 11.13. Моменты распределения пребывания в сети

Метод	Моменты		
	v_1	v_2	v_3
Имитация	6.859	39.772	334.68
Расчет	6.855	39.697	333.41

Результаты расчета свидетельствуют о правильности математических выкладок и программ, реализующих численный метод и имитационную модель соответственно. Они подтверждают правомерность и полезность потокоэквивалентной декомпозиции G-сетей. Изложенный метод может служить источником полезных аналогий для расчета подобных сетей с иными воздействиями отрицательных заявок.

11.16.7. Пространственные процессы

В главе 14 [168] приводится постановка задачи о пространственных процессах прибытия для ячеечной структуры — проблема, актуальнейшая для мобильной телефонии. К сожалению, до анализа процесса обслуживания дело не дошло.

11.17. Оптимизация сетей обслуживания

Задача проектирования сети обслуживания включает в себя:

- выбор структуры сети,
- расчет состава и размещения оборудования,
- выбор режимов его функционирования,
- размещение информационных массивов.

Дадим краткую характеристику этих этапов.

Синтез *структуры* сети начинается с выделения ее будущих узлов как точек (районов) концентрации спроса на обслуживание, оценки величины этого спроса и элементов первичной матрицы интенсивностей обмена. В качестве исходной структуры коммуникационной сети принимается одна из базовых структур, определяемых имеемыми средствами автоматизации проектирования и/или эвристическими соображениями: вырожденный граф (изолированные вершины), полносвязный, звезда, кольцо и т. п. Затем этот граф модифицируется последовательным исключением, добавлением или заменой *одного* ребра. Может быть задано ограничение по минимальной связности (например, граф должен оставаться связным при обрыве любых двух ребер).

При расчете состава и оптимизации *оборудования* выбираются пропускная способность каналов связи, число и производительность процессоров, емкость памяти (по всем иерархическим уровням), количество и тип терминальных устройств и т. п.

Выбор *режимов функционирования* предполагает определение:

- режимов коммутации сети связи (коммутация каналов, сообщений или пакетов) и маршрутизации заявок;
- режимов работы операционных систем вычислительных комплексов (пакетная обработка или квантование времени);
- состава резидентной части операционной системы;
- количества разделов оперативной памяти;
- максимального числа активных пользователей;

- специального режима (с квантованием процессора, порогом включения, с «прогулками», с приоритетами) и назначение параметров выбранного режима.

Размещение информационных массивов производится по узлам сети, иерархическим уровням памяти, отдельным накопителям и цилиндрам дисковой памяти и является основным фактором, определяющим эффективность работы баз данных.

Перечисленные проблемы как правило решаются по критерию минимума средних затрат при фиксированных предельных затратах или иных внешних условиях, определенных на предыдущих этапах проектирования. Этапы охвачены многочисленными обратными связями (например, распределение информации по узлам сети может вызвать изменение потоков запросов, которое потребует корректировки топологии сети) и используют разнообразный математический аппарат [5, 36, 151].

11.17.1. Выбор производительности узлов сети

Пусть сеть состоит из M узлов, моделируемых системами $M/M/1$, интенсивность внешнего потока равна Λ, трудоемкость обслуживания в i-м узле составляет в среднем s_i операций, быстродействие — h_i опер/с. Стоимость аппаратного оснащения узла $z_i = g\sqrt{h_i}$ (закон Гроша). Требуется выбрать такие $\{h_i\}$, чтобы при $\sum\limits_i z_i \le Z$ среднее время пребывания заявки в сети было минимально.

Прежде всего получим математическое выражение целевой функции. Среднее время пребывания заявки в системе $M/M/1$ составляет $(\mu - \lambda)^{-1}$. Ожидаемое число заявок получим умножением на λ. На основании формулы Литтла для сети в целом получаем

$$T = \Lambda^{-1} \sum_i \frac{\lambda_i}{\mu_i - \lambda_i} = \Lambda^{-1} \sum_i \frac{\lambda_i}{h_i/s_i - \lambda_i}.$$

Далее, суммарные затраты

$$L = \sum_i g\sqrt{h_i} = g \sum_i \sqrt{h_i} \le Z,$$

причем минимум времени пребывания достигается на границе допустимой по затратам области. Окончательная формулировка задачи: найти

$$\max_{\{h_i\}} \sum_i (h_i/\sigma_i - 1)^{-1}$$

при условии $\sum\limits_i \sqrt{h_i} = \bar{Z}$, где $\sigma_i = s_i\lambda_i$, $\bar{Z} = Z/g$.

Запишем функцию Лагранжа

$$L = \sum_i (h_i/d_i - 1)^{-1} + f(\sum_i (\sqrt{h_i} - \bar{Z}).$$

Условия оптимальности $\{h_i\}$ имеют вид

$$\begin{aligned} \partial L/\partial h_i &= \frac{f}{2\sqrt{h_i}} - [\sigma_i(h_i/\sigma_i - 1)^2]^{-1} = 0, \qquad i = \overline{1, M}, \\ \partial L/\partial f &= \sum_i \sqrt{h_i} - \bar{Z} = 0. \end{aligned}$$

Их можно свести к системе

$$\begin{aligned} \sqrt{h_i} &= f\sigma_i(h_i/\sigma_i - 1)^2/2 = 0, \qquad i = \overline{1, M}, \\ f &= 2\bar{Z}/\sum_i d_i(h_i/\sigma_i - 1)^2, \end{aligned}$$

которая решается численно при начальном приближении, получаемом из соображений равной стоимости узлов: $M\sqrt{h} = \bar{Z}$, т. е.

$$h_i = (\bar{Z}/M)^2, \qquad i = \overline{1, M}.$$

В процессе уточнений $\{h_i\}$ должно быть обеспечено $h_i > \sigma_i$.

11.17.2. Оптимизация маршрутной матрицы

Постановка задачи и идея алгоритма. Динамичность условий эксплуатации сетей обслуживания определяет особый интерес к *маршрутизации* заявок, изменение которой требует минимальных временны́х и финансовых затрат и организационных усилий. На маршрутную матрицу, помимо общевероятностных, могут быть наложены многочисленные ограничения, диктуемые технологическими и/или организационными соображениями. Даже в случае информационно-вычислительных сетей, узлы которых оснащены однотипными ЭВМ, возможности переброски

заявок из узла в узел ограничиваются наличием в альтернативных узлах дубликатов информационных массивов и обрабатывающих программ, дополнительного оборудования и подготовленного персонала.

Ниже обосновывается и тестируется метод многошагового преобразования исходной матрицы передач. Расчет начинается с определения интенсивностей входящих в узлы потоков из уравнений баланса заявок (11.8.1). Затем для всех узлов должно быть проверено условие отсутствия перегрузки

$$\lambda_i b_{i,1}/n_i < 1,$$

обеспечивающее существование в сети стационарного режима, и — при необходимости — скорректированы исходные данные. Алгоритм оптимизации состоит из двух предварительных шагов (максимизация передач заявок непосредственно в сток и исключение циклических маршрутов) и итерационной части (выравнивания загрузки узлов сети).

Далее мы будем считать, что

- нули исходной матрицы означают запрет соответствующих перемещений;
- вероятности перехода из каждого i-го узла в сток ограничены сверху константами $\{\bar{r}_{i,M+1}\}$.

Расчет узлов сети. Предположение о простейших потоках на входе узлов обычно по обсуждавшимся в главе 2 причинам вполне приемлемо. Однако допущение о показательных (немарковских) распределениях *времени обслуживания* как правило является необоснованным, и порожденные им ошибки могут быть сколь угодно велики — см. комментарии к выводу формулы Полячека—Хинчина. Отмеченные обстоятельства определяют необходимость моделирования узлов сети системами с простейшим входящим потоком и произвольным распределением времени обслуживания. Последнее приходится аппроксимировать параллельно-последовательным набором фаз с экспоненциальной задержкой в каждой. Приемлемую точность (выравнивание трех заданных моментов) обеспечивает, например, гиперэкспоненциальная аппроксимация с двумя составляющими. После такой аппроксимации расчет распределения числа заявок в СМО можно выполнить методами главы 7.

Маршрутизация. На основании формулы Литтла для сети в целом среднее время пребывания заявки в *разомкнутой* сети

$$v = \sum_{i=1}^{M} \bar{k}_i \Big/ \Lambda \tag{11.17.1}$$

(среднее число заявок в сети делится на суммарную интенсивность входящего потока). Поэтому задача минимизации среднего времени пребывания заявки в сети эквивалентна минимизации суммарного числа заявок в узлах.

Прежде всего отметим очевидные общие рекомендации по улучшению маршрутизации:

1) максимизация вероятностей переходов *непосредственно в сток*;

2) исключение петлевых и циклических маршрутов.

Первая из них с учетом отмеченного выше ограничения $r_{i,M+1} \le \bar{r}_{i,M+1}$ приводит к пересчету вероятностей перехода для всех узлов i, смежных со стоком, согласно

$$\begin{aligned} \tilde{r}_{i,M+1} &= \bar{r}_{i,M+1}, \\ \tilde{r}_{i,j} &= r_{i,j}\frac{1-\bar{r}_{i,M+1}}{1-r_{i,M+1}}, \qquad j = \overline{1,M}. \end{aligned}$$

Петлевые маршруты (возврат в текущий узел на повторение обслуживания) можно исключить соответствующим пересчетом длительности обслуживания в узле. Пусть $\beta(s)$ — преобразование Лапласа—Стилтьеса (ПЛС) распределения длительности обслуживания в узле, а p — вероятность возврата на дообслуживание. Тогда ПЛС *суммарной* длительности обслуживания в узле

$$\varphi(s) = \sum_{k=0}^{\infty}(1-p)p^k\beta^{k+1}(s) = (1-p)\beta(s)\sum_{k=0}^{\infty}[p\beta(s)]^k = \frac{(1-p)\beta(s)}{1-p\beta(s)}.$$

Избавимся от знаменателя и разложим обе части полученного уравнения по степеням s. Приравнивая коэффициенты при одинаковых степенях s, получаем выражения для моментов суммарной длительности обслуживания в узле

$$f_1 = \frac{b_1}{1-p},$$

$$f_m \;=\; b_m + \frac{p}{1-p}\sum_{k=0}^{m-1}\frac{m!}{k!\,(m-k)!}f_k b_{m-k}, \quad m = 2,3,\ldots$$

Приведенное распределение соответствует предоставлению заявке всего времени обслуживания в узле *подряд*. Такая замена не влияет на распределение длины очереди в узле.

С помощью формул (11.17.2) легко вывести соотношения между коэффициентами немарковости (см. разд. 1.4) исходного и приведенного распределений:

$$\xi_2^* = (1-p)\xi_2.$$

Таким образом, при устремлении вероятности возврата к единице приведенное распределение стремится к показательному.

При реализации описанного подхода заметно увеличиваются моменты распределения времени ожидания в узлах и (по «замкнутым» типам заявок) несколько уменьшаются интенсивности потоков. Комбинированное влияние этих факторов и изменения матрицы передач на итоговую характеристику (среднее время пребывания заявки в сети) оказывается незначительным. Точность вычисления *высших* моментов распределения времени пребывания в сопоставлении с имитационным моделированием при исключении петель оказывается несколько хуже.

Циклические маршруты порождаются необходимостью возврата на повторное выполнение ранее пройденных этапов обработки и существенно затягивают пребывание заявки в сети. В предельном случае (при возврате с вероятностью p с выхода на вход сети) здесь возникает полная аналогия с «узловым» возвратом.

В хорошо отлаженном процессе обслуживания возвраты должны быть практически исключены. Соответственно вероятность возврата в узел $r_{j,i*}$ для $i^* < j$ должна быть распределена между прочими «преемниками» узла:

$$\tilde{r}_{j,i} = r_{j,i}/(1 - r_{j,i*}). \qquad (11.17.2)$$

Дополнительным полезным эффектом исключения циклов является существенное ограничение переборов при анализе потоков: после переупорядочения узлов (например, с помощью алгоритма Л. Р. Форда) предшественниками j-го узла могут быть только узлы с номерами $i < j$, а преемниками — с $i > j$. Заметим, что для графов с возвратами алгоритм Форда зацикливается.

Возможности дальнейшего улучшения работы сети связаны с уменьшением суммарного количества заявок в узлах. Даже однократный расчет среднего числа заявок в узле методами гл. 7 весьма трудоемок; эта трудоемкость многократно усугубляется при любом алгоритме перераспределения потоков (см., например, [52], где обсуждаются методы пси-преобразования и главных осей Брента). Еще более сложны (но оправданы в своей предметной области) методы выбора оптимальных *маршрутов в сетях передачи данных* [28, гл. 7]. Упомянутые методы опираются на предложенную еще Джексоном и не выдерживающую серьезной критики аппроксимацию узлов одноканальными системами с показательно распределенным обслуживанием. Однако многолетние поиски простой, работающей в широком диапазоне условий и достаточно точной приближенной формулы расчета среднего числа заявок в системе (здесь ключевой момент — оценка среднего времени ожидания) оказались безуспешными.

Реально применима лишь стратегия *последовательного выравнивания коэффициентов загрузки узлов*. Первый вариант этой стратегии (выравнивание загрузки в паре узлов с общим предшественником) сравнивался с выравниванием среднего числа заявок в той же паре. Он приводил практически к тому же результату (разница — в третьем знаке), но исключал из каждой итерации дополнительное четырехкратное обращение к процедуре расчета модели $M/H_2/n$ для узла сети. Ниже обсуждается более общая стратегия выравнивания загрузки, приводящая к уменьшению общего числа итераций:

1) На каждом шаге алгоритма выбирается наиболее загруженный узел $j = \mathrm{Arg}\max_i\{\lambda_i b_{i,1}/n_i\}$.

2) Определяется узел d из числа создающих в нем наибольшую нагрузку $\lambda_d r_{d,j}$ и имеющий не менее двух преемников.

3) Все потоки, выходящие из узла d, посредством изменения соответствующей строки маршрутной матрицы перераспределяются между узлом j и прочими $i^+ \in \Gamma(d)$ (от первого в пользу последних) из соображений выравнивания коэффициентов загрузки.

После каждого такого шага посредством решения уравнений баланса выполняется пересчет интенсивностей $\{\lambda_i\}$ входящих в узлы потоков, рассчитываются узловые СМО и средние количества $\{k_i\}$ заявок в них.

Накопленная сумма последних после деления на Λ дает среднее время пребывания заявки в сети v. Итерации продолжаются до стабилизации v с приемлемой погрешностью.

Обсудим технику выравнивания загрузки узлов. Пусть z_j — доля потока $x = \lambda_d r_{d,j}$, изымаемая из узла j, а z_{i+} — передаваемая в узел i^+. Условие равенства коэффициентов загрузки имеет вид

$$\rho_j - z_j x b_{j,1}/n_j = \rho_{i+} + z_{i+} x b_{i+,1}/n_{i+}, \quad \forall\, i^+. \tag{11.17.3}$$

При этом должно выполняться равенство

$$z_j = \sum_{i+} z_{i+}. \tag{11.17.4}$$

Из уравнения (11.17.3) находим

$$z_{i+} = \left(\rho_j - \rho_{i+} - z_j \frac{x b_{j,1}}{n_{j,1}}\right) \frac{n_{i+}}{x b_{i+,1}} = \frac{n_{i+}}{b_{i+,1}} \left(\frac{\rho_j - \rho_{i+}}{x} - \frac{b_{j,1}}{n_j} z_j\right).$$

Суммируя эти уравнения по всем $\{i^+\}$ и учитывая (11.17.4), находим

$$z_j = \frac{\rho_j \sum\limits_{i+} n_{i+}/b_{i+,1} - \sum\limits_{i+} \rho_{i+} n_{i+}/b_{i+,1}}{x\left(1 + \dfrac{b_{j,1}}{n_{j,1}} \sum\limits_{i+} \dfrac{n_{i+}}{b_{i+,1}}\right)}. \tag{11.17.5}$$

Тестирование алгоритма. Для тестирования алгоритма была выбрана сеть с 6 рабочими узлами и начальной матрицей передач

$$R = \begin{bmatrix} 0.2000 & 0.7000 & 0.1000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.3000 & 0.0000 & 0.3000 & 0.4000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.5000 & 0.2000 & 0.3000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.7000 & 0.3000 & 0.0000 \\ 0.2000 & 0.0000 & 0.0000 & 0.0000 & 0.4000 & 0.2000 & 0.2000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.4000 & 0.6000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{bmatrix}$$

(выведены строки $i = \overline{0, M}$ и столбцы $j = \overline{1, M+1}$). Предельная вероятность перехода в сток была выше указанных в последнем столбце только для четвертой строки (0.3000). Количество каналов по узлам принималось {1, 3, 2, 1, 3, 2}. Все распределения длительности обслуживания

предполагались показательными с единичным средним — эти допущения не снижали общности результатов, поскольку расчеты выполнялись на базе пакета программ МОСТ, допускающего упомянутые в разд. 3 фазовую аппроксимацию распределений. Интенсивность внешнего потока принималась $\Lambda = 3.8$.

Начальным (минус 2-м) шагом алгоритма явилось спрямление маршрутов — в смысле максимизации вероятностей передачи заявок в сток из смежных с ним узлов, где эта вероятность была ниже установленного предела $\bar{r}_{i,M+1}$. Соответственно четвертая строка матрицы передач изменилась на

$$0.1750 \quad 0.0000 \quad 0.0000 \quad 0.0000 \quad 0.3500 \quad 0.1750 \quad 0.3000.$$

На следующем шаге была устранена возвратная дуга из 4-го в 1-й узел, после чего та же строка преобразовалась в

$$0.0000 \quad 0.0000 \quad 0.0000 \quad 0.0000 \quad 0.4242 \quad 0.2121 \quad 0.3636.$$

В таблице 11.14 представлена динамика коэффициентов загрузки узлов.

Таблица 11.14. Коэффициенты загрузки узлов

Итер.	Узлы					
	1	2	3	4	5	6
-2	0.9336	0.9800	0.9250	0.8681	0.9659	0.9439
-1	0.9105	0.9777	0.9233	0.8598	0.9459	0.9197
0	0.7600	0.9627	0.9120	0.8056	0.9297	0.9168
1	0.9120	0.9272	0.8854	0.8299	0.9303	0.9118
2	0.9120	0.9272	0.8854	0.8299	0.9229	0.9185
3	0.9107	0.9106	0.8983	0.8196	0.9222	0.9210
4	0.9107	0.9106	0.8983	0.8196	0.9217	0.9214
5	0.9107	0.9106	0.8983	0.8196	0.9216	0.9215
6	0.9107	0.9106	0.8983	0.8196	0.9216	0.9216
7	0.9107	0.9106	0.8983	0.8196	0.9216	0.9216

Таблица 11.15 показывает перераспределение числа заявок по узлам, среднее время пребывания заявки в сети и его уменьшение от шага к шагу.

