В. Я. РОЗЕНБЕРГ, А. И. ПРОХОРОВ

Что такое теория массового обслуживания

„Советское радио"

В. Я. РОЗЕНБЕРГ, А. И. ПРОХОРОВ

ЧТО ТАКОЕ ТЕОРИЯ МАССОВОГО ОБСЛУЖИВАНИЯ

ИЗДАТЕЛЬСТВО „СОВЕТСКОЕ РАДИО"

МОСКВА — 1962

В книге изложены основные идеи и некоторые методы теории массового обслуживания. Книга снабжена значительным количеством разнообразных примеров, иллюстрирующих применение аппарата теории.

По изложению книга доступна, в первой своей части, широкому кругу читателей. В основном она предназначена для инженерно-технических работников и студентов старших курсов математических и технических специальностей, которые найдут в ней как некоторые пути подхода к задачам массового обслуживания, так и примеры решения таких задач, доведенные до численного решения.

ПРЕДИСЛОВИЕ РЕДАКТОРА

Идеи и методы теории массового обслуживания за последнее время приобретают весьма широкое распространение во многих прикладных областях. Характерно то, что круг практических задач, решаемых методами теории массового обслуживания, непрерывно расширяется. Если ранее в качестве основных областей приложения этой теории указывались задачи телефонии, бытового обслуживания, торговой сети, здравоохранения и так далее, то теперь она находит применение при исследовании динамики функционирования сложных систем автоматического управления, исследовании технологических процессов крупных промышленных предприятий, снабженных средствами комплексной автоматизации и механизации, в области организации и планирования производства и других областях народного хозяйства.

В книге В. Я. Розенберга и А. И. Прохорова «Что такое теория массового обслуживания» рассматриваются основные задачи теории массового обслуживания и приводятся иллюстративные примеры, помогающие усвоить практическое применение излагаемых результатов.

Авторы ставили перед собой две четко разграниченные задачи. Первая из них — это ознакомление широкого круга инженеров с элементарным аппаратом теории массового обслуживания и наиболее распространенными способами его практического применения.

С этой целью в книге изложены формулировки основных задач, вывод соответствующих дифференциальных уравнений и получение главнейших формул как результат решения этих уравнений, а также примеры, иллюстрирующие использование формул для решения прикладных задач, и краткий анализ численных результатов. В конце каждого параграфа приведены сводки основ-

ных формул. Это позволяет читателям, интересующимся только практическими применениями аппарата теории массового обслуживания, переходить сразу от формулировки задач к окончательным результатам, опуская громоздкие выкладки и трудные места в доказательствах теорем.

Вторая, не менее важная, задача книги — на элементарных примерах показать наиболее существенные стороны постановки прикладных задач, связанных с массовым обслуживанием, проиллюстрировать принципы формализации и математического описания таких процессов массового обслуживания, для решения которых, быть может, в настоящее время не представляется возможным рекомендовать конкретный аналитический аппарат. Это позволит инженерам, использовав идеи теории массового обслуживания, приобрести навыки в постановке ряда важных прикладных задач для решения их на электронных цифровых машинах.

Необходимо отметить, что в настоящее время имеются доступные для практического использования методы решения широкого круга задач, связанных с массовым обслуживанием, на электронных цифровых машинах универсального назначения. Хотя эти вопросы и выходят за рамки книги, однако о них кратко идет речь в заключении. Там же приведены ссылки на соответствующую литературу.

Представляется, что эта книга окажется полезной для широкого круга читателей, интересующихся практическими применениями теории массового обслуживания.

Н. П. Бусленко.

ОТ АВТОРОВ

В настоящей книге авторы сделали попытку в относительно популярной форме изложить основные идеи и некоторые методы одного из новых разделов теории вероятностей, который получил название теории массового обслуживания. Являясь до некоторой степени частью теории вероятностей, теория массового обслуживания последнее время становится все более самостоятельной. Это объясняется теми важными практическими задачами, которые она может решать, и спецификой ее математического аппарата. Авторы не ставили перед собой задачу дать исчерпывающее изложение идей и методов этой теории. Основная цель — привлечь к ней внимание широкого круга читателей. Поэтому авторы излагают только основные положения, иллюстрируемые примерами, представляющими интерес для специалистов различного профиля.

Книга написана таким образом, что она может быть полезна как читателям, не владеющим математическим аппаратом, так и инженерно-техническим работникам, интересующимся областями приложений теории массового обслуживания. Кроме того, книга может быть полезна студентам старших курсов математических специальностей, которые смогут найти в ней значительное число примеров использования аппарата теории массового обслуживания при решении различных практических задач.

Читатель, которого интересует только один вопрос — что такое теория массового обслуживания, может ограничиться чтением § 1 гл. 1. Читатель, которого заинтересует вопрос о том, где она может быть применена, должен прочесть всю первую главу. При чтении этой главы он совсем не столкнется с необходимостью знания математики.

Читатель, которого интересует только вопрос о том, как применяется аппарат теории массового обслуживания, найдет в конце каждого параграфа гл. 3 необходимые выводы и подробно рассмотренные примеры.

Читатель, которого, кроме этого, заинтересует вопрос о том, как получены эти выводы, каков подход к решению задач массового обслуживания, должен прочесть всю книгу.

Для полного понимания всего изложенного достаточно знаний в объеме курса высшего технического учебного заведения. Основные сведения по теории вероятностей авторы старались пояснять по ходу изложения.

Нужно заметить, что терминология теории массового обслуживания еще не полностью установилась. Так как исторически возникновение теории массового обслуживания тесно связано с телефонным делом, то в некоторых источниках можно найти термин «поток вызовов» вместо «потока требований», хотя последний более полно отражает природу изучаемых процессов; «линии обслуживания» вместо «систем обслуживания» и т. д.

Если в результате прочтения книги читатель заинтересуется теорией массового обслуживания и у него возникнет желание использовать ее в своей практической деятельности, то авторы будут считать свою задачу выполненной.

В гл. 1, § 1, 2 и 3 написаны авторами совместно, § 4 написан А. И. Прохоровым. Совместно написан также § 3 гл. 2. Глава 3 и § 1, 2 гл. 2 написаны В. Я. Розенбергом.

Авторы считают своим приятным долгом выразить глубокую благодарность Б. В. Гнеденко, Н. П. Бусленко, Я. И. Хургину и Н. Н. Лозинскому, прочитавшим рукопись и сделавшим ряд существенных замечаний.

ВВЕДЕНИЕ

У вас в дороге оторвалась пуговица. Что может помочь вам? Ответ прост: иголка с ниткой, которую можно найти в ателье бытового обслуживания. У вас испортился холодильник. Очевидно, что помощь может оказать мастер по ремонту. Но если сказать, что, кроме иголки с ниткой или мастера по ремонту, в этих случаях может помочь математика, то вы это воспримете как шутку. Каким образом иголку с ниткой или мастера может заменить математика? Какое отношение вообще имеет математика к пришиванию пуговицы? Прочтите эту книгу и вы увидите, что здесь имеется связь, и довольно тесная.

Почти всегда знакомство с новой книгой читатель начинает с изучения ее оглавления и беглого просмотра. И нередко случается так, что, увидев страницы, пестрящие математическими формулами и выкладками, он кладет ее на место.

Причины этого могут быть различными, но основной из них является следующая: читатель, даже изучавший математику в высшем учебном заведении, мало сталкиваясь в своей практической деятельности с необходимостью применения полученных математических знаний, забывает ее. Поэтому такая книга становится для него «вещью в себе» и даже если он в состоянии, приложив некоторые усилия, разобраться в написанном, он все же редко принимается за ее чтение.

А между тем, как много важного и интересного, а подчас поистине увлекательного скрывается за «сухим» языком формул и цифр — языком математики.

Разгадка тайны атомного ядра, создание атомных двигателей, мощных энергосистем, сложнейших агрегатов, создание и запуск космических кораблей обязаны в значительной мере достижениям математики.

С каждым годом математика все глубже вторгается во все области науки и техники. И сейчас гораздо легче перечислить те области науки, где она не применяется, чем те, где она находит широкое применение.

У многих людей сложилось неверное представление о математике: считают, что это окончательно сформировавшаяся и до конца разработанная наука, что в математике уже все сделано; сформулированы и доказаны почти все теоремы, что, в лучшем случае, есть еще несколько теорем, которые пока не доказаны, но они представляют интерес только для математиков.

Может быть, следующее утверждение несколько утрирует существующее положение, но очень часто можно слышать вопрос: «Ну что нового можно сделать в математике, ведь то, что дважды два — четыре, уже доказано! Построено и развито дифференциальное и интегральное исчисление, разработана теория вероятностей и т. д. Что же делать дальше?»

На наш взгляд это положение объясняется рядом причин.

Во-первых, в тех разделах математики, которые изучаются, изложение имеет строгий, четкий, законченный вид. Те задачи, которые предлагаются учащимся, имеют, как правило, решение в простом окончательном виде. Это приводит учащегося к мысли, что в математике есть теоремы и методы, знание которых позволяет решить *любую задачу*.

Во-вторых, до последнего времени при практическом использовании более или менее сложных математических методов сталкивались с трудностями, связанными с вычислительной работой. Практическое использование ряда методов, разработанных теоретически, было немыслимо: необходимость выполнения колоссального объема вычислений при решении тех или иных задач с помощью этих методов делала невозможной получение численных решений.

Например, такие относительно простые задачи, как интегрирование системы дифференциальных уравнений или решение системы линейных алгебраических уравнений высокого порядка, далеко не всегда удавалось довести до численного ответа.

В результате у многих складывалось представление о теоретической оторванности математики от практики,

о том, что дальнейшая разработка математических методов представляет лишь теоретический интерес.

Однако это не так, и каждому, кто в своей деятельности сталкивался с необходимостью решения практических задач, известно, что число задач, для которых еще неизвестны математические методы решения, особенно численные методы, гораздо больше тех, которые могут быть решены, что еще существует огромное, фактически неограниченное, число задач и проблем, при решении которых математика пока бессильна. Следовательно, необходимо разработать методы решения этих задач, а это, в свою очередь, неизбежно приводит к появлению не только новых математических методов, но и новых математических дисциплин.

Новые возможности и вместе с тем новые трудности возникают перед математикой в связи с созданием новых мощных средств производства вычислительных работ, таких, как быстродействующие электронные вычислительные машины.

С одной стороны, электронные вычислительные машины позволяют производить большое количество арифметических операций (порядка сотен тысяч операций в секунду у существующих и миллионов и даже десятков миллионов у перспективных). Здесь речь идет об универсальных электронных вычислительных машинах— машинах дискретного действия.

С другой стороны, эти новые средства вычислений требуют широкого использования существующих и развития новых численных методов решения задач. Поэтому электронные вычислительные машины стимулируют бурный рост целого ряда новых математических дисциплин. За последние годы появились такие новые научные направления, как линейное и динамическое программирование, теория игр, теория поиска и др.

К таким дисциплинам относится и теория массового обслуживания. Эта теория является относительно новой. На начальное ее развитие особое влияние оказал датский ученый А. К. Эрланг (1878—1929 гг.). Его труды в области проектирования и эксплуатации телефонных станций явились толчком к появлению ряда работ в области массового обслуживания.

Особенно возрос интерес к методам теории массового обслуживания за последние годы. Появившаяся возмож-

ность более глубокого изучения количественной стороны различных процессов с помощью электронных вычислительных машин побудила практиков изучить и использовать наиболее совершенные элементы современного математического аппарата, в том числе и теорию массового обслуживания.

Огромную роль в развитии этой теории сыграл выдающийся советский математик А. Я. Хинчин [20—22]. Его книга «Математические методы теории массового обслуживания» явилась первым трудом, в котором строго были сформулированы идеи и методы теории. Большую работу по дальнейшему развитию идей и методов теории массового обслуживания ведет крупнейший советский математик Б. В. Гнеденко [3—7] со своими учениками. Значительный вклад в развитие методов теории сделал А. Н. Колмогоров [13]. Интересные результаты получил Б. А. Севастьянов, который обобщил задачу Эрланга на случай произвольного распределения времени обслуживания. Значительный интерес представляют результаты применения метода статистических испытаний к решению задач массового обслуживания (Н. П. Бусленко [1]).

Теория массового обслуживания, опираясь в основном на аппарат теории вероятностей, занимается изучением процессов, связанных с массовым обслуживанием. Целью ее является не изучение какого-либо конкретного процесса обслуживания, а разработка методов решения типичных задач, пригодных для решения задач из различных областей. Прикладное значение этой теории весьма велико. Насколько широк круг областей, в которых могут быть использованы идеи и методы теории, читатель увидит даже бегло просмотрев примеры, приведенные в книге.

ГЛАВА ПЕРВАЯ

ПРЕДМЕТ ТЕОРИИ МАССОВОГО ОБСЛУЖИВАНИЯ

1. ЧТО ТАКОЕ ТЕОРИЯ МАССОВОГО ОБСЛУЖИВАНИЯ

Уже само название теории в какой-то степени раскрывает ее содержание. Интуитивное, бытовое представление, которое вызывают у нас слова «массовое обслуживание», в значительной мере помогает раскрыть и понять предмет теории. Во всяком случае, даже на первый взгляд ясно, что содержание теории имеет прямое отношение к *обслуживанию,* причем к обслуживанию *массовому*.

Во всех областях человеческой деятельности, или по крайней мере в большинстве их, мы сталкиваемся с процессами, которые имеют характер массового обслуживания. Наиболее простым и достаточно наглядным примером этого является бытовое обслуживание во всех его видах, будь то обслуживание продавцами покупателей в магазинах, продажа билетов в железнодорожных, театральных и других кассах, ремонт различных бытовых предметов в мастерских, обеспечение разговора по телефону с нужным абонентом или оказание медицинской помощи больным в поликлинике и на дому и т. д., словом, бесчисленное множество самых разнообразных процессов, в которых имеет место массовый спрос на обслуживание в прямом или переносном смысле этого слова. В переносном потому, что имеется в виду обслуживание *во всех его видах*. Так, например, обстрел зенитной батареей самолетов противника, равно как и сбрасывание бомб с самолета на цель, можно также считать «обслуживанием».

Естественно, что во всех случаях большое значение имеет степень удовлетворения потребности в обслужива-

нии, или *качество обслуживания*. Очень важно знать, насколько хорошо будет отремонтирован ваш телевизор или ботинки в мастерской, хорошей ли будет слышимость при разговоре по телефону, как скоро избавится больной от своего недуга, будет ли выздоровление окончательным и не возникнут ли рецидивы? Все это, конечно, очень важно и, как правило, ответы на эти вопросы нас интересуют в первую очередь.

Но в тех случаях, когда возникает потребность или необходимость в осуществлении массового обслуживания, организационные вопросы становятся не менее, а в определенных ситуациях и более важными. Если, например, в мастерской прекрасно отремонтируют обувь, но это будет сделано через год после того, как обувь была сдана в починку, то, естественно, такое обслуживание вас мало устроит. Точно так же вряд ли вас удовлетворит разговор по телефону при самой отличной слышимости, если междугородняя соединит вас с нужным абонентом через месяц или хотя бы на следующие сутки после того, как в этом разговоре возникла необходимость. Если в магазине до того, как вы будете обслужены самым внимательнейшим и наилюбезнейшим продавцом, вам придется долго стоять в очереди, вы наверняка, в лучшем случае про себя, будете ругать дирекцию магазина за плохую организацию обслуживания.

Этими примерами мы хотим подчеркнуть, что, кроме *собственно качества* обслуживания, не меньшее значение имеет и то, *как это обслуживание организовано*. Отсюда возникает необходимость изучения организационной стороны процесса обслуживания. Эта сторона процесса обслуживания может характеризоваться самыми различными показателями: временем ожидания начала обслуживания, длиной очереди, возможностью получения отказа в обслуживании (ведь не всегда ваша заявка на обслуживание может быть удовлетворена, например, в кассе может не оказаться билетов, а в мастерской могут отсутствовать детали, необходимые для ремонта вашего телевизора).

Все эти факторы, несомненно, имеют немаловажное значение в тех процессах, где происходит массовое обслуживание. И вряд ли следует доказывать, что все они зависят от самых различных условий, учесть которые не всегда представляется возможным. Так, напри-

мер, длина очереди в железнодорожную кассу зависит от опыта кассира, содержания тех операций, которые он должен выполнять при оформлении билетов, числа пассажиров, количества поездов, на которые одновременно продаются билеты в этой кассе. Кроме того, длина очереди зависит от настроения кассира, от того, насколько четко пассажиры формулируют свои требования, от необходимости отсчитывать каждому пассажиру сдачу и т. д.

Легко понять, что число этих факторов в тех или иных обстоятельствах может быть весьма значительным. Но одновременно ясно, что далеко не все они равнозначны. Естественно, что среди них есть такие, которые являются основными, решающими, есть и второстепенные, есть и такие, которые хотя и оказывают некоторое влияние на длину очереди, однако в определенных условиях их можно и не учитывать. Основными факторами в нашем примере с железнодорожной кассой являются число пассажиров, нуждающихся в билетах, скорость работы кассира и, наконец, организация работы билетных касс на вокзале, т. е. распределение билетов на различные поезда между кассами.

Предположим, что на нас возложена задача улучшить качество обслуживания пассажиров в железнодорожных кассах. Число пассажиров, обратившихся за билетами, от нас не зависит, поэтому повлиять на этот фактор мы не можем. Скорость работы кассира зависит от его натренированности, методов работы. Она может быть увеличена только до определенных пределов, не превосходящих психофизиологических возможностей человека. Поэтому остается только один путь решения поставленной задачи — изменить организацию обслуживания.

Как это сделать? Можно сделать так, чтобы каждый кассир продавал билеты только на один поезд или чтобы все кассиры продавали билеты на все поезда. В последнем случае, по крайней мере, будет гарантия, что любой пассажир, независимо от того, куда он едет, не будет простаивать в очереди больше других.

Такого рода изменения в работе касс можно произвести без особого труда. Но какое влияние они окажут на качество обслуживания пассажиров? В нашем примере можно просто взять и попробовать, а затем оце-

нить полученные результаты и на основании их сравнения выбрать лучшую организацию. Но далеко не во всех случаях это так легко сделать.

Все, что сказано относительно проблем массового обслуживания в быту, имеет прямое отношение ко всем или, по крайней мере, почти ко всем областям человеческой деятельности — экономике, производству, транспорту, военному делу, медицине и т. п. Так, например, в производстве мы постоянно сталкиваемся с процессами типа массового обслуживания. Очевидно, что ремонт различной техники (автомашин, судов, станков, различной аппаратуры и т. п.) с точки зрения организации имеет много общего с ремонтом холодильников, обуви, платья и т. п. в быту с той только разницей, что на производстве масштабы более значительны; следовательно, правильное решение задач подобного типа на производстве может принести большую пользу.

Возьмем организацию ремонта автомашин. Большое значение имеет, конечно, качество ремонта каждого отдельного автомобиля, но не меньшее значение имеет и правильная организация ремонтной базы. Всегда важно знать, какое количество авторемонтных мастерских или заводов нужно иметь для того, чтобы обеспечить своевременный ремонт автомашин.

Если их будет мало, то образуется такая «очередь» автомобилей, ожидающих ремонта, которая, например в масштабе республики, а тем более всей страны, может привести к огромным материальным потерям, и получится так, что простои машин могут оказаться более дорогостоящими, чем расширение ремонтной базы. Но совершенно ясно, что неограниченное ее расширение не только неразумно, но и нереально. Следовательно, нужно найти какое-то оптимальное решение, которое позволило бы решить эту задачу так, чтобы, с одной стороны, обеспечить минимальный простой автомашин в ожидании ремонта, а с другой стороны, развернуть ремонтную базу таким образом, чтобы это развертывание не привело к ненужной трате государственных средств.

Как это сделать?

Можно, конечно, пойти по самому простому пути (например, как в случае с железнодорожными кассами), как чаще всего и поступают, — это ждать, что покажет практика, эксперимент, а затем корректировать ранее

принятые решения. В какой-то мере этот путь неизбежен, но чем точнее выбрано первоначальное решение, тем лучший результат будет получен и в более короткие сроки.

И вот тут-то, при принятии этого первоначального решения, как и других последующих, неоценимую помощь может оказать теория массового обслуживания, потому что для принятия такого обоснованного решения необходимо произвести ряд предварительных расчетов, уметь найти соотношение между вероятным числом неисправных автомашин, средним временем ремонта каждой из них, мощностью ремонтной базы в целом и длиной «очереди» или средним временем ожидания начала ремонта и т. д. — словом, получить ответы на целый ряд вопросов, решение которых как раз и является делом теории массового обслуживания. Даже самый поверхностный экскурс в любую сферу человеческой деятельности, который мысленно может проделать читатель, покажет, что подобных задач возникает довольно много в каждой из них.

Распределение электроэнергии между потребителями, организация снабжения, сопряжение оборудования различной мощности при проектировании автоматических поточных линий и многое другое — все это области, в которых помогает или может помочь теория массового обслуживания.

Велика ее роль и в вопросах теории надежности различных систем. Она позволяет, рассматривая выход из строя элемента системы как потребность в обслуживании, находить количественные показатели степени надежности различных систем, обосновывать выбор элементов этих систем, исходя из необходимой надежности всей системы в целом.

Тем более важно применение теории массового обслуживания в военном деле. Можно рассматривать, например, обстрел вражеских танков нашей артиллерийской батареей тоже как своего рода «обслуживание», от результатов которого будет зависеть успех боя.

Естественно, что во всех этих примерах метод «проб» далеко не всегда может быть использован: иногда такая «проба» может быть сопряжена с огромными материальными затратами, иногда она может быть

сопряжена с опасностью для здоровья и даже жизни людей, а иногда и просто невозможна в силу тех или иных причин.

Но нельзя ли научиться анализировать результаты некоторых экспериментов не производя их?

Оказывается можно. И эту возможность в значительной степени обеспечивает теория массового обслуживания, которая позволяет использовать математический аппарат для оценки явлений и процессов, имеющих характер массового обслуживания, в какой бы форме оно ни протекало.

Прежде чем перейти к изложению ее содержания и метода, а также математического аппарата, который она использует, необходимо усвоить некоторые термины и понятия, применяемые в этой теории.

Итак, в науке, в производстве, в быту, в процессах боевых действий и т. п. часто возникают такие ситуации, когда имеется необходимость в обслуживании большого количества однородных требований. При этом под термином *требование* понимается запрос на удовлетворение какой-либо потребности. Под *обслуживанием* будем понимать удовлетворение потребности.

Из предыдущих примеров ясно, что удовлетворение потребности может происходить как в интересах того, с чьей стороны поступило требование, так и в интересах того, кто удовлетворяет потребность. Такой подход обогащает понятие требования, позволяет, как будет видно из дальнейшего, расширить область применимости теории массового обслуживания.

Нужно заметить, что в дальнейшем термин требование иногда отождествляется с его носителем. Так, например, когда мы говорим «требование нуждается в обслуживании», то тем самым отождествляем требование с его материальным носителем. Требование на обслуживание может поступить со стороны телефонного абонента, владельца неисправной автомашины, покупателя в магазине и т. д., но для сокращения о всех этих процессах мы будем говорить как о требованиях, нуждающихся в обслуживании. Точно так же, имея в виду обеспечение абонента связью с нужным ему номером, ремонт автомашины, обслуживание покупателя и т. п., мы будем говорить как об обслуживании требований. Такое сокращенное обозначение термином

требование как заявки на обслуживание, так и лица (предмета), нуждающегося в обслуживании, в дальнейшем значительно облегчит изложение без ущерба для его точности и понимания.

Таким образом, этот термин обобщает всевозможные виды заявок на обслуживание со стороны самых разнообразных объектов.

Как явствует из всего вышесказанного, по своей природе обслуживание может иметь самый различный характер. При этом каждое поступившее требование нуждается в обслуживании со стороны какого-либо устройства, либо человека, либо группы людей (бригады). В дальнейшем те средства, которые осуществляют обслуживание требований, будем называть *обслуживающими аппаратами* или *обслуживающими устройствами*. Правда, этот термин нельзя признать очень удачным, ибо в некоторых случаях обслуживание производится одним человеком, в некоторых — группой людей, а в некоторых действительно аппаратом, т. е. каким-то техническим устройством. Все это говорит о том, что термин «обслуживающий аппарат» недостаточно удобен, однако трудно, не выдумывая «заумных» слов, найти такой новый термин, который бы достаточно хорошо объединял столь разнородные понятия. На практике очень часто встречаются случаи, когда один человек или группа людей, а также различные технические устройства могут обслуживать сразу несколько требований.

Во избежание всякой путаницы будем считать такие устройства состоящими из нескольких обслуживающих аппаратов (точнее, такого их количества, какое количество требований в состоянии обслужить в одно и то же время данное устройство). Таким образом, под «обслуживающим аппаратом» или обслуживающим устройством впредь мы будем понимать *то*, что способно обслуживать в данный момент только одно требование.

На практике, как правило, приходится иметь дело не с одним обслуживающим аппаратом, а с группой аппаратов, состоящей из ограниченного их числа. Совокупность однородных обслуживающих аппаратов называется *обслуживающей системой*. При этом под *однородными обслуживающими аппаратами* понимаются такие, которые способны удовлетворять одинаковые тре-

бования. Например, парикмахерскую можно рассматривать как обслуживающую систему, состоящую из ограниченного числа однородных обслуживающих «аппаратов» — парикмахеров. Естественно, что каждому мастеру присущи свои индивидуальные качества: один очень быстро бреет, другой стрижет и т. д., но в общем любой из них способен выполнять все операции, которые от него потребуются. Поэтому условие однородности не предусматривает, что все обслуживающие аппараты обладают одинаковыми характеристиками. Может быть, что один из них, в среднем, быстрее осуществляет обслуживание, чем другой. Здесь существенно то, что каждый из них способен удовлетворить поступившее требование.