Таблица 11.15. Перераспределение ожидаемого числа заявок

Итер.	Количество заявок - по узлам						T	ΔT
	1	2	3	4	5	6		
-2	14.0649	50.1720	12.8177	6.5817	29.4308	17.3001	34.3072	
-1	10.1681	44.9671	12.5152	6.1309	18.5583	11.9391	27.4418	6.8654e-0
0	3.1667	26.8761	10.8406	4.1440	14.2873	11.5028	18.6362	8.8056e-0
1	10.3636	13.8061	8.1956	4.8796	14.4208	10.8183	16.4432	2.1930e-0
2	10.3636	13.8061	8.1956	4.8796	13.0403	11.7458	16.3240	1.1922e-1
3	10.2024	11.2465	9.3049	4.5427	12.9228	12.1363	15.8831	4.4088e-1
4	10.2024	11.2465	9.3049	4.5427	12.8422	12.2074	15.8805	2.5248e-3
5	10.2024	11.2465	9.3049	4.5427	12.8230	12.2245	15.8800	5.3549e-4
6	10.2024	11.2465	9.3049	4.5427	12.8184	12.2287	15.8799	1.2446e-4
7	10.2024	11.2465	9.3049	4.5427	12.8173	12.2297	15.8799	2.9638e-5

Конечная матрица передач

$$R_{15} = \begin{bmatrix} 0.2397 & 0.6470 & 0.1133 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.3000 & 0.0000 & 0.3000 & 0.4000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.5000 & 0.2000 & 0.3000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.6864 & 0.3136 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.4242 & 0.2121 & 0.3636 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.4000 & 0.6000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{bmatrix}$$

Неоднородный поток заявок. Предложенный подход можно обобщить на неоднородный поток заявок. Снабдим параметры задачи дополнительным верхним индексом типа заявки $^{(q)}$. Тогда суммарная загрузка j-го узла определится выражением

$$\rho_j = \frac{1}{n_j} \sum_q b_{j,1}^{(q)} \sum_i \lambda_{i,j}^{(q)}.$$

Для наиболее загруженного в этом смысле узла j матрицы передач $\{R^{(q)}\}$ можно корректировать в порядке убывания слагаемых внешней суммы (его загрузки заявками типа q) описанным выше способом — по всем типам заявок или до тех пор, пока его загрузка не опустится ниже следующего по степени загруженности узла.

Из сопоставления результатов с разумными ожиданиями следует, что рассмотренный метод вполне работоспособен и обеспечивает значительное уменьшение среднего времени T пребывания заявки в сети, особенно на начальных шагах.

Применительно к сетям с повышенной ролью временно́го фактора пользователей обычно интересует вопрос о вероятности пребывания заявки в сети долее заданного времени, причем само это время указывается со значительным элементом произвола. Практически в таких случаях нужно указанным выше способом оптимизировать маршрутную матрицу, методами глав 7–8 получить согласно *высшие* моменты распределения времени пребывания заявки в сети и построить по ним аппроксимацию ДФР законом Вейбулла ДФР для значений аргумента, перекрывающих диапазон возможных изменений директивного срока.

Алгоритм коррекции маршрутных матриц прост, эффективен и позволяет легко учесть ограничения на допустимость исправления их элементов. Его программная реализация опирается на надежно тестированные средства пакета МОСТ. Если ограничиться *приближенными* оценками среднего времени пребывания заявок в узлах, субоптимизация маршрутной матрицы и среднего времени пребывания заявки в сети могут быть выполнены обсуждаемыми здесь элементарными средствами.

Глава 12

Пакеты программ для расчета систем и сетей обслуживания

Расчеты систем и сетей обслуживания должны выполняться как на этапе проектирования соответствующих систем, так и в ходе их эксплуатации. В последнем случае необходимость расчетов может порождаться изменением:

- объема нагрузки вследствие подключения дополнительных пользователей;
- характера нагрузки (появление новых задач, смена алгоритмов или технологии решения старых);
- организации работы (введение приоритетов, специализация каналов и узлов, перераспределение потоков в сети, смена операционной системы вычислительной машины);
- конфигурации аппаратной части (вывод в ремонт или на длительную профилактику части устройств, подключение дополнительных, модернизация или замена устаревшей техники);
- требований к качеству функционирования системы (ужесточение их, специализация по видам заявок, предъявление дополнительных требований и т. п.).

Эти расчеты в сущности являются вычислительным экспериментом над математической моделью системы (сети) обслуживания.

Описанные в предыдущих главах методы расчета систем и сетей являются весьма сложными даже в своей идее. Необходимость автоматической генерации ключей микросостояний и матриц интенсивностей переходов, обеспечения требуемой точности, уплотнения информации, учета специфических структур матриц, ускорения сходимости итераций, и т. п. дополнительно усложняют программную реализацию упомянутых методов. Практически как проектировщики, так и менее подготовленный в теории эксплуатационный персонал должны быть вооружены готовыми *пакетами прикладных программ* (ППП) по решению задач теории очередей.

12.1. Структура пакетов прикладных программ

ППП — важнейший инструмент выполнения массовых расчетов в достаточно формализованных областях знания. Каждый такой пакет является комплексом взаимосвязанных прикладных программ, специальных и общих средств системного обеспечения.

Функциональная часть пакета — это совокупность базисных алгоритмов, комбинированием которых может быть получено решение достаточно широкого круга задач предметной области. Модульность функциональной части обеспечивает:

- обозримость, структуризацию и систематизацию ее;
- возможность параллельной разработки;
- сокращение общего объема программного текста;
- гибкость применения;
- уменьшение потребности в оперативной памяти избирательной загрузкой только необходимых в частном случае модулей;
- уменьшение затрат на трансляцию.

Недооценка требований модульности порождает псевдоуниверсального и крайне неповоротливого монстра. В проведенном автором численном эксперименте моделирование системы массового обслуживания на GPSS World потребовало в 80 (!!!) раз больше машинного времени, чем прогон имитационной же модели, записанной на Фортране.

Специальные средства системного обеспечения — это общий для функциональных модулей специфический механизм приведения их в действие. К таким средствам можно отнести: входной язык пакета, транслятор написанных на нем заданий, архивы программ и текстовых вставок, банк исходных данных и результатов счета, монитор интерактивного общения, справочную систему (`help`). Объем и сложность специальных средств системного обеспечения ППП существенно зависят от реализуемого уровня автоматизации его использования. Автоматическое планирование вычислений требует включения в систему в той или иной форме знаний об общих закономерностях предметной области, семантике процедур и их параметров, условиях применения отдельных модулей. Само это планирование может проводиться как на основе встроенной жесткой логики (например, в процессе диалога с пользователем), так и на базе общих методов искусственного интеллекта [142]. Последний подход при всей своей элегантности обычно менее эффективен.

При работе с ППП непосредственно и/или через специальные системные средства используется и *общее* программное обеспечение вычислительной установки: компиляторы, макрогенераторы, редакторы связей, загрузчики, подпрограммы общематематических задач; текстовые редакторы; файловые системы.

Общие и специальные системные средства пакета в идеале [30] должны предоставить пользователю возможности контроля времени счета, объема занимаемой памяти, отслеживания прохождения этапов вычислений, оперативного документирования, привязки листингов к наборам исходных данных и версиям программ.

Пакеты прикладных программ должны удовлетворять общим требованиям, предъявляемым к программному продукту. Из них мы выделим два основных, в совокупности обеспечивающих удобство и безопасность эксплуатации ППП: документированность и проверенность. Проверенность пакета должна демонстрироваться в ходе автономного и комплексного тестирования, на моделях с известным решением и пере-

крестным методом, при фиксированных и случайных входных данных. Она должна быть удостоверена авторитетной комиссией специалистов.

12.2. История ППП по теории очередей

Сложность решения достаточно содержательных задач ТМО сравнительно рано сделала их объектом усилий программистов. Одно из первых упоминаний [272, 1966г.] описывает программу, позволяющую рассчитывать итерационным методом марковские СМО с числом состояний до 5000. В программе были применены координатное задание элементов разреженных матриц и специальные приемы работы с клеточными матрицами.

Бурный рост числа публикаций по *пакетам* программ для задач ТМО отмечается с конца 1970-х гг. Эти работы ведутся независимо и практически одновременно в целом ряде стран: США (Рейзер и Кобаяси — пакет RESQ [249]), Франции (Мерль, Потье и Веран — пакет QNAP [222]), Советском Союзе (Митрофанов [74], Порховник [84], Рыжиков [91], Сайкин [133], Яблокова [150]). Очерк истории создания зарубежных пакетов дается в обзорной статье Сойера и Мак-Нейра [252]. В сборнике [225] приводятся описания ряда более поздних разработок [166, 169, 209, 228, 271, 267]. Впечатляет география появления уже первых пакетов: США, Пуэрто-Рико, Европа, Средний Восток, Япония и Южная Африка [252]. Из более поздних отечественных разработок назовем серию пакетов, выполненных под руководством Ю. И. Митрофанова [74]–[78] и В. М. Вишневского [26], а также обширный комплекс программ, создание которых в масштабе СССР координировалась В. Ф. Матвеевым (МГУ).

Характеризуя состояние и общий уровень этих разработок *в совокупности*, отметим:

- преимущественную ориентацию пакетов на расчет сетей обслуживания, в первую очередь — удовлетворяющих условиям теоремы BCMP;

- сочетание в пакетах аналитических и имитационных моделей [14, 75], большой удельный вес чисто имитационных пакетов ([21, 24, 74, 77, 137] и множество версий GPSS);

- использование широкого круга аппроксимаций и эвристических приемов, в том числе недостаточно проверенных (пакет IQNA — по данным [252], QNA — в части суммирования и преобразования потоков);

- разнообразие языков разработки (Фортран, Алгол-60, ПЛ/1, Паскаль, Симула);

- более или менее развитую системную часть в большинстве пакетов, ориентированных на расчетные применения [84, 26];

- появление интегрированных суперсистем (PERFORMS [209]), соединяющих набор специализированных пакетов с базой данных о характеристиках оборудования вычислительных систем и трудоемкости реализации основных подпрограмм операционной системы;

- тенденцию к переносу пакетов на наиболее современные технические средства (от персональных ЭВМ [1] до суперкомпьютеров семейства Cray [239]).

Сопоставление названных пакетов затруднительно, поскольку опубликованные сведения о них неполны и зачастую имеют рекламный характер, а техническая документация недоступна (в частности, по экономическим причинам). На конференции по характеристикам и надежности вычислительных систем [219] в 1983 г. отмечалось, что нельзя требовать от компьютерных наук того же совершенства, которое достигнуто за сотни лет в «классическом» инженерном деле. М. Ньютс подвел итоги дискуссии следующим образом:

> «Доклады на этой конференции показывают, что когда талантливые люди работают над хорошей проблемой, почти всегда достигается значительный прогресс, хотя, к счастью, конца нашему делу не видно.»

То же самое можно сказать и о *нынешнем* состоянии вопроса.

12.3. Современные пакеты

Ниже дается комментированный разбор описаний нескольких новых ППП (PEPSY, SPNP, MOSES, SHARPE) для расчета систем с очередями (см. [165], где имеются ссылки на первоисточники, примеры использования и скриншоты). Во введении к обзору (с. 571) перечисляются инструменты оценки сетей (Performance Evaluation):

- дискретно-событийное моделирование;
- расчет марковских цепей в дискретном или непрерывном времени;
- точный или приближенный расчет П-сетей;
- приближенный расчет несепарабельных сетей;
- иерархические и частные модели, комбинирующие вышеперечисленное.

Далее декларируется, что сетевые модели «легки для понимания и компактно описывают задачу — в особенности для П-сетей» (тезис более чем сомнителен), и отмечается, что слишком многие ситуации не удовлетворяют условиям BCMP. В частности, подчеркнуто, что «мы нуждаемся в инструментах, которые автоматически генерируют пространство состояний и дают точный или приближенный результат».

12.3.1. PEPSY

Пакет «Performance Evaluation and Prediction System» (Нюрнбергский университет) разработан под Unix (для Windows реализовано *подмножество* алгоритмов) и предлагает свыше 30 алгоритмов для открытых, замкнутых и смешанных сетей. Сеть задается текстом или графически. Пакет включает стандартный блок спецификаций: число узлов и типов заявок, распределение времени обслуживания в каждом узле (задается двумя моментами или интенсивностью и коэффициентом вариации), маршрутизация. Имеются контрольные файлы, в которые сведены системные ограничения. Ввод — интерактивный. Программный селектор по информации о сети выдает список методов, которые могут быть применены. Алгоритмы делятся на шесть групп:

- CONV — для П-сетей;

- MVA — для П-сетей;
- приближенные алгоритмы — для П-сетей;
- приближенные алгоритмы — для несепарабельных сетей;
- автоматическая генерация и стационарные решения для цепей Маркова в непрерывном времени;
- дискретно-событийное имитационное моделирование.

Выходной файл автоматически получает префикс, указывающий на примененный метод. Выводятся средние показатели для сети *в целом* и для каждого узла — входящий поток, частота посещений, коэффициент загрузки, среднее время пребывания, число заявок.

Модернизированная версия пакета XPEPSY имеет улучшенный графический интерфейс.

12.3.2. SPNS

Ciardo, Muppala и Trivedi разработали «Stochastic Petri Net Package», ориентированный преимущественно на задачи надежности. Язык описания модели — надмножество C. Имеется возможность определять дополнительные средства (подобно языку PLUS в GPSS World). С помощью SPNS могут быть рассчитаны характеристики стационарные, переходные, усредненные по времени, до поглощения. Есть анализ чувствительности к параметрам. Имеются проверка корректности разметки графа и ряд дополнительных «примочек». Пакет позволяет моделировать «процессы с доходами».

12.3.3. MOSES

Пакет «Modeling, Specification, and Evaluation Systems» разработан в Эрлангенском университете. Пакет основан на специальном языке описания MOSEL — инструменте описания цепей Маркова с непрерывным временем. MOSES принимает описание системы и создает генератор матрицы переходов Q. Имеются пять методов решения систем линейных уравнений (все они описаны в [165]). Программа содержит декларативную часть (в том числе VECTOR — состояния, RULES — правила перехода, RESULT — требуемые показатели).

12.3.4. SHARPE

Пакет «Symbolic Hierarchical Automated Reliability Performance Evaluator» разработан в 1986 г. в Duke University. Он написан на языке C и работает на всех платформах. В отличие от PEPSY и SPNP, здесь пользователь может сам выбрать тип модели. Модели могут быть иерархическими. Постановка задачи — текстовая, недавно появилась графическая оболочка. Возможен интерактивный ввод. Достижимые результаты зависят от типа модели. Можно рассчитывать переходные режимы. Решение уравнений баланса осуществляется описанным в [165] методом сверхрелаксации или по Гауссу—Зейделю. Описаны несколько моделей: центральный вычислитель и подсистема ввода—вывода, M/M/5/100, M/M/1/K (замкнутая с поломкой и восстановлением каналов). Для каждой из них приведены схема и входной файл с комментариями. В цитируемой книге отмечается, что «в этом контексте невозможно продемонстрировать полную мощь SHARPE».

На с. 601 [165] проведено сопоставление вышеперечисленных и еще нескольких пакетов со ссылками на их описания.

12.4. Пакет МОСТ

Совокупность моделей, реализующих методы теории очередей, чрезвычайно разнообразна. Ключом к разработке эффективных и удобных в применении программ является продуманное решение вопросов организации и упаковки информации. Перекладывать эти проблемы на плечи пользователей означает гарантированную неудачу. М. Ньютс [233] полагал, что «application of the matrix methods requires almost always a careful examination of the specific model at hand», и на этом основании отклонял просьбы о создании пакета программ для применения матричных методов.

Мы считаем, однако, что квалифицированному (знающему базовые понятия теории очередей и основы программирования) пользователю помочь можно и должно. Ниже дано описание истории, состава, возможностей и технологии применения разработанного автором пакета МОСТ по состоянию на конец 2011 г.

12.4.1. История пакета

Изложенные в главах 1-11 теоретические результаты и расчетные методики позволили разработать пакет прикладных программ для расчета систем и сетей обслуживания при весьма общих предположениях об элементах математической модели.

Пакет предназначен для решения возникающих в рамках перечисленных во введении и им подобных приложений исследовательских задач анализа проектируемых систем обслуживания, а также для оценки влияния на работу действующих систем изменения входящего потока, числа и производительности каналов обслуживания, его трудоемкости, введения приоритетных дисциплин и т. п. Пакет может применяться в *учебном процессе* вузов и курсов повышения квалификации инженеров при проведении учебно-исследовательских работ по дисциплинам, связанным с ТМО, выполнении курсовых, дипломных и диссертационных работ.

Первая версия пакета на Алголе 60 (точнее, Альфа-языке) в составе 18 процедур была разработана в 1977 г. [91, 92]. Затем он был переведен на язык ПЛ/1 и существенно расширен. Эти работы в декабре 1983 г. докладывались на заседании Научного Совета по прикладным проблемам при Президиуме АН СССР. В марте 1987 г. пакет успешно прошел межведомственные испытания и под названием МОСТ (Массовое Обслуживание — СТационарные задачи) был передан для тиражирования и распространения в Государственный фонд алгоритмов и программ (Таллиннский НУЦ — [83]). В эту версию входили 83 функциональные процедуры (6000 операторов ПЛ/1) и около 50 тестов (2000 операторов). По состоянию на 1988 г. пакет эксплуатировался приблизительно в 30 организациях.

На первом этапе дальнейшего развития пакет был дополнен процедурами расчета

- систем фазового типа с ограниченной очередью,
- некоторых систем с ограниченной надежностью,
- дополнительной функции распределения на базе ДФР Вейбулла с поправочным многочленом,
- временны́х характеристик замкнутых систем.

Возможности этой версии МОСТа суммированы в Руководстве программиста к ней [54].

Второй этап развития состоял в разработке:

- быстродействующих вариантов расчета сложных многофазных систем методом матрично-геометрической прогрессии;
- новых процедур расчета временны́х характеристик систем (в частности, реализующих метод пересчета);
- декомпозиционных алгоритмов расчета сетей;
- средств учета преобразования потоков в сетях обслуживания — как автономных (суммирование и случайное прореживание), так и встроенных (выходящий из узла поток);
- новых процедур расчета систем с квантованием времени и циклических, более общей схемы смешанного приоритета, задачи с динамическими приоритетами.

Кроме того, была учтена выявленная в процессе решения сетевых задач целесообразность исключения из всех процедур любых выдач информации, кроме аварийных.

Базовая версия пакета МОСТ-2 была реализована на ЕС-1066, программно совместимой с IBM/370, и работала под управлением VM/SP. Пакет использовался в системе виртуальных машин (ВМ) с разделенными файлами, обеспечивающей древовидную структуризацию информации. Создание, просмотр, редактирование и удаление файлов и компиляция программ обеспечивались стандартными средствами ПДО [135].