В том случае, когда число обслуживающих аппаратов в обслуживающей системе ограниченно, особенно ясно вырисовываются те задачи, которые могут возникнуть при изучении таких систем.

Всем известна, например, такая ситуация, когда число обслуживающих аппаратов мало, поэтому, пожалуй, нет необходимости в том, чтобы слишком подробно останавливаться на вопросе: что значит «мало»? Для тех, кому приходится стоять в очереди в парикмахерской, магазине, у газетного киоска и т. д., ожидая обслуживания, это и так ясно. При всех обстоятельствах любая очередь является убедительным доказательством того, что в данный момент система не справляется с обслуживанием.

Но является ли наличие очереди достаточной причиной для того, чтобы требовать увеличения числа обслуживающих аппаратов?

Можно ли лишь по одному такому признаку, как наличие очереди, судить о том, что система вообще плохо справляется с обслуживанием? Очевидно, нет.

Если речь идет, например, о такой наиболее близкой всем форме обслуживания, как, скажем, обслуживание покупателей в магазине, то ясно, что о том, насколько хорошо продавцы справляются со своими основными обязанностями — обслуживанием покупателей, нельзя судить по наличию очереди покупателей, стоящих в этом магазине в данный момент. Нужен подробный анализ работы магазина за длительный промежуток времени с учетом всех основных факторов, влияющих на нее, для того чтобы иметь возможность сделать обоснованный

вывод о необходимости увеличения числа продавцов. Причем этот анализ должен быть проведен прежде всего с точки зрения основной цели, которая преследуется деятельностью данного магазина. Как правило, одним из основных требований является максимальное удовлетворение запросов покупателей и минимальные накладные расходы. Поэтому анализ работы данной обслуживающей системы должен производиться в первую очередь с точки зрения стоимости обслуживания. Следующим, и не менее важным фактором, определяющим качество работы магазина, является время, затрачиваемое покупателем в ожидании обслуживания.

Итак, далеко не полный разбор этого простого примера показывает, что даже при беглом взгляде на процессы, связанные с обслуживанием, возникает целый ряд весьма интересных и очень важных задач, решение которых может принести большую практическую пользу. Так, например, в ситуации, когда обслуживающая система недостаточно загружена, будут иметь место непроизводительные расходы на ее содержание, следовательно, решение задачи о рентабельности данной системы позволит получить определенную экономию материальных средств.

Приведенный пример с обслуживанием покупателей в магазине является типичным. Все вышесказанное можно повторить и для обслуживания клиентов в парикмахерской, в столовой; можно провести аналогию и с обслуживанием абонентов на телефонной станции. Немало подобных ситуаций можно найти и в области промышленного производства. Так, например, если рассмотреть процесс обслуживания группы ткацких станков одним рабочим, то он мало чем отличается от ситуации, возникающей при обслуживании покупателей в магазине. Здесь аналогом «покупателя», т. е. требования, нуждающегося в обслуживании, является обслуживаемый станок, а роль «продавца» — обслуживающего аппарата, играет рабочий, причем проблемы здесь возникают такие же: какое количество рабочих необходимо иметь для работы на данном количестве станков, как добиться наилучшего обслуживания, т. е. такого, которое обеспечит наиболее высокую производительность труда и минимальную себестоимость выпускаемой продукции, и т. д.

Без особого труда каждый читатель сможет предложить не один десяток подобных задач, с которыми он сталкивается или в процессе своей трудовой и научной деятельности, или в быту. Некоторые примеры подобного рода будут приведены и детально рассмотрены ниже.

Что же является общим как во всех вышеприведенных примерах, так и во всех других ситуациях и процессах, связанных с массовым обслуживанием? Прежде всего это то, что в каждой такой ситуации всегда имеется группа обслуживающих единиц, которые (как мы условились выше) во всех случаях могут быть названы обслуживающими аппаратами. При этом обслуживающие аппараты могут быть объединены в обслуживающую систему. Во-вторых, то, что любой процесс массового обслуживания протекает по одному и тому же принципу: заявки (требования) на обслуживание поступают в обслуживающую систему, обрабатываются обслуживающими аппаратами и покидают ее. И, наконец, общей целью изучения всех процессов массового обслуживания является задача обеспечения *успешной работы* обслуживающей системы. Здесь мы сталкиваемся с понятием *успешная работа.* Естественно, что это понятие в каждом отдельном случае будет иметь свой конкретный смысл, однако независимо от того, какой смысл в это понятие вкладывается, можно заранее утверждать, что для обеспечения успешной работы обслуживающей системы недостаточно использовать только одни качественные методы.

Бóльше того, решение задачи наилучшей организации функционирования обслуживающей системы одними качественными методами вообще невозможно и неизбежно требует применения количественных методов решения этих задач. Только количественные методы могут позволить обоснованно судить о том, что данный способ организации лучше или хуже другого, что данная обслуживающая система справляется с обслуживанием лучше всех возможных и т. д. В рассмотренных выше задачах были использованы такие количественные характеристики, как стоимость обслуживания, время ожидания начала обслуживания, длина очереди.

Таким образом, мы приходим к очень важному выводу: изучение процессов массового обслуживания требует привлечения количественных методов анализа этих

процессов, а так как различные процессы массового обслуживания содержат очень много общего, то нет необходимости разрабатывать свои количественные методы для каждого конкретного случая. Достаточно выявить типичные задачи, абстрагироваться от конкретного процесса обслуживания и разработать методы решения, которые были бы приемлемы и для решения частных задач, связанных с массовым обслуживанием. С этой целью должна быть рассмотрена некоторая абстрактная система обслуживания, отвлеченный поток требований, поступающих в эту систему, и обслуживание этих требований, не связанное ни с какими конкретными его формами.

Приведенные выше рассуждения позволяют нам теперь более четко представить, что такое теория массового обслуживания, чем она занимается, а также сформулировать ее предмет и цели, которые она преследует.

Предметом теории массового обслуживания является количественная сторона процессов, связанных с массовым обслуживанием.

Целью теории является разработка математических методов для отыскания основных характеристик процессов массового обслуживания для оценки качества функционирования обслуживающей системы.

Все задачи массового обслуживания имеют вполне определенную структуру, которая схематически может быть изображена так, как это показано на рис. 1.

Последовательность событий будем называть *потоком*. Поток, состоящий из требований на обслуживание, назовем *потоком требований*.

Поток требований, нуждающихся в обслуживании и поступающих в обслуживающую систему, называется *входящим потоком*. Поток требований, покидающих обслуживающую систему, называется *выходящим потоком*. При этом требования, поступающие в обслуживающую систему, могут покидать ее и будучи необслуженными. Так, например, зайдя в магазин, вы можете не обнаружить необходимый вам товар, нужный вам абонент может оказаться занятым, вышедший из строя станок долгое время, в силу тех или иных причин, может оставаться неисправным и т. д. Следовательно, выходящий поток может состоять или только из одних обслу-

женных, или как из обслуженных, так и необслуженных требований.

Входящий поток, функционирование обслуживающей системы и, как результат обслуживания, выходящий поток подлежат количественному описанию. Более подробное изложение методов количественного описания входящего потока читатель найдет в § 1 гл. 2.

Функционирование обслуживающей системы в целом определяется, в первую очередь, ее организацией.

На практике часто обслуживание одного требования осуществляется последовательно несколькими обслужи-

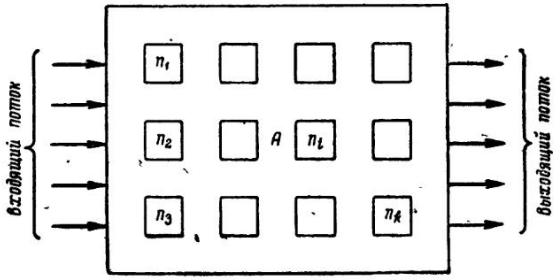

Рис. 1. Схематическое изображение системы массового обслуживания $n_1, n_2, \ldots, n_i, \ldots n_k$ — обслуживающие аппараты, A — обслуживающая система.

вающими аппаратами. При этом, как правило, очередной обслуживающий аппарат начинает работу по обслуживанию требования после того, как предыдущий закончил свою работу. Таким образом, процесс обслуживания носит многофазовый характер.

Многофазовые процессы представляют значительный интерес, однако решение задач, связанных с ними, сопряжено со значительными трудностями. Последние успехи в этой области связаны с использованием метода статистических испытаний (метод Монте-Карло). В дальнейшем мы ограничимся рассмотрением только однофазовых процессов обслуживания.

На практике могут встречаться самые разнообразные виды организации обслуживающих систем, наиболее типичными из которых являются следующие:

— все обслуживающие аппараты системы равноправны,

— обслуживающие аппараты, входящие в данную систему, неравноправны.

В первом случае вновь поступившее требование поступает на обслуживание в один из свободных аппаратов, причем предпочтения при этом не отдается ни одному из них. Такая организация системы обслуживания носит название *неупорядоченной*. Примером неупорядоченной системы обслуживания может служить обслуживание в магазине, когда покупатель подходит к любому из свободных продавцов.

Во втором случае, когда обслуживающие аппараты неравноправны, все они, как правило, пронумерованы и новое требование обслуживается только первым аппаратом, если он свободен. Если же первый аппарат занят обслуживанием ранее поступившего требования, то новое требование поступает во второй аппарат; если занят второй, то в третий и т. д.

Таким образом, новое требование обслуживается тем свободным аппаратом, который имеет наименьший номер. Такие системы обслуживания, следуя А. Я. Хинчину, будем называть *упорядоченными*. Примером такой системы может служить обслуживание абонентов автоматической телефонной станцией.

Возможны и другие способы организации обслуживающей системы. Так, например, обслуживающие аппараты могут загружаться только в порядке очереди. Освободившийся обслуживающий аппарат становится в очередь и не загружается до тех пор, пока не получат работу все аппараты, освободившиеся ранее него.

Естественно, что функционирование обслуживающей системы характеризуется не только ее организацией, но и качеством работы каждого обслуживающего аппарата, однако решение вопроса о качестве работы каждого обслуживающего аппарата выходит за рамки теории массового обслуживания и является предметом других методов исследования. В теории массового обслуживания работа каждого обслуживающего аппарата характеризуется временем, затрачиваемым им на обслуживание одного требования.

Выходящий поток характеризуется различно в зависимости от организации взаимодействия входящего потока и обслуживающей системы. Это взаимодействие находится в тесной связи с характером задачи массового

обслуживания. Так, нередко могут существовать такие ситуации, когда поступившее в обслуживающую систему требование, найдя все обслуживающие аппараты занятыми, становится в очередь и ждет, пока один из них не освободится. В таких случаях выходящий поток будет состоять целиком из обслуженных требований. Бывают и такие ситуации, когда требование, поступившее в систему, найдя все обслуживающие аппараты занятыми, покидает ее не дожидаясь обслуживания и, следовательно, выходящий поток будет состоять как из обслуженных, так и из необслуженных требований.

В первом случае возникают такие вопросы, как определение длины очереди, времени ожидания начала обслуживания и т. д. Во втором случае возникают такие задачи, как определение числа необслуженных требований, степени загруженности обслуживающей системы и т. д.

Эти два случая не исчерпывают всех возможных способов «поведения» требования, поступившего в систему в момент, когда все обслуживающие аппараты заняты. Возможны такие положения, когда требование может находиться в системе обслуживания не больше определенного времени, после чего оно покидает систему независимо от того, начато обслуживание или нет.

Системы массового обслуживания, а в соответствии с этим и задачи могут различаться в зависимости от порядка принятия требований на обслуживание в том случае, когда образуется очередь.

При этом возможны следующие основные случаи:

— требования поступают на обслуживание в порядке очереди, освободившийся аппарат принимает на обслуживание требование, поступившее ранее других;

— освободившийся аппарат принимает на обслуживание требование, которое в кратчайшее время должно покинуть систему;

— требования поступают на обслуживание в случайном порядке, в соответствии с заданными вероятностями.

Как правило, в большинстве задач массового обслуживания входящий поток зависит не от воли человека, а от целого ряда случайных факторов, и поэтому очень трудно регулировать количество поступающих требований или точно определить, какое число требований по-

ступит за данный промежуток времени. Поэтому входящий поток обычно описывается с помощью вероятностных характеристик, о чем подробнее будет рассказано ниже. От воли человека зависит организация обслуживающей системы — каким образом распределить поступающие требования между обслуживающими аппаратами, какое количество обслуживающих аппаратов выделить, как сгруппировать аппараты для обслуживания? От того, насколько успешно будут решены эти вопросы, зависит *качество функционирования обслуживающей системы.*

Здесь под качеством функционирования системы понимается не то, насколько хорошо выполнено обслуживание, а то, насколько полно загружена система обслуживания, не простаивает ли оборудование, не образуется ли очередь. Конечно, от воли человека зависит и работа отдельного обслуживающего аппарата, но улучшение качества или скорости его работы не является задачей теории массового обслуживания.

Задачей теории массового обслуживания является отыскание функциональных зависимостей величин, характеризующих качество функционирования обслуживающей системы, от характеристик входящего потока, параметров, характеризующих возможности одного обслуживающего аппарата, и способов организации всей обслуживающей системы в целом. Качество функционирования системы существенно зависит от того, как организовано управление процессом обслуживания, поэтому задача отыскания количественных характеристик организации управления является очень важной.

Эти зависимости могут иметь как детерминированный, так и вероятностный характер, но в обоих случаях они позволяют определить, насколько хорошо будет работать данная обслуживающая система при данных значениях входящих параметров. Вопрос выбора количественных характеристик является особенно трудным в том случае, когда они имеют вероятностный характер. Обоснование выбора этих характеристик может быть произведено только в рамках конкретных задач и скорее относится к области исследования операций, чем к массовому обслуживанию.

После выбора количественных характеристик возникает не менее трудная задача по определению такого

набора значений параметров, при котором обслуживающая система будет функционировать наилучшим образом. Но и эта задача не является задачей теории массового обслуживания.

Всякая задача массового обслуживания будет считаться решенной, если удастся найти количественные характеристики качества функционирования обслуживающей системы и выразить их через величины, характеризующие входящий поток и обслуживающую систему. На этом, собственно, задача массового обслуживания заканчивается и дальнейшее исследование должно использовать любые другие методы, такие, например, как линейное или квадратическое программирование, динамическое программирование или известные классические методы отыскания наибольшего значения тех или иных характеристик.

В каких же областях человеческой деятельности может найти применение аппарат теории массового обслуживания? Мы уже говорили, что ситуации типа массового обслуживания можно обнаружить в подавляющем большинстве областей деятельности человека.

В первую очередь, широкое поле деятельности для нее открывается в области промышленного производства вообще и производственного планирования в особенности. В области промышленного производства приходится очень часто сталкиваться с задачами массового обслуживания. Так, например, массовое обслуживание имеет место при обеспечении заводами-поставщиками и фабриками предприятий-потребителей и торговой сети своей продукцией. Обеспечение заводов и фабрик сырьем также носит характер массового обслуживания. Аналогичный характер имеет распределение электроэнергии между различными предприятиями.

Организация взаимодействия между цехами в пределах одного завода представляет собой пример задач того же типа, только на более низкой ступени. Продолжая этот путь и переходя к масштабам цеха, мы внутри него также обнаружим ряд подобных проблем, начиная со снабжения цеха сырьем и кончая организацией обслуживания и ремонта станков. Важность и интерес решения подобных задач бесспорны, и вряд ли есть необходимость говорить о том, какое значение имеет улучшение качества обслуживания в любом из этих процессов.

Большое количество проблем, связанных с массовым обслуживанием, должно быть решено при производственном планировании. Представляется несомненным, что затрата усилий на детальное изучение количественной стороны процессов планирования даст большой экономический эффект и позволит более глубоко проникнуть в природу этих процессов. Немалую пользу может оказать теория массового обслуживания на стадии технического проектирования. При проектировании любого промышленного предприятия весьма важным является вопрос о степени загруженности оборудования. Так, еще в процессе технического проектирования необходимо уметь определять нужное количество оборудования и его мощность исходя из объема работ, которые должны выполняться с помощью этого оборудования. При решении этой задачи необходимо учитывать такие случайные факторы, как время обслуживания, выход из строя отдельных устройств за счет поломок и время, требуемое для их устранения, а также ряд факторов, от которых будет зависеть эксплуатация этого проектируемого оборудования. Поэтому такая задача имеет более сложный характер, чем, скажем, деление произведения трудоемкости изготовления одной детали и их требуемого числа на мощность станка для определения нужного количества станков.

Использование методов теории может помочь также в осуществлении выбора лучшего, наиболее экономичного проекта из нескольких возможных.

Исключительно широкое применение должны найти методы теории в области бытового обслуживания населения. Так, например, правильно организовать сеть пунктов по ремонту телевизоров, радиоприемников и т. д. можно лишь с учетом потока ожидаемых заявок на обслуживание путем использования методов теории.

Много задач, при решении которых могут быть использованы методы теории, встречается в различных областях военного дела. В военное время исключительно важное значение приобретает правильная организация ремонтной базы оружия и боевой техники, обеспечение пропускной способности различных военных систем.

В процессе проектирования новых образцов оружия и научно-исследовательской работы по обоснованию тре-

бований к его характеристикам также приходится сталкиваться с большим количеством задач типа массового обслуживания. Целью каждого вида оружия является наиболее эффективное его использование в общей системе сил и средств. Использование сил при нанесении удара по противнику может быть истолковано как «обслуживание» противника. Конечно, это весьма своеобразное обслуживание, но такой подход позволяет использовать аппарат теории массового обслуживания для отыскания различных характеристик эффективности новых образцов оружия, используемых в общей системе вооружения, еще на стадии их проектирования.

Используя аппарат теории массового обслуживания, можно решать задачи по сравнительной оценке различных образцов оружия и выбирать лучшие из них еще до практического создания этих образцов. Правила артиллерийской стрельбы, торпедной стрельбы, бомбометания, методики использования тех или иных боевых и технических средств могут рассматриваться как организация соответствующей обслуживающей системы, состоящей из таких обслуживающих аппаратов, как, например, ракеты, артиллерийские орудия, торпедные аппараты, самолеты-бомбардировщики и т. д. и, следовательно, применение теории массового обслуживания может оказать значительную помощь в проверке качества этих правил и методик.

Методы теории могут найти применение и при планировании боевых действий. Решение оперативно-тактических задач требует в настоящее время использования нового, более совершенного математического аппарата. Одним из элементов этого аппарата является теория массового обслуживания. Она может оказать существенную помощь при оценке качества различных оборонительных систем. Ее можно использовать при решении некоторых задач, связанных с планированием наступательных действий. Широкое поле деятельности открывается перед методами этой теории в области тыловых задач, так как все задачи по снабжению и обеспечению имеют характер задач массового обслуживания.

В связи с тем, что все, или почти все, процессы, связанные с медицинским обслуживанием населения, будь то снабжение населения, больниц и клиник медикаментами, медицинским инструментом и оборудованием или

собственно медицинское обслуживание, заключающееся в оказании врачебной помощи людям в поликлинике, больнице и на дому — есть не что иное, как процессы массового обслуживания, то можно с уверенностью сказать, что применение теории массового обслуживания окажет неоценимую услугу нашим врачам и специалистам по организации здравоохранения.

2. ТЕОРИЯ МАССОВОГО ОБСЛУЖИВАНИЯ И НАРОДНОЕ ХОЗЯЙСТВО

Рассмотрим теперь некоторые, наиболее типичные, задачи из области промышленного производства, ремонта техники, снабжения, организации приемных пунктов по обслуживанию населения и т. п., в решении которых может оказать существенную помощь аппарат теории массового обслуживания. При этом главное внимание будем уделять не строгости формулировки этих задач, а тому конкретному содержанию, которое отражает эти задачи.

Первый пример. В промышленном производстве часто встречается задача об обслуживании рабочим группы станков. Естественно, что раз речь идет о группе станков, обслуживаемых одним рабочим, то, следовательно, имеется в виду промышленная эксплуатация группы автоматов или полуавтоматов. Предположим, для большей наглядности, что в нашем случае эксплуатируемыми являются ткацкие автоматы, которые обслуживаются одной ткачихой. Функции ткачихи при обслуживании этой группы станков заключаются в том, что она устраняет всевозможные неисправности, которые могут возникнуть при их работе. Под неисправностью будем понимать не только такие, как обрыв нити, различные неполадки в работе автоматов, но и необходимость остановки станка для закладки новой порции сырья, снятия готовой продукции и т. п. Эти операции выполняются ткачихой без посторонней помощи. Но могут быть и более сложные неисправности, требующие посторонней помощи.

Все эти операции относятся к обслуживанию, причем каждая из них требует различного времени на ее выполнение, что особенно относится к действительным

неисправностям. Поэтому время обслуживания, изменяясь в довольно значительном диапазоне, вообще говоря, есть величина случайная. Предположим, что станки работают исправно, нити не обрываются и, следовательно, потребность в обслуживании некоторое время отсутствует,— ткачиха свободна. Таким образом, это время ткачихой тратится целиком непроизводительно. Естественно, что для производства выгодно добиваться такого положения, при котором это время будет минимальным; однако из этого не следует, что выгодным является положение, когда станки часто выходят из строя.

Но вот на одном из станков оборвалась нить — поступило требование на обслуживание. Ткачиха должна заняться обслуживанием этого станка. Пока она связывает нить или устраняет какую-то другую неисправность на первом станке, из строя может выйти еще один, но так как ткачиха в этот момент занята, то второй станок ждет, пока она не освободится, т. е. пока не будет закончено обслуживание первого станка. Но за это время могут потребовать внимания третий, четвертый и т. д. станки и может образоваться очередь станков, ожидающих обслуживания. Теперь уже непроизводительным является время, которое простаивает станок в ожидании обслуживания и, естественно, чем меньше этот промежуток времени, тем выгоднее для производства.

Таким образом, перед нами возникает очень интересная и, на первый взгляд, кажущаяся противоречивой задача. С одной стороны, если за ткачихой закреплено мало станков, то она будет недостаточно загружена, и периоды, в течение которых станки не потребуют ее внимания, будут велики. Следовательно, вся зарплата ткачихи за это время ляжет дополнительной нагрузкой на выработанную продукцию, причем чем больше будет время, в течение которого ткачиха простаивает, тем большая доля ее зарплаты ложится непроизводительным бременем на каждый метр выработанной ткани.

Если же за ткачихой закрепить очень много станков, то она не сможет уделить каждому из них достаточно внимания и возможно такое положение, когда очередь из станков, нуждающихся в обслуживании, будет образовываться довольно часто; ткачиха перестанет справ-

ляться с их обслуживанием, производительность оборудования будет низка и стоимость продукции резко возрастет за счет времени простоя станков в ожидании начала обслуживания.

Итак, если мы закрепим за одним рабочим слишком много станков — плохо, если мы закрепим за ним слишком мало станков, тоже плохо. Следовательно, нужно закрепить за ним не слишком много, но и не слишком мало станков, т. е. такое их количество, которое было бы оптимальным со всех точек зрения. А сколько? Где та золотая середина, которая обеспечит наилучшую эксплуатацию оборудования, наибольшую производительность труда и минимальную стоимость продукции? Наши рассуждения доказывают существование такого оптимального числа станков, обслуживаемых одним рабочим, которое обеспечит все эти требования, но эти рассуждения ничего не говорят о том, каким образом найти это оптимальное число.

Очевидным является утверждение, что никакими качественными рассуждениями этой задачи не решить. Естественным представляется следующий путь решения этой задачи. Выделим рабочему один станок и в течение нескольких дней будем фиксировать результаты его работы. Затем закрепим за ним два станка и проделаем то же самое. Будем продолжать этот эксперимент до тех пор, пока себестоимость метра ткани будет убывать. Как только она начнет возрастать, прекратим эксперимент и будем считать, что за рабочим выгоднее всего закрепить такое количество станков, при котором в данном эксперименте или стоимость одного метра ткани была минимальной или объем продукции—наибольшим. Таким способом, по-видимому, можно решить эту задачу, но назвать этот способ универсальным трудно. Он может быть вполне хорош, если мы решаем эту задачу только для одного типа станков. Но ведь число различных типов станков чрезвычайно велико. Кроме того, приемы и способы обслуживания непрерывно меняются и каждое такое изменение потребует повторения всего этого громоздкого эксперимента. Таким образом, мы приходим к выводу, что вышеописанный способ пригоден лишь для единичного использования и не может быть рекомендован как универсальный способ решения задач такого типа.

Другой путь решения этой задачи, как и всех задач такого типа, заключается в отыскании новых количественных методов описания подобных процессов, создании математических моделей этих процессов и изучении этих моделей. Как мы уже говорили раньше, целью теории массового обслуживания является построение количественных методов описания процессов типа массового обслуживания. То, что рассмотренная нами задача относится к числу задач массового обслуживания, несомненно. Обслуживающим аппаратом здесь является рабочий. Поток требований на обслуживание здесь состоит из станков, требующих внимания рабочего. Обслуживание заключается в связывании оборванной нити, устранении неисправностей и пуске станка.

Нужно заметить, что предсказать заранее момент, когда тот или иной станок потребует внимания рабочего, невозможно. Поток требований на обслуживание является случайным.

Постараемся найти пути, по которым должно идти решение поставленной задачи. Количественный метод ее решения требует умения описывать количественно все элементы изучаемого процесса. Одним из таких элементов является время обслуживания. Как уже отмечалось выше, время устранения неисправности является случайной величиной. Следовательно, необходимо уметь количественно описывать эту случайную величину. Способом количественного описания ее является, в частности, закон распределения вероятностей этой случайной величины или какие-либо его характеристики.

Вторым, очень важным элементом является поток требований. Для описания потока требований необходимо знать, какова вероятность того, что за данный промежуток времени выйдет из строя один, два, . . ., n станков. Следовательно, количественно поток требований может быть описан случайной функцией.

Для получения характеристик времени обслуживания и потока требований необходимо собрать и обработать статистический материал по эксплуатации станков. Казалось бы, эта работа ничем не отличается от той, которую мы описывали выше в забракованном методе решения задачи. Действительно, как там, так и тут необходимо проведение экспериментального изучения работы станков и рабочего как обслуживающего аппарата. Но

здесь уже достаточно изучить не весь процесс в целом, а лишь отдельные его элементы, что является гораздо более простой задачей.

После того как изучены основные элементы процесса, возникает задача о получении количественных характеристик, позволяющих оценить течение процесса и сделать вывод о результатах обслуживания. Одной из таких характеристик является среднее время ожидания станками начала обслуживания. Это время является функцией количества станков, закрепленных за одним рабочим. Если найти количественный метод получения среднего времени ожидания начала обслуживания, то можно без дальнейших экспериментов, только путем вычислений с помощью этого показателя, определять качество того или иного способа организации.

Помимо среднего времени ожидания начала обслуживания, значительный интерес представляет и такая величина, как среднее время, затрачиваемое рабочим на обслуживание станков. Знание этих характеристик позволит определить основные экономические показатели способа организации, например, стоимость одного метра ткани, потери за счет простоя станков и т. д.

При этом стоимость производимой продукции является, пожалуй, самым главным и наиболее полным критерием, определяющим качество организации эксплуатации станков. В этот критерий обязательно войдет и время простоя станка и заработная плата рабочего. Следовательно, основной задачей, которая должна быть решена в дальнейшем, является получение этих величин как функций продолжительности времени обслуживания.

Этот метод (способ) решения задачи является гораздо более перспективным в связи с тем, что те количественные характеристики, которые будут получены при этом, равно как и методы их получения, пригодны не только для данной задачи, но и для всех задач подобного типа.

Все эти рассуждения позволяют нам перейти к следующей количественной формулировке рассматриваемой задачи. Дано n станков, обслуживание которых возложено на одного рабочего. Известны характеристики времени обслуживания одного станка и потока требований на обслуживание. Необходимо найти среднее

время простоя станков в ожидании начала обслуживания, среднее время обслуживания станков и стоимость продукции.

Подробное решение этой задачи и методы подхода к решению задач подобного типа будут изложены в третьей главе.

Рассмотрим теперь пример, который является простым и естественным обобщением предыдущего и не менее часто встречается на практике.

Второй пример. Пусть ситуация в основном такая же, как и в предыдущем случае, т. е. по-прежнему имеется n станков, которые нуждаются в обслуживании, и все характеристики обслуживания, приведенные для первой задачи, сохраняются. Отличие же будет состоять только в том, что теперь станки обслуживаются не одним рабочим, а группой рабочих (бригадой), состоящей из m человек. Естественным является предположение, что число обслуживаемых станков больше числа рабочих в бригаде, т. е. $n>m$. На первый взгляд этот способ организации обслуживания станков обладает некоторым преимуществом по сравнению с первым.

Действительно, если обслуживания потребовал не один станок, а сразу несколько и количество станков, потребовавших обслуживания, меньше числа рабочих, входящих в состав бригады, то все они могут быть обслужены одновременно. В результате время ожидания начала обслуживания сократится фактически до нуля и, следовательно, сократится время простоя станков. А если при этом выбрать число станков, обслуживаемых бригадой, достаточно большим, то можно добиться довольно полной загрузки всех членов бригады.

В том случае, когда заявки на обслуживание поступят от такого количества станков, которое превосходит количество рабочих в бригаде, то образуется очередь, в которой будет столько станков, на сколько общее количество станков, нуждающихся в обслуживании, будет больше числа рабочих, входящих в состав бригады. При этом не без оснований можно предполагать, что время ожидания начала обслуживания станков, попавших в эту очередь, будет значительно меньше, чем если бы все эти станки обслуживались одним рабочим. Однако этих рассуждений недостаточно для того, чтобы утверж-

дать, что бригадный способ обслуживания лучше. Кроме того, даже если считать, что этот способ обслуживания группы станков лучше, то возникает вопрос о том, какое количество станков нужно закрепить за бригадой.

Таким образом и при решении этой задачи неизбежно применение количественных методов. При этом помимо таких вопросов, как определение среднего времени ожидания начала обслуживания, среднего времени, которое затрачивает каждый рабочий на обслуживание станков и т. д., возникает и ряд новых вопросов: выбор рационального числа членов бригады, количества станков, которое целесообразно закрепить за данной бригадой, и т. п.

Нетрудно заметить, что первая задача фактически является частным случаем только что разобранной и все ответы на нее могут быть получены из решения второй задачи при $m=1$.

Третий пример. Хотя этот пример и дается в несколько шутливой форме, однако это не помешает читателю увидеть за ним целый ряд больших и практически важных примеров аналогичной природы. Представим, что на вокзале находится мастерская бытового обслуживания, выполняющая, в частности, срочный ремонт одежды в присутствии заказчика. Пусть в мастерской работает несколько мастеров. Ясно, что число пассажиров, обратившихся за помощью к ним, может превзойти число мастеров и, следовательно, образуется очередь. Предположим, что в этот момент в мастерскую зашел пассажир, брюки или пиджак которого нуждаются в штопке. Если поезд, на котором он едет, отходит в ближайшее время и у пассажира нет времени ожидать своей очереди, то он, пожалуй, предпочтет остаться с дыркой в брюках, нежели опоздать на поезд, и поэтому покинет мастерскую не дожидаясь обслуживания. Представляет интерес ряд вопросов.

Во-первых, нас должна интересовать характеристика полноты удовлетворения потребностей пассажиров, ибо если большинство их останется неудовлетворенным, то вряд ли есть смысл вообще иметь мастерскую на вокзале. Такой характеристикой может быть вероятность того, что пассажир не будет обслужен. Кроме того, с точки зрения рентабельности мастерской, полезно знать и степень загрузки каждого мастера. Эта задача

не является простой, как это может показаться на первый взгляд, так как число пассажиров, обращающихся за помощью в мастерскую, вообще говоря, является случайным — нельзя точно предугадать, какому количеству пассажиров и в какое время потребуется помощь мастерской. Кроме того, время обслуживания одного клиента есть величина случайная, так как оно зависит от характера обслуживания, в котором нуждается пассажир.

При формулировании этого примера мы сознательно сделали оговорку о том, что пассажир спешит и остается в мастерской только при условии, что его обслуживание начнется немедленно и закончится не позже определенного момента, т. е. выходящий поток требований из данной системы обслуживания может состоять как из удовлетворенных, так и из неудовлетворенных заявок на обслуживание.

С точки зрения настоящего примера это условие является несколько искусственным, так как, вообще говоря, не все пассажиры опаздывают на поезд, и большинство из них имеет возможность подождать. Однако это условие является весьма важным, ибо в других задачах подобного типа оно может выполняться более строго. Например, если вместо рассмотренной системы обслуживания заниматься изучением такой, как, скажем, центр связи, состоящий из группы радиоприемников, и предположить, что роль пассажиров играют поступающие в этот центр сообщения от различных источников, то есть все приемники будут заняты, очередное сообщение не будет принято. Следовательно, в этом случае условие потери требования из-за занятости всех обслуживающих аппаратов (приемников) выполняется совершенно строго.

Четвертый пример. Телефонная станция может одновременно обеспечивать разговоры ограниченного числа абонентов. Если количество заявок на обслуживание превысит число линий связи, по которым одновременно могут вестись телефонные разговоры, то каждый очередной абонент, который обратится на станцию в этот момент, будет получать отказ. Любому из читателей приходилось сталкиваться с явлением, когда в часы «пик» в телефонной трубке слышатся частые гудки уже после набора первой цифры или буквы. Это означает,

что станция полностью загружена, хотя тот абонент, с которым нужно связаться, в этот момент свободен.

Специалистов, разрабатывающих и эксплуатирующих автоматические телефонные станции, очень часто интересует, какова вероятность того, что очередной абонент получит отказ, и какова степень загрузки всех линий. В этом примере существенным является то, что поток вызовов, т. е. требований на обслуживание, является случайным. Нельзя заранее точно указать, когда и сколько абонентов будут нуждаться в разговоре по телефону. Временем обслуживания здесь будет продолжительность разговора. Она также не является величиной постоянной и совершенно очевидно, что ее нужно рассматривать как случайную. Следует заметить, что эта задача является одной из первых, решение которых привело к созданию теории массового обслуживания.

Пятый пример. В масштабе очень большого автохозяйства или даже в масштабе автохозяйства всей страны общее число автомашин весьма велико. В процессе эксплуатации по тем или иным причинам автомашины выходят из строя и требуют ремонта. Кроме того, правила эксплуатации автомашин предусматривают проведение профилактических осмотров и ремонтов. Поэтому для решения многих задач, связанных с планированием, важно знать среднее число машин, которые будут находиться в ремонте. Этот пример частично рассматривался в первом параграфе, поэтому здесь мы дополнительно уточним только некоторые положения.

Чтобы определить среднее число машин, нуждающихся в ремонте в данный момент, нужно, во-первых, установить, какое число автомашин выходит из строя за определенный промежуток времени. Эта величина не является постоянной. Она зависит от времени года, от состояния дорог в данном районе, от квалификации водителей, соблюдения графиков планово-предупредительных осмотров и ремонтов и многих других, в том числе и чисто случайных факторов. Поэтому нужно определить ее вероятностные характеристики. Во-вторых, нужно определить время ремонта. Оно также не является величиной постоянной и зависит от многих факторов: от характера поломки или аварии, от оснащенности ремонтной базы, от опытности мастеров и рабочих, производящих ремонт, наличия запасных частей и т. д.

В общем это типичная задача массового обслуживания. Знание потока требований (количества машин, нуждающихся в ремонте и осмотре) и времени обслуживания (ремонта) позволяет решить более строго задачу об определении среднего числа автомашин, находящихся в ремонте. В этом примере существенным является то, что число «требований» (неисправных машин) может быть очень большим.

Шестой пример. Предположим, что в некоторый пункт поступают срочные сообщения. Например, на телеграф — телеграммы «молнии». Правила работы телеграфа предусматривают, что немедленно по получении такой телеграммы почтальон отправляется в путь для доставки ее адресату. Моменты поступления телеграмм заранее не известны, поэтому поток телеграмм можно рассматривать как случайный. Если телеграф большой (скажем Московский Центральный), то число телеграмм может быть очень большим. Практически его можно считать неограниченным, так как в праздники, например, существует возможность такого положения, когда телеграммы из разных мест отправит большое количество людей.

Ясно, что все эти телеграммы должны быть доставлены, при этом время доставки зависит от того, на каком удалении от телеграфа живет адресат, какими видами транспорта может пользоваться почтальон, на каком этаже находится квартира адресата и т. д. Необходимо определить вероятность того, что в процессе доставки одновременно будет находиться данное число телеграмм k, а следовательно, одновременно будет занято k почтальонов. И хотя на телеграфе обычно работает почтальонов гораздо меньше, чем поступает телеграмм, значение этой величины позволит более точно определить действительно необходимое количество почтальонов и тем самым более качественно решить задачу своевременной доставки корреспонденции адресатам. Ясно, что эта задача является также задачей массового обслуживания. Требованием на обслуживание здесь является принятие телеграммы, а обслуживанием — их доставка. В роли обслуживающих аппаратов выступают почтальоны, доставляющие телеграммы.

За этим примером очень легко увидеть ряд подобных с гораздо более глубоким и серьезным содержанием.

Если, например, телеграммы заменить важными сообщениями, а почтальонов — средствами их передачи, то возникает задача о необходимой мощности средств передачи сообщений. Методы решения всех задач такого типа общие, и одна из них будет подробнее разобрана в § 2 гл. 3. Существенным в задачах этого вида является предположение о том, что телеграмма срочная (сообщение важное), и поэтому доставка (передача) ее должна быть начата немедленно, и что общее число сообщений (телеграмм) может быть очень большим.

Седьмой пример. Предположим, что на полевом стане работает некоторое количество сельскохозяйственных машин. Полевой стан находится на значительном удалении от центральных ремонтных мастерских. Пусть мастерские достаточно мощные и не только имеют средства доставки неисправных машин для ремонта, но и обладают возможностью выделить подвижную мастерскую для ремонта вышедших из строя машин прямо в поле. Нужно определить, какой способ организации ремонта лучше выбрать — везти ли неисправные машины в центральные мастерские или развернуть передвижную мастерскую в поле и ремонтировать машины на месте?

На первый взгляд может показаться, что лучше везти неисправные машины в центральную мастерскую: такие мастерские обладают несравненно большей мощностью, чем передвижные, и, следовательно, ремонт каждой поступившей машины будет начат сразу же, как только ее доставят, причем можно не сомневаться в том, что благодаря возможностям центральных мастерских ремонт этих машин будет произведен не только качественно, но, пожалуй, и в более сжатые сроки, чем в полевых условиях. Однако при этом будет затрачено время на доставку машины в мастерскую и обратно. Если же ремонтировать ее в передвижных мастерских, то ремонт в полевых условиях не потребует затрат времени на перевозку, но в силу недостаточной мощности такой мастерской может произойти ее перегрузка и часть неисправных машин будет ожидать, пока закончится ремонт машин, ранее вышедших из строя. Какой же способ ремонта выбрать?

Ясно, что в период уборочной или посевной, как, впрочем, и при выполнении других работ в поле, экономия времени является одним из важнейших факторов,

от которых зависит успех всей работы. Поэтому предпочтительней будет тот способ организации ремонта, при котором время простоя машин будет наименьшим. Эта задача распадается на две задачи типа массового обслуживания, в которых поток требований состоит из неисправных машин, требующих ремонта. И здесь, как и в предыдущем примере, поток требований является случайным, так как нельзя заранее предсказать, сколько машин выйдет из строя в данный момент.

Обслуживание в обеих задачах состоит в устранении поломки. Но в одной задаче обслуживание (ремонт) начинается сразу же, как только требование на обслуживание поступило в обслуживающую систему (в центральную мастерскую), а во второй задаче возможно положение, при котором образуется очередь из неисправных машин. Если внимание сосредоточить не на машинах, а на сложившейся ситуации, то нетрудно понять, что этот пример является частным случаем проблемы о централизованном и децентрализованном обслуживании, основным вопросом которой является определение наивыгоднейшего способа организации.

Восьмой пример. Рассмотрим пример, связанный с ремонтом кораблей (судов). Пусть для ремонта группы кораблей или судов выделено определенное количество доков. Сами корабли (суда) несут определенную службу в заданном районе (лов рыбы, дозор и т. д.).

Время от времени эти корабли нуждаются в ремонте. Но теперь нас интересует не среднее время ремонта или время ожидания каждым кораблем момента начала обслуживания, как это было в ранее рассмотренных примерах, а количество кораблей, которые будут находиться в исправном состоянии. Этот пример также является частным случаем задачи массового обслуживания. Поток требований здесь — это поток неисправных кораблей, а обслуживающая система — ремонтные доки. Число исправных кораблей можно определить, если знать, сколько их находится в ремонте, а это нетрудно, так как если известно число занятых доков (для простоты полагаем, что каждый док в состоянии принять только один корабль), то, стало быть, известно и число ремонтируемых кораблей; прибавив сюда число кораблей, ожидающих ремонта, и вычтя эту сумму из общего числа кораблей, мы получим число исправных кораблей.

Существенным в этой задаче является то, что число неисправных кораблей может превзойти число доков, и тогда образуется очередь из кораблей, ожидающих ремонта, а общее число неисправных кораблей ограниченно (оно не может быть больше общего числа кораблей, которое практически не очень велико). Подробнее эту задачу мы рассмотрим в § 3 гл. 3.

Девятый пример. Предположим, что в небольшом городе имеется только одна мастерская по ремонту телевизоров. Естественно, что у жителей этого города в случае, если телевизор вышел из строя, только один выход — обратиться в эту мастерскую. Общее число телевизоров в городе может быть достаточно велико и, следовательно, очередь из неисправных телевизоров может оказаться весьма значительной. Возникает вопрос о том, сколько времени владелец неисправного телевизора будет ждать окончания ремонта? Это время зависит от потока неисправных телевизоров, характера неисправностей, времени ремонта каждого из них, числа мастеров в мастерской и т. д.

Существенным в этой задаче является то, что число неисправных телевизоров может быть довольно большим, а мастерская одна, т. е. отремонтировать телевизор больше негде. Поэтому, как бы ни была длинна очередь, неисправный телевизор может покинуть мастерскую только после окончания ремонта.

Десятый пример. Автотранспортная контора принимает заявки на перевозку грузов. Так как число заявок на обслуживание (перевозку грузов) может быть очень большим, иной раз гораздо бо́льшим, чем могут обслужить все транспортные средства, принадлежащие данной конторе, то естественно, что необходимо чем-то ограничить прием заявок. Иногда таким ограничением, причем ограничением совершенно естественным, может служить длина очереди, т. е. число заявок, которые ожидают удовлетворения. Если очередная заявка поступает в тот момент, когда очередь достигла определенной величины, она не принимается. На языке теории массового обслуживания она *теряется* или *получает отказ* в обслуживании.

Если осуществляется доставка срочных грузов и, следовательно, заказчик не может ждать, пока его заявка будет удовлетворена, то эта заявка потеряна для конто-

ры. Заказчик постарается изыскать какие-то другие возможности доставки своих грузов по назначению в нужные сроки и, в частности, попытается обратиться в другую контору. Интерес представляет вопрос: какова вероятность того, что при данной мощности автопарка конторы и данном потоке заявок, который, вообще говоря, является случайным, очередная заявка не будет принята на обслуживание? Эта задача усложняется тем, что время доставки грузов, заявки на перевозку которых поступили раньше, является также случайной величиной. Это время зависит от многих факторов: от того, откуда и куда нужно перевести грузы, какое время будет затрачено на погрузку и разгрузку, т. е. от качества груза (габаритов, тары и т. д.) и как осуществляется погрузка, от времени суток и времени года, от качества дороги, от характера движения на ней и т. д. Поэтому нельзя точно указать заранее моменты, когда освободятся машины.

Подробнее эта задача для частного случая решена в § 3 гл. 3.

* * *

Итак, мы познакомили читателя с несколькими примерами из области народного хозяйства. Как видно будет из дальнейшего, при решении этих задач с успехом может быть применена теория массового обслуживания и ее аппарат. Мы рассмотрели всего десять примеров, однако число конкретных задач данного типа неограниченно. Мы специально не уточняли условия задач и ограничивались лишь их общей формулировкой, так как решение большинства из них будет дано в гл. 3, после того как будут рассмотрены основные понятия теории и методы подхода к решению задач.

3. ТЕОРИЯ МАССОВОГО ОБСЛУЖИВАНИЯ И ТЕХНИЧЕСКОЕ ПРОЕКТИРОВАНИЕ

Дальнейший научный и технический прогресс в значительной степени зависит от правильного и высококачественного проектирования технических устройств и различных систем для промышленности. Однако иногда бывают случаи, когда в результате недоучета тех или иных параметров, влияющих на функционирование этих устройств и систем, проектируются устройства и систе-

мы, эксплуатация которых не приносит ожидаемого результата, что приводит к весьма значительным непроизводительным затратам, а в отдельных случаях и к полной их непригодности.

Теория массового обслуживания может оказать большую помощь при проектировании технических устройств и различных систем, имеющих целью удовлетворение массовых потребностей в самом широком смысле этого слова. Правильный выбор параметров таких систем позволит избежать многих узких мест в производственных потоках, неполной загрузки отдельных звеньев обслуживающих систем, обеспечит значительную экономию материальных ресурсов. Применение методов теории позволяет установить, какие результаты могут быть достигнуты при эксплуатации проектируемого устройства еще задолго до его создания. Как и в предыдущем параграфе, рассмотрим некоторые конкретные задачи.

Одиннадцатый пример. При проектировании автоматической телефонной станции (АТС) выбор ее мощности тесно связан с тем потоком вызовов, который данная станция должна будет обслуживать. Пропускная способность АТС всегда ограниченна, поэтому АТС может обеспечить связь лишь между определенным числом абонентов. Если все линии связи АТС заняты, то вновь поступивший вызов получит отказ, даже если вызываемый абонент свободен.

Если запроектировать станцию недостаточной мощности, то пропускная способность ее окажется малой и такая станция будет плохо обслуживать абонентов: будет велика вероятность того, что в момент вызова все линии будут заняты. Если же проектировать АТС слишком большой мощности, то это приведет к возрастанию ее стоимости, усложнению эксплуатации, снижению надежности и т. д. Следовательно, при проектировании АТС необходимо всегда так рассчитывать ее мощность, чтобы вероятность отказа не превосходила заданной, и в то же время АТС «укладывалась» бы в другие требования, которые к ней предъявляются при проектировании.

В сущности эта задача мало чем отличается от рассмотренной в четвертом примере, естественным развитием которого она является. Там задача заключалась в оценке качества функционирования существующей си-

стемы по известным ее параметрам, а здесь задача заключается в определении параметров проектируемой обслуживающей системы исходя из требований к качеству обслуживания. В этом смысле данный пример является обратным предыдущему. Для решения этой задачи необходимо оценить ожидаемый поток вызовов на обслуживание, определить его характеристики. После этого необходимо задаться допустимой вероятностью того, что абонент в момент вызова найдет все линии связи занятыми, т. е. получит отказ. Это условие и является оценкой качества обслуживания. Затем нужно установить зависимость между этой вероятностью и пропускной способностью (мощностью) АТС. А из нее уже без особого труда можно будет найти и необходимую мощность АТС.

Двенадцатый пример. При проектировании энергосистем возникает весьма важный вопрос: какой суммарной мощностью должна обладать энергосистема для того, чтобы успешно обслуживать всех потребителей данного района? Как решить эту задачу?

На первый взгляд она не кажется сложной. Нужно просуммировать мощности всех потребителей электроэнергии, находящихся в данном районе, и получится суммарная мощность энергосистемы. Но правильно ли это? Конечно, нет. Ведь каждый из потребителей не все время потребляет одинаковую мощность. Например, городской транспорт (трамвай, метро, троллейбус) наибольшую мощность потребляет в часы «пик», днем, а вечером движение постепенно стихает и в течение определенного промежутка времени фактически прекращается совсем. Многие заводы и фабрики ночью не работают. Зато вечером электроэнергия необходима для освещения квартир и улиц.

Примеров неравномерного потребления электроэнергии можно привести очень много. Поэтому решать задачу определения необходимой суммарной мощности проектируемой системы простым суммированием необходимых мощностей каждого потребителя нельзя, так как они не обязательно нуждаются в ней одновременно. В зависимости от географического расположения района, для которого проектируется данная энергосистема, характера производственного цикла, предприятия и другие потребители будут подавать заявки на электроэнер-

гию в различное время. При этом время, в течение которого каждый потребитель будет пользоваться электроэнергией, также не является величиной постоянной. Читатель спросит: причем же здесь теория массового обслуживания? Где же здесь обслуживающие аппараты? Что является обслуживающей системой и как, в каком виде в нее поступают заявки на обслуживание?

Действительно, на первый взгляд эта задача имеет лишь косвенное отношение к теории массового обслуживания. Не видно «обслуживающих аппаратов», не вырисовывается четко структура обслуживающей системы. Однако и здесь можно попытаться применить аппарат теории массового обслуживания и, как будет видно из дальнейшего изложения, небезуспешно. Ведь заявкой на обслуживание можно считать требование на электроэнергию, т. е. необходимость в ее получении тем или иным потребителем.

Условно обслуживающим аппаратом можно считать ту часть мощности, которая идет на удовлетворение этой заявки. Если условиться, что все заявки удовлетворяются одной и той же мощностью электроэнергии, то число обслуживающих аппаратов можно считать равным отношению всей мощности энергосистемы к той мощности, которая необходима для удовлетворения одной заявки. При больших мощностях энергосистемы это число может быть очень большим. Один из случаев решения этой задачи более подробно будет рассмотрен в § 2 гл. 3.

Тринадцатый пример. При проектировании технических устройств и систем немаловажное значение имеет правильный выбор элементов, из которых состоят эти системы и устройства. При этом правильный выбор элементов не является тривиальной задачей. На первый взгляд может показаться, что при решении этой задачи всегда выгоднее брать более надежные элементы. Но может оказаться, что такой подход не оправдывает себя с экономической точки зрения. Ведь наиболее надежные элементы обычно являются самыми дорогими, поэтому может получиться так, что тот эффект, который обеспечат эти элементы, не окупит произведенных затрат. И, наоборот, бывают случаи, когда надежность проектируемой системы в процессе ее последующей эксплуатации имеет настолько важное значение, что проектиров-

щики вынуждены идти на довольно значительные материальные затраты. Так, например, при проектировании электронных вычислительных машин выбор элементов с высокой степенью надежности имеет особое значение.

В электронной вычислительной машине выход из строя одного элемента арифметического устройства или устройства управления означает, как правило, выход из строя всей машины. Это влечет за собой значительные затраты времени на отыскание и устранение неисправного элемента, поэтому необходимо обеспечивать возможно большую надежность каждого элемента. В связи с этим возникает необходимость в обосновании экономической целесообразности производства значительных затрат на повышение надежности элементов.

Эту задачу можно также отнести к классу задач типа массового обслуживания. Требованием на обслуживание здесь является выход из строя элемента. «Обслуживающей системой» является группа инженеров и техников, осуществляющая эксплуатацию электронной вычислительной машины (смена). Практика показывает, что, как правило, в случае поломки вся смена занимается отысканием и устранением ее, поэтому можно считать, что одновременно такая «обслуживающая система» обрабатывает только одно требование, и, следовательно, она состоит из одного обслуживающего аппарата. Поток требований, равно как и время обслуживания, здесь, так же как и в предыдущем случае, является случайной величиной. Один из примеров решения задач такого типа более подробно будет рассмотрен в § 2 гл. 3.