Главная программа записывалась на языке ПЛ/1 [144] и запускалась процедурой `APL` для записи результата в файл или `APT` — для непосредственного вывода на терминал. Подробное описание технологии работы с ЕС-версией пакета, цикла из 8 лабораторных работ на его базе (см. разд. 13.12) и каталог пакета приведены в Руководстве [110].

Перенос пакета МОСТ на ПЭВМ сдерживался отсутствием в распространенных системах программирования для них встроенных средств работы с динамическими массивами и комплексными переменными. В версию 5.0 Фортрана 77 фирмы Microsoft [19] такие средства (и ряд других черт Фортрана 90 [6, 113, 116]) включены, что и позволило сравнительно легко решить задачу перевода. Предварительно (и отчасти в процессе перевода) были разработаны таблица соответствия конструкций ПЛ—Фортран [46], методика «безбумажного» ручного перевода и

несколько технологических программ (например, выравнивания промежуточного продукта по стандартным позициям Фортрана [111]). В процессе перевода ряд программ был серьезно переработан и по существу:

1. Благодаря установлению связи между высшими моментами периодов непрерывной занятости (в том числе с разогревом) и моментами активного времени удалось исключить трудоемкую и неустойчивую технологию численного дифференцирования из процедур расчета приоритетных систем (кроме RW).

2. В процедуре PRTYRW радикально изменена технология численного интегрирования (см. разд. 9.4.5), что резко уменьшило трудоемкость счета.

3. Изменение способа вычисления неполной гамма-функции позволило улучшить работу процедуры FLAGC и по-иному подойти к задаче вычисления частичных моментов распределений, на которую опираются PRICONT, RRG1, SET.

Кроме того, имея в виду перспективу автоматизации применения пакета, были упрощены имена процедур и унифицированы обозначения некоторых параметров.

12.4.2. Общая характеристика МОСТа

Пакет имеет *единые теоретические основы* — аппроксимацию непрерывных распределений по методу моментов и законы сохранения ТМО.

Пакет построен по *модульному* принципу, облегчающему его разработку, тестирование и модификацию и увеличивающему гибкость применения, а также уменьшающему общий его объем. Отдельными модулями независимо от кратности их использования оформлялись внутренние части некоторых алгоритмов, сводящиеся к решению стандартных математических задач (GAUSS, ROOTPOL). Однотипные вычисления, выполняемые над данными с разными атрибутами (типичный случай REAL — COMPLEX) оформлялись как разные модули (например, INTR и INTRC).

Пакет *автономен* том смысле, что не использует других библиотек стандартных процедур. Его «самообеспеченность» вызвана необходимостью согласования формулировок стандартных задач с принятыми

в пакете — по атрибутам переменных, пределам изменения индексов, упаковке матриц специальной структуры и т. п.

Пакет *избыточен* по своему составу. Ряд процедур имеет совпадающие или почти совпадающие области применения (для примера сошлемся на три процедуры многократного численного дифференцирования в нуле). Некоторые процедуры являются частными случаями других (в этом смысле $\mathtt{MG1} \subset \mathtt{EG1}$, $\mathtt{GM1} \subset \mathtt{GE1}$, $\mathtt{GM1} \subset \mathtt{GMN}$). Для других комбинаций (например, EG1 и GE1) области применения пересекаются частично. Структурная избыточность пакета позволяет:

- организовать взаимное тестирование процедур на пересечении областей их применения;
- оценить сравнительную эффективность различных подходов к решению задачи и учесть ее тонкую специфику;
- решать каждую задачу средствами минимально необходимой общности, т. е. с наименьшими затратами памяти и процессорного времени;
- при «отказе» одной процедуры попытаться (с реальными шансами на успех) применить альтернативную.

Набор процедур расчета параметров аппроксимирующих распределений по моментам исходных дополнен процедурами, восстанавливающими моменты по упомянутым параметрам. Это позволяет контролировать один из ключевых аспектов технологии применения пакета.

Пакет *эффективен* по использованию машинных ресурсов. Для процедур, решающих частные задачи, реализованы специализированные алгоритмы. В программах, связанных с большим объемом промежуточных результатов, применены различные виды плотной упаковки матриц и специальные процедуры работы с ними, исключающие вычисление заведомо нулевых элементов[1]. Из циклов вынесены постоянные части матричных произведений. Обратные матрицы как правило получаются на месте прямых. Все функциональные модули хранятся в объектном виде, что исключает необходимость их повторной трансляции.

[1] В связи с улучшением технических характеристик ПЭВМ в процедурах, созданных в последние годы, такая оптимизация не проводилась.

Пакет ориентирован на *компетентного* пользователя — программиста, знающего основы теории вероятностей и фундаментальные соотношения теории массового обслуживания на уровне основных идей данной книги. Такой пользователь, реализуя технологию *сборочного программирования*, самостоятельно пишет главную программу на входном языке пакета, включающую в себя:

- заголовок,
- объявления переменных,
- препроцессорные операторы интерфейса вызываемых модулей,
- задание исходных данных,
- операторы вызова модулей пакета,
- дополнительные нестандартные фрагменты,
- вывод результатов.

Этим обеспечиваются: полное и рациональное использование возможностей пакета; построение условных и циклических вариантов счета; задание специфических целевых функций; включение расчета систем и сетей обслуживания в контур охватывающей оптимизационной задачи; выбор и сопоставление результатов альтернативных расчетных схем; полная свобода управления выводом.

Доступ на *внутренний* уровень пакета (служебные и вспомогательные процедуры) существенно облегчает работу по решению новых типов задач.

Пакет *дружественно* настроен к пользователю. В его состав входит структурированный электронный каталог, содержащий информацию о назначении каждой процедуры, смысле и формате параметров, рекомендации по применению и выбору свободных параметров. Все процедуры имеют мнемонические имена. Порядок следования параметров унифицирован в соответствии со схемой Кендалла $A/B/n/R$. Предусмотрен (насколько допускает система программирования) автоматический контроль согласования атрибутов фактических и формальных параметров. Обширная библиотека тестов и результатов их выполнения содержит богатый набор примеров логики и техники программирования задач теории очередей и является источником полезных аналогий.

Процедуры пакета выдают диагностические сообщения в случае некорректного задания исходных распределений (отрицательная дисперсия, ненормированное распределение объема пачки) и перегрузки рассчитываемых систем. В документации к полным версиям пакета даются рекомендации по построению цепочек и составлению программ и содержатся Руководства к лабораторным работам на базе пакета, которые могут и должны использоваться при самостоятельном изучении пакета и обучении студентов и иных категорий пользователей работе с ним. Выполнение работ поможет:

- лучше усвоить теорию очередей,
- оценить широту возможностей пакета,
- убедиться в его надежности,
- приобрести необходимые для самостоятельной работы технические навыки.

12.5. Профессиональный МОСТ

12.5.1. Перечень процедур

По состоянию на сентябрь 2011 г. пакет МОСТ насчитывает около 180 процедур, записанных на Фортране 90 и разбитых на следующие функциональные группы:

- APPR — аппроксимация распределений;
- BASE — основные процедуры расчета систем;
- ADVANCED — продвинутые модели;
- FLOWS — операции с потоками;
- IMIT — имитационное моделирование;
- TIME — расчет временных характеристик;
- PRTY — приоритетные дисциплины обслуживания;
- NETW — сети обслуживания;

- MATR — формирование матриц интенсивностей переходов;
- MATH — общематематические процедуры;
- SERV — вспомогательные процедуры.

APPR. В эту группу входят процедуры расчета параметров аппроксимаций распределений гамма-плотностью с поправочным многочленом, ДФР Вейбулла с поправочным многочленом, а также фазовых аппроксимаций Кокса по трем моментам, гиперэкспоненты по трем и произвольному числу моментов, гиперэрланговой аппроксимации. Имеются также процедуры восстановления моментов по параметрам аппроксимации, табулирования плотностей и дополнительных функций распределения.

BASE содержит процедуры расчета систем с вышеупомянутыми фазовыми аппроксимациями интервалов между заявками и длительностей обслуживания. Для этих процедур имеются версии с бесконечной и с ограниченной очередью, а в случае простейшего входящего потока — и «замкнутые» варианты. Сюда же относятся процедуры, реализующие классический подход Кроммелина (`MD1, EDN`), метод вложенных цепей Маркова (`MG1, GM1, GMN`) и его комбинации с методом фаз (`EG1` и `GE1`).

ADVANCED включает процедуру `MMVN` расчета системы с неоднородными каналами, обобщения `MG1` и `MHN` на поток групповых заявок с пачками случайного объема, процедуры расчета систем с квантованием времени — циклических и многоуровневых с относительными приоритетами уровней.

FLOWS содержит процедуру `FILTR` случайного прореживания рекуррентного потока, `FLOW` — суммирования двух таких потоков; три процедуры расчета распределения числа заявок, прибывающих за случайный интервал времени.

IMIT объединяет программы имитационных моделей, использованных для обоснования и верификации численных методик — для проверки законов сохранения, случайного выбора заявок из очереди, многоуровневых систем с квантованным обслуживанием, многоканальных приоритетных систем. В эту группу входят также процедура `RANDOM` генерации равномерно распределенных псевдослучайных чисел и процедура ускоренного получения произвольного члена последовательности по заданному номеру (для формирования датчиков непересекающихся серий псевдослучайных чисел).

TIME — это группа процедур, обеспечивающих переход от распределения числа заявок в очереди к моментам распределения времени ожидания начала обслуживания при различных предположениях об аппроксимации интервалов между заявками входящего потока: экспоненциальная, двухфазные эрлангово и гиперэкспоненциальное распределения, произвольная. Большая их часть обеспечивает решение основного интегрального уравнения сохранения стационарной очереди. Процедуры различаются и по методам решения этого уравнения, часть которых рассмотрена в тексте книги. Процедура RTIME реализует метод пересчета через коэффициенты немарковости входящего потока.

PRTY обеспечивает расчет одноканальных систем с различными видами приоритетов. PRMIX позволяет реализовать схему классов (с относительными приоритетами внутри классов и абсолютными — между классами); PRDIFF — также смешанные приоритеты в зависимости от *разности* номеров классов. Как предельные варианты этих дисциплин реализуются обычные схемы относительного приоритета и абсолютного приоритета с прерыванием и дообслуживанием.

Процедуры PRS1 и PRW1 реализуют дисциплины с обслуживанием прерванных заявок заново — соответственно с новой и прежней реализациями случайной длительности обслуживания. Имеются также аналоги перечисленных процедур, позволяющие получать три начальных момента распределений ожидания/пребывания заявки в системе. Процедура DYNPR1 рассчитывает среднее время ожидания для системы с динамическими приоритетами, линейно растущими по времени ожидания. Имеются две процедуры, реализующие полуэмпирические методы расчета средних времен ожидания/пребывания для *многоканальных* систем: NPNWAIT для относительного приоритета и PRNSOJ — для абсолютного.

NETW объединяет процедуры, необходимые для расчета сетей обслуживания на основе потокоэквивалентной декомпозиции их. Это прежде всего доработки процедур группы BASE, дополнительно реализующие для узлов сети расчет выходящего потока. Они построены на коксовых и гиперэкспоненциальных аппроксимациях исходных распределений и основаны преимущественно на быстродействующем методе расчета узлов методом матрично-геометрической прогрессии.

Процедура NWSTIME позволяет через моменты распределения времени пребывания заявки в каждом из узлов при *однократном* посещении и маршрутную матрицу вероятностей переходов между

узлами получить моменты времени пребывания заявки в сети. Собственно сетевые процедуры позволяют рассчитывать сети однородные и неоднородные; замкнутые, разомкнутые и смешанные.

MATH включает процедуры решения общематематических задач, согласованные по формату данных с функциональными процедурами пакета. Это процедура вычисления логарифма гамма-функции; две процедуры решения алгебраических уравнений и процедура решения нелинейных уравнений методом Вегстейна; группа программ для решения задач линейной алгебры — расчета определителя и решения систем линейных уравнений с вещественными и комплексными коэффициентами; решения сильвестровой системы линейных уравнений; обращения матриц — общего и треугольного вида, упакованных в линейный массив и распакованных. Из процедур анализа укажем процедуры DIFNDIF, DIFNEWT, DIFSTIR многократного численного дифференцирования в нуле таблично заданной функции и SIMFAST — ускоренного численного интегрирования для подсчета усеченных моментов распределений. Ряд процедур предназначен для вероятностных задач: это процедуры CONV и CONVC свертки двух распределений в моментах; FASTCONV — быстрой многократной свертки распределения с самим собой (полезна, например, при регулярном прореживании потока); DISCONV — свертки дискретных распределений; MFACT — расчета факториальных моментов; GENERW — вычисления производящей функции.

SERV служит для решения вспомогательных задач, связанных с численными методами теории очередей. Это расчет преобразований Лапласа и моментов распределения периодов непрерывной занятости, в том числе с разогревом; вычисление стартовых вероятностей в методе Кроммелина; задание начальных приближений к векторам условных вероятностей микросостояний в итерационном методе расчета многофазных систем и итоговая нормировка вероятностей в том же методе.

12.5.2. Применение МОСТа

Подготовка к работе с ППП МОСТ состоит из следующих этапов:

- составление цепочки (выбор основной процедуры и звеньев для подготовки исходных данных и расчета производных показателей);
- согласование обозначений параметров;

- выбор значений свободных параметров;
- оформление главной программы;
- ввод задания в машину, его выполнение и анализ результатов.

В Руководстве даются подробные рекомендации по каждому этапу со схемами выбора решений. Наиболее поучителен первый этап, который мы здесь проиллюстрируем.

Пусть необходимо определить влияние коэффициента вариации v распределения времени обслуживания при фиксированном среднем на распределение времени пребывания заявки в системе $M/G/3$ для значений $v = 0,\ 0.5,\ 0.7,\ 1.0$. Прежде всего очевидно, что случай $v = 0$ соответствует схеме $M/D/n$. Поскольку эта схема сохраняет принцип «первый пришел — первый обслужен», для нее формулы (3.2.4) верны по отношению к распределениям числа заявок и времени пребывания. Имеем цепочку

$$\text{MDN} \rightarrow \text{MFACT} \rightarrow \text{MTIME} \rightarrow \text{FCWEIB}$$

(по распределению числа заявок в $M/D/n$ вычисляем его факториальные моменты; через `MTIME` получаем моменты распределения времени пребывания в системе; по ним строим таблицу ДФР).

Рассмотрим выбор выбор математических моделей для $v > 0$. Поскольку предполагаются известными лишь два момента распределения G, искать его аппроксимацию можно в классе гамма-распределений, причем формула $\alpha = 1/v^2$ для перечисленных выше ненулевых значений v дает $\alpha = 4, \approx 2$ и 1 соответственно. Целочисленные α указывают на применимость распределений Эрланга соответствующего порядка, но при $\alpha = 4$ предпочтительнее воспользоваться H_2-аппроксимацией (в этом случае число микросостояний системы окажется много меньше). Следовательно, при $v = 0.5$ нужно применить цепочку

$$\text{HYPER3} \rightarrow \text{MHN} \rightarrow \text{MFACT} \rightarrow \text{MTIME} \rightarrow \text{CONV} \rightarrow \text{FCWEIB}$$

Здесь соответствующей настройкой процедуры `MFACT` получаем моменты распределения числа заявок, ожидающих начала обслуживания; через `MTIME` переходим к моментам распределения времени ожидания, а с помощью `CONV` вычисляем моменты свертки этого распределения с распределением чистой длительности обслуживания. Эта цепочка может быть применена и как универсальный подход (возможно, кроме $v = 0.7$).

Вариант $v = 0.7$ дает $\alpha = 2$ — это особый случай, исключающий H_2-аппроксимацию. Правда, в процедуре HYPER3 предусмотрена его обработка с принудительным изменением второго момента на 10%. Если же выбрать эрланговскую аппроксимацию, то получим

MEKN → MFACT → MTIME → CONV → FCWEIB

Наконец, при $v = 1$ имеем $\alpha = 1$ и цепочку

MMN → MFACT → MTIME → CONV → FCWEIB

Указанная совокупность цепочек обеспечит решение поставленной задачи с наименьшей затратой процессорного времени. Разумеется, цепочки могут быть и условными. Конечное звено цепочек FCWEIB может быть заменено процедурой FLAGC. Программная реализация цепочек предполагает должную передачу информации через параметры вызываемых процедур.

В качестве второй задачи рассмотрим возможные подходы к расчету моментов распределения интервалов между заявками при регулярном прореживании потока (поочередное назначение поступающих заявок одному из m узлов). Регулярное прореживание потока эквивалентно операции свертки исходного распределения интервалов между смежными заявками. Задача может быть решена:

- непосредственно в моментах $(m-1)$-кратным последовательным применением процедуры CONV или однократным обращением к FASTCON (в последнем случае свертки организуются на основе двоичного представления кратности);
- по цепочке GLAG → LAPLAG → DIFNEWT (распределение аппроксимируется гамма-плотностью с поправочным многочленом; для $s = 0, h, \ldots, nh$ строится таблица m-й степени преобразования Лапласа; к таблице применяется процедура численного дифференцирования интерполяционного многочлена Ньютона; знак нечетных производных инвертируется).

В последнем варианте можно вместо DIFNEWT использовать процедуру DIFNDIF безразностного численного дифференцирования или DIFSTIR — дифференцирования интерполяционного многочлена Стирлинга. Для применения DIFSTIR значения s должны задаваться по обе стороны от нуля.

Реализация всех перечисленных подходов обеспечивает взаимное тестирование использованных процедур.

12.5.3. Состав пакета

Пакет МОСТ/F90 включает в себя:

- статическую библиотеку из 180 объектных модулей (3.2 Мбайта);
- каталог пакета;
- библиотеку `INTERFACE`-блоков;
- библиотеку из 120 тестов на Фортране 90;
- библиотеку результатов тестирования;
- файл с прототипами операторов вызова.

Каждый расчет выполняется в виде *проекта* системы Фортран 90. В проект *включаются* вызывающая программа и упомянутая статическая библиотека. После компиляции и линкования («Building») система формирует 32-разрядное приложение, запуск которого дает результаты счета. В документацию к этой версии пакета предполагается включить описание его теоретических основ (данную книгу) и «Руководство по расчету систем с очередями», представляющее собой «фортранизированный» вариант [110].

12.5.4. Учебный МОСТ

Продукт МОСТ/FPS1 [128] опирается на библиотеку исходных модулей Фортрана PowerStation 1.0 объемом 20 тыс. строк текста. Он сохраняет идеологию профессиональной версии и усеченные ее возможности, но реализован на Фортране PowerStation 1.0, соединяющем удобства Windows-интерфейса с умеренными требованиями к дисковой памяти (до 15 Мбайт) и процессору (Pentium не обязателен).

12.5.5. «Автоматизированный» МОСТ

Расширение круга пользователей МОСТа вовлечет в него лиц с программистской и теоретической подготовкой, недостаточной для работы на профессиональном уровне. Для таких пользователей С. В. Кокориным

разработана автоматизированная версия пакета — МОСТ/А. Она позволяет рассчитывать разомкнутые и замкнутые системы и сети обслуживания по исходным данным, вводимым в процессе диалога, с учетом трех моментов исходных распределений (с согласия пользователя МОСТ/А по двум моментам подберет третий). Допускается прямое указание типов конкретных распределений (показательное, эрланговское, детерминированное) — в этих случаях объем вводимых данных уменьшается. В состав пакета МОСТ/А входят:

- ведущая программа, организующая диалог с пользователем и ввод исходных данных;
- четыре генератора программ, формирующих главные Фортран-программы в соответствии с характером задачи и исходными данными конкретного варианта;
- файлы программных заготовок-фрагментов, необходимых для этого формирования;
- вырезки из библиотек объектных модулей и интерфейсов профессиональной версии — 50 процедур, в том числе 21 непосредственно включаемая в цепочки и 29 вызываемых транзитивно (список последних был построен автоматически);
- четыре теста — примеры диалога с системой, выводящие на все группы вариантов, и соответствующие результаты счета.

Ввод исходных данных производится с контролем корректности информации (неотрицательность дисперсии распределений, отсутствие перегрузки систем, нормировка к единице сумм строк матрицы передач при расчете сетей). Результаты заносятся в стандартный ответный файл в форме, удобной для использования современными графическими средствами типа `GRAPHER` или `GNUPLOT`. Руководство к этой версии пакета содержит описание его возможностей и технологии и обзорный вариант данной книги «Задачи и методы расчета систем с очередями» (около 20 страниц).

В стадии предварительной проработки находится расширение круга задач, решаемых в автоматизированном режиме (МОСТ/FP2).

В заключение хочется подчеркнуть уникальные возможности, открываемые *профессиональным* уровнем МОСТ/F90:

- расчет обширного класса систем методом вложенных цепей Маркова с учетом произвольного числа моментов временны́х распределений;
- точный расчет систем с регулярным обслуживанием;
- расчет систем с циклическим обслуживанием и с квантованием времени;
- расчет сетей с учетом преобразования потоков в узлах;
- возможность сопоставления альтернативных методов расчета, оценки влияния типов аппроксимации и разницы в высших моментах;
- возможность нестандартного использования модулей внутреннего уровня пакета для самостоятельных исследований;
- наличие учебника, руководства к лабораторным работам и программного обеспечения этих работ, в совокупности позволяющих обучаться и обучать современным методам расчета систем и сетей обслуживания (для примера укажем такую полезную работу, как экспериментальная проверка законов сохранения на имитационной модели);
- наличие библиотеки тестов, содержащих примеры использования всех процедур пакета и облегчающих такое обучение.

Глава 13

Тестирование МОСТа

13.1. Вводные положения

Известно, что любое количество правильных прогонов программы не гарантирует ее корректности. Излагаемые здесь выборочно принципы тестирования МОСТа реализуют программу-минимум: продемонстрировать работоспособность его процедур в условиях, допускающих проверку правильности их функционирования [100].

Составной частью МОСТ'а является набор из 120 тестов, хранящихся в директории `TST`. Все они оформлены как главные Фортран-программы. Изучение и прогон этих тестов:

- позволяют убедиться в правильности инсталляции пакета сравнением результатов с эталонными (директория `RES`);
- доставляют примеры техники программирования на Фортране 90 с вызовом процедур МОСТ'а;
- свидетельствуют о правильности алгоритмов, положенных в основу функционирования МОСТ'а, и их программной реализации;
- дают примеры многовариантного решения сходных задач, которые могут быть использованы при практическом применении МОСТ'а.

Некоторые файлы упомянутой директории `RES` (например, `BUSIES1`, `TINBATCH, TIMITSET`) объединяют результаты прогона разных процедур в целях большей наглядности их сопоставления.

В связи с техникой программирования на Фортране 90 отметим тест OPNTIME, активно использующий «псевдопроцедуры». В тесте TMODBATCH можно увидеть использование COMMON-блока. С точки зрения расчета *сетей обслуживания* поучителен тест TMODNW, проверяющий неотрицательность дисперсий временны́х распределений, нормировку строк матрицы передач и коэффициенты загрузки узлов (аналогичные проверки *встроены* во все сетевые процедуры).

13.2. Принципы тестирования

Тестирование процедур МОСТа основывается на одном или нескольких из следующих принципов:

- сопоставление альтернативных подходов;
- сравнение результатов численных алгоритмов и имитационного моделирования (при тестировании программ с динамическим приоритетом и многоуровневым обслуживанием эта возможность была единственной);
- сохранение инвариантов (средняя интенсивность входящего и выходящего потока; формула Литтла для разомкнутых и замкнутых систем и сетей в целом и по заявкам каждого типа в отдельности; постоянство взвешенной суммы средних времен ожидания);
- расчет частных вариантов с известным решением (например, формулами для расчета моментов периодов непрерывной занятости);
- соответствие результатов качественным ожиданиям (например, при увеличении нагрузки на одноканальную систему — приближение коэффициентов немарковости выходящего потока к аналогичным коэффициентам входящего);
- решение взаимообратных задач (параметры аппроксимаций по исходным моментам и восстановление моментов по этим параметрам; решение уравнений и подстановка корней; обращение упакованных матриц и вычисление произведения обратной матрицы на прямую);

- правильное решение сложной задачи, подалгоритмом которой является проверяемая процедура (примеры — расчет матриц интенсивностей переходов, начальных вероятностей в системах с регулярным обслуживанием).

Если прогон теста требует вспомогательных процедур, таковые включались в один файл с главной программой. При тестировании имитационных моделей это были датчики псевдослучайных чисел, при численном интегрировании — подынтегральная функция, при решении нелинейного уравнения методом Вегстейна — его правая часть.

Полнота обсуждаемого здесь тестирования соответствует требованиям этапа приемо-сдаточных испытаний: проверяются основные ветви и потенциальные особые случаи (на этапе разработки проводилось более детальное тестирование). Все тесты ориентированы на корректные наборы исходных данных. Действия пользователя при работе с собственными программами и данными обсуждаются в разделе Руководства, посвященном отладке.

Ниже приводится обоснование тестирования функциональных групп процедур МОСТа. Рекомендуется выполнять тестирование в порядке их последующего перечисления. Для каждой процедуры указывается непосредственно вызывающий ее тест или одна из более общих процедур, в составе которых работает данная — именно та, которая вызывает максимальное число не имеющих автономных тестов вспомогательных.

Реализация представляемого набора тестов позволяет уверенно рекомендовать пакет МОСТ/FP для практического использования, но, разумеется, не гарантирует отсутствия в нем ошибок и ситуаций, с которыми процедуры пакета не справляются.

13.3. Математические процедуры

Процедуры `DET`, `GAUSS`, `GAUSSC`, `INTR`, `INTRC`, `INVERT`, `INVJ`, `INVJC` служат для решения задач линейной алгебры. Работа `DET` проверяется тестом `TDET` на вычислении детерминанта с известным значением:

$$\begin{vmatrix} 1 & 1 & 1 \\ 2 & 3 & 4 \\ 3 & 5 & 6 \end{vmatrix} = -1.$$

Процедура GAUSS решения систем линейных уравнений с вещественными коэффициентами тестируется посредством TGAUSS на совпадение найденных корней с эталонными значениями, которые предварительно использовались для расчета столбца свободных членов. Аналогичная ей GAUSSC отличается только комплексными атрибутами и тестируется в составе MNODE.

Процедура MULT перемножения комплексных матриц тривиальна и тестируется работой в составе MNODE.

Процедуры обращения вещественных матриц общего вида INVJ и верхних треугольных INVERT, INVJ проверяются тестом INVERS посредством вывода произведения исходной матрицы на обратную. Их аналоги для работы с комплексными матрицами INTRC и INVJC контролируются надлежащей работой в составе модулей HPNR и MNODE.

Процедуры многократного численного дифференцирования в нуле DIFNEWT (дифференцируется стандартный интерполяционный многочлен Ньютона), DIFNDIF (коэффициенты многочлена Ньютона строятся по безразностным формулам) и DIFSTIR (дифференцируется интерполяционный многочлен Стирлинга, построенный по центральным разностям) тестируются совместно программой ALLDIF. Здесь по таблице преобразования Лапласа-Стилтьеса (ПЛС) распределения времени пребывания заявки $\nu(s) = (\mu - \lambda)/(\mu - \lambda + s)$ в системе $M/M/1$ определяются производные, которые должны с точностью до знака совпадать с моментами показательного распределения $v_k = k!/(\mu - \lambda)^k$. Тест выводит теоретические значения производных и погрешности численного дифференцирования для каждой из тестируемых процедур и порядка производной $k = \overline{1,8}$ при шаге построения таблиц $10^{-2}, 10^{-3}, 10^{-4}$. В ходе тестирования было определено рациональное значение шага для построения таблицы ПЛС $h \approx 10^{-3}\mu$.

Процедуры GENERW и MFACT расчета производящей функции и факториальных моментов дискретного распределения вероятностей соответственно проверялись на модели $M/M/1$, для которой получены явные выражения результатов

$$\begin{aligned} P(z) &= \frac{1-\rho}{1-\rho z}, \\ f_{[k]} &= k!\Big(\frac{\rho}{1-\rho}\Big)^k, \qquad k = 1, 2, \ldots \end{aligned}$$

Правильность работы остальных процедур (и прежде всего процедур

формирования матриц интенсивностей переходов между микросостояниями систем фазового типа) гарантируется успешным выполнением вызывающих их процедур.

BUSYTEST испытывает три процедуры расчета ПЛС периода занятости. Полученные таблицы дифференцируются в DIFSTIR, и найденные моменты сопоставляются с известными результатами для модели $M/M/1$:

$$\begin{aligned} \pi_1 &= (\mu-\lambda)^{-1}, \\ \pi_2 &= 2\mu/(\mu-\lambda)^3, \\ \pi_3 &= 6\mu(\mu+\lambda)/(\mu-\lambda)^5 \end{aligned}$$

и модели $M/D/1$:

$$\begin{aligned} \pi_1 &= (b^{-1}-\lambda)^{-1}, \\ \pi_2 &= b^{-1}/(b^{-1}-\lambda)^3, \\ \pi_3 &= b^{-1}(2\lambda+b^{-1})/(b^{-1}-\lambda)^5. \end{aligned}$$

Первая модель используется для проверки BUSYLAG и BUSYHYP, вторая — для BUSYDET и BUSYHYP.

В тесте TBUSYM аналогично организовано тестирование процедуры BUSYMOM, непосредственно вычисляющей моменты распределения периода занятости.

Тест TWARMUP проверяет совпадение результатов BUSYMOM и процедуры WARMUP расчета моментов периода непрерывной занятости системы $M/G/1$ с разогревом в случае, когда распределение длительности разогрева совпадает с распределением обслуживания заявок, образующих основной период занятости.

Процедуры свертки CONV и быстрой многократной свертки FASTCON тестировались на гамма-распределении, для которого свертка порождает также гамма-распределение с параметром формы, умноженным на кратность свертки.

Процедура SIMFAST вычисления усеченных моментов

$$f_n = \int_a^b x^n f(x)\,dx$$

проверяется посредством SIMTEST на задаче вычисления для $n = \overline{1,5}$

$$\begin{aligned} s_n &= \int x^n \sin x\,dx &&= -x^n\cos x + n\int x^{n-1}\cos x\,dx, \\ c_n &= \int x^n \cos x\,dx &&= x^n\sin x - n\int x^{n-1}\sin x\,dx. \end{aligned}$$

Для $a = 0$ и $b = \pi/2$ имеем

$$\begin{aligned} s_0 &= 1 = {}_0; \\ s_n &= nc_{n-1}, \\ c_n &= (\pi/2)^n - ns_{n-1}, \qquad n = 1, 2, \ldots \end{aligned}$$

Тест TWEG проверяет процедуру WEG решения уравнения методом итераций по Вегстейну на задаче $x = x^3 - 1$. Контроль решения осуществляется сопоставлением левой и правой частей. Здесь необходимо включение в тест функции FWEG.

Работа ROOTPOL контролируется тестом TROOTPL применительно к многочлену, коэффициенты которого получены по заданным корням:

$$\begin{aligned} P_5(x) &= (x-1)(x-2)(x-3)(x-4)(x-5) \\ &= x^5 - 15x^4 + 85x^3 - 225x^2 + 274x - 120. \end{aligned}$$

Аналогично TALLRTS посредством ALLROOTS вычисляет все корни многочлена

$$\begin{aligned} P_5(x) &= (x-1)(x+1)(x-3)(x^2+4x+5) \\ &= x^5 + x^4 - 8x^3 - 16x^2 + 7x + 15. \end{aligned}$$

Процедура LOGAM вычисления логарифма гамма-функции контролируется тестом TLOGAM по выполнению $\Gamma(0.5) = \sqrt{\pi}$ и функционального соотношения $\Gamma(k+1) = k\Gamma(k)$ — конкретно, равенства $\Gamma(1.5) = \sqrt{\pi}/2$. В процессе тестирования был подобран оптимальный сдвиг аргумента в асимптотическом разложении логарифма гамма-функции [152].

13.4. Аппроксимационные процедуры

Процедуры этого класса, связанные с аппроксимацией распределений посредством гамма-плотности и распределения Вейбулла с поправочным многочленом, проверяются на треугольном распределении в тесте TRIANG. В частности, в последнем сопоставляются:

- моменты распределений, полученные по ответным параметрам процедур аппроксимации, с исходными моментами (взаимообратные пары GLAG — MOMLAG, PARMWB — MOMWB, COX3 — MOMCOX);
- плотности по DENSLAG и DENSWEIB— с эталонной

$$f(t) = \begin{cases} l - 1 + t, & 1 - l \le t < 1, \\ l + 1 - t, & 1 \le t < 1 + l, \\ 0 & ; \end{cases}$$

- ДФР по FLAGC, FCWEIB— с эталонной

$$\bar{F}(t) = \begin{cases} 1, & t < 1-l, \\ 1-(l-1+t)^2/(2l^2), & 1-l \le t < 1, \\ (l+1-t)^2/(2l^2), & 1 \le t < 1+l, \\ 0 & ; \end{cases}$$

- полученные по ALAG значения вероятностей

$$a_j = \int_{1-l}^{1+l} \frac{(\lambda t)^j}{j\,!} e^{-\lambda t} f(t)\,dt$$

с теоретическими

$$a_j = \Big\{ \Big(l - 1 + \frac{j+1}{\lambda} \Big) \sum_{i=0}^{j} \frac{[\lambda(1-l)]^i}{i!} e^{-\lambda(1-l)} + \frac{1}{\lambda} \frac{[\lambda(1-l)]^{j+1}}{j\,!} e^{-\lambda(1-l)}$$
$$+ \Big(-l - 1 + \frac{j+1}{\lambda} \Big) \sum_{i=0}^{j} \frac{[\lambda(1+l)]^i}{i!} e^{-\lambda(1+l)} + \frac{1}{\lambda} \frac{[\lambda(1+l)]^{j+1}}{j\,!} e^{-\lambda(1+l)}$$
$$+ 2\Big(1 - \frac{j+1}{\lambda} \Big) \sum_{i=0}^{j} \frac{\lambda^i}{i!} e^{-\lambda} - 2\frac{\lambda^j}{j!} e^{-\lambda} \Big\} \Big/ (\lambda l^2), \qquad j = 0, 1, \ldots$$

- отдельно вероятность

$$a_0 = (e^{-\lambda(1-l)} + e^{-\lambda(1+l)} - 2e^{-\lambda l})/(\lambda l)^2$$

с доставляемой процедурой-функцией LAPLAG.

Процедура HYPER3 построения H_2-аппроксимации по трем моментам и функционально обратная ей MOMHYP контролируются тестом HYPT на совпадение исходных и восстановленных моментов. Заметим, что третий момент распределения Эрланга второго порядка восстанавливается увеличенным на 10% (процедура HYPER3 соответственно корректирует исходные данные, чтобы избежать деления на нуль). Программа THYPG контролирует восстановление моментов, выравниваемых более общей процедурой HYPERG.

Тест THERL выполняет аналогичный контроль правильности расчета гиперэрланговской аппроксимации процедурой HERL.

Для сопоставления качества гиперэкспоненциальной и гамма-аппроксимаций при вычислении ПЛС функции LAPLAG и LAPHYP совместно тестируются в LAPTEST на треугольном распределении. Здесь же LAPHYP дополнительно тестируется на вырожденном распределении.

Взаимообратные по своим функциям процедуры PARMWB, MOMWB проверяются дополнительно тестом WBT.

Задачи, требующие расчета усеченных моментов аппроксимируемых гамма-плотностью с поправочным многочленом распределений (SPT, RRG1, PRICONT), требуют обращения к процедуре FKLAG. Тест последней TFKLAG основан на том, что для распределений Эрланга α — целое, поправочный многочлен — тождественная единица)

$$\int_0^t x^k \frac{\mu(\mu x)^{\alpha-1}}{\Gamma(\alpha)} e^{-\mu x}\, dx = \frac{(\alpha+k-1)!}{(\alpha-1)!\mu^k} \int_0^{\mu t} \frac{u^{\alpha+k-1}}{(\alpha+k-1)!} e^{-u}\, du$$
$$= \frac{(\alpha+k-1)\,!}{(\alpha-1)\,!\mu^k}\Big(1 - e^{-\mu t} \sum_{i=0}^{\alpha+k-1} \frac{(\mu t)^i}{i\,!}\Big).$$

Более общие случаи могут быть проверены (для $k = 0$) сопоставлением результатов счета по FCWEIB и FLAGC. Такое сопоставление, в частности, проводится в тесте TRIANG.

13.5. Служебные процедуры

Тест TPRINT процедуры PRINTAR должен проверить возможность печати произвольной вырезки из линейного массива по столбцам (в пять столбцов). Для контроля такой последовательности удобно задать распечатку части массива натуральных чисел. Возможность и правильность построения столбцов неравной длины проверяется заданием числа элементов вывода, не кратным пяти; возможность произвольного выбора выводимой части массива — смещением параметров JMIN, JMAX относительно границ выводимого массива. Конкретно в упомянутом тесте выводятся элементы со второго по 24-й. Процедура PRINTARF работает аналогично PRINTAR, но с выводом результата в связанный с ее последним параметром файл.

Тест MFACT сопоставляет факториальные моменты распределения числа заявок для модели $M/M/1$, вычисляемые в MFACT, с их теоретическими значениями

$$\begin{aligned} f_{[k]}^{(n)} &= (1-\rho)\sum_{i=k}^{\infty} i(i-1)\cdots(i-k+1)\rho^{i+n} = k\rho^n(1-\rho)\sum_{i=k}^{\infty}\binom{i}{k}\rho^i \\ &= k\rho^n(1-\rho)\rho^k/(1-\rho)^{k+1} = k\rho^n[\rho/(1-\rho)]^k, \end{aligned}$$

где n, равное числу каналов, соответствует характеристикам очереди, а $n=0$ — системе в целом.