Четырнадцатый пример. При проектировании производственных поточных линий и предприятий в целом, одной из характерных задач является задача по ликвидации «узких мест» производственного цикла. Почти всегда при сочленении различных устройств возникает задача правильной организации технологического процесса.

Ясно, что если одно звено производит продукции больше, чем может переработать следующее, или, наоборот, второе звено недогружено в связи с недостаточной производительностью первого, то процесс организован не наилучшим образом. Необходимо так организовать процесс, чтобы устройства во всех звеньях были загружены полностью и равномерно.

В более сложном случае возникает вопрос о том, какое количество устройств одного типа необходимо закрепить за устройствами другого типа, чтобы обеспечить полную их загрузку. Так, например, при проектировании заводов, производящих радиоэлектронную аппаратуру, т. е. продукцию, нуждающуюся в контроле на испытательных стендах, важно правильно определить необходимое число таких стендов. Если завод большой, то поток готовой продукции, даже при конвейерном производстве, в какой-то степени является случайной величиной.

Часто бывает так, что нельзя точно предсказать моменты поступления готовой продукции и даже точного числа готовых изделий. Время контроля этих изделий иногда является случайной величиной, которая зависит от результатов контроля. Например, если завод производит радиоприемники, то в зависимости от разброса параметров радиодеталей после того, как приемники сошли с конвейера, необходимо «довести» эти параметры до установленных величин, т. е. настроить каждый приемник, если его характеристики не соответствуют техническим условиям.

Этот процесс по времени может быть различным, так как зависит не только от разброса тех или иных параметров, но и от квалификации рабочего, выполняющего настройку, его знаний, опыта, сноровки и даже настроения. Следовательно, предсказать, сколько потребуется времени на выполнение этой операции, невозможно. Таким образом, здесь возникает типичная задача массового обслуживания, в которой испытательные стенды играют роль обслуживающих аппаратов, а готовая продукция образует поток требований на обслуживание.

Если неправильно определить число обслуживающих аппаратов — испытательных стендов, то это или создаст задержку готовой продукции и ухудшит ее контроль в случае, когда их мало, или приведет к излишним затратам на их установку и содержание в случае, когда их много. Поэтому это число всегда необходимо определять таким образом, чтобы время ожидания начала испытания не превосходило определенной величины (было наименьшим) и чтобы при этом загрузка испытательных стендов была достаточно полной. Нетрудно заметить, что эта задача имеет много общего с задачами сопряжения

групп разнотипных станков в производственных линиях, планированием связи цехов и заводов-поставщиков с цехами и заводами-потребителями и т. д.

Более подробно один из примеров такого типа будет разобран в § 2 гл. 3.

Пятнадцатый пример. При проектировании различных управляющих систем и, в частности, при проектировании электронных вычислительных машин для управления различными процессами большое значение имеет правильное определение параметров этих систем, т. е. их быстродействия, разрядности, объема памяти и т. п. Если с помощью электронной вычислительной машины осуществляется управление конкретным процессом, то в нее непрерывно поступают все необходимые сообщения о ходе этого процесса. Часть этих сведений сразу же перерабатывается машиной, а часть поступает в ее память и направляется на переработку несколько позднее. В зависимости от того, насколько велик объем этой информации, будет определяться и объем запоминающего устройства вычислительной машины.

Если память машины будет больше, чем нужно, то это приведет к повышению потребляемой энергии, увеличению габаритов машины и ее стоимости, а если она будет по своему объему меньше, то такая машина просто не будет справляться с возложенной на нее задачей. Поэтому в процессе проектирования необходимо определить, какой объем памяти нужно иметь для запоминания и хранения всех поступающих в нее сообщений. Если, например, машина может одновременно обрабатывать только одно сообщение, то все последующие сообщения, поступившие в то время, когда машина занята, должны быть записаны в буферную память, откуда потом они будут поступать в машину.

В ряде процессов каждое сообщение может содержать больше или меньше сведений и, следовательно, будет иметь различное число знаков, поэтому время обработки каждого поступившего сообщения будет различным. Моменты поступления сообщений также могут быть случайными. Возникает задача о том, какого объема буферную память необходимо запроектировать для того, чтобы при данном потоке сообщений и известном времени обработки каждого из них вероятность потери сооб-

щения из-за ограниченного объема памяти была бы мала (не больше заданной или допустимой величины).

Очевидно, что эта задача также относится к классу задач массового обслуживания. Для ее решения необходимо уметь определять длину очереди, состоящей из сообщений, и вероятность того, что эта длина не превзойдет определенной величины.

Интерес могут представлять некоторые видоизменения этой задачи. Так, например, некоторые сообщения представляют ценность только в течение определенных отрезков времени. Если в течение определенного времени такое сообщение не было использовано, то его можно отбросить, так как его содержание обесценилось. Предположим, что электронная машина управляет посадкой космического корабля. Тогда все сообщения о его состоянии, обработка которых задержалась больше, чем на определенный промежуток времени, теряют цену. Ну а те, которые не были переработаны до момента посадки, могут быть отброшены совсем.

Исходя из этого, можно определить такой параметр, как скорость работы машины при переработке необходимой информации, т. е. ее быстродействие. Это быстродействие должно быть таким, чтобы вероятность обесценивания сообщения из-за того, что машина не успеет его переработать, была бы мала. Ответить на этот вопрос можно в том случае, когда выявлено соотношение между временем ожидания начала переработки сообщения и скоростью работы электронной вычислительной машины. Подробнее один из примеров такого типа будет рассмотрен в § 2 гл. 3.

Шестнадцатый пример. При проектировании различных производственных агрегатов, а также при планировании их работы нередко возникает задача, решение которой связано с обеспечением различными видами обслуживания потока требований, поступающего на группу последовательно расположенных обслуживающих аппаратов.

Рассмотрим следующий пример. В упаковочный цех поступает готовая продукция. Упаковку осуществляют автоматы, расположенные вдоль одного конвейера. Первое готовое изделие поступает в первый упаковочный автомат. Если второе изделие поступает сразу за первым и упаковка первого еще не закончена, то оно поступает

на следующий по порядку автомат и так далее... Короче говоря, каждое очередное готовое изделие поступает на ближайший свободный упаковочный автомат, т. е. мы имеем дело с упорядоченной организацией системы обслуживания, в которой роль обслуживающих аппаратов играют упаковочные автоматы. При этом «требование», т. е. изделие, нуждающееся в упаковке, не может покинуть систему не будучи удовлетворенным (все изделия обязательно должны быть упакованы).

Следовательно, когда все автоматы заняты, изделие, поступившее в этот момент, нужно снова вернуть на упаковочную линию в том случае, если нет специального устройства, автоматически останавливающего движение конвейера с готовой продукцией. При проектировании таких линий большой интерес представляет вопрос о полной загрузке всех автоматов. Общее их число нужно выбирать таким, чтобы обеспечить быструю упаковку готовой продукции. Если, например, окажется, что при выбранной организации потребуется много автоматов, то, возможно, нужно искать другую организацию: расположить автоматы параллельно и распределять изделия между ними равномерно.

В этом случае изделие, поступившее в момент, когда данный автомат занят, будет ждать своей очереди. При такой организации можно обеспечить экономию в оборудовании за счет увеличения времени ожидания начала обслуживания (упаковки). Можно прибегнуть и к другим способам организации, но при этом всегда нужно так выбирать число автоматов, чтобы они обеспечивали полное обслуживание готовой продукции.

Определение необходимых характеристик при решении этих задач, таких как вероятность того, что все автоматы будут заняты в первом случае и среднего времени ожидания начала упаковки во втором — можно осуществить с помощью методов теории массового обслуживания. Следовательно, обе эти задачи являются задачами массового обслуживания.

4. ИСПОЛЬЗОВАНИЕ ТЕОРИИ МАССОВОГО ОБСЛУЖИВАНИЯ В ВОЕННОМ ДЕЛЕ

Рассмотренные в предыдущих параграфах примеры не могут не натолкнуть военного читателя на мысль, что теория массового обслуживания может найти довольно широкое применение и в военном деле. Действительно,

военные специалисты без особого труда за каждым примером § 2 и 3 могут увидеть соответствующие военные ситуации, в которых можно использовать аналогичный подход к решению многих задач, с которыми они сталкиваются в практической деятельности.

Не вдаваясь в излишние подробности, остановимся на основных областях военного дела, в которых применение теории массового обслуживания может оказаться полезным. Имеется три основных направления приложения этой теории, представляющих весьма существенный интерес для военных специалистов:

1. Проблемы, связанные с организацией всевозможных военных систем обслуживания в целях построения оптимальной системы в каждом конкретном случае.

2. Проблемы, связанные с организацией управления силами в бою.

3. Проблемы, связанные с использованием методов теории массового обслуживания при математическом моделировании процессов боевых действий.

Рассмотрим коротко каждое из этих направлений.

Задачи, связанные с организацией систем обслуживания, в целях построения оптимальной системы

К таким задачам относятся способы организации обслуживания боевой техники (как в процессе ведения боевых действий, так и в мирное время), способы построения систем ремонтных (судоремонтных) мастерских, организация систем медицинского обслуживания в боевых условиях и т. п. Подобные проблемы часто встречаются не только в вопросах, связанных с использованием оружия, обслуживанием боевой техники и вспомогательных средств, но и в задачах, связанных с планированием боевых действий. Так, например, задачу отражения налета бомбардировочной авиации силами противовоздушной обороны можно рассматривать как задачу такого типа. Действительно, зенитные, артиллерийские или ракетные установки, как правило, обладают ограниченными возможностями при отражении налета, так как число их ограниченно и зачастую меньше числа бомбардировщиков, участвующих в налете.

Процесс стрельбы по бомбардировщикам можно рассматривать как процесс обслуживания, а поступление

данных о целях (бомбардировщиках) — процессом поступления требований на обслуживание.

Аналогичный характер имеет задача противоракетной обороны. Совершенно очевидно, что оценка качества функционирования такой обслуживающей системы, как система ПВО, представляет очень большой интерес и имеет не меньшее практическое значение. Но оценка качества функционирования, т. е. способности отражать налет бомбардировщиков или ракет, не будет достаточно полной, если она будет производиться только чисто качественным путем. Поэтому при решении подобной задачи неизбежно привлечение количественных методов.

При количественном изучении этого процесса возникает ряд очень интересных и важных вопросов, имеющих как практическое, так и теоретическое значение. Так, например, большое практическое значение имеют вопросы о том, каким образом организовать систему обороны так, чтобы при данном качестве отдельной зенитной установки добиться наибольшей защищенности охраняемых объектов; какое число бомбардировщиков (ракет) при данной организации не будет уничтожено; какое число бомбардировщиков (ракет) не подвергнется воздействию сил обороны и т. д. При получении этих характеристик большую помощь могут оказать методы теории массового обслуживания.

К числу задач подобного типа относятся и вопросы, связанные с посадкой группы самолетов на один или несколько аэродромов. Обслуживанием в этом случае будет обеспечение требования на посадку самолета. При проектировании аэродромов, в зависимости от планируемой загрузки, нужно определить необходимое количество посадочных полос. Их число должно быть определено из условия, что время ожидания момента посадки не превзойдет определенной величины. При эксплуатации военных и гражданских аэродромов как в мирное, так особенно в военное время, всегда необходимо уметь при данном числе посадочных полос так спланировать полеты, чтобы загрузка посадочных полос была наибольшей, а очередь самолетов, ожидающих посадки, и, следовательно, среднее время ожидания посадки, наименьшими. Изучить эти вопросы и получить нужные количественные характеристики можно с помощью методов теории массового обслуживания.

В остальных задачах, таких, как организация ремонта боевой техники, обеспечение связи и т. п., использование методов теории массового обслуживания почти ничем не будет отличаться от применения их в таких «гражданских» задачах, как ремонт автомашин, организация связи на АТС и других. Так, например, совершенно аналогичные задачи по сравнению с некоторыми типами задач, приведенных в § 2 и 3, возникают на судоремонтных заводах и в ремонтных мастерских при ремонте кораблей, оружия и боевой техники; на аэродромах при профилактическом осмотре и ремонте самолетов. Для того чтобы обеспечить постоянную боеготовность оружия и техники, необходимо иметь достаточно мощные ремонтные мастерские, заводы. Мощность таких мастерских будет находиться в прямой зависимости от количества техники, нуждающейся в ремонте, и скорости ее выхода из строя.

Как правило, перед ремонтными заводами (мастерскими) стоит задача обеспечить нужное количество исправных единиц. Если мощность мастерских окажется недостаточной, то они не будут справляться с ремонтом и образуется очередь из неисправной техники со всеми вытекающими из этого положения последствиями, а если мощность их будет избыточной, то они не будут полностью загружены и, следовательно, будут иметь место излишние затраты средств. Задача состоит в правильном определении мощности ремонтных заводов (мастерских). Ясно, что она ничем не отличается от подобных задач, возникающих в промышленности.

Можно еще указать, что аналогичные задачи возникают и в полевых госпиталях, санитарных пропускниках и обмывочных пунктах. При этом возникает потребность так организовать систему обслуживания (систему полевых госпиталей, санпропускников, обмывочных пунктов и т. п.), чтобы в допустимые сроки пропустить через эти системы всех, нуждающихся в помощи. Особую роль, в связи с возможностью применения ядерного оружия и, следовательно, радиоактивного заражения, приобретает вопрос о правильной организации системы обмывочных пунктов. Эвакуация раненых и пораженных БРВ из зоны очага массового поражения также может рассматриваться как своего рода массовое обслуживание. Во всяком случае совершенно очевидно, что все эти задачи

относятся к классу задач массового обслуживания и несомненно, что применение количественных методов при их решении позволит значительно улучшить качество обслуживания и получить определенный тактический и оперативный эффект при существенном сокращении материальных и других затрат.

Следует также отметить, что решение многих задач, связанных с организацией всевозможных систем массового обслуживания в военном деле, в целях построения оптимальной системы в каждом конкретном случае, необходимо не только при использовании уже готовых образцов оружия и боевой техники, но и на стадии проектирования новых образцов оружия, а также при разработке новых организационных форм управления применительно к новым условиям ведения военных действий.

Проблемы, связанные с организацией управления силами в бою

К этим проблемам относятся задачи, связанные с организацией систем по переработке информации. При более глубоком исследовании оказывается, что даже такой сложный динамический процесс, как бой, в какой-то степени содержит элементы массового обслуживания. Если, например, требованием на обслуживание считать появление тех или иных сил противника, а обслуживанием — нанесение соответствующего удара по этим силам, то возникает ситуация типа массового обслуживания. Обслуживающей системой в этом случае являются наши силы. При таком, на первый взгляд, неожиданном подходе можно ожидать некоторых успехов в количественном описании боя на основе идей и с помощью аппарата теории массового обслуживания.

Как известно, процесс управления силами в бою основывается на целенаправленной переработке информации о силах противника, его тактике, данных о своих силах, их возможностях и данных об обстановке, в которой будет протекать бой.

В настоящее время все большее значение приобретают количественные методы, используемые при обработке этой информации, и, в частности, количественные методы оценки качества управления, которые требуют

разработки специальных критериев эффективности. Эти критерии должны обеспечить выбор из различных способов управления наилучших или оптимальных способов из всех возможных. Теория массового обслуживания может оказать большую помощь при построении таких критериев эффективности.

Для задач, связанных с передачей и переработкой информации (какими, по сути дела, являются задачи управления), при рассмотрении их с точки зрения теории массового обслуживания, требованием на обслуживание можно условно считать поступление новой информации, а обслуживанием — ее передачу или переработку. Эти задачи имеют большое значение в связи с передачей ряда функций по переработке информации, для улучшения качества управления, автоматическим устройствам, в основном быстродействующим электронным вычислительным машинам. При этом возникает задача о создании рациональных систем * по переработке информации на различных ступенях управления, которая также может быть отнесена к классу задач массового обслуживания и успешно решена при использовании методов и аппарата теории массового обслуживания.

Проблемы, связанные с использованием методов теории массового обслуживания при математическом моделировании процессов боевых действий

Оценка различных тактических приемов при использовании новых видов оружия и боевой техники, при взаимодействии разнородных сил, и выявление закономерностей в процессах боевых действий возможны различными способами. Можно, например, проводить эту исследовательскую работу в ходе войны, но тогда эта «наука» потребует ненужных жертв. Можно также проводить эти исследования с помощью специальных военных игр и учений, но это требует больших материальных затрат и поэтому эксперименты не могут быть достаточно массовыми, а следовательно, не дадут ожидаемого эффекта в полной мере.

---

* Здесь под рациональной понимается такая система машин, которая обеспечивает надежную переработку всего объема поступающей информации в допустимые сроки и не является при этом избыточной (т. е. нерентабельной, работающей вхолостую, не на полную мощность).

Новый путь, который за последнее время находит все более широкое применение в армиях и флотах многих стран благодаря появлению быстродействующих электронных вычислительных машин, — это путь математического моделирования процессов боевых действий с последующим многократным розыгрышем на электронных вычислительных машинах. Этот путь обеспечивает получение нужных характеристик изучаемого процесса.

Коротко содержание математического моделирования заключается в следующем. Определяются основные параметры, от которых зависит окончательный результат боя. Таких параметров может быть в каждом конкретном случае довольно много. К ним относятся, например, тактико-технические данные оружия и боевой техники своих сил и сил противника, географические и метеорологические условия, в которых ведутся боевые действия, и многие другие. Затем определяются основные законы, которым подчиняется изменение всех или, по крайней мере, основных параметров. После этого устанавливается зависимость результатов боя от этих величин и способов действий (тактики). Все это описывается с помощью соответствующего математического аппарата (т. е. строится математическая модель исследуемого процесса), программируется и многократно проигрывается на электронных вычислительных машинах. Результаты каждого розыгрыша фиксируются, и после этого все полученные результаты соответствующим образом обрабатываются и анализируются. Одним из преимуществ этого метода является его низкая стоимость. Что же касается точности этого метода, т. е. степени приближения полученных результатов моделирования к реальным результатам, то она зависит от того, насколько точно будут определены основные параметры, влияющие на течение процесса, законы, которым они подчиняются, и степень влияния их на изучаемый процесс.

Этот короткий и далеко не полный перечень задач, связанных с использованием боевой техники, организацией вспомогательных органов, с управлением силами в бою и моделированием боевых действий, показывает, что область применения методов теории массового обслуживания в военном деле весьма обширна, поэтому изучение этой теории несомненно окажет пользу военным специалистам в их практической деятельности.

ГЛАВА ВТОРАЯ

ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ МАССОВОГО ОБСЛУЖИВАНИЯ

1. ВХОДЯЩИЙ ПОТОК (ПОТОК ТРЕБОВАНИЙ)

Целью функционирования всякой обслуживающей системы является удовлетворение заявок (требований) на обслуживание. Поэтому *поток требований* является одним из основных и наиболее важных понятий теории массового обслуживания. Как уже было отмечено в § 1 гл. 1, потоком требований (входящим потоком) называется совокупность заявок на обслуживание, поступающих в обслуживающую систему.

Изучение потока требований является первой задачей, которая неизбежно возникает как при теоретической разработке проблем массового обслуживания, так и при практическом применении ее методов к решению конкретных задач. Это нетрудно понять, если представить себе мысленно любую задачу типа массового обслуживания. Какая бы цель перед нами ни стояла, во всех случаях для того чтобы предпринять какие-то конкретные шаги по реорганизации обслуживающей системы с целью улучшения качества ее функционирования, мы всегда должны сначала самым тщательным образом изучить поток требований, поступающих в эту систему. Тем более важно уметь описывать поток требований количественно. Целью дальнейшего изложения и является отыскание математических методов, которые позволяют это делать.

Если выбрать некоторый момент времени $t_0=0$ за начальный, то в ряде процессов нельзя или, по крайней мере, довольно трудно точно предсказать момент поступления следующего требования, а также моменты поступ-

ления всех следующих за ним требований. Так, например, если рассматривать заявки на ремонт, поступающие в автомастерскую от владельцев неисправных автомобилей, и за начальный момент ($t_0=0$) брать момент открытия мастерской, то ясно, что если и случится так, что первая заявка всегда будет поступать в момент открытия, то моменты поступления всех последующих заявок сегодня, завтра, послезавтра не будут совпадать. Иначе говоря, время поступления каждой заявки, как и количество заявок в течение дня (автомашины ведь не выходят из строя точно по графику), есть величины случайные, т. е. такие, которые под влиянием случайных обстоятельств могут принимать различные значения.

Процесс поступления заявок на обслуживание есть случайный процесс. Поток требований может быть описан некоторой функцией $X(t)$, определяющей число требований, нуждающихся в обслуживании за промежуток времени $(0, t)$. Функция $X(t)$ есть случайная величина для каждого значения t. Действительно, если мы выберем промежутки времени даже одинаковой продолжительности, то и в этом случае мы не можем быть уверены, что в каждый из этих промежутков времени поступит одинаковое число требований. Ведь за данный промежуток времени $(0, t)$ может и не поступить ни одного требования, а может поступить 1, 2, ..., n требований. Но какой бы продолжительности промежутки времени мы ни выбирали, не может быть такого положения, что в течение этого промежутка поступит 1,5 требования; 2,3 требования и т. д.

Таким образом, особенностью случайной величины, описываемой функцией $X(t)$, для всякого значения t является то, что она может принимать только целочисленные значения 0, 1, 2, ..., k, где k — целое число.

Очевидно, что число требований, поступивших за промежуток времени $(0,t)$, зависит от величины этого промежутка, т. е. от значения t. Так, весьма вероятно, что, например, за минуту в мастерскую не поступит ни одного заказа. Но одновременно очень мала вероятность того, что за час, и тем более за весь день, также не поступит ни одного заказа на обслуживание. Поэтому функция $X(t)$, определяющая число требований, поступающих за время t, зависит от параметра t и, следовательно, является однопараметрическим семейством слу-

чайных величин. Читателям, знакомым с теорией случайных функций, известно, что такое семейство является *случайной функцией*. Эта случайная функция принимает только целые неотрицательные значения при любых значениях t (t не может быть меньше нуля) и с возрастанием t не убывает. Действительно, число требований, нуждающихся в обслуживании и поступающих

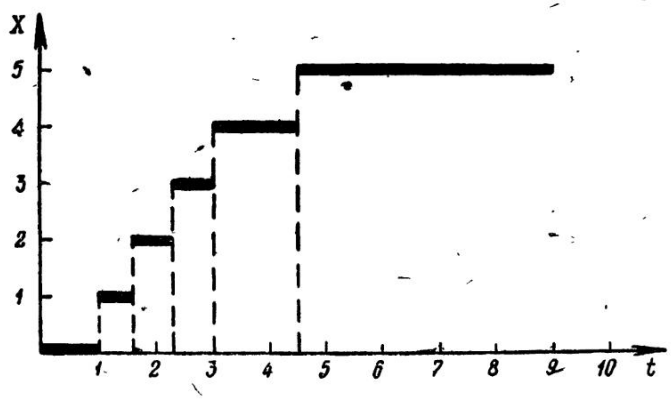

Рис. 2. График реализации случайной функции $X(t)$.

в некоторую систему, может быть только целым положительным и с течением времени не может убывать.

Если осуществить несколько опытов и в каждом регистрировать значения $X(t)$, то полученные при этом функции, как правило, не будут совпадать. Пусть $X_0(t)$ — функция, образованная значениями $X(t)$ в данном опыте. Эта функция уже не является случайной. Она называется *реализацией случайной функции* $X(t)$ в данном опыте.

На рис. 2 изображена одна из реализаций случайной функции $X(t)$. Так, если считать, что на рис. 2 изображен график поступления заказов в мастерскую (на оси t отложено время в часах, а на оси X — число заказов), то этот график означает следующее. После открытия мастерской в течение часа не поступило ни одного заказа. За второй час поступило два заказа, один вначале, а второй минут через тридцать. За третий час поступил один заказ. В начале четвертого часа работы поступил еще один заказ. Последний, пятый, заказ поступил в середине пятого часа работы. Это, конечно,

не означает, что по такому закону заказы будут поступать каждый день. Поэтому функция и называется реализацией случайной функции.

Говоря более строго, в данном случае реализацией случайной функции является «неслучайная» функция одного аргумента времени. Для полного описания случайной функции практически невозможно определить все ее реализации, так как их может быть бесчисленное множество. Поэтому используют другой способ ее задания. Случайная функция $X(t)$ будет полностью определена, если для любых положительных промежутков времени $t_1, t_2, \ldots, t_n$, мы можем указать число требований, поступивших за каждый из этих промежутков. Но как было указано выше, число требований, поступивших за любой из этих отрезков времени, есть величина случайная. Следовательно, нужно уметь характеризовать случайные величины. Как известно, полная характеристика случайной величины дается законом распределения*. Но нам нужно знать одновременно поведение функции $X(t)$ за промежутки времени продолжительностью $t_1, t_2, \ldots, t_n$. Поэтому необходимо дать характеристику группы случайных величин $X(t_1), X(t_2), \ldots, X(t_n)$. Такой характеристикой является n-мерный закон распределения группы случайных величин

$$X(t_1),\ X(t_2),\ X(t_3), \ldots,\ X(t_n).$$

Но функция $X(t)$ может принимать только целые положительные значения, поэтому она может быть задана более просто. Для полного определения потока требований достаточно знать, какова будет вероятность того, что за время $(0, t_1)$ поступит k_1 требований, за время $(0, t_2)$ поступит k_2 требований и т. д. Если эта вероятность будет известна для любой группы целых положительных $k_1, k_2, \ldots, k_n$ и положительных $t_1, t_2, \ldots, t_n$, то поток требований будет полностью описан. Эту вероятность мы будем обозначать через

$$P\{X(t_1)=k_1,\ X(t_2)=k_2, \ldots,\ X(t_n)=k_n\}.$$

---

* Законом распределения некоторой случаной величины γ, точнее интегральным законом, является функция $F(t)$, показывающая, какова вероятность того, что $\gamma<t$, где t — произвольное число, т. е. $F(t)=P\{\gamma<t\}$. Здесь $P\{\gamma<t\}$ обозначает вероятность неравенства $\gamma<t$.