Процедура TCONV испытывает процедуры CONV и FASTCON свертки распределений в моментах для гамма-распределений с одинаковым масштабным параметром μ. В этом случае моменты k-го порядка

$$f_k = \alpha(\alpha+1)\cdots(\alpha+k-1)/\mu^k,$$

а моменты свертки двух распределений с параметрами α_1 и α_2 будут подсчитываться по той же формуле при $\alpha = \alpha_1 + \alpha_2$. Для проверки CONV эта идея реализуется на двух различных распределениях. «Комплексный» аналог процедуры (CONVC) тестируется в составе MNODE.

Процедура FASTCON служит для быстрого построения многократной свертки распределения с самим собой, управляемого разложением требуемой кратности по степеням двойки. В тесте для нее обрабатывается набор кратностей с различной расстановкой нулей и единиц — в частности, четные и нечетные.

Процедура DISCONV свертки дискретных распределений вероятностей по формуле

$$d_j = \sum_{i=0}^{j} a_i b_{j-i}, \quad j = 0, 1, \ldots$$

тестируется на двух пуассоновских распределениях числа заявок простейшего потока с одинаковой интенсивностью и разными интервалами времени. В качестве контрольного используется то же распределение для суммарного интервала. Расхождения начинаются при превышении индексом результата минимальной из верхних границ сворачиваемых распределений.

Тест TGENERW испытывает процедуру GENERW вычисления производящей функции

$$\Pi^{(n)}(z) = \sum_{j=0}^{\infty} z^j p_{j+n},$$

которая для системы вида $M/M/1$ с распределением числа заявок $p_k = (1-\rho)\rho^k, \quad k = 0, 1, \ldots$ сводится к

$$\Pi^{(0)}(z) = (1-\rho)/(1-\rho z), \quad \Pi^{(1)}(z) = \rho\Pi^{(0)}(z),$$

для значений $n = 0$ и $n = 1$ соответственно.

BUSYTEST испытывает три процедуры расчета ПЛС периода занятости. Полученные таблицы дифференцируются посредством DIFSTIR, и найденные моменты сопоставляются с известными результатами для модели $M/M/1$:

$$\begin{aligned} \pi_1 &= (\mu-\lambda)^{-1}, \\ \pi_2 &= 2\mu/(\mu-\lambda)^3, \\ \pi_3 &= 6\mu(\mu+\lambda)/(\mu-\lambda)^5 \end{aligned}$$

и модели $M/D/1$:

$$\begin{aligned} \pi_1 &= (b^{-1}-\lambda)^{-1}, \\ \pi_2 &= b^{-1}/(b^{-1}-\lambda)^3, \\ \pi_3 &= b^{-1}(2\lambda+b^{-1})/(b^{-1}-\lambda)^5. \end{aligned}$$

Первая модель используется для проверки BUSYLAG и BUSYHYP, вторая — для BUSYDET и BUSYHYP.

В тесте TBUSYM аналогично организовано тестирование процедуры BUSYMOM, непосредственно вычисляющей моменты распределения периода занятости.

Тест TWARMUP проверяет совпадение результатов BUSYMOM и процедуры WARMUP расчета моментов периода непрерывной занятости системы $M/G/1$ с разогревом в случае, когда распределение длительности разогрева совпадает с распределением обслуживания заявок, образующих основной период занятости.

Тест TXLIM проверяет процедуру XLIM расчета предельного отношения вероятностей в разомкнутых СМО. Результатом XLIM является ответный параметр CC вызываемой тестом процедуры HHN. Он сопоставляется с отношением компонент ответного массива W(JMAX-1)/W(JMAX-2), а также с расчетом по приближенной формуле $x = \rho^{2/(v_A^2+v_B^2)}$. Поскольку значение CC — предельное при $j \to \infty$, здесь следует ожидать лишь приблизительного согласия.

Остальные вспомогательные процедуры тестируются в составе базовых процедур пакета, успешная отработка которых подтверждает их правильность. Укажем примеры их применения:

- EXPOM — при расчете выходящего из узла потока (MNODE и др.);
- EXPROOT и LAGROOT — в EG1;
- NORMW — в базовых процедурах с неограниченной очередью (MHN);
- NORMWR — в расчете систем с ограниченной очередью и замкнутых (HPNR);
- PST — в составе EDN;
- START, STARTH — в базовых процедурах итерационного расчета (в зависимости от типа распределения обслуживания — MEN, MHN).

Отметим особенности тестирования процедуры PST. В процедуре EDN при $kn = 1$ она формирует один вещественный корень вспомогательной системы уравнений. В случае $kn > 1$ при нечетном kn дополнительно определяются пары комплексных сопряженных корней, а при четном — еще один вещественный. Значит, должен быть проверен по крайней мере набор моделей $M/D/1$, $M/D/3$, $M/D/2$, $E_4/D/2$.

13.6. «Матричные» процедуры

К этой группе относятся процедуры, формирующие матрицы интенсивностей переходов между микросостояниями систем, марковизируемых методом фиктивных фаз. Сложность формирования исходных данных для них оправдывает их тестирование только в составе базовых процедур. В качестве минимального покрытия множества матричных процедур мы рекомендуем набор {MEN, MHN, HPNR, MMVN}. Этот выбор согласован с процессом испытания служебных процедур (см. разд. 13.5). Конкретно, процедуры KEY, HMATRA, HMATRA1, HMATRC проверяются в составе MHN; EMATRA и EMATRB — в составе MEN; EMATRBC, PA1M, PCM — в составе HPNR и, наконец, VMATRA, VMATRA1 — при вызове MMVN.

13.7. Базовые процедуры

Базовые процедуры расчета изолированных СМО тестируются на соответствие результатов, полученных для одной и той же модели. В основу проверки этих процедур положен принцип *взаимности* [99] —

подбор таких исходных данных, при которых сравниваемые процедуры должны давать одинаковые или близкие результаты. В то же время должны быть проверены весь диапазон функций каждой процедуры и значения входных параметров, априорно вызывающие опасения. Такие параметры обсуждались в разделе об аппроксимации распределений. Эти соображения могут быть конкретизированы следующим образом:

1) H-аппроксимация тестируется
   - — показательным случаем (M);
   - — эрланговским E_k при $k \geq 3$ или регулярным D (проверяется работа модели с комплексными параметрами);
   - — случаем Γ-распределения с показателем степени $\alpha = 1.5$ («патологический» вариант с отрицательной вещественной вероятностью).

2) C_2-аппроксимация тестируется на распределениях с коэффициентом вариации из интервала (1,2) (парадоксальная область), и распределениях класса E_3 (комплекснозначные параметры).

3) E_k-аппроксимация обязательно тестируется М-вариантом и общим случаем $k > 1$.

4) Гиперэрлангова аппроксимация проверяется с целым и с дробным числом фаз;

5) G-аппроксимация (с непосредственным заданием входных моментов) тестируется на моделях M, E и (при наличии «взаимности») на D (случай малых коэффициентов вариации).

6) Все «многоканальные» процедуры должны быть проверены для случаев $n = 1$ и $n > 1$.

7) Частные результаты, получаемые с помощью простейших методик, должны согласовываться с их аналогами для более «мощных» процедур.

8) Все «неоднородные» модели при фактической однородности должны давать тот же результат, что и соответствующая однородная модель.

9) Каждая тестовая ситуация для возможности локализации ошибки должна обсчитываться минимум тремя программными модулями.

Теория построения рациональных схем взаимного тестирования модулей функционально избыточных программных систем описана в [99]. Построенный на ее основе с учетом вышеприведенных рассуждений сокращенный вариант схемы взаимного тестирования моделей разомкнутых СМО с бесприоритетным обслуживанием показан на рис. 13.1.

Исходные данные подбирались из расчета коэффициента загрузки $\rho = 0.7$ и J_{max} =19. Моменты исходных распределений считались как моменты гамма-распределения с параметром формы α через первый момент:

$$f_i = f_{i-1} \cdot f_1 \cdot (1 + (i-1)/\alpha), \qquad i = 2, 3.$$

Для унификации программирования расчетов вырожденное (D)-распределение обсчитывалось так же при $\alpha = 10^9$. Числа на пересечении имен модели и модуля указывают ожидаемую длину очереди — этот показатель компактен, инвариантен к масштабу времени и более чувствителен к способу расчета, чем ожидаемое число заявок в системе.

Сравнительная эффективность методов Такахаси—Таками и матрично-геометрической прогрессии при большом числе каналов определялась тестами `TMHBIGN` и `TMHBIGG` соответственно. В связи с большим временем счета для отличения ситуации от зависания эти тесты выводят на экран текущие значения n, а в ответный файл записывается время обсчета каждого варианта. При этом времена, меньшие 0.01 (порог чувствительности подпрограммы контроля), заменяются нулями.

Проверенные вышеуказанными способами базовые процедуры могут использоваться как эталоны при тестировании имитационных моделей. В этом случае достаточно однократно убедиться в правильности логики моделий, так что прогонять последние для различных комбинаций исходных распределений не было необходимости.

Процедуры расчета систем с ограниченной очередью тестировались при тех же исходных данных и `JMAX=7` (для демонстрации заметного влияния ограниченности очереди). Системы замкнутые тестировались при `JMAX=7`, зависимости интенсивности входящего потока от числа заявок в системе $\lambda(k) = 0.3n(7-k)$ и средней длительности обслуживания $b = 1$.

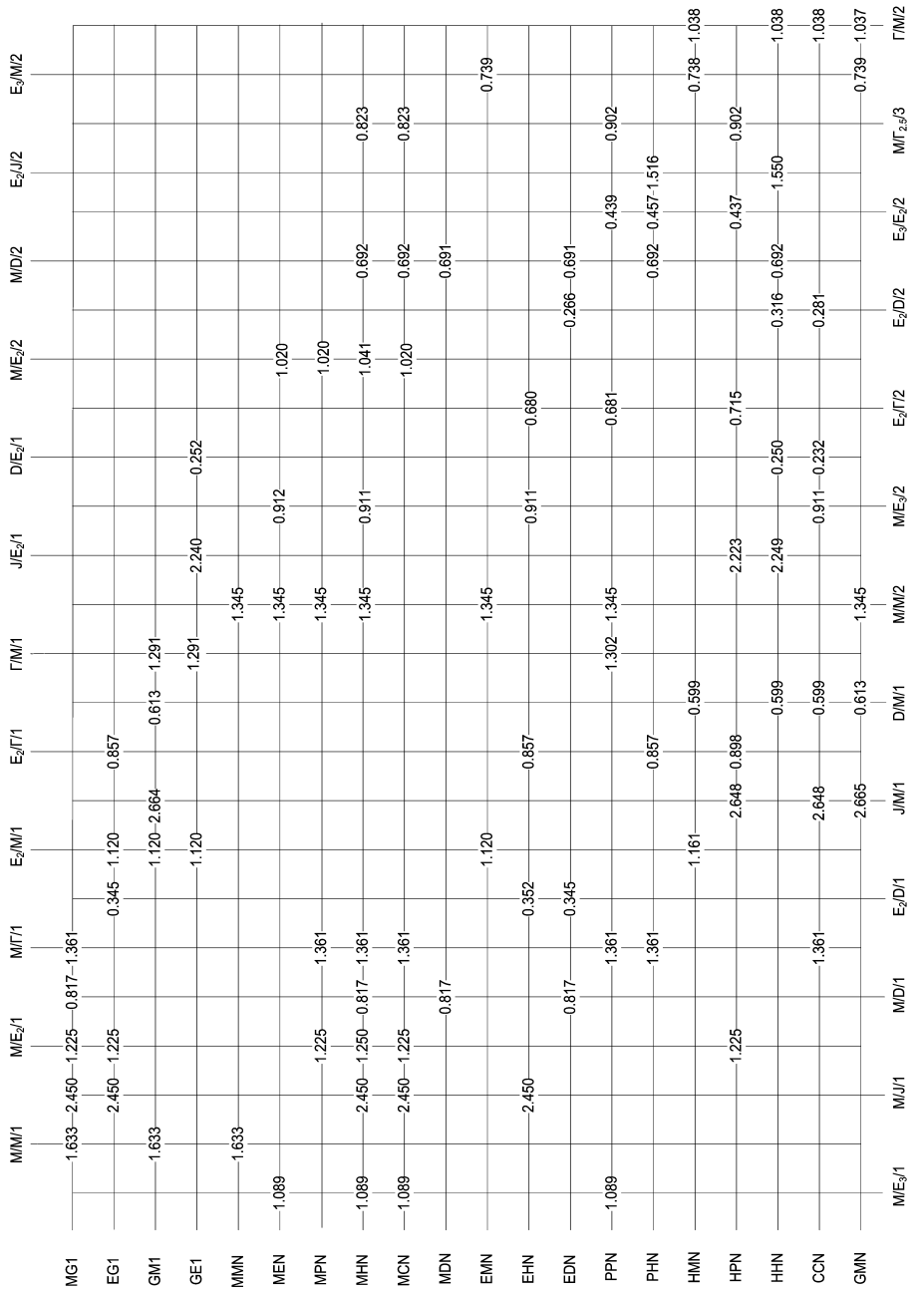

Рис. 13.1. Тестирование программ расчета разомкнутых систем

13.8. Временны́е процедуры для FCFS

Для облегчения сопоставления результатов расчета временны́х характеристик систем обслуживания процедуры расчета последних проверяются двумя комплексными тестами. Тест OPNTIME (разомкнутые системы) проверяет процедуры E2TIME, H2TIME, HETIME, GLTIME, XLTIME, RTIME и WFCFS на моделях систем $M/E_3/1$, $E_2/M/2$, $\Gamma_{1.5}/M/1$. В качестве эталонных используются моменты распределения времени ожидания, получаемые из процедур GE1 и GMN (разд. 5.8.2 и 5.7.2 соответственно). Их сравнение с результатами более общих численных методов проведено в разд. 8.7. Несовместимые комбинации «модель — процедура» не обрабатываются.

Тест CLOTIME применяет процедуру MHNC к «замкнутым» моделям $\hat{M}/E_3/1$, $\hat{M}/M/3$ и $\hat{M}/D/3$. Для контроля используются средние времена ожидания, получаемые по формуле Литтла.

13.9. «Приоритетные» процедуры

В указанную группу собраны процедуры расчета систем с дисциплиной обслуживания, отличной от стандартной FCFS. В основу тестирования собственно приоритетных процедур положены следующие соображения:

- первые моменты распределения времени пребывания заявки в системе, получаемые согласно «полным» процедурам (PRTYNP, PRTYPR, PRTYMIX, PRTYRS, PRTYRW) должны совпадать с получаемыми из их аналогов, дающих только первые моменты (PRDIFF, PRMIX, PRS1 и PRW1);

- процедуры для смешанного приоритета при соответствующих режимах их использования должны выводить на варианты «чистых» приоритетов; например, PRDIFF при разности типов DP=0 должна давать обсчитываемые моделью PRMIX варианты с чисто абсолютным приоритетом, при DP=K — с относительным);

- результаты расчета моделей $M/M/1$ по PRTYPR и PRTYRS, а также $M/D/1$ по PRTYRS и PRTYRW, должны совпадать;

- результаты расчета средних времен ожидания (пребывания) в системе с помощью «многоканальных» процедур NPNWAIT и PRNSOJ при $n = 1$ должны совпадать с результатами их «одномоментных» аналогов, а при $n > 1$ — с результатами имитационного моделирования.

Из средних характеристик наиболее чувствительны к технике расчета и дисциплинам обслуживания средние времена пребывания в системе заявок младшего приоритетного типа (в примерах — третьего). Эти результаты и представлены на рис. 13.2. Сравнение данных на этом рисунке нужно проводить по горизонталям.

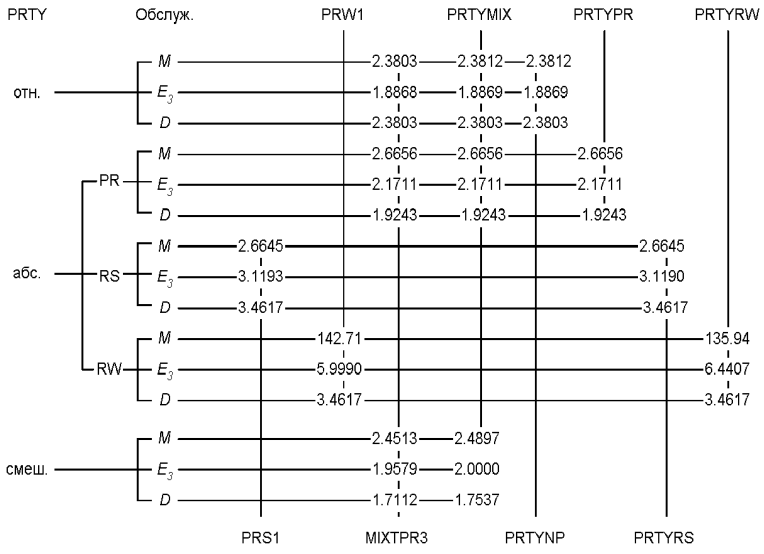

Рис. 13.2. Тестирование процедур расчета приоритетных режимов

Отметим, что результаты тестирования подтверждают уменьшение средних времен ожидания и пребывания заявок в системе при уменьшении вариации длительностей обслуживания и ранее установленный для бесприоритетных систем эффект дробления производительности — при обсчете многоканальных приоритетных систем.

Процедуру PRICONT (непрерывная шкала приоритетов) в принципе можно сопоставить с настроенной на соответствующий приоритет PRMIX после разбиения плотности распределения на достаточное число интервалов и замены непрерывного распределения неоднородным потоком с

усредненными по интервалам характеристикам обслуживания. Процедура DYNPRN (динамические приоритеты, линейно возрастающие по времени ожидания) должна сопоставляться с имитационной моделью DYNIMIT, а в одноканальном случае — и с DYNPR1. Здесь в связи вынужденным применением в DYNPRN аппроксимаций распределений по двум моментам следует ожидать незначительных расхождений.

Процедура RRG1 расчета циклической системы $M/G/1$ (тест TRR) тестировалась на модели $M/M/1$, для которой вероятность возврата на дообслуживание постоянна (решение известно только в средних).

Результаты процедуры SET (Shortest Elapsed Time) сопоставляются с имитационной моделью SETIMIT.

13.10. Сетевые процедуры

Процедура FILTR случайного прореживания рекуррентного потока тестируется в TFILTR вычислением моментов распределения интервалов между выходящими заявками. Средствами контроля служат:

- сопоставление моментов из FILTR с полученными численным дифференцированием ПЛС искомого распределения;
- обратная пропорциональность между первым моментом и вероятностью z сохранения заявки;
- теоретически доказанная пропорциональность коэффициента немарковости ξ_2 и вероятности z сохранения заявки;
- стремление коэффициентов немарковости к нулю (т. е. потока — к простейшему) при $z \to 0$.

Тест TSUMFL проверяет процедуру FLOWSUM на задаче последовательного суммирования n потоков, $n = \overline{1,10}$. Критериями правильности работы FLOWSUM являются:

- обратная пропорциональность среднего интервала между заявками сумарного потока числу слагаемых;
- стремление коэффициентов немарковости ξ_2 и ξ_3 суммарного потока к нулю при увеличении n (т. е. его приближение к простейшему);

- сохранение $\xi_2 = \xi_3 = 0$ при суммировании *простейших* потоков.

Процедуры «сетевой» группы, отвечающие за расчет отдельных узлов, в своей основной части копируют базовые и нуждаются в отдельной проверке только в связи с анализом выходящего потока. Правильность расчета последнего подтверждается:

- равенством среднего интервала интервалу между заявками выходящего потока;

- совпадением коэффициентов немарковости для исследуемых моделей при существенно различных алгоритмах анализа выходящего потока;

- близостью этих коэффициентов к соответствующим значениям для регулярного потока $(\xi_2 = -1, \quad \xi_3 = -5)$ в модели $M/D/1$ при $\rho = 0.9$;

- их уменьшением по абсолютной величине при увеличении n и уменьшении ρ.

- согласием результатов при гиперэкспоненциальной и коксовой аппроксимациях одной и той же модели.

Дополнительная проверка GNODE проводится тестом TRANFLOW, демонстрирующим близкую к линейной зависимость коэффициентов немарковости XD выходящего потока от аналогичных коэффициентов XA входящего. Опорная зависимость строится по точкам для гамма-распределения с параметром $\alpha =$ 0.3 и 3.0. Относительные отклонения от линейности по ξ_2 и ξ_3 (DD2 и DD3 соответственно) вычисляются при $\alpha = 0.5,\ 1.0,\ 10^9$; последнее значение соответствует регулярному потоку. Предположение об упомянутой линейности существенно используется в процедурах OPNETHH и OPNETWR.

Схема сопоставления результатов расчета среднего времени пребывания заявки в *сети обслуживания* представлена на рис. 13.3.

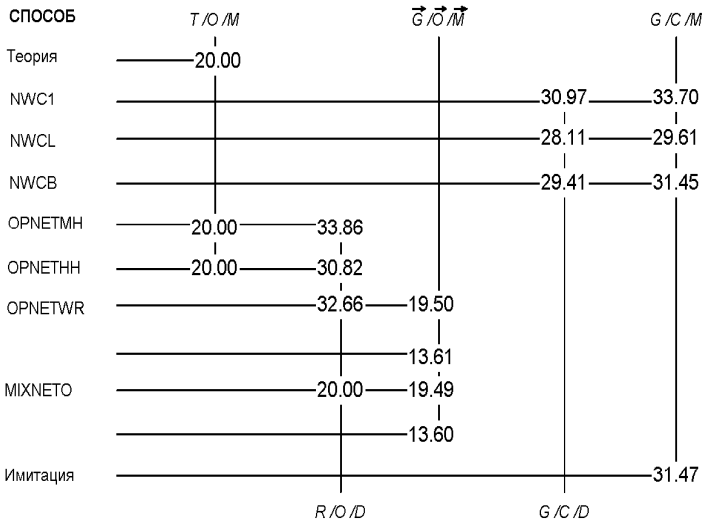

Рис. 13.3. Тестирование «сетевых» процедур

Здесь применены трехэлементные обозначения сетевых моделей вида $S/F/B$, где

- S — структура (T — тандем, W — дерево, R — кольцо, G — общий случай),
- F — тип потоков (O — открытый, C — замкнутый, X — смешанный),
- B — общий для всех узлов тип распределения обслуживания по стандартной нотации Кендалла.

Для неоднородного потока ($\vec{G}/\vec{O}/\vec{M}$) приведены данные по двум типам заявок. В указанных вариантах заявки каждого типа обходили один из узлов сети, что позволило проверить правильность работы со списками посещаемых вершин.

Тесты `TNWC1`, `TNWCL` и `TNWCB` проверяют соответствующие процедуры расчета *среднего* времени пребывания заявки в замкнутой однородной сети и показатели загрузки узлов при различных предположениях об интенсивностях входящих в узлы потоков (не зависящая от числа заявок, линейно убывающая, поток от экспоненциальной сбалансированной сети). Контроль осуществляется по относительному согласию этих

результатов между собой и улучшению этого согласия с ростом популяции от 8 до 14. Результаты свидетельствуют о монотонном уменьшении средних времен пребывания заявок в сети при уменьшении вариации распределений обслуживания (росте α) и возрастании — при увеличении популяции.

Тест TOPNWMH контролирует процедуру OPNETMH расчета открытой сети с простейшими входящими потоками, а TOPNWHH — OPNETHH с рекуррентными входящими потоками и пересчетом высших моментов распределений интервалов между заявками. Эталонной задачей является тандем из систем $M/M/1$ (см. пояснения к тесту NWSTIME), обсчитываемый по полной схеме. Кроме того, в обоих случаях рассчитывается циклическая сеть с регулярным обслуживанием.

TOPNWR тестирует процедуру OPNETWR расчета открытой неоднородной сети. Контрольной задачей здесь является псевдонеоднородный вариант упомянутой циклической сети. Считается также неоднородная марковская сеть.

TMIXNWO тестирует процедуру MIXNETO расчета смешанной сети. Варианты прогона подобраны для обеспечения сопоставимости результатов с полученными по частным алгоритмам:

- открытая псевдонеоднородная — OPNETHH,
- открытая с двумя типами заявок — OPNETWR,
- замкнутая однородная — NWC1.

Расчет общего случая (два «замкнутых» и два «открытых» типа) в процессе разработки сопоставлялся с результатом прогона имитационной модели. Маршрутные матрицы заданы с учетом переменного списка посещаемых узлов по типам заявок и многообразия «классификации» узлов (чисто открытые, замкнутые и смешанные).

Процедура NWSTIME расчета высших моментов распределения времени пребывания заявки в сети тестировалась на тандеме из N систем вида $M/M/1$ — в этом случае распределение времени пребывания заявки в сети есть N-кратная свертка показательного закона с параметром $\mu - \lambda$, т. е. распределение Эрланга N-го порядка с моментами

$$v_k = N(N+1)\dots(N+k-1)/(\mu-\lambda)^k, \qquad k = 1, 2, \dots$$

Кроме того, NWSTIME проверялась для разомкнутых сетей произвольной структуры на согласие получаемых через нее первых моментов с вычисляемыми согласно (11.15.1).

13.11. Выводы по тестированию МОСТа

Систематичность, глубина, полнота и разнообразие методов тестирования, использованных при разработке МОСТа, и положительные результаты тестирования позволяют считать теоретические основы пакета правильными, а их программную реализацию — достаточно надежной.

13.12. Лабораторные работы с МОСТом

На базе последовательных версий МОСТа автор более 30 лет проводил (в составе различных дисциплин учебного плана) лабораторные работы по теории очередей. Перечислим несколько конкретных работ и соответствующие методические установки.

Преобразование потоков. Исследуется влияние входных параметров на процессы просеивания, суммирования и преобразования потоков в сетях обслуживания. Экспериментально (по вычисляемым значениям коэффициентов немарковости) устанавливается *асимптотическое* приближение к простейшему просеиваемых потоков — при увеличении вероятности выбрасывания заявки, суммарных — при увеличении числа составляющих. Выявляется приближение регулярно просеиваемого потока к регулярному, выходящего из системы $A/B/1$ при $\rho \to 1$ — к рекуррентному с распределением B интервалов между заявками. Обнаруживается линейная связь между коэффициентами немарковости выходящего и входящего потоков.

Проверка законов сохранения. В каждом варианте этой работы для заданных числа каналов и типа распределения длительности обслуживания стандартная имитационная модель доставляет распределение числа заявок в системе и статистические моменты распределения времени ожидания начала обслуживания. Тип и параметры исходных распределений задаются в сменных внешних датчиках случайных чисел — процедурах пользователя. Проверяются закон сохранения заявок (3.1.2), обе формулы Литтла для средних значений временных характеристик и

формулы (3.2.4) для высших моментов. Выводы о справедливости различных форм законов сохранения делаются на основе относительной невязки левой и правой частей соответствующих равенств и сопоставляются с предсказанными теорией.

Влияние высших моментов. Для одноканальных систем $M/G/1$ и $GI/M/1$ исследуется влияние числа учтенных моментов немарковского распределения на выбор оптимального объема восстанавливаемого ЗИП и относительное увеличение ожидаемых затрат, а также на относительную погрешность вероятности превышения допустимого времени пребывания заявки в системе. Здесь прежде всего ожидается вывод о недопустимости замены показательным законом распределений, заметно от него отличающихся (равномерное, треугольное, вырожденное, Вейбулла с большим коэффициентом вариации).

Дробление производительности и масштабный эффект. Исследуется влияние числа каналов $n = \overline{1,3}$ на среднее время ожидания и ДФР времени пребывания в системе при фиксированном коэффициенте загрузки $\rho = 0.8$. Устанавливается, что при увеличении числа каналов, т. е. при дроблении производительности, среднее время ожидания уменьшается, а среднее время пребывания возрастает. Обучаемый подводится к выводу, что с учетом необходимости ремонта и профилактического обслуживания оптимален выбор $n = 2$.

В других вариантах той же работы исследуется изменение средних длительностей ожидания и пребывания в системе при одновременном k-кратном увеличении интенсивностей входящего потока и обслуживания. Обнаруживается, что при этом упомянутые средние уменьшаются в k раз — независимо от числа каналов и вида распределения длительности обслуживания.

В Руководстве к профессиональному МОСТу приводятся описания 9 лабораторных работ. В учебной версии их четыре. Их количество может быть легко увеличено. Имеется реальная возможность выдать всем обучаемым индивидуальные задания. В целях экономии дисплейного времени рекомендуется бригадная организация. С учетом времени набора предварительно подготовленной программы, отладки программы, оформления и защиты отчета работа укладывается в четырехчасовое занятие. В ходе проведения цикла лабораторных работ наблюдалось заметное повышение интереса обучаемых к дисциплине и улучшение понимания основных ее положений.

Литература

[1] *Аветисян Л. А. и др.* Пакет программ расчета характеристик сетей массового обслуживания. // Тезисы докладов III Всесоюзного совещания по распределенным автоматизированным системам массового обслуживания. М.: 1990. — С. 189–190.

[2] *Амосова Н. Н.* Введение в теорию массового обслуживания: Учебное пособие. — СПб.: Академия гражданской авиации, 2004. — 65 с.

[3] *Артамонов Г. Т.* Цифровая вычислительная машиа как система массового обслуживания // В сб. «Массовое обслуживание в системах передачи информации». — М.: Наука, 1969. — с. 59–81.

[4] *Ахиезер Н. И.* Классическая проблема моментов и некоторые вопросы анализа, связанные с нею. — М.: Физматгиз, 1961. — 310 с.

[5] *Балыбердин В. А.* Методы анализа мультипрограммных систем. — М.: Радио и связь, 1992. — 152 с.

[6] *Бартеньев О. В.* Современный Фортран. — М.: Диалог–МИФИ, 1998. — 397 с.

[7] *Башарин Г. П., Громов А. И.* Матричный метод нахождения стационарных распределений для некоторых нестандартных систем массового обслуживания // АиТ. — 1978. — № 1. — С. 29–38.

[8] *Башарин Г. П., Толмачев А. Л.* Некоторые результаты теории сетей массового обслуживания // В сб. [72], с. 52–64.

[9] *Башарин Г. П., Толмачев А. Л.* Теория сетей массового обслуживания и ее приложение к анализу информационно-вычислительных систем //В кн. «Итоги науки и техники. Теория вероятностей. Математическая статистика. Теоретическая кибернетика». — М.: ВИНИТИ, 1983. — т. 21. С. 3–119.

[10] *Башарин Г. П., Бочаров П. П., Коган Я. А.* Анализ очередей в вычислительных сетях. Теория и методы расчета. — М.: Наука, Физмат, 1989. — 336 с.

[11] *Безкоровайный М. М., Костогрызов А. И., Львов В. М.* Инструментально-моделирующий комплекс для оценки качества функционирования информационных систем «КОК»: Руководство системного аналитика. — М.: СИНТЕГ, 2000. — 116 с.

[12] *Беллман Р.* Введение в теорию матриц / Пер. с англ. — М.: Наука— Физмат, 1969. — 368 с.

[13] *Беляков В. Г., Митрофанов Ю. И., Ярославцев А. Ф.* Пакет прикладных программ для математического моделирования сетевых систем // Тезисы докладов 11-го Всесоюзного семинара по вычислительным сетям, ч. 3. — М.– Рига: 1986. — С. 145–150.

[14] *Берсенев Г. Б., Шкатов П. Н., Морозов А. А.* Построение и анализ моделей вычислительных систем аналитическими, имитационными и аналитико-имитационными методами // Сб. научных трудов Тульского политехнического ин-та. — Тула: 1981. — С. 3–17.

[15] *Богуславский Л. Б., Коган Я. А.* Асимптотический анализ производительности вычислительных структур большой размерности // АВТ. — 1983. — № 5. — С. 73–79.

[16] *Боровков А. А.* Асимптотические методы в теории массового обслуживания. — М.: Наука, Физмат, 1980. — 382 с.

[17] *Бочаров П. П.* О приближенной оценке времени пребывания в неэкспоненциальных сетях массового обслуживания // Системы массового обслуживания и информатика. — М.: Изд-во ун-та им. П. Лумумбы, 1987. — С. 21-30.

[18] *Бочаров П. П., Печинкин А. В.* Теория массового обслуживания: Учебник. — М.: Изд-во ун-та им. П. Лумумбы, 1995. — 529 с.

[19] *Брич З. С., Капилевич Д. В., Клецкова Н. А.* Фортран-77 для ПЭВМ ЕС: справочное издание. — М.: Финансы и статистика, 1991. 286 с.

[20] *Бронштейн О. И., Духовный И. М.* Модели приоритетного обслуживания в информационно-вычислительных системах. — М.: Наука, 1976. — 220 с.

[21] *Бутомо И. Д., Хижняк И. П., Дадиомов А. М.* PSP — пакет программ моделирования систем массового обслуживания // Проблемы системотехники и АСУ: межвузовский сборник. — Л.: СЗПИ, 1981. — вып.4. — С. 123–128.

[22] *Быкадоров А. В.* Многоканальная система с эрланговским входящим потоком и постоянным временем обслуживания // Изв. АН СССР, Техн. кибернетика. — 1972. — № 3. — С. 121–126.

[23] *Вентцель Е. С.* Исследование операций. — М.: Сов. радио, 1972. — 551 с.

[24] *Вишневский В. М., Белокриницкая Л. Б., Шеленков В. Л.* Пакет прикладных программ для исследования интегральных характеристик телеавтоматических систем массового обслуживания // М.: Сб. трудов ин-та проблем управления. — 1980. Вып. 22. — С. 38–42.

[25] *Вишневский В. М., Герасимов А. И.* Об одном подходе к исследованию вычислительных сетей // 6-я Всесоюзная школа-семинар по вычислительным сетям. /Научный совет по комплексной проблеме «Кибернетика» АН СССР. — М.: 1981. С. 22–30.

[26] *Вишневский В. М.* Принципы построения и реализации пакета программ анализа и синтеза сетей массового обслуживания (ПЕГАС) // Тезисы докладов 11-го Всесоюзного семинара по вычислительным сетям. — М.–Рига: 1986. — С. 64–67.

[27] *Вишневский В. М., Круглый З. Л.* Оптимизация замкнутых стохастических сетей // АиТ. — 1987. — № 2. — С. 41–53.

[28] *Вишневский В. М.* Теоретические основы проектирования компьютерных сетей. — М.: Техносфера, 2003. — 512 с.

[29] *Гнеденко Б. В.* Беседы о теории массового обслуживания. — М.: Знание, 1973. — 63 с.

[30] *Горбунов-Посадов М. М., Корягин Д. А., Мартынюк В. В.* Системное обеспечение пакетов прикладных программ / Под ред. А. А. Самарского. — М.: Наука, 1990. — 208 с.

[31] *Демидович Б. П., Марон И. А.* Основы вычислительной математики. — М.: Физматгиз, 1963. — 659 с.

[32] *Джейсуол Н. К.* Очереди с приоритетами / Пер. с англ. — М.: Мир, 1973. 279 с.

[33] *Емельянов А. А.* Имитационное моделирование в управлении рисками. СПб.: ГИЭА, 2000. — 375 с.

[34] *Ермаков С. М., Кривулин Н. К.* Элементы теории массового обслуживания: Учеб. пособие. — СПб.: Изд-во СПб Университета, 1998. — 86 с.

[35] *Ершов А. Т.* Введение в теорию случайных процессов и теорию массового обслуживания: Учеб. пособие. — М.: Гос. Ун-т управления, 2004. — 36 с.

[36] *Жожикашвили В. А., Вишневский В. М.* Сети массового обслуживания. Теория и применение к сетям ЭВМ. — М.: Радио и связь,1988. — 192 с.

[37] *Задорожный В. Н.* Аналитико-имитационные исследования систем и сетей массового обслуживания: Монография. — Омск: Изд-во ОмГТУ, 2010. 280 с.

[38] *Захаров Г. П.* Методы исследования систем передачи данных. — М.: Радио и связь, 1982. — 208 с.

[39] *Золотухина Л. А.* Теория массового обслуживания в приложении к задачам судостроения: Учеб. пособие. — Л.: Ленингр. кораблестроительный институт, 1989. — 106 с.

[40] *Ивлев В. В.* Определение параметров обобщенных многочленов Лагерра в задачах надежности // Изв. АН СССР, Техн. кибернетика. — 1981. — №6. — С. 68–72.

[41] *Ивницкий В. А.* Сети массового обслуживания и их применение в ЭВМ // Зарубежная радиоэлектроника. — 1977. — № 7. — С. 33–70.

[42] *Ивницкий В. А.* Теория сетей массового обслуживания. — М.: Физматлит, 2004.

[43] *Ивченко Г. И., Каштанов В. А., Коваленко И. Н.* Теория массового обслуживания. — М.: Высшая школа, 1982. — 256 с.

[44] *Икрамов Х. Д.*. Численное решение матричных уравнений. — М.: Наука, Физмат, 1984. — 192 с.

[45] *Ирвин Дж., Харль Д.*. Передача данных в сетях: инженерный подход / Пер. с англ. — СПб.: БХВ–Петербург, 2003. — 448 с.

[46] Информационный бюллетень №82. Фортран-77 для персональных ЭВМ. Методические рекомендации по переносу ПЛ-программ на ПЭВМ. — СПб.: ВИККА им. А. Ф. Можайского, 1993. — 27 с.

[47] *Кендалл М. Дж., Стьюарт А.* Теория распределений / Пер. с англ. — М.: Наука, 1966. — 587 с.

[48] *Кениг Д., Штойян Д.* Методы теории массового обслуживания / Пер. с нем. — М.: Радио и связь, 1981. — 127 с.