Очевидно, что эта вероятность может быть отлична от нуля только в том случае, если при $t_1<t_2<\ldots<t_n$ величины k_i $(i=1, 2, 3,\ldots, n)$ удовлетворяют условию $k_1 \leqslant k_2 \leqslant \ldots \leqslant k_n$.

Это утверждение вытекает из того, что функция $X(t)$ не убывает с возрастанием t. Знание функции

$$F(t_1, t_2, \ldots, t_n; k_1, k_2, \ldots, k_n) = \\ = P\{X(t_1)=k_1, X(t_2)=k_2, \ldots, X(t_n)=k_n\}$$

для любых $t_1, t_2, \ldots, t_n$ и $k_1, k_2, \ldots, k_n$ полностью определяет поток требований. Зная эти вероятности, мы всегда сможем ответить на любой вопрос о потоке требований и определить любую его характеристику. В частности, можно определить вероятность того, что за промежуток времени $(0, t)$ поступит точно k $(k=0, 1, 2, \ldots, n)$ требований. Вероятность этого будет равна

$$F(t, k) = P\{X(t)=k\}.$$

Так, например, вероятность того, что за время t не поступит ни одного требования, равна

$$F(t, 0) = P\{X(t)=0\}.$$

Можно определить, например, вероятность того, что в течение суток в исследуемую систему обслуживания каждый час будет поступать только одно требование. Эта вероятность равна

$$F(1, 2, 3, \ldots, 24; 1, 2, 3, \ldots, 24) = \\ = P\{X(1)=1, X(2)=2, \ldots, X(24)=24\}.$$

Напомним, что все это можно определить при условии, что функция $F(t_1, t_2, \ldots, t_n; k_1, k_2, \ldots, k_n)$ известна. Но задача отыскания такой функции в общем случае является весьма трудной.

Таким образом, принципиально может быть описан любой поток требований. Но, как видно, это описание не является простым и достаточно удобным. Изучение процессов массового обслуживания при таком описании потока требований является весьма трудной задачей.

Часто на практике встречаются потоки, обладающие свойствами, позволяющими найти более простые способы

их описания. Так, многие потоки требований обладают свойством *стационарности. Стационарными* являются потоки, для которых вероятность поступления определенного количества требований в течение определенного промежутка не зависит от начала отсчета времени, а зависит от длины промежутка. Строго говоря, поток называется стационарным *, если закон распределения группы случайных величин

$$X(t_1),\ X(t_2),\ldots,\ X(t_n)$$

совпадает с законом распределения

$$X(t_1+a)-X(a),\ X(t_2+a)-X(a),\ldots,\ X(t_n+a)-X(a),$$

т. е. распределение случайных величин зависит от t_1, t_2, ..., t_n и не зависит от величины a, где a — любой произвольный отрезок времени. Как частный случай из этих рассуждений вытекает, что для стационарных потоков

$$P\{X(t)=k\}=P\{X(t+a)-X(a)=k\},$$

где $(k=0,\ 1,\ 2,\ldots,\ n)$, т. е. вероятность того, что ровно k требований будет получено за промежуток времени $(0, t)$, равна вероятности получения k требований за промежуток времени $(a,\ a+t)$ при любом значении a.

Нетрудно понять, что наличие свойства стационарности значительно облегчает изучение потока требований. Действительно, если известен характер потока требований, поступающих в обслуживающую систему с некоторого начального момента $t_0=0$, то для того чтобы получить характеристики потока начиная с момента $t=a$, нет необходимости изучать этот поток заново. Можно воспользоваться характеристиками потока, полученными ранее. Число требований, поступающих в систему обслуживания после момента $t=a$, т. е. $X(t+a)-X(a)$, при наличии свойства стационарности будет подчиняться тому же закону, что и $X(t)$.

---

* Стационарный поток требований, рассматриваемый в теории массового обслуживания, отличается от стационарного процесса в том смысле, в котором он определен в теории случайных функций. С точки зрения теории случайных функций стационарный поток является процессом со стационарными приращениями.

Свойством стационарности обладают многие реальные потоки требований. Свойством стационарности может обладать, например, поток требований, образованный станками, требующими внимания обслуживающего их рабочего. Стационарным можно считать и поток вызовов, поступающих на автоматическую телефонную станцию. Правда, в течение суток режим работы АТС может меняться в значительных пределах, поэтому поток вызовов следует считать стационарным лишь на отдельных отрезках времени в течение суток.

В некоторых реальных потоках число требований, поступивших в систему после произвольного момента времени t, не зависит от того, какое число требований поступило в систему до момента t. Это свойство независимости характера потока требований от числа ранее поступивших требований и моментов времени их поступления носит название *отсутствия последействия*. Более строго, поток требований называется потоком *без последействия* в тех случаях, когда закон распределения группы $X(t_i+a)-X(a)$ $(i=0, 1, 2, \ldots, n)$ при $t_i>0$ и любом $a>0$ не зависит от значений величины $X(t)$ при $t<a$. В частности, условная вероятность поступления k требований за промежуток времени $(a, a+t)$ при предположении, что количество требований, поступивших в систему до a, будет любым, совпадает с безусловной вероятностью этого события.

Свойством отсутствия последействия обладают многие реальные потоки. Так, например, поток вызовов на АТС является потоком без последействия ибо, как правило, очередной вызов поступает независимо от того, когда и сколько вызовов было до этого момента.

Можно считать, что этим свойством обладает и поток требований на ремонт неисправного оборудования (например, станков). Правда, в связи с тем, что количество станков, как правило, ограниченно, такой поток обладает свойством отсутствия последействия только условно. При большом числе единиц оборудования это свойство можно считать хорошо выполняющимся, при малом — оно теряет силу.

В некоторых случаях и поток вызовов на АТС теряет свойство отсутствия последействия. Может оказаться, что один звонок, содержащий какой-нибудь важный приказ, повлечет за собой множество других звонков. Или,

скажем, такие приятные сообщения, как запуск нового космического корабля, переданные из одной инстанции в другую, вызывают целую лавину телефонных звонков. Но, как правило, в потоке вызовов на АТС заметное последействие отсутствует.

Обозначим $P\{X(t)=k\}=V_k(t)$, где $k=0, 1, 2, \ldots$ Иными словами, $V_k(t)$ есть вероятность того, что за промежуток времени $(0, t)$ при $t>0$ поступит точно k требований.

Оказывается, стационарный поток без последействия обладает одним важным свойством, заключающимся в том, что его можно полностью охарактеризовать системой функций $V_k(t)$, где $k=1, 2, 3, \ldots$ Для доказательства достаточно показать, что

$$P\{X(t_1)=k_1;\ X(t_2)=k_2;\ldots;\ X(t_n)=k_n\},$$

можно выразить через $V_k(t)$, $(k=0, 1, 2, \ldots, n)$. Это и будет означать, что стационарный поток требований без последействия полностью характеризуется функциями $V_0(t)$, $V_1(t)$, ..., $V_n(t)$.

Если за отрезки времени t_1, t_2, ..., t_n поступило соответственно k_1, k_2, ..., k_n требований, то это равносильно тому, что за время t_2-t_1 поступило k_2-k_1 требований, а за время t_3-t_2 поступит k_3-k_2 требований и т. д. Поэтому вероятность того, что $X(t_i)=k_i$ $(i=1, 2, \ldots, n)$ равна вероятности того, что $X(t_i)-X(t_{i-1})=k_i-k_{i-1}$, где $i=1, 2, \ldots$ Для того чтобы обеспечить общность записи, мы здесь приняли, что $X(t_0)=k_0=0$. Следовательно,

$$\mathcal{P}=P\{X(t_1)=k_1,\ X(t_2)=k_2, \ldots,\ X(t_n)=k_n\}=$$

$$=P\{X(t_1)-X(t_0)=k_1-k_0,\quad X(t_2)-X(t_1)=k_2-k_1, \ldots,$$

$$X(t_n)-X(t_{n-1})=k_n-k_{n-1}\}.$$

Но промежутки времени (t_0, t_1), $(t_1, t_2), \ldots, (t_{n-1}, t_n)$ не пересекаются, поэтому события поступления k_1-k_0, $k_2-k_1, \ldots, k_n-k_{n-1}$ требований за эти промежутки времени независимы.

Применив теорему умножения вероятностей*, получим

$$\mathcal{P} = \prod_{i=1}^{n} P\{X(t_i) - X(t_{i-1}) = k_i - k_{i-1}\}^{**}.$$

По условию поток требований стационарный:

$$P\{X(t_i) - X(t_{i-1}) = k_i - k_{i-1}\} = \\ = P\{X(t_i - t_{i-1}) = k_i - k_{i-1}\}.$$

Следовательно,

$$\mathcal{P} = \prod_{i=1}^{n} P\{X(t_i - t_{i-1}) = k_i - k_{i-1}\}.$$

Но $P\{X(t_i - t_{i-1}) = k_i - k_{i-1}\}$ есть вероятность того, что за время $t_i - t_{i-1}$ поступит точно $s_i = k_i - k_{i-1}$ требований, а она равна $V_{s_i}(t_i - t_{i-1})$. Поэтому

$$\mathcal{P} = \prod_{i=1}^{n} V_{s_i}(t_i - t_{i-1}),$$

где $s_i = k_i - k_{i-1}$.

Таким образом, указанное выше свойство действительно имеет место, и для того чтобы описать стационарный поток без последействия, достаточно получить систему функций $V_k(t)$, где $k = 0, 1, 2, \ldots$ и $t > 0$. Это свой-

---

* Произведением событий $A_1, A_2, \ldots, A_n$ называется событие A, заключающееся в одновременном появлении всех событий A_i ($i = 1, 2, \ldots, n$). Теорема умножения вероятностей позволяет вычислять вероятность произведения конечного числа событий. Для произведения группы *независимых* событий она формулируется следующим образом: вероятность произведения событий равна произведению вероятностей этих событий, т. е.

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_n).$$

** Здесь символ $\prod_{i=1}^{n} a_i$ обозначает произведение группы сомножителей от a_1 до a_n, т. е. $\prod_{i=1}^{n} a_i = a_1 \cdot a_2 \cdot \ldots \cdot a_n$.

ство в значительной степени упрощает изучение таких потоков и облегчает их описание.

В целом ряде случаев, когда мы имеем дело с конкретной системой обслуживания, характер потока требований таков, что в любой момент времени может поступить только одно требование. Потоки, обладающие этим свойством, называются *ординарными*. Свойство ординарности имеет важное значение. Оно показывает, что в таких потоках невозможно, или почти невозможно, одновременное появление двух или большего числа требований.

Если обозначить через $\psi(t)$ вероятность появления за промежуток времени $(0, t)$ не меньше двух требований, то можно более строго сформулировать свойство ординарности следующим образом.

Поток требований называется ординарным, если

$$\lim_{t \to 0} \frac{\psi(t)}{t} = 0,$$

или $\psi(t) = o(t)$ * при $t \to 0$. Иными словами, вероятность того, что появится больше одного требования за малый промежуток времени t, есть бесконечно малая величина более высокого порядка, чем t. Это и означает, что почти невероятно поступление двух или нескольких требований за малый промежуток времени. В некоторых реальных потоках это свойство является очевидным, а в некоторых интуитивно очевидным или, по крайней мере, справедливым с достаточно хорошим приближением к действительности. Разумеется, на практике встречается немало потоков и не обладающих свойством ординарности (например, тот же поток вызовов на АТС), но о таких потоках будет сказано ниже.

Простейший поток. Особый интерес представляют так называемые простейшие потоки. Потоки такого типа или близкие к ним часто встречаются на практике.

Простейшим потоком требований называется поток, одновременно обладающий свойствами стационарности, ординарности и отсутствия последействия.

Описание простейших потоков имеет наиболее простой вид. Для них значительно проще получаются аналитические решения задач массового обслуживания.

* Через $o(t)$ обозначается величина бесконечно малая по сравнению с t.

Задачи с такими потоками лучше изучены, поэтому такие потоки избраны тем иллюстративным материалом, на котором демонстрируются задачи и примеры, изложенные ниже.

Поскольку простейший поток является стационарным и у него отсутствует последействие, для его полного описания вполне достаточно знать систему функций $V_0(t)$, $V_1(t)$, $V_2(t)$, ... Прежде чем приступить к выводу выражения для функций $V_k(t)$, где k может принимать любые положительные целочисленные значения ($k=0, 1, 2, \ldots$), докажем следующую теорему, формулирующую основное свойство стационарного потока требований.

Теорема 1. *Для любого стационарного потока существует предел*

$$\lim_{t \to 0} \frac{W(t)}{t} = \lambda > 0,$$

где

$$W(t) = 1 - V_0(t).$$

Величина λ называется *параметром потока.* Она может быть неограниченно большой.

Функция $W(t) = 1 - V_0(t)$ является вероятностью того, что за время t в систему поступит по крайней мере одно требование. Действительно, так как

$$\sum_{k=0}^{\infty} V_k(t) = 1, \text{ то } 1 - V_0(t) = \sum_{k=1}^{\infty} V_k(t).$$

Но

$$W(t) = 1 - V_0(t),$$

т. е.

$$W(t) = \sum_{k=1}^{\infty} V_k(t).$$

Следовательно, $W(t)$ есть вероятность того, что за время t в систему поступит на обслуживание по крайней мере одно требование.

Перейдем теперь к доказательству теоремы *1.* Оно основано на следующей лемме.

Лемма 1. Если $f(x) \geqslant 0$ *есть функция, неубывающая на отрезке* $0 < x \leqslant a$ *и* $f(x+y) \leqslant f(x) + f(y)$ *при*

x, y и $x+y \in (0, a)$, то отношение $\frac{f(x)}{x}$ при $x \to 0$ *стремится к пределу, который либо является конечным числом, либо неограниченно возрастает, и равен нулю только в случае, когда* $f(a) = 0$*.

Опираясь на лемму, для доказательства теоремы *1* достаточно показать, что функция $W(t)$ удовлетворяет условиям леммы. Действительно, уже из определения функции $W(t)$ следует, что $W(t) \geqslant 0$. Если же выбрать промежуток a достаточно большим, то вероятность появления в системе хотя бы одного требования за этот промежуток $(0, a)$ будет отлична от нуля, т. е. $W(a) > 0$. Исключением может быть только случай, когда поток требований вообще отсутствует, что для нас не представляет никакого интереса. Из определения функции $W(t)$ вытекает также, что она является возрастающей функцией. Теперь осталось показать справедливость неравенства

$$W(t_1 + t_2) \leqslant W(t_1) + W(t_2).$$

Оно вытекает из теоремы сложения вероятностей **, справедливой для суммы любого конечного числа событий. Действительно, левая часть неравенства $W(t_1+t_2)$ есть вероятность суммы двух событий, состоящих в появлении хотя бы одного требования за промежуток времени $(0, t_1)$ или (t_1, t_1+t_2), поэтому

$$W(t_1 + t_2) = W(t_1) + W(t_2) - W(t_1)\,W(t_2).$$

Но произведение $W(t_1)\,W(t_2)$ неотрицательно, поэтому

$$W(t_1 + t_2) \leqslant W(t_1) + W(t_2).$$

Таким образом, все условия леммы *1* выполнены, откуда следует справедливость теоремы *1*.

---

* Доказательство этой леммы дано в Приложении.

** Суммой событий B и C называется событие A, состоящее в наступлении хотя бы одного из событий (B или C). Сумма событий обозначается так: $A = B + C$. Для событий B и C теорема сложения вероятностей формулируется следующим образом. Вероятность наступления хотя бы одного из двух событий равна сумме вероятностей появления этих событий без вероятности их совместного появления, т. е. $P(A) = P(B) + P(C) - P(B \cdot C)$. Здесь $P(B \cdot C)$ есть вероятность одновременного появления событий B и C. Если эти события не могут появиться одновременно, т. е. эти события несовместны, то $P(B \cdot C) = 0$ и тогда $P(A) = P(B) + P(C)$.

Теперь вернемся к простейшему потоку и определим вид функций $V_k(t)$ при $k=0, 1, 2, \ldots$

Предположим, что в некоторую обслуживающую систему поступают заявки на обслуживание. Система начинает функционировать с некоторого момента $t=0$. Рассмотрим отрезок времени $(0, t+\Delta t)$ и определим, какова вероятность того, что за это время поступит точно k требований. Точно k требований может поступить одним из $k+1$ различных несовместимых способов, представленных в следующей таблице.

Промежутки времени	Количество требований, поступивших за данный промежуток времени							
$(0, t)$	k	$k-1$	$k-2$	$k-3$	. . .	2	1	0
$(t, t+\Delta t)$	0	1	2	3	. . .	$k-2$	$k-1$	k

Таким образом, если за промежуток времени $(0, t)$ в систему поступит точно k требований, а за промежуток $(t, t+\Delta t)$ ни одного требования, то за весь промежуток $(0, t+\Delta t)$ поступит k требований. А если за промежуток $(0, t)$ поступит $(k-1)$ требование, а за промежуток $(t, t+\Delta t)$ — только одно требование, то все равно в сумме за время $(0, t+\Delta t)$ поступит k требований и т. д. Последним возможным событием будет такое событие, когда за промежуток времени $(0, t+\Delta t)$ не поступит ни одного требования, а за промежуток $(t, t+\Delta t)$ поступит k требований. Следовательно, эти $k+1$ способов для рассматриваемого нами случая исчерпывают все возможные случаи поступления точно k требований за промежуток времени $(0, t+\Delta t)$.

Свойство отсутствия последействия, которое здесь имеет место, поскольку мы условились, что исследуемый поток требований является простейшим, позволяет, используя теорему умножения вероятностей, вычислить вероятность каждого из событий, т. е. каждого из этих способов появления точно k требований на обслуживание в данной системе. Так, вероятность того, что за время $(0, t)$ появится k требований, а за время $(t, t+\Delta t)$ не появится ни одного, равна

$$V_k(t) \cdot V_0(\Delta t).$$

В силу отсутствия последействия первое событие не зависит от второго, поэтому можно применить теорему умножения вероятностей независимых событий. Вероятность того, что за промежуток $(0, t)$ появится $k-1$ требований, а за промежуток $(t, t+\Delta t)$ — только одно, будет равна

$$V_{k-1}(t)\cdot V_1(\Delta t).$$

Продолжая аналогичные рассуждения, можно вычислить и вероятности всех остальных возможных случаев появления точно k требований за время $(0, t+\Delta t)$. Так как все эти случаи попарно несовместны, т. е. никакие два не могут произойти одновременно, то, используя теорему сложения вероятностей, получаем вероятность появления k событий $V_k(t+\Delta t)$, которая равна

$$V_k(t+\Delta t)=\sum_{i=0}^{k} V_i(t)\,V_{k-i}(\Delta t). \qquad (2.1)$$

Оказывается, что все слагаемые суммы, стоящей в правой части этого равенства, кроме последних двух, являются бесконечно малыми величинами более высокого порядка, чем само приращение времени Δt. Более того, оказывается, что их сумма есть бесконечно малая величина этого же порядка. Покажем это.

Во-первых, очевидно, что

$$R_k=\sum_{i=0}^{k-2} V_k(t)\cdot V_{k-i}(\Delta t)\leqslant\sum_{i=0}^{k-2} V_{k-i}(\Delta t)=\sum_{i=2}^{k} V_i(\Delta t).$$

Справедливость этого вытекает из того, что $0\leqslant V_k(t)<1$ при $k=0, 1, 2, \ldots$.

Во-вторых, очевидно, что если в сумме $\sum_{i=2}^{k} V_i(\Delta t)$ продолжать суммирование неограниченно, то она не уменьшится, т. е.

$$R_k\leqslant\sum_{i=2}^{k} V_i(\Delta t)\leqslant\sum_{i=2}^{\infty} V_i(\Delta t).$$

Теперь покажем, что даже последняя сумма есть бесконечно малая величина по сравнению с Δt, т. е.

$$\sum_{i=2}^{\infty} V_i(\Delta t) = o(\Delta t).$$

За время Δt может не поступить ни одного требования, может поступить только одно или только два и т. д., и это взаимно исключающие друг друга события. Поэтому очевидна справедливость равенства

$$\sum_{i=0}^{\infty} V_i(\Delta t) = 1.$$

Если первые два слагаемых суммы, стоящей в левой части этого равенства, перенести направо, то оно преобразуется в следующее:

$$\sum_{i=2}^{\infty} V_i(\Delta t) = 1 - V_0(\Delta t) - V_1(\Delta t).$$

Сумма, стоящая в этом новом равенстве справа, есть не что иное, как вероятность поступления не менее двух требований за время Δt. Сумма $V_0(\Delta t) + V_1(\Delta t)$ есть вероятность непоступления ни одного требования или поступления точно одного требования за время Δt. Следовательно, $1-[V_0(\Delta t) + V_1(\Delta t)]$ есть вероятность поступления двух, трех или более требований за время Δt, т. е. вероятность поступления не меньше двух требований за это время. Но выше мы эту вероятность обозначили $\psi(\Delta t)$ и указали, что свойство ординарности, которым обладает простейший поток, заключается в том, что $\psi(\Delta t) = o(\Delta t)$, т. е. $\psi(\Delta t)$ есть бесконечно малая величина более высокого порядка, чем Δt.

Поэтому

$$\sum_{t=2}^{\infty} V_i(\Delta t) = \psi(\Delta t) = o(\Delta t) \quad \text{при } \Delta t \to 0.$$

Так как $R_k \leqslant \psi(\Delta t)$, то тем более

$$R_k = o(\Delta t) \quad \text{при } \Delta t \to 0,$$

что и требовалось доказать. При этом не следует удивляться тому, что мы записали $R_k = o(\Delta t)$, а не $R_k \leqslant o(\Delta t)$. Ведь $o(\Delta t)$ это не какая-то определенная величина, а лишь обозначение того, что некоторая величина есть бесконечно малая более высокого порядка, чем Δt. Поэтому выражение $R_k \leqslant o(\Delta t)$ равносильно утверждению, что R_k по сравнению с Δt—бесконечно малая более высокого порядка, т. е.

$$R_k = o(\Delta t).$$

Теперь преобразуем правую часть равенства (2.1). Сумма всех слагаемых, кроме последних двух, есть бесконечно малая более высокого порядка, чем Δt, так как $R_k = o(\Delta t)$. Поэтому

$$V_k(t+\Delta t) = V_k(t)\,V_0(\Delta t) + V_{k-1}(t)\,V_1(\Delta t) + o(\Delta t). \quad (2.2)$$

Найдем значения $V_0(\Delta t)$ и $V_1(\Delta t)$. Из теоремы *1* следует, что вероятность поступления по крайней мере одного требования за время Δt отличается от $\lambda \Delta t$ на бесконечно малую величину более высокого порядка, чем Δt, т. е.

$$W(t) = \lambda \Delta t + o(\Delta t),$$

но

$$W(\Delta t) = 1 - V_0(\Delta t),$$

где $V_0(\Delta t)$—вероятность того, что ни одно требование не поступит в систему за время Δt, поэтому

$$\lambda \Delta t + o(\Delta t) = 1 - V_0(\Delta t),$$

откуда

$$V_0(\Delta t) = 1 - \lambda \Delta t + o(\Delta t). \quad (2.3)$$

Теперь найдем $V_1(\Delta t)$. Так как вероятность поступления в систему хотя бы одного требования за время Δt есть сумма вероятностей поступления одного, двух, трех и т. д. требований, то

$$W(\Delta t) = V_1(\Delta t) + V_2(\Delta t) + \ldots = V_1(\Delta t) + o(\Delta t).$$

Из этого равенства можно определить значение искомой величины $V_1(\Delta t)$:

$$V_1(\Delta t) = W(\Delta t) + o(\Delta t).$$

Но мы уже знаем, что

$$W(\Delta t) = \lambda \Delta t + o(\Delta t),$$

поэтому

$$V_1(\Delta t) = \lambda \Delta t + o(\Delta t) - o(\Delta t) = \lambda \Delta t + o(\Delta t).^* \quad (2.4)$$

Полученные выражения (2.3) для $V_0(\Delta t)$ и (2.4) для $V_1(\Delta t)$ подставим в (2.2). В результате получим уравнение, связывающее V_k, V_{k-1} и λ,

$$V_k(t+\Delta t) = V_k(t) - \lambda V_k(t)\,\Delta t + \lambda V_{k-1}(t)\,\Delta t + V_k(t)\,o(\Delta t) + V_{k-1}(t)\,o(\Delta t).$$

Но так как $V_k(t) \leqslant 1$ и $V_{k-1}(t) \leqslant 1$, то согласно замечанию (см. сноску) $V_k(t)\,o(\Delta t) + V_{k-1}(t)\,o(\Delta t)$ есть бесконечно малая более высокого порядка, чем Δt, т. е.

$$V_k(t)\,o(\Delta t) + V_{k-1}(t)\,o(\Delta t) = o(\Delta t).$$

Поэтому

$$V_k(t+\Delta t) = V_k(t) - \lambda V_k(t)\,\Delta t + \lambda V_{k-1}(t)\,\Delta t + o(\Delta t).$$

Перенеся $V_k(t)$ влево и разделив на Δt, получим

$$\frac{V_k(t+\Delta t) - V_k(t)}{\Delta t} = -\lambda V_k(t) + \lambda V_{k-1}(t) + \frac{o(\Delta t)}{\Delta t}.$$

Заметим, что предел левой части при $\Delta t \to 0$, если он существует, равен производной от $V_k(t)$ по t. Тогда, переходя к пределу, который существует, так как существует предел правой части, получаем

$$\frac{dV_k(t)}{dt} = -\lambda V_k(t) + \lambda V_{k-1}(t) \quad (k = 1, 2, 3, \ldots) \quad (2.5)$$

Предел $\frac{o(\Delta t)}{\Delta t}$ равен нулю, так как $o(\Delta t)$ — бесконечно малая более высокого порядка, чем Δt.

---

* Не следует удивляться, что $o(\Delta t) - o(\Delta t) \neq 0$, а равно $o(\Delta t)$. С величиной $o(\Delta t)$ нельзя оперировать по формальным алгебраическим правилам, так как $o(\Delta t)$—не определенная постоянная или переменная величина, а обозначение того, что $\frac{o(\Delta t)}{\Delta t} \to 0$ при $\Delta t \to 0$, т. е. является символическим обозначением того, что $o(\Delta t)$ есть бесконечно малая величина высшего порядка по сравнению с Δt. Поэтому справедливы следующие равенства: $o(\Delta t) \pm o(\Delta t) = o(\Delta t)$ и т. п., если величину $o(\Delta t)$ понимать в указанном смысле.

Таким образом, для определения $V_k(\Delta t)$ получена бесконечная рекуррентная система линейных однородных дифференциальных уравнений (2.5). Но так как $V_k(t)$ выражено через $V_{k-1}(t)$, то необходимо найти еще одно уравнение для определения $V_0(t)$. Это уравнение можно получить из условия

$$V_0(t+\Delta t)=V_0(t)V_0(\Delta t),$$

которое вытекает из свойства отсутствия последействия и теоремы умножения вероятностей. Последнее условие означает, что вероятность отсутствия требований за время $(0, t+\Delta t)$ равна произведению вероятности отсутствия требований за время $(0, t)$ на вероятность отсутствия требований за время $(t, t+\Delta t)$. Подставляя в это уравнение вместо $V_0(\Delta t)$ выражение (2.3), получаем

$$V_0(t+\Delta t)=V_0(t)[1-\lambda\Delta t+o(\Delta t)].$$

Откуда

$$V_0(t+\Delta t)-V_0(t)=-\lambda V_0(t)\Delta t+V_0(t)o(\Delta t).$$

Поделив правую и левую части этого уравнения на Δt и перейдя к пределу при $\Delta t\to 0$, получим

$$\lim_{\Delta t\to 0}\frac{V_0(t+\Delta t)-V_0(t)}{\Delta t}=\lim_{\Delta t\to 0}\left[-\lambda V_0(t)-V_0(t)\frac{o(\Delta t)}{\Delta t}\right],$$

то есть

$$\frac{dV_0(t)}{dt}=-\lambda V_0(t).$$

Это — дифференциальное уравнение с разделяющимися переменными. Разделим переменные и проинтегрируем его

$$\frac{dV_0(t)}{V_0(t)}=-\lambda dt,$$

$$\ln V_0(t)=-\lambda t+\ln C.$$

Отсюда, потенцируя, получаем, что

$$V_0(t)=Ce^{-\lambda t}. \qquad (2.6)$$

Для определения произвольной постоянной C воспользуемся равенством (2.3), из которого следует, что

$V_0(0)=1$. Это условие имеет простой физический смысл: оно означает, *что вероятность отсутствия требований за промежуток времени, равный нулю, равна единице.* Подставляя в (2.6) $t=0$, получаем

$$V_0(0)=C=1.$$

Поэтому

$$V_0(t)=e^{-\lambda t}. \tag{2.7}$$

Но величина $V_0(t)$ равна вероятности того, что за промежуток времени $(0, t)$ не поступит ни одного требования. Следовательно, эта формула показывает, что с возрастанием t вероятность того, что в систему не поступит ни одного требования, быстро убывает, причем скорость убывания будет тем больше, чем больше λ. Величина λ носит название параметра потока. О том, какой смысл имеет λ, мы расскажем несколько ниже.

Вернемся теперь к системе (2.5). Произведем подстановку

$$V_k(t)=e^{-\lambda t}u_k(t) \quad (k=0,\ 1,\ 2,\ldots), \tag{2.8}$$

где $V_0(t)=e^{-\lambda t}$, т. е. $u_0(t)=1$.

Вместо функций $V_k(t)$ $(k=1,\ 2,\ldots)$ будем искать функции $u_k(t)$ $(k=1,\ 2,\ldots)$. Если мы их найдем, то легко будет найти функции $V_k(t)$. Для функций $u_k(t)$ получим следующую систему уравнений:

$$\frac{du_k(t)}{dt}=\lambda u_{k-1}(t) \quad (k=1,\ 2,\ldots). \tag{2.9}$$

Условия для отыскания произвольных постоянных определим следующим образом. Из (2.4) вытекает, что $V_1(0)=0$. Таким образом, имеем

$$V_0(0)=1 \quad \text{и} \quad V_1(0)=0,$$

но

$$\sum_{k=0}^{\infty}V_k(t)=1,$$

следовательно,

$$V_0(t)+V_1(t)+\sum_{k=2}^{\infty}V_k(t)=1.$$

При $t=0$ получаем

$$1+\sum_{k=2}^{\infty} V_k(0)=1.$$

Отсюда

$$\sum_{k=2}^{\infty} V_k(0)=0.$$

Но величины $V_k(t)\geqslant 0$ $(k=0,\ 1,\ 2,\ldots)$, следовательно, их сумма может быть равна нулю только в случае, когда каждая из них равна нулю, т. е. $V_k(0)=0$ $(k=2,\ 3,\ldots)$ и $V_1(0)=0$. Эти условия имеют простой физический смысл. Они означают, что вероятность появления k требований $(k=1,\ 2,\ldots)$ за промежуток времени, не превышающий нуля, равна нулю.

Итак, для определения произвольных постоянных имеем следующие условия:

$$V_k(0)=u_k(0)=0 \quad (k=1,\ 2,\ldots).$$

Из (2.7) и (2.8) следует, что $u_0(t)=1$. Подставив $u_0(t)$ в первое уравнение (2.9), получим

$$\frac{du_1(t)}{dt}=\lambda,$$

и после интегрирования

$$u_1(t)=\lambda t+C_1.$$

Но $u_1(0)=0=C_1$, поэтому $u_1(t)=\lambda t$. Подставив $u_1(t)$ во второе уравнение (2.9), получим

$$\frac{du_2(t)}{dt}=\lambda^2 t,$$

откуда после интегрирования получаем

$$u_2(t)=\frac{\lambda^2 t^2}{2}+C_2.$$

Но при $t=0$ $u_2(0)=0=C_2$, поэтому

$$u_2(t)=\frac{(\lambda t)^2}{2}.$$

Аналогично получаем

$$u_3(t)=\frac{(\lambda t)^3}{2\cdot 3}=\frac{(\lambda t)^3}{3!}.$$

По индукции легко показать, что

$$u_k(t)=\frac{(\lambda t)^k}{k!} \quad (k=1,\ 2,\ \ldots).$$

Возвращаясь к $V_k(t)$, из (2.8) получаем

$$V_k(t)=\frac{(\lambda t)^k}{k!}\,e^{-\lambda t} \quad (k=0,\ 1,\ 2,\ldots). \tag{2.10}$$

Таким образом, для простейшего потока число требований в промежутке времени t распределено по закону Пуассона с параметром λt.

Простейший поток полностью определяется системой функций (2.10). Функции $V_k(t)$, если не считать t, зависят только от параметра потока λ.

Напомним, что $V_k(t)$ есть вероятность поступления точно k требований за время $(0,\ t)$. Следовательно, для того чтобы дать полную характеристику простейшего потока, достаточно знать только одну величину — *параметр потока.*

Рассмотрим физический смысл λ. Легко показать, что для простейшего потока параметр λ равен *математическому ожиданию* числа требований*, поступивших в си-

---

* Математическое ожидание случайной величины по смыслу близко к среднему ее значению, поэтому его иногда называют средним значением случайной величины. Если случайная величина x принимает значения $x_1,\ x_2,\ldots,x_n$ с вероятностями $p_1,\ p_2,\ldots,p_n$, т. е. вероятность того, что случайная величина принимает значение x_i, равна p_i, то ее среднее значение или математическое ожидание равно

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

Для непрерывной случайной величины x с законом распределения $F(t)$ математическое ожидание равно

$$M[x]=\int_{-\infty}^{\infty} t\,dF(t).$$

стему за единицу времени. Чтобы доказать это, вычислим математическое ожидание числа требований, поступивших за промежуток времени $(0, t)$. Оно равно

$$M_t[k] = \sum_{k=1}^{\infty} kV_k(t) = \sum_{k=1}^{\infty} k \frac{(\lambda t)^k}{k!} e^{-\lambda t} = e^{-\lambda t} \lambda t \sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!}.$$

Но сумма $\sum_{k=1}^{\infty} \frac{(\lambda t)^{k-1}}{(k-1)!}$ является разложением в ряд функции $e^{\lambda t}$ по степеням λt, поэтому

$$M_t[k] = \lambda t e^{-\lambda t} e^{\lambda t} = \lambda t.$$

Следовательно, математическое ожидание числа требований за единицу времени, которое получается из выражения для $M_t[k]$ при $t = 1$, равно

$$M_1[k] = \lambda.$$

Таким образом, *если анализ показывает, что изучаемый поток является простейшим, т. е. стационарным, ординарным и в нем отсутствует последействие, то для его полного описания достаточно вычислить математическое ожидание числа требований, поступивших за единицу времени.*

Легко заметить, что простейший поток обладает еще одним очень интересным свойством. Для простейшего потока вероятность получения в течение промежутка времени длительности t точно k требований достигает наибольшего значения для $t = \frac{k}{\lambda}$ $(k = 0, 1, 2, \ldots)$. В частности, при $\lambda = 1$ максимумы будут достигаться в моменты времени, равные 0, 1, 2, ..., n единиц времени (рис. 3).

Рассмотрим пример. В первом примере § 2 гл. 1 в качестве потока требований рассматривался поток, состоящий из заявок на обслуживание станков. Станок остановился — поступила заявка на обслуживание. Обслуживание состоит в устранении причины остановки станка. Если станки находятся примерно в одинаковом состоянии, то можно предполагать, что поток требований обладает свойством стационарности.

Если вероятность остановки одного станка не очень велика, что, как правило, имеет место на практике, то вероятность того, что два станка остановятся в один момент времени, будет еще меньше. Поэтому можно считать, что этот поток обладает свойством ординарности.

Наличие свойства отсутствия последействия также интуитивно ясно для тех случаев, когда число обслужи-

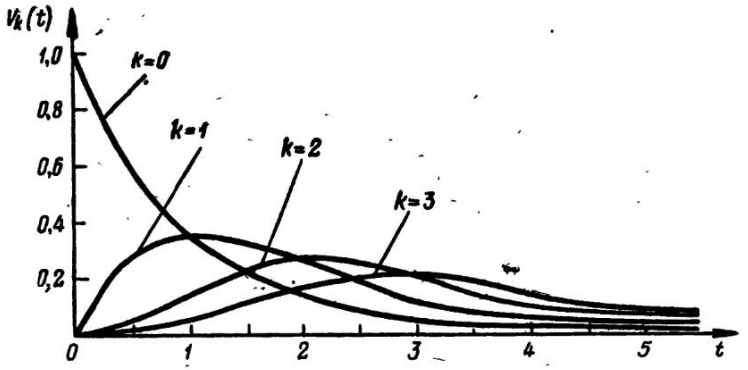

Рис. 3. График вероятности получения точно k требований (k=0, 1, 2, 3) при простейшем потоке требований.

ваемых станков не мало, а среднее время обслуживания мало по сравнению со средним промежутком времени между смежными требованиями.

Нужно заметить, что эти рассуждения, обосновывающие наличие определенных свойств у потока требований, не являются доказательными. Это только качественные рассуждения, которые могут подсказать путь, по которому нужно итти, чтобы отыскать количественное описание потока требований. Обосновать все эти свойства потока можно только путем статистической обработки результатов наблюдений изучаемого потока требований. Методы статистической обработки результатов наблюдений относятся к области математической статистики, поэтому на них мы не будем останавливаться.

Однако если предположить, что наши рассуждения имеют силу доказательства, то для описания этого потока требований достаточно найти одну постоянную величину — математическое ожидание числа станков, нуждающихся в обслуживании за единицу времени.

Пусть среднее число станков, требующих обслуживания за час, равно трем. Тогда, используя описание простейшего потока, можно ответить на любые вопросы, которые могут возникнуть при изучении этого процесса обслуживания. Так, из формулы (2.10) следует, что вероятность того, что за час потребуют обслуживания k станков, равна

$$V_k(1) = \frac{3^k}{k!} e^{-3} = \frac{3^k}{k!} \cdot 0{,}0498.$$

Задаваясь различными значениями k, можно составить таблицу вероятностей остановки станков, которая будет иметь следующий вид:

k	0	1	2	3	4	5	6
$V_k(1)$	0,0498	0,1494	0,2241	0,2241	0,1680	0,1008	0,0504

Если, например, среднее время обслуживания одного станка составляет 10 минут, то представляет интерес вопрос о том, какова вероятность, что за 10 минут остановится больше чем один станок. Ответ на этот вопрос имеет весьма важное значение при определении необходимого количества рабочих для обслуживания заданного количества станков. Для определения этой вероятности найдем вероятность того, что за 10 минут остановятся два, три, четыре станка и т. д.

Эти вероятности приведены в следующей таблице:

k	2	3	4	5
$V_k\left(\frac{1}{6}\right)$	0,076	0,013	0,003	0,000

Здесь

$$V_k\left(\frac{1}{6}\right) = \frac{1}{2^k \cdot k!} e^{-0{,}5} \approx \frac{0{,}606}{2^k \cdot k!}.$$

Вероятность того, что за 10 минут ($^1/_6$ часа) остановится более 5 станков, равна нулю с погрешностью не большей 0,0005. Следовательно, интересующая нас веро-

ятность того, что за 10 минут остановится не меньше двух станков, с точностью до $5 \cdot 10^{-4}$ равна

$$\sum_{k=2}^{4} V_k\left(\frac{1}{6}\right)=0{,}092.$$

Поэтому за 7-часовой рабочий день в среднем 4 раза рабочий встретится с таким положением, когда в течение 10 минут потребуется обслуживать не меньше двух станков. А вероятность того, что за 10 минут потребуют обслуживания не меньше трех станков, будет равна

$$\sum_{k=3}^{4} V_k\left(\frac{1}{6}\right)=0{,}016.$$

С таким положением рабочий столкнется реже двух раз в три дня. Рассмотрение этого примера показывает, что умение количественно описывать поток требований является весьма полезным, так как позволяет найти ряд важных характеристик процесса обслуживания.

Поток с ограниченным последействием. В начале этого параграфа был изложен общий способ описания потока требований. Сейчас мы рассмотрим еще один способ, который использует такие свойства функции $X(t)$, как неотрицательность и монотонность.

Пусть t_0, t_1, ..., t_n есть моменты поступления последовательных требований потока. Величина $t_0=0$ — начальный момент потока. Обозначим

$$z_i=t_i-t_{i-1} \quad (i=1,2,\ldots,n).$$

Очевидно, что $z_i \geqslant 0$, так как $t_{i-1}<t_i$ $(i=1, 2, \ldots, n)$ Величина $z_1=t_1$, а величины $z_i (i \geqslant 2)$ — промежутки времени между $(i-1)$-м и i-м требованиями. Очевидно, что *поток требований будет задан в том случае, если будут известны все моменты поступления требований или промежутки между смежными требованиями.* Поскольку $z_i (i=1, 2, \ldots, n)$ — случайные величины, то для определения потока требований нужно задать закон распределения этих случайных величин.

Не останавливаясь на подробностях, заметим, что можно доказать эквивалентность рассмотренных выше двух способов описания потоков требований.

Дадим теперь определение потока с ограниченным последействием. *Поток требований называется потоком с ограниченным последействием, если для него последовательность $z_1, z_2, \ldots, z_n$ есть последовательность взаимно независимых величин.*

Требование ограниченности последействия является более широким, чем требование отсутствия последействия. Если поток обладает свойством отсутствия последействия, то для него $z_1, z_2, \ldots, z_n$ есть последовательность взаимно независимых величин. *Обратное утверждение не имеет места.* Если последовательность состоит из взаимно независимых величин, то из этого не вытекает отсутствие последействия в потоке.

Обобщением простейшего потока является стационарный ординарный поток с ограниченным последействием. Если для полного определения простейшего потока достаточно задания одной постоянной величины, то для определения стационарного ординарного потока с ограниченным последействием достаточно знания одной функции $\varphi_0(t)$, определением которой мы и займемся.

Введем функцию $h_0(\tau, t)$, равную вероятности отсутствия требований за время t, при условии, что за предшествующий промежуток времени τ поступило по крайней мере одно требование. По предположению τ есть предшествующий t и смежный с ним промежуток времени. Тогда отношение $\frac{h_0(\tau, t)}{W(\tau)}$ есть условная вероятность отсутствия требований за время t, при условии, что за промежуток времени τ поступило хотя бы одно требование. Напомним, что $W(\tau)$ есть вероятность поступления по крайней мере одного требования за время τ. Тогда функцию $\varphi_0(t)$ можно определить как предел отношения $\frac{h_0(\tau, t)}{W(\tau)}$ при $\tau \to 0$, т. е.

$$\varphi_0(t) = \lim_{\tau \to 0} \frac{h_0(\tau, t)}{W(\tau)}.$$

Эта функция, введенная Пальмом, носит название функции Пальма.

Таким образом, функция $\varphi_0(t)$ равна вероятности того, что за время t не поступит ни одного требования,

при условии, что в начальный момент поступило по крайней мере одно требование. Эта функция отличается от $V_0(t)$ тем, что для $V_0(t)$ ничего не известно о том, когда поступило предыдущее требование, а для $\varphi_0(t)$ известно. Очевидно, что $\varphi_0(t)$ невозрастающая функция. Эта функция полностью определяет стационарный ординарный поток с ограниченным последействием. Как отмечалось выше, ограниченное последействие заключается в независимости случайных величин $z_1, z_2, \ldots, z_n$. Поэтому для полного определения потока достаточно знать законы распределения z_k, которые мы обозначим

$$F_k(x) = P\{z_k < x\} \qquad (k = 1, 2, 3, \ldots).$$

Если показать, что все $F_k(x)$ выражаются через $\varphi_0(t)$, то этим будет доказано, что стационарный ординарный поток с ограниченным последействием полностью определяется функцией $\varphi_0(t)$.

Докажем следующую теорему.

Теорема 2. *Для стационарного ординарного потока с ограниченным последействием функции распределения* $F_k(t)$ имеют следующий вид:

$$F_1(t) = \lambda \int_0^t \varphi_0(u)\, du, \tag{2.11}$$

$$F_k(t) = 1 - \varphi_0(t) \quad (k \geqslant 2), \tag{2.12}$$

где λ — параметр потока, имеющий конечное значение.

Рассмотрим два смежных интервала τ и t. Событие отсутствия требований за время t может иметь место двумя независимыми способами. При отсутствии требований за время t может быть, что и за время τ требования не поступали, а может быть, что за время τ поступило хотя бы одно требование. Поэтому, используя определение функции $h_0(\tau, t)$ и теорему сложения вероятностей, получим, что вероятность отсутствия требований за время t равна

$$V_0(t) = V_0(t+\tau) + h_0(\tau, t).$$

Отсюда

$$V_0(t+\tau) = V_0(t) - h_0(\tau, t).$$

Преобразуем это равенство следующим образом:

$$V_0(t+\tau)-V_0(t)=-h_0(\tau,\ t)\frac{W(\tau)}{W(\tau)}.$$

Разделим обе части равенства на τ:

$$\frac{V_0(t+\tau)-V_0(t)}{\tau}=-\frac{h_0(\tau,\ t)}{W(\tau)}\frac{W(\tau)}{\tau}$$

и перейдем к пределу при $\tau\to 0$.

Так как поток стационарный, то по теореме *1*

$$\lim_{\tau\to 0}\frac{W(\tau)}{\tau}=\lambda>0,$$

причем по условию теоремы *2* λ — конечно. Следовательно, предел правой части существует и мы имеем дифференциальное уравнение

$$\frac{dV_0(t)}{dt}=-\lambda\varphi_0(t).$$

После интегрирования и использования для определения произвольной постоянной условия $V_0(0)=1$, получим

$$V_0(t)=1-\lambda\int_0^t\varphi_0(u)\,du.$$

Но $F_1(t)=P\{t_1<t\}$ есть вероятность того, что за время $(0,\ t)$ поступит хотя бы одно требование. Поэтому

$$F_1(t)=1-V_0(t)=1-\left[1-\lambda\int_0^t\varphi_0(u)\,du\right]=$$

$$=\lambda\int_0^t\varphi_0(u)\,du.$$

Таким образом, первая часть теоремы доказана:

$$F_1(t)=\lambda\int_0^t\varphi_0(u)\,du,$$

т. е. закон распределения $F_1(t)$ выражается через функцию $\varphi_0(t)$ и параметр потока λ.

Нужно заметить, что для стационарного ординарного потока с ограниченным последействием λ не равно математическому ожиданию числа требований за единицу времени, как это имело место для простейшего потока. Для определения λ может быть использовано соотношение

$$\lambda = \lim_{\tau \to 0} \frac{W(\tau)}{\tau}.$$

Кроме того, параметр потока может быть определен и из соотношения (2.11) через $\varphi_0(t)$. Так как вероятность поступления хотя бы одного требования за время $(0, \infty)$ равна единице (считается, что поток имеет по крайней мере одно требование), то из (2.11) следует

$$F_1(+\infty) = 1 = \lambda \int_0^\infty \varphi_0(u)\, du.$$

Поэтому для определения λ может быть использовано уравнение

$$\lambda \int_0^\infty \varphi_0(u)\, du = 1.$$

Перейдем к доказательству второй части теоремы, т. е. покажем, что

$$F_k(t) = 1 - \varphi_0(t).$$

Начнем с $k=2$. Событие отсутствия требований за время t при наличии требований за предшествующий промежуток времени может осуществиться следующим образом:

1) За время τ поступило одно требование, а за время $\tau - t_1 + t$, ни одного, т. е. $\tau - t_1 + t < z_2$ (рис. 4).

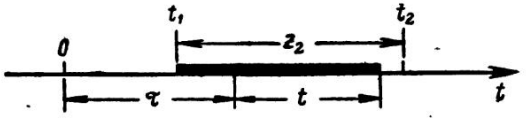

Рис. 4. На отрезке $(0, \tau)$ поступил один вызов, а на $(t, \tau - t_1 + t)$ ни одного.

2) За время τ поступило два требования, а за время $\tau - t_2 + t$ ни одного, т. е. $\tau - t_2 + t < z_3$, и т. д. Так как вероятность неравенства $z_k > \tau - t_{k-1} + t$ $(k=2, 3,\ldots)$ равна

$$P\{z_k > \tau - t_{k-1} + t\} \leqslant P\{z_2 > t\} = 1 - P\{z_2 < t\} = \\ = 1 - F_2(t),$$

то

$$h_0(\tau, t) = V_1(\tau) P\{z_2 > \tau - t_1 + t\} + \\ + V_2(t) P\{z_3 > \tau - t_2 + t\} + \ldots \leqslant$$

$$\leqslant V_1(\tau)[1 - F_2(t)] + \sum_{k=2}^{\infty} V_k(\tau)[1 - F_k(t)] \leqslant$$

$$\leqslant V_1(\tau)[1 - F_2(t)] + \sum_{k=2}^{\infty} V_k(\tau).$$

Обозначим

$$\psi_m(\tau) = \sum_{k=m}^{\infty} V_k(t) \qquad (m = 2, 3, \ldots).$$

Тогда

$$h_0(\tau, t) \leqslant V_1(\tau)[1 - F_2(t)] + \psi_2(\tau).$$

С другой стороны, если за время τ поступило точно k требований, а $z_{k+1} > \tau + t$, то за время t не поступит ни одного требования, поэтому

$$h_0(\tau, t) \geqslant V_1(\tau) P\{z_2 > t + \tau\} + V_2(t) P\{z_3 > t + \tau\} + \ldots \geqslant \\ \geqslant V_1(\tau)[1 - F_2(t + \tau)].$$

Следовательно, $h_0(\tau, t)$ заключено в пределах

$$V_1(\tau)[1 - F_2(t + \tau)] \leqslant h_0(\tau, t) \leqslant V_1(\tau)[1 - F_2(t)] + \psi_2(\tau).$$

Разделим все части этого неравенства на $W(\tau)$ и перейдем к пределу при $\tau \to 0$:

$$\frac{V_1(\tau)}{W(\tau)}[1 - F_2(t + \tau)] \leqslant \frac{h_0(\tau, t)}{W(\tau)} \leqslant \frac{V_1(\tau)}{W(\tau)}[1 - F_2(t)] + \frac{\psi_2(\tau)}{W(\tau)}. \tag{2.13}$$

Вычислим предел отношения $\frac{V_1(\tau)}{W(\tau)}$ при $\tau \to 0$, при этом напомним, что из (2.4) нам известно, что

$$V_1(\tau) = \lambda\tau + o(\tau) \quad (\tau \to 0).$$

Поэтому

$$\lim_{\tau \to 0} \frac{V_1(\tau)}{W(\tau)} = \lim_{\tau \to 0} \frac{\lambda + \frac{o(\tau)}{\tau}}{\frac{W(\tau)}{\tau}} = \frac{\lambda}{\lambda} = 1.$$

Следовательно, предел левой части (2.13) равен $1 - F_2(t)$.

По определению

$$\lim_{\tau \to 0} \frac{h_0(\tau, t)}{W(\tau)} = \varphi_0(t).$$

Для вычисления предела правой части (2.13) воспользуемся следующей леммой.