[49] *Клейнрок Л.* Теория массового обслуживания / Пер. с англ. — М.: Машиностроение, 1979. — 432 с.

[50] *Клейнрок Л.* Вычислительные системы с очередями / Пер. с англ. — М.: Мир, 1979. — 600 с.

[51] *Климов Г. П.* Стохастические системы обслуживания. — М.: Наука, 1966. 243 с.

[52] *Кокорин С. В., Рыжиков Ю. И.* Оптимизация параметров сетей массового обслуживания на основе комбинированного использования аналитических и имитационных моделей // Приборостроение. — 2010. — № 11. — С. 61–66.

[53] *Кокс Д. Р., Смит В. Л.* Теория восстановления /Пер. с англ. — М.: Сов. радио, 1967.

[54] Комплекс программ МОСТ по решению задач теории массового обслуживания. Руководство программиста. — МО СССР, 1986. — 113 с.

[55] *Конвей Р. В., Максвелл В. Л., Миллер Л. В.* Теория расписаний / Пер. с англ. — М.: Наука, 1975. — 359 с.

[56] *Корн Г., Корн Т.* Справочник по математике для научных работников и инженеров / Пер. с англ. — М.: Наука, 1973. — 832 с.

[57] *Кульгин М.* Теория очередей и расчет параметров сети // BYTE - Россия, ноябрь 1999. — С. 26–33.

[58] *Кульгин М.* Технологии корпоративных сетей: Энциклопедия. — СПб.: Питер, 2000. — 704 с.

[59] *Кульгин М.* Практика построения компьютерных сетей: Для профессионалов. — СПб.: Питер, 2001. — 320 с.

[60] *Ланс Дж.* Численные методы для быстродействующих вычислительных машин / Пер. с англ. — М.: Иностранная литература, 1962. — 208 с.

[61] *Липаев В. В., Яшков С. Ф.* Эффективность методов организации вычислительного процесса в АСУ. — М.: Статистика, 1975. — 255 с.

[62] *Литвинов М. Л.* Метод Кендалла и определение характеристик моделей обслуживания // Изв. АН СССР, Техн. кибернетика. — 1975. — № 6. — С. 74–79.

[63] *Лифшиц А. Л., Мальц Э. А.* Статистическое моделирование систем массового обслуживания. — М.: Сов. радио, 1978. — 247 с.

[64] *Ливщиц Б. С., Пшеничников А. П., Харкевич А. Д.* Теория телетрафика: Учебник. — М.: Связь, 1979. — 163 с.

[65] *Лысенкова В. Т.* Исследование многолинейных систем массового обслуживания с ограниченным накопителем и приоритетами // Автореферат канд. дисс. — М.: ИППИ. — 1973.

[66] *Мартин Дж.* Системный анализ передачи данных, т. 2 / Пер. с англ. — М.: Мир, 1975. — 431 с.

[67] Математика для экономистов. Т. 6. Теория массового обслуживания / В. П. Чернов, В. Б. Ивановский. — М.: ИНФРА-М, 2000. — 156 с.

[68] *Матвеев В. Ф., Ушаков В. Г.* Системы массового обслуживания: Учеб. пособие. — М.: МГУ, 1984. — 239 с.

[69] Материалы Всероссийских научно-практических конференций «Опыт практического применения языков и программных систем имитационного моделирования в промышленности и прикладных разработках». — СПб.: ЦНИИ технологии судостроения, 2003, 2005, 2007, 2009 гг.

[70] *Меткалф М., Рид Дж.* Описание языка программирования Фортран 90 / Пер. с англ. — М.: Мир, 1995. — 302 с.

[71] Методическое руководство по расчету систем массового обслуживания на ЕС ЭВМ // Инф. бюллетень № 18. — МО, 1980. — 55 с.

[72] Методы развития теории телетрафика. — М.: Наука, 1979. — 208 с.

[73] *Мину М.* Математическое программирование. Теория и алгоритмы / Пер. с франц. — М.: Наука, 1990. — 488 с.

[74] *Митрофанов Ю. И., Иванов А. Н.* КИМДС — комплекс процедур имитационного моделирования обобщенных дискретных систем // Программирование. — 1978. — № 5. — С. 74–83.

[75] *Митрофанов Ю. И.* Пакеты программ для аналитического и имитационного моделирования сетей вычислительных комплексов: Препринт. — М.: 1981. — 42 с.

[76] *Митрофанов Ю. И. и др.* Методы и программные средства аналитического моделирования сетевых систем: Препринт. — М.: 1982. — 67 с.

[77] *Митрофанов Ю. И., Зобенков В. А., Чернавин Н. С.* Система СИМПЛ имитационного моделирования дискретных систем. — Саратов: СГУ, 1989. — 72 с.

[78] *Митрофанов Ю. И., Табояков А. А.* Пакет прикладных программ АМОС для аналитического моделирования дискретных систем. — Саратов: СГУ, 1989. — 32 с.

[79] *Мова В. В., Пономаренко Л. А., Калиновский А. М.* Организация приоритетного обслуживания в АСУ. — Киев: Техніка, 1977. — 160 с.

[80] Научная записка о возможности использования методов теории массового обслуживания в работе Госплана СССР. — М.: НИИ планирования и нормативов, 1967. — 45 с.

[81] *Никифоров А. Ф., Уваров В. Б.* Основы теории специальных функций. М.: Наука, 1974. — 303 с.

[82] Основы теории вычислительных систем / Под ред. С. А. Майорова. — М.: Высшая школа, 1978. — 408 с.

[83] Пакет прикладных программ МОСТ для расчета стационарных режимов в системах массового обслуживания. — Эстонское НПИ ВТИ. — 1988.

[84] *Порховник В. А.* Аналитическая компонента пакета прикладных программ для вычисления значений вероятностных характеристик систем массового обслуживания // Автореферат канд. дисс. — М.: МЭИ, 1976.

[85] *Поттгоф Г.* Теория массового обслуживания / Пер. с нем. — М.: Транспорт, 1979. — 144 с.

[86] Приоритетные системы обслуживания. — М.: МГУ, 1973. — 447 с.

[87] *Риордан Дж.* Вероятностные системы обслуживания / Пер. с англ. — М.: Связь, 1966. — 164 с.

[88] *Романовский В.* Математическая статистика. — Ташкент, 1961–1963.

[89] *Рыжиков Ю. И.* Управление запасами. — М.: Наука, Физмат, 1969. 344 с.

[90] *Рыжиков Ю. И.* О методах расчета оптимального набора запасных частей // Экономика и математические методы. — 1971. — т. VII. — вып. 2. С. 207–212.

[91] *Рыжиков Ю. И.* Комплекс программ для расчета систем массового обслуживания повышенной сложности // Программирование. — 1978. № 4. С. 87–91.

[92] *Рыжиков Ю. И.* Машинные методы расчета систем массового обслуживания. — Л.: ВИКИ им. А. Ф. Можайского, 1979. — 177 с.

[93] *Рыжиков Ю. И.* Алгоритм расчета многоканальной системы с эрланговским обслуживанием // АиТ. — 1980. — № 5. — С. 30–37.

[94] *Рыжиков Ю. И., Хомоненко А. Д.* Итеративный метод расчета многоканальных систем с произвольным распределением времени обслуживания // Проблемы управления и теории информации. — 1980. № 3. — С. 32–38.

[95] *Рыжиков Ю. И.* К расчету неоднородных вычислительных систем // АВТ. — 1981. — № 6. — С. 81–86.

[96] *Рыжиков Ю. И., Никифоров Г. К.* Итерационный метод расчета обслуживания неоднородного потока в многоканальной марковской системе // Кибернетика. — 1983. — № 2. — С. 109–113.

[97] *Рыжиков Ю. И.* Метод расчета распределения времени реакции интерактивных систем // АВТ. — 1985. — № 3. — С. 65–69.

[98] *Рыжиков Ю. И.* Рекуррентный расчет многоканальных систем обслуживания с неограниченной очередью // АиТ. — 1985. — № 6. С. 88–93.

[99] *Рыжиков Ю. И.* Тестирование функционально избыточных пакетов программ // Программирование. —1986. — № 1. — С. 22–29.

[100] *Рыжиков Ю. И.* Тестирование комплекса программ по расчету систем массового обслуживания // УСиМ. — 1986. — № 2. — С. 84–89.

[101] *Рыжиков Ю. И.* К расчету замкнутых сетей обслуживания // АВТ. — 1987. — № 6. — С. 29–31.

[102] *Рыжиков Ю. И., Демиденко Ю. А.* Определение моментов распределения времени пребывания заявки в вычислительной сети // АВТ. — 1988. № 1. — С. 24–27.

[103] *Рыжиков Ю. И., Хомоненко А. Д.* Расчет разомкнутых немарковских сетей с преобразованием потоков // АВТ. — 1989. — № 3. — С. 15–24.

[104] *Рыжиков Ю. И.* Имитационное моделирование систем массового обслуживания. — Л.: ВИККИ им. А. Ф. Можайского. — 1991. — 111 с.

[105] *Рыжиков Ю. И.* Численная реализация преобразования Лапласа и его обращения в задачах массового обслуживания // Тезисы докладов 7-й Белорусской школы-семинара по теории массового обслуживания. — Минск, 1991. — С. 111-112.

[106] *Рыжиков Ю. И.* Три метода расчета временных характеристик разомкнутых систем массового обслуживания // АиТ. — 1993. — № 3. — С. 127–133.

[107] *Рыжиков Ю. И.* Расчет систем фазового типа методом матрично-геометрической прогрессии // В сб. «Сети связи и сети ЭВМ. Анализ и применение». — Минск, 1992. — С. 95-96.

[108] *Рыжиков Ю. И.* Расчет циклической системы с произвольным обслуживанием на основе усредненной вероятности возврата // В сб. «Системные проблемы связи и управления». — МО РФ. — 1994. С. 35.

[109] *Рыжиков Ю. И.* Пакет прикладных программ для анализа систем и сетей обслуживания // Тезисы докладов 47-й НТК СПб Гос. Ун-та телекоммуникаций. — СПб, 1994. — С. 80.

[110] *Рыжиков Ю. И.* Руководство по расчету систем с очередями. — СПб: ВИККА им. А. Ф. Можайского, 1995. — 73 с.

[111] *Рыжиков Ю. И.* Алгоритм перевода ПЛ-программ на Фортран-77 для ПЭВМ. // Сб. алгоритмов и программ №13. — МО РФ, 1995.

[112] *Рыжиков Ю. И.* Пакет прикладных программ для ПЭВМ по расчету систем и сетей обслуживания // Тезисы докладов 2-й НТК училища радиоэлектроники ПВО. — Пушкин, 1995. — С. 88.

[113] *Рыжиков Ю. И.* Программирование на Фортране PowerStation для инженеров. — СПб.: КОРОНА принт, 1999. — 160 с.

[114] *Рыжиков Ю. И.* Решение научно-технических задач на персональном компьютере. — СПб.: КОРОНА принт, 2000. — 272 с.

[115] *Рыжиков Ю. И.* Теория очередей и управление запасами. — СПб.: Питер, 2001. — 384 с.

[116] *Рыжиков Ю. И.* Современный Фортран. — СПб.: КОРОНА принт, 2004. 288 с.

[117] *Рыжиков Ю. И.* Имитационное моделирование. Теория и технологии. СПб.: КОРОНА принт, 2004. — 380 с.

[118] *Рыжиков Ю. И.* Вычислительные методы. — СПб.: БХВ-Петербург, 2007. — 400 с.

[119] *Рыжиков Ю. И.* Оценка системы моделирования GPSS World // Информационно-управляющие системы. — 2003. — № 2–3. — С. 30–38.

[120] *Рыжиков Ю. И.* ИММОД-2003 — аналитический обзор. // Информационно-управляющие системы. — 2004. — № 6(13). — С. 45–56.

[121] *Рыжиков Ю. И.* Расчет многоуровневой системы очередей. // Информационно-управляющие системы. — 2005. — № 5(18). — С. 31–34.

[122] *Рыжиков Ю. И.* Полный расчет системы обслуживания с распределениями Кокса // Информационно-управляющие системы. — 2006. — № 2(21). — С. 38–46.

[123] *Рыжиков Ю. И.* Средние времена ожидания и пребывания в многоканальных приоритетных системах // Информационно-управляющие системы. — 2006. — № 6(25). — С. 43–49.

[124] *Рыжиков Ю. И.* Расчет систем обслуживания с групповым поступлением заявок // Информационно-управляющие системы. — 2007. — № 2(27). С. 39–49.

[125] *Рыжиков Ю. И.* Расчет систем со случайным выбором на обслуживание // Информационно-управляющие системы. — 2007. — № 3(28). — С. 56–59.

[126] *Рыжиков Ю. И.* Компьютерное моделирование систем с очередями: курс лекций. — СПб.: ВКА им. А. Ф. Можайского, 2007. — 164 с.

[127] *Рыжиков Ю. И.* Имитационное моделирование: курс лекций. — СПб.: ВКА им. А. Ф. Можайского, 2007. — 125 с.

[128] *Рыжиков Ю. И.* Руководство по расчету систем с очередями на базе пакета МОСТ/FPS1: учебно-методическое пособие. — СПб.: ВКА им. А. Ф. Можайского, 2007. — 92 с.

[129] *Рыжиков Ю. И.* Пакет программ для расчета систем с очередями и его тестирование // Труды СПИИРАН. Вып. 7. — СПб.: Наука, 2008. — С. 265–284.

[130] *Рыжиков Ю. И., Уланов А. В.* Опыт расчета сложных систем массового обслуживания // Информационно-управляющие системы. — 2009. — № 32(39). — С. 56–62.

[131] *Рыжиков Ю. И.* Многоканальные системы с динамическим приоритетом // Проблемы информатики и управления (Киев). — 2009. — № 4. — С. 106-110.

[132] *Саати Т. Л.* Элементы теории массового обслуживания и ее приложения / Пер. с англ. — М.: Сов.радио, 1965. — 510 с.

[133] *Сайкин А. И.* Разработка и исследование методов расчета характеристик вычислительных систем на основе стохастических сетевых моделей // Автореферат канд. дисс. — Л.: ЛИТМО, 1979. — 19 с.

[134] *Севастьянов Б. А.* Эргодическая теорема для марковских процессов и ее приложение к телефонным системам с отказами // Теория вероятностей и ее применения. — 1957. — т. 2, вып.1. — С. 106–116.

[135] Система виртуальных машин для ЕС ЭВМ: Справочник / Под ред. Э. В. Ковалевича. — М.: Финансы и статистика, 1985. — 360 с.

[136] *Сотский Н. М., Чуркин Е. А.* Стационарное распределение длины очереди в многоканальной системе массового обслуживания с приоритетами // АиТ. — 1985. — № 1. — С. 69–76.

[137] *Сочнев А. В.* Пакет прикладных программ анализа информационно-вычислительных систем // Изв. ЛЭТИ им. В. И. Ульянова-Ленина, вып. 288. — Л.: 1981. — С. 41.

[138] *Степанов С. Н.* Численные методы расчета систем с повторными вызовами. — М.: Наука, 1983. — 229 с.

[139] *Таранцев А. А.* Инженерные методы теории массового обслуживания. — СПб.: Наука, 2007. — 175 с.

[140] *Тихоненко О. М.* Модели массового обслуживания в системах обработки информации. — Минск: «Университетское», 1990. — 191 с.

[141] *Томашевский В. Л.* Многоканальные приоритетные системы массового обслуживания // Автореферат канд. дисс. — М.: МГУ, 1986. — 14 с.

[142] *Тыугу Э. Х.* Концептуальное программирование. — М.: Наука, 1984. — 255 с.

[143] *Уолрэнд Дж.* Введение в теорию сетей массового обслуживания / Пер. с англ. — М.: Мир, 1993. — 335 с.

[144] *Фролов Г. Д., Олюнин В. Ю.* Практический курс программирования на языке ПЛ/1. — М.: Наука, 1983. — 384 с.

[145] *Хомоненко А. Д.* Вероятностный анализ приоритетного обслуживания с прерываниями в многопроцессорных системах // АВТ. — 1990. — № 2. — С. 55–61.

[146] *Хомоненко А. Д.* Распределение времени ожидания в системах массового обслуживания типа $GI_q/H_k/n/R \leq \infty$ // АиТ. — 1990. № 8. — С. 91–98.

[147] *Хомоненко А. Д.* Численные методы анализа систем и сетей массового обслуживания. — МО СССР, 1991. — 195 с.

[148] *Шнепс М. А.* Численные методы теории телетрафика. — М.: Связь, 1974. 232 с.

[149] *Штойян Д.* Качественные свойства и оценки стохастических моделей / Пер. с нем. — М.: Мир, 1979. — 268 с.

[150] *Яблокова Т. Л.* Пакет прикладных программ «Аналитические модели систем массового обслуживания». — Киев: КПИ, 1979. — 264 с.

[151] *Янбых Г. Ф., Столяров Б. А.* Оптимизация информационно-вычислительных сетей. — М.: Радио и связь, 1987. — 232 с.

[152] *Янке Е., Эмде Ф., Леш Ф.* Специальные функции. — М.: Наука, 1968. 344 с.

[153] *Яшков С. Ф.* Анализ очередей в ЭВМ. — М.: Радио и связь, 1989. 216 с.

[154] Advances in Queueing Theory and Network Applications / Yue W., Takahashi Y., Takagi H. (eds). — Springer, 2009. — 309 pp.

[155] *Akar N., Arican E.* A Numerically Efficient Method for the MAP/D/1/K Queue via Rational Approximation // Queueing Systems: Theory and Applications (QUESTA). — v. 22 (1996). — P. 97–120.

[156] *Allen A. O.* Queuing Models of Computer Systems // Computer. — 1980. — v.13, no. 4. — P. 13–24.

[157] Applied Probability — Computer Science: the Interface. — v.2. — Boston: Birkhäuser, 1982.

[158] *Artalejo J. R.* G-networks: A Versatile Approach for Work Removal in Queuing Networks // European J. of Operational Research. — 2000, v. 126. P. 233–249.

[159] *Balbo G., Serazzi G.* Multi-Class Product-Form Closed Queuing Networks under Heavy Loading Conditions // in [225, P.283–295].

[160] *Balsamo S., Iazzeola G.* Product-Form Synthesis of Queuing Networks // IEEE Trans. on Software Engineering. — 1985. — v. SE-11, no. 2. P. 194–199.

[161] *Bard Y.* An Analytic Model of the VM/370 System // IBM J. of R&D. — 1978. — v. 22. — P. 498–508.

[162] *Bard Y.* Some Extensions to Multiclass Queuing Network Analysis // IBM J. of R&D. — 1978. — v. 22. — P. 51–62.

[163] *Baskett F., Chandy K. M., Muntz R. R., Palacios J. G.* Open, Closed, and Mixed Networks of Queuing with Different Classes of Customers // J. of the ACM. — 1975. — v. 22, no. 2. — P. 248–260.

[164] *Bhat U. N.* An Introduction to Queueing Theory Modeling and Analysis in Applications. — Boston etc.: Birkhauser, 2008. — 272 pp.