Лемма 2. *Для стационарного ординарного потока с ограниченным последействием и любого* $m > 1$

$$\lim_{\tau \to 0} \frac{\psi_{m+1}(\tau)}{\psi_m(\tau)} = 0,$$

где

$$\psi_m(\tau) = \sum_{k=m}^{\infty} V_k(\tau),$$

т. е. $\psi_m(\tau)$ *есть вероятность поступления за время* τ *не менее* m *требований*.*

Поэтому

$$\lim_{\tau \to 0} \frac{\psi_2(\tau)}{W(\tau)} = \lim_{\tau \to 0} \frac{\psi_2(\tau)}{\psi_1(\tau)} = 0,$$

так как

$$W(\tau) = \psi_1(\tau).$$

Следовательно, предел правой части неравенства (2.13) равен

$$1 - F_2(\tau).$$

---

* Доказательство этой леммы приведено в Приложении.

В пределе при $\tau \to 0$ из (2.13) получаем:

$$1 - F_2(t) \leqslant \varphi_0(t) \leqslant 1 - F_2(t),$$

т. е.

$$\varphi_0(t) = 1 - F_2(t)$$

и

$$F_2(t) = 1 - \varphi_0(t).$$

Осталось доказать справедливость этого соотношения при любом $k > 2$. Из стационарности потока вытекает, что закон распределения временных интервалов между первым и вторым требованиями, следующими за моментом $t = a > 0$, такой же, как и закон распределения интервалов между первыми двумя требованиями, следующими за моментом 0:

$$F_k(t) = F_2(t) = 1 - \varphi_0(t) \qquad (k > 2).$$

Таким образом, показано, что полное описание потока требований, обладающего свойствами стационарности, ординарности и ограниченного последействия, требует знания функции $\varphi_0(t)$, определяющей вероятность отсутствия требований за время t, при условии, что в начальный момент требование поступило.

Предположим, что в качестве функции $\varphi_0(t)$ выбрана функция e^{-bt} $(b > 0)$. Эта функция может быть равной $\varphi_0(t)$, так как $e^{-bt} \leqslant 1$, при $0 \leqslant t < \infty$, и монотонно убывает. В этом случае

$$F_1(t) = \lambda \int_0^t e^{-bu} du = \lambda \left[-\frac{1}{b} e^{-bu} \right]_0^t = \frac{\lambda}{b} [1 - e^{-bt}].$$

Определим параметр потока λ из уравнения

$$\lambda \int_0^\infty \varphi_0(u)\, du = 1:$$

$$\lambda \int_0^\infty e^{-bu} du = 1, \quad \lambda \left[-\frac{1}{b} e^{-bu} \right]_0^\infty = 1,$$

$$\frac{\lambda}{b} = 1, \quad b = \lambda.$$

Следовательно,

$$F_1(t) = 1 - e^{-\lambda t},$$

но функции

$$F_k(t) = 1 - \varphi_0(t) = 1 - e^{-\lambda t} \qquad (k = 2,\ 3, \ldots).$$

Таким образом, закон распределения $F_1(t)$ совпадает с законом распределения $F_k(t)$ при любом $k \geqslant 2$. Это означает, что в потоке требований отсутствует последействие, но поскольку поток обладает, кроме того, свойствами стационарности и ординарности, он превращается в простейший. Следовательно, *при* $\varphi_0 = e^{-\lambda t}$ *стационарный ординарный поток с ограниченным последействием превращается в простейший поток.*

2. ВРЕМЯ ОБСЛУЖИВАНИЯ

Перейдем к рассмотрению еще одного весьма важного понятия теории массового обслуживания, которое играет большую роль при анализе, постановке и решении задач обслуживания. Это — *время обслуживания.*

Время обслуживания есть прежде всего характеристика функционирования каждого отдельного аппарата обслуживающей системы. Оно показывает, сколько времени затрачивается на обслуживание одного требования данным обслуживающим аппаратом. Необходимо помнить, что этот показатель обслуживания ничего общего не имеет с оценкой качества обслуживания, а характеризует лишь пропускную способность одного обслуживающего аппарата. При этом предполагается, что если обслуживание требования, поступившего в систему, закончилось, то заявка на обслуживание удовлетворена полностью.

В силу самых различных причин время обслуживания может меняться от одного требования к другому. Эти причины, в первую очередь, связаны с тем, что поступающие требования не будут полностью идентичными. Телевизоры или радиоприемники, поступающие в мастерскую для ремонта, как правило, имеют самые разнообразные неисправности, и даже в тех случаях, когда неисправности идентичны, время, требуемое для их устранения, может быть различным, если ремонтируются телевизоры или радиоприемники различных марок.

Другой причиной изменения времени обслуживания является состояние и возможности самих обслуживающих аппаратов. Очевидно, что если обслуживание производится человеком, то время обслуживания даже абсолютно идентичных требований будет различно не только у разных людей, но даже при осуществлении обслуживания одним человеком (в зависимости от его сноровки при выполнении тех или иных операций, усталости, настроения и т. п.). А если обслуживающими аппаратами являются машины, то время обслуживания — в зависимости от их эксплуатационных характеристик, марки машины и т. п. — может меняться в значительных пределах. Поэтому в общем случае время обслуживания является *случайной величиной* и, следовательно, может быть описано законом распределения.

Если обозначить время обслуживания через γ, то полной его характеристикой будет закон распределения

$$F(t)=P\{\gamma<t\} \quad (t\geqslant 0).$$

Здесь функция $F(t)$ определяет вероятность того, что время обслуживания γ будет меньше некоторого наперед заданного значения t. Так как время обслуживания не может быть отрицательной величиной, то

$$F(t)=0 \quad \text{при } t<0.$$

Иначе говоря, функция $F(t)$, как всякая функция распределения, должна быть положительной монотонно возрастающей функцией и не должна превосходить единицу.

О том, какой конкретный вид имеет функция распределения $F(t)$, ничего нельзя сказать заранее без детального изучения функционирования обслуживающего аппарата. Даже в одной обслуживающей системе время обслуживания различных обслуживающих аппаратов может характеризоваться различными функциями распределения. Однако в дальнейшем для простоты будем рассматривать системы, состоящие из однотипных обслуживающих аппаратов, характеризуемых общим законом распределения времени обслуживания. Естественно, что знание функций распределения времени обслуживания как случайной величины имеет для нас весьма существенное значение, так как позволяет получить ответы на ряд важных вопросов. Убедимся в этом.

Допустим, что в результате анализа функционирования обслуживающей системы мы определили вид функции распределения времени обслуживания для обслуживающих аппаратов этой системы. Пусть она имеет вид

$$F(t)=1-\frac{1}{(t+1)^2},$$

где t — время в минутах.

Функция $F(t)$ действительно может иметь такой вид, так как

$$0\leqslant 1-\frac{1}{(t+1)^2}<1$$

при условии, что $0\leqslant t<\infty$

и $$F'(t)=\frac{2}{(t+1)^3}>0 \quad \text{при } t>0,$$

т. е. $F(t)$ монотонно возрастает.

Тогда, зная вид функции $F(t)$, можно ответить на целый ряд вопросов. Например, можно определить, какова будет вероятность того, что время обслуживания не превзойдет 10 минут. Эту вероятность мы получим, подставив $t=10$ в функцию $F(t)$:

$$F(10)=1-\frac{1}{(10+1)^2}\approx 0{,}99.$$

Можно также вычислить и среднее время обслуживания одного требования. Оно будет равно

$$M[\gamma]=\int_0^\infty t\,dF(t)=\int_0^\infty\frac{2dt}{(t+1)^3}=2\int_1^\infty\frac{z-1}{z^3}\,dz=$$

$$=2\left[-\frac{1}{z}+\frac{1}{2z^2}\right]_1^\infty=1 \text{ мин.}$$

Вероятность того, что за это время обслуживание очередного требования будет закончено, равна

$$F(1)=1-\frac{1}{4}=0{,}75,$$

т. е. при данном виде функции распределения времени обслуживания из 100 требований в среднем 75 будут обслужены за время, не превышающее одной минуты.

Показательный закон распределения времени обслуживания. Как в теоретических исследованиях, так и во многих практических расчетах большое значение имеет показательный закон распределения времени обслуживания, при котором функция распределения времени обслуживания $F(t)$ имеет вид

$$F(t) = 1 - e^{-\nu t}.$$

Параметр ν, входящий в показательный закон распределения, имеет простой физический смысл. Величина, обратная ν $\left(\frac{1}{\nu}\right)$, является средним временем обслуживания (математическим ожиданием времени обслуживания). Действительно

$$M[\gamma] = \int_0^\infty t dF(t) = [-te^{-\nu t}]_0^\infty + \int_0^\infty e^{-\nu t} dt =$$

$$= 0 - \frac{1}{\nu}[e^{-\nu t}]_0^\infty = \frac{1}{\nu}.$$

На рис. 5 показаны графики показательного закона распределения при $\nu = 1$, 2 и 3. Эти графики показы-

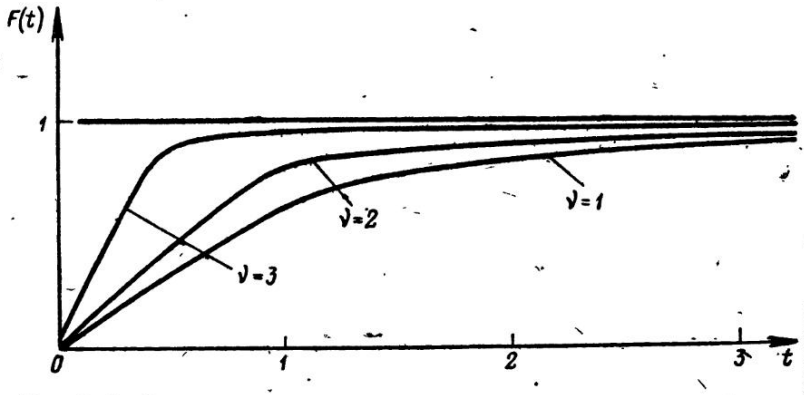

Рис. 5. График показательного закона распределения времени обслуживания $F(t)$ при $\nu = 1, 2, 3$.

вают, что при показательном законе распределения времени обслуживания вероятность того, что обслуживание закончится вскоре после его начала, велика.

На практике, когда мы имеем дело с реальными процессами обслуживания, могут встречаться положения,

при которых это свойство не имеет места. Поэтому, несмотря на то, что процессы массового обслуживания с показательным законом распределения времени обслуживания до сих пор привлекают внимание многих исследователей, теоретический и практический интерес представляют и другие законы распределения времени обслуживания.

Разработка методов решения задач массового обслуживания с произвольным временем обслуживания пока еще встречает немалые трудности и представляет широкое поле деятельности для желающих заняться не только изучением теории массового обслуживания и ее применением в различных областях, но и дальнейшим развитием этой теории.

Следует заметить, что значительные успехи в последнее время достигнуты благодаря использованию метода статистического моделирования (Монте-Карло). Использование этого метода существенно расширило круг задач теории массового обслуживания, эффективное решение которых может быть получено с помощью электронных цифровых вычислительных машин. В частности, метод Монте-Карло позволяет получать решение задач массового обслуживания с любым законом распределения времени обслуживания.

Следует специально остановиться еще на одном важном свойстве показательного закона распределения времени обслуживания. Оно заключается в том, что при показательном законе распределения времени обслуживания закон распределения оставшейся части времени обслуживания не зависит от того, сколько оно уже длится.

Действительно, если обозначить через $f_a(t)$ вероятность того, что обслуживание, которое уже длилось в течение времени a, продлится еще не менее t, то

$$f_0(t) = 1 - F(t) = e^{-\nu t}$$

и

$$f_0(a+t) = e^{-\nu(a+t)}.$$

Здесь $f_0(t)$ — вероятность того, что время обслуживания γ будет не меньше t. Так как $F(t) = P\{\gamma < t\}$, то $f_0(t)$, равное $P\{\gamma \geqslant t\}$, в сумме с $F(t)$ равно единице. Поэтому

$$f_0(t) = 1 - P\{\gamma < t\} = 1 - F(t).$$

С другой стороны, по теореме умножения вероятностей, вероятность того, что обслуживание продлится не меньше чем $a+t$, равна произведению вероятности того, что обслуживание продлится не меньше чем a, умноженной на вероятность того, что оно продлится не менее t, при условии, что оно уже длится в течение промежутка времени a, т. е.

$$f_0(a+t)=f_0(a)\,f_a(t).$$

Поэтому

$$f_0(a)\,f_a(t)=e^{-\nu(a+t)}$$

и

$$f_a(t)=\frac{1}{f_0(a)}\,e^{-\nu(a+t)}=e^{\nu a}e^{-\nu(a+t)}=e^{-\nu t}.$$

Таким образом, получается, что

$$f_a(t)=e^{-\nu t}=f_0(t),$$

так как из ранее сказанного видно, что

$$f_0(t)=e^{-\nu t}.$$

Получается, что условная вероятность $f_a(t)$ совпадает с вероятностью $f_0(t)$ и, следовательно, закон распределения не зависит от длины промежутка времени $(0,\ a)$, в течение которого уже длится обслуживание данного требования.

Для иллюстрации всего сказанного рассмотрим один пример, заимствованный у Б. В. Гнеденко. Предположим, что мы имеем некоторую систему обслуживания. Пусть в момент поступления очередного требования в эту систему к его обслуживанию немедленно приступают n свободных обслуживающих аппаратов, причем каждый аппарат действует независимо от других. Обслуживание будет закончено, как только его закончит один обслуживающий аппарат.

Обслуживающих систем с такой организацией обслуживания может быть очень много. Достаточно обратить внимание на некоторые ситуации, имеющие место в военном деле. Обстрел цели группой установок (например, стрельба торпедами по кораблю, налет самолетов на

корабль или какой-нибудь важный объект противника, стрельба зенитных орудий по самолету противника и т. п.) является классическим примером вышеуказанной организации системы обслуживания. Естественно, что стрельба торпедами, снарядами или бомбометание будет продолжаться до тех пор, пока цель не будет уничтожена. Следовательно, «обслуживание» в такой системе заканчивается с уничтожением объекта.

Можно привести и другие примеры, иллюстрирующие обслуживающие системы с вышеуказанной организацией, например прием радиограмм группой радиоприемных станций. Обслуживание заканчивается, как только одна из станций примет радиограмму. Или поиск корабля, потерпевшего крушение, группой самолетов. Задача будет решена, как только один из самолетов найдет корабль. Можно продолжить примеры, однако вернемся к первому.

Пусть закон распределения времени обслуживания данной «заявки» (объекта противника) каждой установкой (обслуживающим аппаратом) — показательный, со средним временем обслуживания соответственно для каждой установки

$$\frac{1}{\nu_1}, \frac{1}{\nu_2}, \ldots, \frac{1}{\nu_n}.$$

Необходимо найти закон распределения времени обслуживания всеми n аппаратами.

Предположим, что $\gamma_1, \gamma_2, \ldots, \gamma_n$ есть время обслуживания поступившего требования соответственно 1-м, 2-м,..., n-м аппаратом. Тогда вероятность того, что время обслуживания γ окажется больше t, будет равна

$$P\{\gamma > t\} = P\{\min(\gamma_1, \gamma_2, \ldots, \gamma_n) > t\},$$

потому что обслуживание будет окончено, как только его закончит один аппарат. Очевидно, что

$$P\{\min(\gamma_1, \gamma_2, \ldots, \gamma_n) > t\} = P\{\gamma_1 > t, \gamma_2 > t, \ldots, \gamma_n > t\}.$$

Последняя вероятность без особого труда может быть вычислена по теореме умножения вероятностей. Так как

в нашем примере $\gamma_1, \gamma_2, \ldots, \gamma_n$ не зависят друг от друга, то

$$P\{\gamma_1 > t, \gamma_2 > t, \ldots, \gamma_n > t\} = \prod_{i=1}^{n} P\{\gamma_i > t\}.$$

Но (по условию) закон распределения времени обслуживания $\gamma_i (i = 1, 2, \ldots, n)$ — показательный. Поэтому

$$P\{\gamma_i > t\} = 1 - F_i(t) = e^{-\nu_i t} \quad (i = 1, 2, \ldots, n).$$

Подставляя это значение величины $P\{\gamma_i > t\}$ в предыдущее равенство, получаем

$$P\{\gamma > t\} = \prod_{i=1}^{n} P\{\gamma_i > t\} = e^{-(\nu_1 + \nu_2 + \ldots + \nu_n) t}$$

Обозначим сумму $\nu_1 + \nu_2 + \ldots + \nu_n$ через β; тогда

$$P\{\gamma > t\} = e^{-\beta t},$$

иными словами, закон распределения времени обслуживания требования, поступившего в рассматриваемую нами систему, всеми n аппаратами, есть показательный закон.

Математическое ожидание времени обслуживания равно

$$\frac{1}{\beta} = \frac{1}{\nu_1 + \nu_2 + \ldots + \nu_n},$$

т. е. оно будет тем меньше, чем большее количество обслуживающих аппаратов примет участие в обслуживании. Таким образом, для нашего военного примера математически подтверждается одно из основных положений тактики о роли массированного комбинированного удара по противнику (использования сил).

Если все обслуживающие аппараты одного типа и время обслуживания каждым аппаратом требования, поступившего в систему, подчиняется одному и тому же закону распределения (например, если «обслуживаю-

щей» системой является батарея однотипных орудий и т. п.), т. е.

$$\nu_1 = \nu_2 = \ldots = \nu_n = \nu,$$

то

$$P\{\gamma > t\} = e^{-n\nu t},$$

а математическое ожидание времени обслуживания $M[\gamma]$ будет

$$M[\gamma] = \frac{1}{\beta} = \frac{1}{n\nu},$$

т. е. среднее время обслуживания уменьшается в n раз по сравнению с обслуживанием одним аппаратом. Так, например, если обслуживание осуществляется одновременно пятью аппаратами, для каждого из которых среднее время обслуживания равно 0,5 минуты, т. е. $n=5$ и $\nu=2$ (так как $\frac{1}{\nu}$ есть среднее время обслуживания, то $\frac{1}{\nu}=0{,}5$ и $\nu=2$), то среднее время обслуживания пятью аппаратами равно

$$M[\gamma] = \frac{1}{5 \cdot 2} = 0{,}1 \text{ минуты}.$$

Покажем, что при этом уменьшается и дисперсия *, т. е. степень рассеивания (разброса) времени обслуживания около его математического ожидания. Дисперсия

---

* Дисперсией случайной величины называется математическое ожидание квадрата ее отклонения от математического ожидания случайной величины (M_x). Для дискретной случайной величины, принимающей значения $x_1, x_2, \ldots, x_n$ с вероятностями $p_1, p_2, \ldots, p_n$ дисперсия равна

$$D = \sum_{i=1}^{n} (x_i - M_x)^2 p_i.$$

Для непрерывной случайной величины x с плотностью распределения $f(x)$ дисперсия равна

$$D = \int_{-\infty}^{+\infty} (x - M_x)^2 f(x)\, dx.$$

показательного закона распределения времени обслуживания равна

$$D[\gamma] = M[\gamma]^2 - (M[\gamma])^2 = \frac{1}{\nu^2}.$$

Следовательно, при групповом обслуживании требования дисперсия $D[\gamma]$ равна $\frac{1}{(n\nu)^2}$, т. е. уменьшается в n^2 раз по сравнению с обслуживанием, которое осуществляется одним аппаратом. В частности, в рассмотренном примере ($\nu = 2$) при обслуживании требования одним аппаратом дисперсия равна 0,25, а при обслуживании одновременно пятью аппаратами всего 0,01.

Итак, мы разобрали основные понятия теории массового обслуживания, ознакомились с их физическим смыслом и математическим описанием, научились определять некоторые из них. Теперь рассмотрим основные типы систем массового обслуживания и показатели эффективности их функционирования.

3. ОСНОВНЫЕ ТИПЫ СИСТЕМ МАССОВОГО ОБСЛУЖИВАНИЯ И ПОКАЗАТЕЛИ ЭФФЕКТИВНОСТИ ИХ ФУНКЦИОНИРОВАНИЯ

Все примеры задач массового обслуживания, которые мы рассматривали выше (главным образом, во втором и третьем параграфах первой главы), объединяет их общая структура.

Протекание процесса обслуживания, в общих чертах, во всех случаях одинаковое — поток требований поступает и обслуживается системой обслуживания. Однако это сходство обнаруживается только в самых общих чертах. Каждой из систем обслуживания свойственна определенная организация. В соответствии с этой организацией определяется и характер задач массового обслуживания. Поэтому дальнейшие рассуждения о типах задач основаны прежде всего на отличиях, присущих той или иной организации обслуживающей системы. Фактически во всех приведенных ранее примерах существует немало признаков, по которым можно было бы разделить все задачи на группы и как-то их классифицировать, что в значительной степени облегчило бы их решение. С этой целью в настоящем параграфе, проанализи-

ровав эти примеры, мы выделим основные типы задач и укажем основные критерии, характеризующие протекание процесса.

Системы обслуживания с потерями и без потерь. Первым признаком, позволяющим разбить задачи массового обслуживания на группы, является поведение требования, поступившего в систему в момент, когда все обслуживающие аппараты заняты.

Первая группа задач характеризуется тем, что требование *не может ждать начала обслуживания,* или, что фактически то же самое, система обслуживания отказывает требованию, поступившему в тот момент, когда все обслуживающие аппараты заняты. Ясно, что подобное свойство может иметь место только в системах с ограниченным числом обслуживающих аппаратов. Это свойство означает, что полностью отсутствуют условия для образования очереди. Если требование, поступившее в систему, получило отказ, то, следовательно, оно покидает систему необслуженным. Оно потеряно для обслуживания в этой системе. Поэтому часто подобные обслуживающие системы называются *системами с потерями,* а соответствующие задачи массового обслуживания называются *задачами обслуживания в системах с потерями.* К числу таких задач из рассмотренных в § 2 и 3 гл. 1 относятся третий, четвертый и одиннадцатый примеры. В частности, в третьем примере, по условию, «пассажир очень спешит» и не может ждать, пока освободится какой-нибудь из мастеров. В этом случае потеря требования, т. е. уход из мастерской пассажира, нуждающегося в обслуживании, происходит по инициативе, исходящей от самого пассажира (требования). В четвертом и одиннадцатом примерах абонент, обратившийся на автоматическую телефонную станцию, получает отказ, если в момент поступления его вызова (заявки) нужная линия связи занята разговором. В этом случае инициатива исходит от обслуживающей системы: АТС дает частые гудки и необслуженное требование теряется.

Вторая группа задач характеризуется тем, что требование, поступившее в систему обслуживания, может ее покинуть только тогда, когда оно полностью обслужено. Требований, ожидающих обслуживания, может оказаться довольно много. В этом случае совокупность таких требований, поступивших в систему в тот момент, когда

все обслуживающие аппараты заняты, образует *очередь*. Поэтому такие системы обслуживания получили название *систем с ожиданием или систем без потерь*. Соответствующие задачи массового обслуживания называются *задачами обслуживания в системах с ожиданием*. Примерами систем обслуживания без потерь являются системы обслуживания самолетов при посадке на аэродром (здесь «потеря» требования равносильна гибели самолета); системы ремонта неисправной техники (неисправная техника не может быть использована без предварительного ремонта). Потеря требования в последнем случае невозможна, и вся неисправная техника образует очередь.

Примерами задач массового обслуживания в системах с ожиданием из числа приведенных в § 2 и 3 гл. 1 являются первый, второй, седьмой, четырнадцатый и пятнадцатый.

В первом и во втором примерах заявкой на обслуживание является неисправность в работе станка. Ясно, что неисправный станок не может продолжать работу, пока не будут устранены неполадки. Поэтому он не может покинуть обслуживающую систему до того, пока не будет произведено полное его обслуживание. Следовательно, в этих примерах мы имеем дело с задачами обслуживания в системах с ожиданием.

Аналогичное положение имеет место в седьмом примере (неисправная сельскохозяйственная машина не может продолжать работу до окончания ремонта), в четырнадцатом примере (готовая продукция не может покинуть завода без контроля ее качества на испытательных стендах: готовые изделия будут ждать своей очереди), в пятнадцатом примере (поступившая информация будет находиться в буферной памяти электронной машины до тех пор, пока не освободятся устройства, предназначенные для переработки этой информации).

Задачи массового обслуживания в системах с потерями и в системах без потерь (или системах с ожиданием) не исчерпывают всех типов задач теории массового обслуживания. Большое количество задач лежит между задачами этих двух типов, так как организация систем массового обслуживания может быть различной. В задачах первой группы требование ни при каких условиях не остается в обслуживающей системе, если в мо-

мент его поступления все аппараты заняты; в задачах второго типа требование всегда ждет, пока не будет начато и закончено его обслуживание; для задач третьей группы характерно *наличие некоторых промежуточных условий.*

Требование, поступившее в систему обслуживания в момент, когда все обслуживающие аппараты заняты, необязательно должно покинуть систему, но и необязательно будет ждать конца обслуживания. Требование покинет систему, если будут выполнены некоторые дополнительные условия. При этом в различных задачах условия, при которых требование должно покинуть обслуживающую систему, могут быть самыми разнообразными. Из числа рассмотренных в § 2 и 3 гл. 1 примеров только десятый пример относится к этой группе задач. В этом примере заявка на доставку грузов может быть принята и в том случае, когда все автомашины в момент ее поступления заняты. Условием, при котором она не будет принята, является определенная длина очереди. (Заявка будет отклонена, если в очереди уже стоит определенное число ранее принятых заказов.)

В других задачах могут ставиться дополнительно самые разнообразные условия, при которых требование покидает систему. Так, например, в некоторых задачах массового обслуживания таким условием является ограниченное *время пребывания* требования в системе обслуживания. Если суммарное время пребывания требования в системе обслуживания, которое складывается из времени ожидания начала обслуживания и времени обслуживания, превзойдет определенную величину, то требование покидает обслуживающую систему независимо от того, начато его обслуживание или нет, а если обслуживание начато, то независимо от того, закончено оно или нет. К такого рода задачам относится, например, текущий осмотр проходящих поездов на железнодорожной станции. Если за время стоянки поезда осмотр не закончен, то поезд отправляется дальше, конечно, при условии, что никаких существенных неисправностей как за время осмотра, так и за время движения не обнаружено.