[165] *Bolch S., Greiner S., Meer de,H., Trivedi K. S.* Queueing Networks and Markov Chains Modeling and Performance Evaluation with Computer Science Application. — N. Y. etc, Wiley & Sons, 1998. — 726 pp.

[166] *Booyens M., Kristzinger P. S., Krzesinski A. E. et al.* SNAP: an Analytic Multiclass Queuing Network Analyzer // in [226, P. 67–80].

[167] *Breuer L.* Spatial Queues: Thesis. — Univ. of Trier. — 128 pp.

[168] *Breuer L., Baum D.* An Introduction to Queueing Theory and Matrix-Analytic Methods. — Springer, 2005. — 278 pp.

[169] *Bruel S. C., Balbo G., Ghanta S., Afshari P. V.* A Mean Value Analysis Based Package for the Solution of Product-Form Queuing Network Models // in [226, P. 99–108].

[170] *Brumelle S. L.* A Generalization of $L = \lambda W$ to Moments of Queue Length and Waiting Times // Operat. Res. — 1972. — v. 20, no. 6. — P. 1127–1136.

[171] *Burke P. J.* The Output of a Queuing System. // Operat. Res. — 1956. no. 4. — P. 699–704.

[172] *Burke P. J.* Output Processes and Tandem Queues // Proc. of the Symp. on Computer Communications. — N. Y., 1972. — P. 419.

[173] *Bux W.* Token-Ring Local-Area Networks and their Performance // Proc. of the IEEE. — 1989. — v. 77, no. 2. — P. 238–256.

[174] *Buzen J. P.* Computational Algorithms for Closed Queuing Networks with Exponential Servers // CACM. — 1973. — v. 16, no. 9. — P. 527–531.

[175] *Buzen J. P.* Fundamental Operational Laws of Computer System Performance // Acta Informatica. — 1976. — v. 7, no. 2. — P. 167–182.

[176] *Carter G. M., Cooper R. B.* Queues with Service in Random Order // Operat. Res., v. 20, 1972, no. 2. — P. 389–407.

[177] *Chandy K. M., Herzog U., Woo L.* Approximate Analysis of Queuing Networks // IBM J. of R&D. — 1975. — v. 19, no. 1. — P. 43–49.

[178] *Chandy K. M., Sauer C. H.* Approximate Methods for Analyzing Queuing Network Models of Computer Systems // Computing Surveys. 1978. — v. 10, no. 3. — P. 281–318.

[179] *Chandy K. M., Sauer C. H.* Computational Algorithms for Product-Form Queuing Networks // CACM. — 1980. — v. 23, no. 10. — P. 573–583.

[180] *Chandy K. M., Sauer C. H.* Approximate Solution of Queuing Models // Computer. — 1980. — v. 13, no. 4. — P. 25–32.

[181] *Chandy K.M. , Martin J.* Characterization of Product Form Queueing Networks // J. of ACM. — 1983. — v. 30, no. 2. — P. 286—289.

[182] *Chatelin F.* Iterative Aggregation-Disaggregation Methods // in [219, P. 199–208].

[183] *Coffman E. G., Denning P. J.* Operating Systems Theory. — Englewood Cliffs, NJ: Prentice-Hall, 1973. — 331 pp.

[184] Computer Networking and Performance Evaluation / T. Hasegava, H. Takagi, Y. Takahashi (eds). — Amsterdam: Elsevier Science Publ., 1986.

[185] Computer Networks and Simulation / S. Shoemaker (ed). — Amsterdam: North-Holland Publ. Co, 1986. — 412 pp.

[186] *Conway A. E., Georganas N. D.* RECAL — A New Efficient Algorithm for the Exact Analysis of Multi-Chain Queuing Networks // J. of ACM. — 1986. v. 33. — P. 768–791.

[187] *Courtois P. J.* Decomposability. Queuing and Computer System Application. — N.Y.: Academic Press, 1977. — 201 pp.

[188] *Cox D. R.* A Use of Complex Probabilities in the Theory of Stochastic Processes // Proc. of the Cambridge Phil. Soc. — 1955. — P. 313.

[189] *Crommelin C. D.* Delay Probability Formulae when the Holding Times are Constant // Post Office Electrical Engineer's J. — 1932. — P. 41–50.

[190] *Daigle J. N.* Queuing Theory with Applications to Packet Telecommunication. — Boston: Springer, 2005. — 326 pp.

[191] *Dantzig D.,van.* Sur la Méthode des Fonctions Génératrices // Colloques Internationaux du CNRS. — 1948. — v. 13. — P. 29–45.

[192] *Devroye L.* Non-Uniform Random Variate Generation. — N.Y.: Springer Verlag, 1986. — 843 pp.

[193] *Evans R. D.* Geometric Distribution in Some Two Dimensional Queuing Systems // Operat. Res. — 1967. — v. 15, no. 5. — P. 830–846.

[194] *Gail H. R., Hantler S. L., Taylor B. A.* Analysis of a Non-Preemptive Priority Multiserver Queue // Advances in Applied Prob. — 1988. — v. 20. P. 852.

[195] *Gall Le, P.* Les Systèmes avec ou sans Attente et Prosessus, v. 1. — Paris: Dunod. — 1962.

[196] *Ganesh A., O'Konnell N., Wichik D.* Big Queues. — Berlin: Springer, 2004. — 260 pp.

[197] *Gelenbe E., Mitrani I.* Analysis and Synthesis of Computer Systems. — N.Y.–London: Academic Press, 1980. — 233 pp.

[198] *Gerasimov A. I.* Analysis of Queuing Networks by Polynomial Approximation // Problems of Control and Information Theory. — 1983. v. 12, no. 3. — P. 219–228.

[199] *Gordon W. J., Newell G. F.* Closed Queuing Systems with Exponential Servers // Operat. Res. — 1967. — v. 15, no. 2. — P. 254–265.

[200] *Grassman W. K., Zhao Y. Q.* Heterogeneous Multiserver Queues wityh General Input // INFOR, 1997, v. 35, n0. 3, p. 208–224.

[201] *Harrison P. G.* An Enhanced Approximation by Pair-Wise Analysis of Servers for Time Delay Distributions in Queuing Networks // IEEE Trans. on Computers. — 1986. v. C-35, no. 1. — P. 54–61.

[202] *Haverkort B. R.* Approximate analysis of Networks of PH/PH/1/K Queues: Theory and Tool Support.

[203] *Herzog U., Hoffman W.* Synchronization Problems in Hierarchically Organized Multiprocessor Computer Systems // in [237, P. 29–48].

[204] *Kasten H., Runnenburg J. T.* Priority in Waiting Line Problems. Amsterdam: Mathematishe Centrum, Dec.1956.

[205] *Keilson J., Nunn W. R.* Laguerre Transformation as a Tool for the Numerical Solution of Integral Equations of Convolution Type // Appl. Math. Comput. — 1979. — v. 5. — P. 313–359.

[206] *Kelly F. P.* Networks of Queues with Different Types // J. Appl. Prob. 1975. v. 12, no. 3. — P. 542–554.

[207] *Kingman J. F. C.* On Queues in which Custemers Are Served in Random Order // Proc. Cambr. Phil. Soc. — 1962, v. 58, no. 1. — P. 79–91.

[208] *Kino I., Morita S.* PERFORMS — a Support System for Computer Performance Evaluation // in [226, P. 119–138].

[209] *Kino I.* A Computational Algorithm for Mixed Queuing Networks // NEC R&D. — 1984. — no. 73. — P. 106–112.

[210] *Kraemer W., Langenbach-Beltz M.* Approximate Formulae for General Single-Server Systems with Single and Batch Arrivals // Angewandte Informatic. — 1978. — P. 396–402.

[211] *Krakowski M.* Conservation Methods in Queuing Theory // Revue Française d'Automatique, Informatique et Récherhe Opérationnelle. 1973. — V–1. — P. 63–84.

[212] *Kuehn P. J.* Approximate Analysis of General Queuing Networks by Decomposition // IEEE Trans. on Communications. — 1979. — v. COM-27, no. 1. — P. 113–126.

[213] *Lam S., Lien Y. L.* A Tree Convolution Algorithm for the Solution of Queuing Networks // CACM. — 1983. — v. 26, no. 3. — P. 203–215.

[214] *Latouche G., Ramaswami V.* A logarithmic Reduction Algorithm for Quadi-Birth-and-Death process // J. Appl. Prob. — 1993. — v. 30. — P. 650–674.

[215] *Lazowska E. D., Addison C. A.* Selecting Parameter Values for Servers of the Phase Type // in [237, P. 407–420].

[216] *Lee A. M.* Applied Queuing Theory. — London: McMillan, 1966. — 244 pp.

[217] *Maaløe E.* Approximation Formulae for Estimation of Waiting Time in Multiple-Channel Queuing Systems // Mgmt. Sci. — 1973. — v. 19. — P. 703–710.

[218] *Marie R.* Méthodes Itératives de Résolution de Modèles Mathématiques de Systèmes Informatiques // Revue Française d'Automatique, Informatique et Récherche Opérationnelle. Informatique. — 1978. v. 12, no. 2. — P. 107–122.

[219] Mathematic Computer Performance and Reliability. // Proc. of the International Workshop, Piza, 1983. — Amsterdam: North-Holland Publ. Co, 1984. — 429 pp.

[220] *Matloff N. S.* Probability Modeling and Computer Simulation: An Integrated Introduction with Applications to Engineering and Computer Science. — Boston: PWS–Kent, 1988. — 358 pp.

[221] *Medhi J.* Stochastic Models in Queueing Theory. — Elsevier, 2003. — 450 pp.

[222] *Merle D., Potier D., Veran M.* A Tool for Computer System Performance Analysis // in [238, P. 195–213].

[223] *Miller D. R.* Steady-State Algorithmic Analysis of M/M/c Two Priority Queues with Heterogeneous Rates // in [157, P. 207–222].

[224] *Mitrani I.* Fixed-Point Approximations for Distributed Systems // in [219, P. 245–258].

[225] Modelling Technique and Performance Evaluations // Proc. of the International Workshop /S. Fdida, G. Pujolle (eds). — Amsterdam: North-Holland Publ. Co., 1987. — 340 pp.

[226] Modelling Technique and Tools for Performance Analysis'85. // Proc. of the Internat. Conf. — Amsterdam: North-Holland Publ. Co., 1986. — 365 pp.

[227] *Morse P. M.* Queues, Iinventories, and Maintenance. — N. Y.: Wiley, 1958. 202 pp.

[228] *Mueller B.* NUMAS: a Tool for the Numerical Modelling of Computer Systems // in [226, P. 141–154].

[229] *Nelson R.* Matrix-Geometric Solutions in Markov Models. A Mathematical Tutorial // Yorktown Heights: IBM Research Division, 1991. — P. 1–25.

[230] *Neuman D. P., Sobel M. J.* Stochastic Models in Operations Research. — N. Y.: McGraw-Hill, 1982. — 548 pp.

[231] *Neuse D., Chandy K. M.* HAM — Heuristic Aggregation Method // Perform. Eval. Review. — 1983. — v. 11, no. 4. — P. 195–212.

[232] *Neuts M. F.* Matrix-Geometric Solutions in Stochastic Models: an Algorithmic Approach. — Baltimore: J. Hopkins Univ. press, 1981.

[233] *Neuts M. F.* Matrix-Analytic Methods in Queuing Theory // Eur. J. of Opns. Res. — 1984. — v. 15. — P. 2–12.

[234] *Newell G. F.* Apprroximate Stochastic Behavior of n-Service Systems with Large n. — Berlin etc.: Springer, 1973. — 118 pp.

[235] *Palm C.* Research on Telephone Traffic Carried Probabilities of a GI/G/c Queuing System in a Generalby Full Availability Groups // Tele, v. 1. — P. 107.

[236] *Papoulis A.* A New Method of Inversion of the Laplace Transform // Quart. of Appl. Math. — 1957. — v. 14. — P. 405–414.

[237] Performance of Computer Systems / M. Arato, A. Butrimenko, E. Gelenbe (eds). — Amsterdam: North-Holland Publ. Co, 1979. — 565 pp.

[238] Performance of Computer Installations. Proc. of the Internat. Conf. / D. Ferrari (ed). — Amsterdam: North-Holland Publ. Co, 1978.

[239] *Potier D., Veran M.* The Markovian Solver of QNAP2 and Examples // in [184, P. 259–279].

[240] *Potier D.* The Modelling Package QNAP2 and Applications to Computer Networks Simulation // in [185, P. 235–265].

[241] *Ramakrishnan K. G., Mitra D.* An Overview of PANACEA, a Software Package for Analysing Markovian Queuing Networks // Bell System Techn. J. 1982. — v. 61. — P. 2849–2872.

[242] *Ramaswami V., Lucantoni D. M.* Algorithms for the Multi-Server Queue with Phase-Type Service // Commun. Statist. — Stochastic Models. — 1985. v. 1, no. 3. — P. 393–417.

[243] *Reiser M., Kobayashi H.* Queuing Networks with Multiple Closed Chains. Theory and Computational Algorithms // IBM J. of R&D. — 1975. — v. 19. P. 283–294.

[244] *Reiser M.* Mean Value Analysis of Queuing Networks, a New Look at an Old Problem // in [237, P. 63–77].

[245] *Reiser M., Lavenberg S. S.* Mean Value Analysis of Closed Multichain Queuing Networks // J. of ACM. — 1980. — v. 27, no. 2. — P. 313–322.

[246] *Robertazzi T. G.* Computer Networks and Systems: Queueing Theory and Performance Evaluation. — N. Y. etc: Springer, 1990. — 306 pp.

[247] *Roode J. D.* Multiclass Operational Analysis of Queuing Networks // in [237, P. 339–352].

[248] *Rosenlund S. I.* The Random Order Service $G/M/m$ Queue // Naval Res. Logistics, v. 27, no. 2. — P. 207–215.

[249] *Sauer C. H., Reiser M., McNair E. A.* RESQ — a Package for Solution of Generalized Queuing Network // AFIPS Conf. Proc, v. 46. Montwale, N. J., 1977. — P. 977–986.

[250] *Sauer C. H.* Approximate Solution of Queuing Networks with Simultaneous Resource Possession // IBM J. of R&D. — 1986. — v. 25, no. 6. — P. 894–903.

[251] *Sauer C. H., Chandy K.M.* Computer Systems. Performance Evaluation. Englewood Cliffs, NJ: Prentice–Hall, 1981. — 352 pp.

[252] *Sauer C. H., McNair E. A.* The Evolution of the Research Queuing Package // in [226, P. 5–24].

[253] *Schwetman H.* Some Computational Aspects of Queuing Network Models // in [157, P. 135–155].

[254] *Schweitzer P. J.* Aggregation Methods for Large Markov Chains // in [219, P. 275–286].

[255] *Seelen L. P.* An Algorithm for Ph/Ph/c Queues // Eur. J. of Operational Research. — 1986. — v. 23. — P. 118–127.

[256] *Serfozo R.* Basics of Applied Stochastic Processes.— Berlin: Springer, 2009. 443 pp.

[257] *Shin Y. W., Moon D. H.* Sensitivity and Approximation of M/G/c Queue: Numerical Experiments // 8-th Internat. Symp. on Operations Research and Its Applications. — China, 2009. — P. 140–147.

[258] *Shortle J. F., Mark B. L., Gross D.* Reduction of Closed Queueing Networks for Efficient Simulation // ACM Trans. on Modeling and Computer Simulation. v. 19, 2009, no. 3, article 10.

[259] *Smith J. H. A., de.* A Numerical Solution for the Multi-Server Queue with Hyperexponential Service Times // Opns. Res. Letters. — 1983. — v.2. — P. 217–224.

[260] *Souza e Silva E., Lavenberg S. S., Mitrani R. R.* A Perspective on Iterative Methods for the Approximate Analysis of Closed Queuing Networks // in [219, P. 225–244].

[261] *Souza e Silva E., Lavenberg S. S., Muntz R. R.* A Clustering Approximation Technique for Queuing Network Models with a Large Number of Chains // IEEE Trans. on Computer. — 1986. — v. C–35. — P. 419.

[262] *Sumita U., Keilson J.* Waiting-Time Distribution Response to Traffic Surges via Laguerre Transform // in [157, P. 109–130].

[263] *Suri R.* Robustness of Queuing Network Formulas // J. of ACM. — 1983. v. 30. — P. 584–594.

[264] *Tàkacs L.* Introduction to the Theory of Queues. — N.Y.: Oxford Univ. Press, 1962.

[265] *Takahashi Y., Takami Y.* A Numerical Method for the Steady-State Probabilities of a GI/G/c Queuing System in a General Class // J. of the Operat. Res. Soc. of Japan. 1976. — v. 19, no. 2. — P. 147–157.

[266] *Takahashi Y.* Asymptotic Exponentiality of the Tail of the Waiting Time Distribution in a Ph/Ph/c Queue // Adv. in Applied Probability. — 1981. — v. 13. — P. 619–630.

[267] *Thareja A. K., Buzen J. P., Agraval S. C.* BEST/1-SNA: a Software Tool for Modelling and Analysis of IBM SNA Networks. // in [226, P. 81–98].

[268] *Tian N., Li Q.-L., Cao J.* Conditional Stochastic Decompositions in the M/M/c Queue with Server Vacations // Commun. Statist. — Stochastic Models, 1999, v. 15(2), P. 367–377.

[269] *Tucci S., McNair E. A.* Implementation of Mean Value Analysis for Open, Closed and Mixed Queuing Networks // Computer Performance. 1982. — v. 3. — P. 233–329.

[270] *Vaulot A. E.* Extension des Formules d'Erlang au Cas où les Durées des Conversations Suivent une Loi Quelconque // Revue Gen. Élec. — 1927. — v. 2. — P. 1164–1171.

[271] *Veran M., Potier D.* QNAP2: a Portable Environment for Queuing Systems Modelling // in [184, P. 25-66].

[272] *Wallace W. L., Rosenberg R. S.* Markovian Models and Numerical Analysis of Computer System Behaviour
// AFIPS Joint Computer Conf. Proc. — 1966. — P. 141–148.

[273] *Wallace W. L.* Algebraic Techniques for Numerical Solution of Queuing Networks // in Math. Methods of Queuing Theory. Proc. of a Conference at Western Michigan Univ. — Berlin–N.Y.: Springer, 1974. — P. 295–306.

[274] *Walstra R. J.* Fixed Point Approximations in Models of Networks of Queues // in [184, P. 229–244].

[275] *Whitt W.* The Queuing Network Analyser // Bell Systems Techn. J. — 1983. v. 62. — P. 2779–2815.

[276] *Willig A.* A Short Introduction to Queueing Theory // Techn. report. — Berlin: Tecn. univ, 1999. — 42 pp.

[277] *Whishart D. M. G.* A Queuing System with χ^2 Service Time Distribution // Ann. Math. Statist. — 1956. — v. 27. — P. 768–779.