Другим примером условий может быть ограничение времени ожидания начала обслуживания. Если время ожидания очередным требованием начала обслуживания превзойдет определенную величину, то требование поки-

нет обслуживающую систему, но если обслуживание будет начато, то оно будет закончено независимо от того, какое время будет затрачено на его обслуживание.

Примером процессов такого типа может быть обслуживание абонента на пункте междугородней телефонной связи. Абоненту должен быть предоставлен разговор в течение часа с назначенного момента. Если за это время разговор не состоялся, то, как правило, абонент покидает телефонную станцию. Но если разговор был начат до истечения часа, то он (разговор) будет закончен даже если суммарное время ожидания и разговора превзойдет один час. Такого рода системы называются *смешанными системами обслуживания*, чем подчеркивается их промежуточное положение между системами с потерями и системами с ожиданием.

Обслуживающие системы с ограниченным и неограниченным числом обслуживающих аппаратов. Кроме деления задач массового обслуживания на три группы по характеру поведения требования в обслуживающей системе, они могут различаться и по числу обслуживающих аппаратов в обслуживающей системе. Так, все задачи массового обслуживания могут быть разбиты на два типа: задачи обслуживания в системах с *ограниченным числом обслуживающих аппаратов* и задачи обслуживания в системах с *неограниченным числом обслуживающих аппаратов*. К первому типу относятся задачи примеров 1, 2, 3, 4. Ясно, что реально неограниченного числа обслуживающих аппаратов ни в одной системе быть не может, однако могут быть системы, в которых число обслуживающих аппаратов настолько велико, что их можно относить к системам с неограниченным числом обслуживающих аппаратов. При очень большом количестве обслуживающих аппаратов ряд задач может быть с достаточной точностью и гораздо проще решен, если рассматривать эту систему как систему с неограниченным числом обслуживающих аппаратов. К числу таких задач относятся задачи, приведенные в примерах 5, 6, 12, 13 гл. 1. В пятом примере при большом количестве автомашин, например, в масштабе всей страны, гораздо проще решать задачу, если считать, что число автомашин неограниченно. В этом можно убедиться при изучении методов решений, изложенных в § 2 гл. 3. Аналогичное явление имеет место при определении надежности

электронных вычислительных машин, решение задачи по определению мощности энергосистемы и т. п.

Задачи с ограниченным и неограниченным потоком требований. Задачи массового обслуживания различаются еще по одному признаку — по числу требований, которые могут одновременно находиться в обслуживающей системе. В некоторых задачах число требований, одновременно находящихся в обслуживающей системе, принципиально не может быть больше определенного числа. Это задачи с ограниченным числом требований. Сюда относится задача обслуживания группы станков (число одновременно ремонтируемых станков не может превзойти общего числа обслуживаемых станков) и ряд других задач, связанных с ремонтом техники.

Однако далеко не во всех процессах массового обслуживания имеет место это свойство. В ряде задач число требований, находящихся одновременно на обслуживании, может быть очень большим, настолько большим, что практически можно и удобно рассматривать поток требований как неограниченный. Здесь под словом «удобно» мы понимаем метод вычислений, а слово «можно» означает, что степень точности вычислений при этом будет достаточной для практического использования получаемого результата.

Рассматривать поток требований как неограниченный удобно также в тех случаях, когда заранее нельзя предопределить, насколько много требований может поступить. К числу задач, в которых поток требований можно условно считать неограниченным, относятся задачи, рассматриваемые в девятом, четырнадцатом и пятнадцатом примерах первой главы. Так, например, в больших городах, таких, как Москва или Ленинград, при огромном числе телевизоров, которое сейчас выпускается, нельзя заранее точно предсказать, сколько из них может потребовать ремонта. Поэтому практически удобно считать такой поток требований неограниченным. Правда не нужно забывать, что при этом вероятность того, что ремонта потребуют сразу очень много телевизоров, мала и будет тем меньше, чем больше это число. Поэтому начиная с некоторого достаточно большого числа телевизоров, требующих ремонта, вероятность этого события станет настолько малой, что ею практически можно будет пренебречь. Это и позво-

ляет рассматривать такой поток как неограниченный.

Аналогичная картина имеет место с информацией, поступающей в управляющую электронную вычислительную машину (пятнадцатый пример). Число сообщений, поступивших в машину и ожидающих в буферной памяти переработки, может быть очень большим, но каким именно — заранее предсказать нельзя. Поэтому такой поток удобно рассматривать как неограниченный. Здесь так же, как в предыдущем случае, вероятность того, что за данный промежуток времени поступит очень много требований, будет мала. Следовательно, результаты, полученные при вычислениях для неограниченного потока, будут близки к тем, которые мы получили бы, рассматривая поток состоящим из ограниченного числа требований.

Упорядоченные и неупорядоченные системы обслуживания. Укажем еще на один признак, по которому различаются обслуживающие системы. Если аппараты обслуживающей системы расположены последовательно (пронумерованы) и очередное требование поступает сначала на первый из них и лишь только в том случае, если он занят, передается второму аппарату и т. д., то, следуя А. Я. Хинчину, такую систему будем называть упорядоченной. Следовательно, в упорядоченной системе требование поступит на обслуживающий аппарат с номером n только в том случае, если в момент его поступления аппараты с номерами 1, 2, 3, . . ., $(n-1)$ заняты. Такой системой является система, рассмотренная в шестнадцатом примере. В этом примере группа упаковочных автоматов расположена последовательно и готовое изделие поступает на упаковку в первый свободный автомат.

Все остальные системы обслуживания, в которых требования распределяются между обслуживающими аппаратами по любому другому принципу, относятся к числу неупорядоченных систем.

Критерии эффективности для процессов массового обслуживания

При решении задач, связанных с массовым обслуживанием, большое значение имеет правильный выбор критериев, характеризующих изучаемый процесс. Одна и та же система обслуживания может характеризовать-

ся с различных точек зрения различными критериями эффективности. Выбор критерия эффективности — очень важный этап исследования, поскольку от того, насколько правильно выбран критерий, зависит оценка самой системы обслуживания. Выбор того или иного критерия должен производиться в каждом конкретном случае исходя из тех задач, которые ставятся перед системой. Перечислить все критерии, которые могут или могли бы быть полезными во всех задачах массового обслуживания, невозможно, поэтому ограничимся указанием наиболее существенных и наиболее часто используемых критериев.

Учитывая, что выбор критерия зависит от типа исследуемой задачи, будем рассматривать их применительно к трем основным группам задач массового обслуживания.

Критерии эффективности в системах обслуживания с потерями. При изучении систем с потерями важнейшей характеристикой таких систем является *вероятность отказа* в обслуживании (вероятность потери требования). Так как отказ от обслуживания требования происходит тогда, когда все обслуживающие аппараты заняты, то вероятность отказа равна вероятности того, что все обслуживающие аппараты окажутся занятыми. Очевидно, что этот критерий пригоден для систем с ограниченным числом обслуживающих аппаратов. Вероятность отказа определяет, в какой степени данная система обслуживания способна удовлетворить поступающий поток требований. Нужно отметить, что этот критерий (вероятность отказа) не связан с качеством обслуживания внутри системы тех требований, которые были приняты на обслуживание. Он дает только «внешнюю» оценку способности системы приступить к обслуживанию поступившего требования.

Степень загрузки обслуживающей системы может характеризоваться таким критерием, как *среднее число занятых аппаратов.* Этот критерий пригоден для любых систем, независимо от числа обслуживающих аппаратов. Более полно загрузка системы может характеризоваться законом распределения количества занятых аппаратов. В некоторых случаях для систем с потерями является полезным такой критерий, как среднее количество потерянных требований за определенный промежуток вре-

мени. Эти основные критерии позволяют в конкретных задачах получить ряд других показателей, необходимых для более полной характеристики процесса обслуживания. Они позволяют делать выводы о степени износа оборудования, рентабельности обслуживающей системы и т. д.

Критерии эффективности в системах обслуживания без потерь. Напомним, что к этой группе относятся задачи, в которых поступившее требование будет находиться в системе обслуживания до тех пор, пока не закончится его обслуживание. Исходя из этого можно указать основные критерии эффективности функционирования таких систем обслуживания. Это, прежде всего, длина очереди, которая определяется числом требований, ожидающих обслуживания. Длина очереди зависит от ряда факторов: от того, когда и сколько требований поступило в систему, сколько времени затрачено на обслуживание поступивших требований и т. д. Поэтому, вообще говоря, длина очереди является случайной величиной.

В качестве показателя длины очереди можно использовать ее математическое ожидание. Математическое ожидание длины очереди характеризует, какие потери будут из-за простаивания в очереди в ожидании обслуживания. Однако этот критерий не является достаточно полным. С точки зрения обслуживания и чистых потерь за счет стояния в очереди, большее значение имеет такой критерий *как время ожидания начала обслуживания.* Время ожидания начала обслуживания (β) является случайной величиной, которая зависит от количества требований, находящихся в данный момент в очереди, времени окончания обслуживания всех предыдущих требований и т. д. Поэтому наиболее полной характеристикой времени ожидания будет закон распределения β. Закон распределения времени ожидания дает полную картину того, как протекает ожидание начала обслуживания. Он позволяет также ответить на все вопросы, относящиеся к ожиданию обслуживания. Так, например, можно найти вероятность того, что обслуживание очередного требования будет начато немедленно при условии, что закон распределения времени ожидания известен.

Заметим, что эта вероятность совпадает с вероятностью того, что в обслуживающей системе есть хоть один

свободный аппарат. Закон распределения времени ожидания удается найти далеко не всегда. В этих случаях приходится пользоваться более простыми критериями, такими, например, как *среднее время ожидания начала обслуживания (математическое ожидание времени начала обслуживания).*

В некоторых задачах более удобным критерием является все же длина очереди, полная характеристика которой может быть задана *законом распределения длины очереди.* Заметим, что если время ожидания β может принимать любые положительные значения, то длина очереди является целым положительным числом.

С точки зрения оценки степени загруженности обслуживающей системы значительный интерес представляет такой показатель, как *среднее число занятых обслуживающих аппаратов.* Число занятых аппаратов является случайной величиной, которая зависит от потока поступающих требований и времени обслуживания каждого. Поэтому наиболее полно степень загруженности обслуживающей системы может быть описана *законом распределения числа занятых аппаратов.*

Кроме перечисленных критериев, системы обслуживания без потерь могут характеризоваться таким критерием, как вероятность иметь не более n единиц в очереди в момент t при заданном начальном состоянии системы. Для примера ремонтных мастерских, обслуживающих различную технику, этот критерий является весьма существенным, ибо он позволяет установить гарантийное число исправных единиц из общего числа имеющихся. Все эти основные критерии позволяют в конкретных процессах массового обслуживания найти все необходимые характеристики протекания процессов.

Критерии эффективности обслуживающих систем смешанного типа. При решении задач смешанного типа могут иметь место самые различные условия, от которых соответственно будут зависеть и критерии эффективности.

Критерии, характеризующие протекание процесса обслуживания в системах смешанного типа, в основном совпадают с теми, которые были перечислены для задач первой и второй групп. В зависимости от специфики условий каждой задачи могут появиться и свои специфические критерии. Так, например, в процессах обслужива-

ния, в которых требование не может находиться в обслуживающей системе больше заданного времени и покидает систему, как только истечет это время, независимо от того, начато его обслуживание или нет, может представлять интерес такой критерий, как *время, затраченное на обслуживание тех требований, которые покинут систему до окончания обслуживания.*

В этих системах требование покидает систему в двух случаях: когда обслуживание его закончено или когда истекло допустимое время пребывания в системе. Во втором случае обслуживание требования может быть или не начато совсем, или же произведено частично. Если частичное обслуживание относится к числу непроизводительных затрат, то все время, затраченное на обслуживание этих требований, фактически расходуется впустую. Поэтому большое значение имеет такой критерий, как *время, затраченное каждым аппаратом на обслуживание таких требований, и суммарное время, затраченное всеми аппаратами системы.* Если отношение этого времени ко всему времени функционирования системы велико, то это означает, что обслуживание организовано плохо и производительность системы низка. В других задачах могут быть свои частные критерии.

Еще раз подчеркнем, что нами рассмотрены *основные* критерии функционирования систем массового обслуживания. Частные критерии в зависимости от специфики изучаемых конкретных процессов могут быть получены из этих основных с учетом особенностей каждого процесса.

ГЛАВА ТРЕТЬЯ

НЕКОТОРЫЕ ЗАДАЧИ МАССОВОГО ОБСЛУЖИВАНИЯ И ИХ РЕШЕНИЕ

В этой главе будут рассмотрены аналитические методы решения некоторых задач массового обслуживания и приложения этих методов к решению конкретных задач. Знание этих методов очень важно, так как такие задачи могут встречаться в самых различных областях деятельности.

Однако не для всех задач типа массового обслуживания аналитические методы разработаны настолько, что ими можно с успехом пользоваться. Больше того, существует немало задач, решение которых такими методами пока невозможно. Тем не менее это нисколько не снижает роли и значения этих методов, а наоборот, подчеркивает важность и необходимость их дальнейшего развития.

Как уже отмечалось выше, электронные цифровые вычислительные машины, используемые для решения задач массового обслуживания методом статистических испытаний, практически позволяют решить многие задачи, аналитическое решение которых неизвестно.

Данная глава посвящена разбору аналитических методов решения, главным образом, тех задач, которые перечислены нами в § 2 и 3 гл. 1. Однако эти методы с успехом могут быть использованы и для решения многих других задач подобного типа.

1. ЗАДАЧИ ОБСЛУЖИВАНИЯ В СИСТЕМАХ С ПОТЕРЯМИ

Как уже отмечалось выше, первая группа задач массового обслуживания характеризуется тем, что требова-

ние, поступившее в момент, когда все обслуживающие аппараты системы заняты, получает отказ в обслуживании и покидает систему. К двум примерам задач такого типа, приведенным в § 3 гл. 1, можно добавить еще неограниченное количество подобных задач. Например, к числу процессов такого типа относится прием (перехват) сообщений по радио, который осуществляется группой приемников. Если в момент передачи очередного сообщения все радиоприемники заняты приемом сообщений, передача которых была начата раньше, то очередное сообщение будет потеряно (частично или полностью).

Ясно, что решение каждой конкретной задачи отдельно вызвало бы много трудностей. Поэтому нужно рассмотреть абстрактную систему обслуживания, отвлеченный поток требований, не связанный ни с каким реальным потоком, и получить необходимые характеристики для этого процесса, а затем из этого общего решения могут быть получены частные решения всех задач этого типа. Так мы и поступим. Найдем общее решение задачи, а затем покажем, каким образом можно его использовать для конкретных задач. При решении ограничимся рассмотрением случая простейшего потока требований и показательного закона распределения времени обслуживания. Начнем с подробной формулировки задачи.

Постановка задачи. Рассмотрим обслуживающую систему, состоящую из n аппаратов. Предположим, что она относится к числу систем с потерями, т. е. требование, поступившее в момент, когда все обслуживающие аппараты заняты, покидает систему. Если в системе в момент поступления требования есть хоть один свободный аппарат, то он немедленно приступает к обслуживанию требования. Каждый аппарат может одновременно обслуживать только одно требование. Для рассматриваемой задачи безразлично, относится система к числу упорядоченных или нет. Время обслуживания одного требования одним аппаратом подчинено показательному закону с параметром ν. Напомним, что это означает следующее: вероятность того, что время обслуживания γ меньше t, равна

$$P\{\gamma < t\} = F(t) = 1 - e^{-\nu t},$$

а $\frac{1}{\nu}$ есть математическое ожидание времени обслуживания (гл. 2, § 2).

В систему на обслуживание поступает простейший поток требований с параметром λ. Напомним, что это значит следующее: поток стационарный, ординарный, без последействия. Вероятность поступления точно k требований за время t равна

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

где λ — математическое ожидание числа требований за единицу времени (гл. 2, § 1).

Основным критерием функционирования такой системы обслуживания является *вероятность отказа*, т. е. вероятность того, что в момент поступления очередного требования все обслуживающие аппараты заняты. Кроме того, в некоторых прикладных задачах может представлять определенный интерес и такой критерий, как *среднее число аппаратов, занятых обслуживанием.* Первый критерий характеризует соотношение потока обслуженных требований и входящего потока, т. е. полноту обслуживания входящего потока. Второй критерий характеризует степень загрузки обслуживающей системы. Конечной целью решения этой задачи является вывод формул для вычисления вероятности отказа и математического ожидания числа занятых обслуживающих аппаратов.

Прежде чем приступить к составлению уравнений, из которых можно получить решение задачи, определим те основные состояния, в которых может находиться обслуживающая система. В каждый момент времени она может находиться в одном из следующих состояний:

— все обслуживающие аппараты свободны,
— занят один обслуживающий аппарат,
— заняты два обслуживающих аппарата и т. д.,
— заняты все n обслуживающих аппаратов.

Ясно, что всех возможных состояний $n+1$. Обозначим через $N(t)$ число аппаратов, занятых в момент времени t при условии, что в начальный момент занято k аппаратов. Функция $N(t)$ есть случайная величина, которая определяется: моментами освобождения тех k ап-

паратов, которые были заняты в начальный момент, моментами поступления новых требований на обслуживание, моментами окончания обслуживания этих новых требований. Если в некоторый момент времени занят k_1 аппарат, т. е. $N(t_1)=k_1$, то легко видеть, что дальнейшее течение процесса обслуживания не зависит от того, что было до момента t_1. Действительно, моменты освобождения k_1 занятых аппаратов не зависят от того, что было до t_1, так как закон распределения времени обслуживания — показательный. Это свойство показательного закона распределения времени обслуживания было доказано в § 2 гл. 2.

Моменты появления новых требований не зависят от того, что было до момента t_1, так как поток требований — простейший и, следовательно, обладает свойством отсутствия последействия. Независимость окончания обслуживания новых требований от t_1 очевидна, поэтому все величины, определяющие $N(t)$, не зависят от того, что было до момента t_1. Функция $N(t)$ является случайной функцией. Такие случайные процессы, течение которых не зависит от прошлого, относятся к очень важному типу процессов, называемых *марковскими процессами, или процессами Маркова* * Следовательно, случайный процесс $N(t)$ является марковским процессом.

Обозначим через $P_k(t)$ вероятность того, что в момент времени t занято k аппаратов системы, т. е.

$$P_k(t)=P\{N(t)=k\} \quad (0\leqslant k\leqslant n),$$

а через $P_{ik}(t)$ — условную вероятность того, что через время t будет занято k аппаратов, если вначале было занято i аппаратов. В дальнейшем часто будет использоваться довольно очевидная формула

$$P_k(t_1+t_2)=\sum_{i=0}^{n}P_i(t_1)\,P_{ik}(t_2), \tag{3.1}$$

---

* Процессом Маркова, или марковским процессом называется такой случайный процесс, течение которого после некоторого момента t не зависит от того, что было до этого момента. Точнее, вероятность перехода из состояния P_i в момент t_1 в состояние P_j в момент t_2 зависит от t_1, t_2, i, j и не зависит от тех состояний, которые были до момента t_1.

которая означает, что если в момент времени t_1 в системе занято i аппаратов с вероятностью $P_i(t_1)$ ($i=1, 2, 3, \ldots, n$), а условная вероятность перехода к k занятым аппаратам через промежуток времени t_2 равна $P_{ik}(t_2)$ ($i=0, 1, 2, \ldots, n$), то полная вероятность того, что в момент t_1+t_2 занято k аппаратов, равна сумме произведений

$$P_i(t_1)\cdot P_{ik}(t_2)$$

по всем возможным i ($i=0, 1, 2, \ldots, n$). В частности, если система состоит из одного обслуживающего аппарата, то

$$P_0(t+\Delta t)=P_0(t)\,P_{00}(\Delta t)+P_1(t)\,P_{10}(\Delta t)$$

есть вероятность того, что этот аппарат свободен. Если в начальный момент он был свободен, то $P_1(t)=0$, т. е.

$$P_0(t+\Delta t)=P_0(t)\,P_{00}(\Delta t).$$

Величина $P_{00}(\Delta t)$ есть условная вероятность того, что за время Δt аппарат не будет занят; но эта вероятность равна тому, что за это время не поступит ни одного требования. Так как поток простейший, то из формулы (2.3) следует, что эта вероятность равна

$$P_{00}(\Delta t)=1-\lambda\Delta t+o(\Delta t).$$

Подставляя $P_{00}(\Delta t)$, получаем

$$P_0(t+\Delta t)=P_0(t)\,[1-\lambda\Delta t+o(\Delta t)],$$

откуда

$$\frac{P_0(t+\Delta t)-P_0(t)}{\Delta t}=-\lambda P_0(t)+\frac{o(\Delta t)}{\Delta t}.$$

Переходя к пределу при $\Delta t\to 0$, получаем

$$\frac{dP_0(t)}{dt}=-\lambda P_0(t).$$

Интегрируя это дифференциальное уравнение с разделяющимися переменными, получаем

$$P_0(t)=C\,e^{-\lambda t}.$$

Из начального условия $P_0(0)=1$ следует, что $C=1$, т. е.

$$P_0(t)=e^{-\lambda t}.$$

Таким образом, если в начальный момент аппарат был свободен, то вероятность того, что он будет свободен через время t, равна

$$P_0(t)=e^{-\lambda t},$$

т. е. она тем больше, чем меньше λ.

Для многих процессов имеет место очень важное свойство, которое заключается в том, что $P_k(t)$ при $t\to\infty$ стремятся к конечным пределам p_k, где p_k — постоянные числа, не зависящие от начального положения, в котором находилась обслуживающая система. Наличие этих пределов для рассматриваемой задачи вытекает из следующей теоремы.

Теорема Маркова. *Если существует такое* $t>0$, что $P_{ik}(t)>0$* $(i=0, 1, 2, \ldots, n;\ k=0, 1, 2, \ldots, n)$, *то для марковского процесса предел*

$$\lim_{t\to\infty} P_{ik}(t)=p_k \quad (k=0,\ 1,\ 2,\ldots,n)$$

существует и не зависит от i.

Доказательство этой теоремы можно найти в книге А. Я. Хинчина «Математические методы теории массового обслуживания».

Заметим, что здесь существенным является строгая положительность $P_{ik}(t)$. Так как $P_{ik}(t)$ есть условная вероятность перехода от i занятых аппаратов в системе обслуживания к k занятым аппаратам, то $P_{ik}(t)>0$ означает, что при достаточно большом промежутке времени возможен переход от i занятых аппаратов к k. Но так как это неравенство справедливо при любых i и k, то, следовательно, переход к k занятым аппаратам возможен из любого начального состояния. Очевидно, что для рассматриваемой задачи это свойство справед-

* Если есть $t>0$ такое, что все $P_{ik}(t)>0$ $(0\leqslant i, k\leqslant n)$, то марковский процесс называется *транзитивным*. Это свойство означает, что возможен переход системы из любого состояния за конечное время.

ливо. Действительно, система, состоящая из n обслуживающих аппаратов, может из любого состояния перейти со временем, если взять это время достаточно большим, в любое другое возможное состояние.

Теперь нетрудно показать, что вероятность того, что занято k аппаратов системы, имеет предел при $t \to \infty$. Это вытекает из следующей теоремы.

Теорема. *Для того, чтобы вероятность* $P_k(t)$ при $t \to \infty$ $(k=0, 1, 2 \ldots)$ *имела предел* p_k *независимо от начальных состояний, необходимо и достаточно, чтобы к этим же пределам стремились условные вероятности* $P_{ik}(t)$ *при любом начальном состоянии* i.

Доказательство этой теоремы приведено в приложении. Вместе с теоремой Маркова эта теорема доказывает наличие стационарного решения в рассматриваемой задаче. Эти замечания, которые сохраняют свою силу для всех процессов, рассмотренных в книге, позволяют значительно упростить дальнейшие вычисления, как это будет видно ниже. Для всех рассматриваемых процессов можно будет отыскивать предельные решения, которые достаточно полно описывают установившийся процесс.

Дифференциальные уравнения решения задачи. Теперь перейдем к составлению уравнения для определения $P_k(t)$. При выводе будем пользоваться обозначениями и результатами, полученными в § 1 и 2 гл. 2. В частности, напомним следующие результаты, полученные нами ранее:

— вероятность того, что за время Δt поступит хотя бы одно требование,

$$W(\Delta t) = \lambda \Delta t + o(\Delta t) \quad (\Delta t \to 0),$$

— вероятность того, что за время Δt поступит не меньше двух требований:

$$\psi(\Delta t) = o(\Delta t) \quad (\Delta t \to 0).$$

Напомним еще раз, что здесь $o(\Delta t)$ обозначает бесконечно малую величину более высокого порядка, чем Δt: т. е.

$$\lim \frac{o(\Delta t)}{\Delta t} = 0, \quad \Delta t \to 0.$$

Для вывода уравнений, определяющих $P_k(t)$, воспользуемся уравнением (3.1)

$$P_k(t_1+t_2)=\sum_{i=0}^{n}P_i(t_1)\,P_{ik}(t_2)\ (k=0,\ 1,\ 2,\ldots).$$

Положим в нем $t_1=t$, а $t_2=\Delta t$, тогда

$$P_k(t+\Delta t)=\sum_{i=0}^{n}P_i(t)P_{ik}(\Delta t)=P_0(t)\,P_{0k}(\Delta t)+$$

$$+P_1(t)\,P_{1k}(\Delta t)+\ldots+P_n(t)\,P_{nk}(\Delta t)\quad (k=0,\ 1,\ 2,\ldots,n). \tag{3.2}$$

В этом равенстве величины $P_{ik}(\Delta t)$ есть условные вероятности того, что за время Δt обслуживающая система перейдет от i занятых аппаратов к k. Вычислим эти вероятности. Положим $k=0$ и определим $P_{00}(\Delta t)$ — вероятность того, что в системе через время Δt не будет ни одного требования при условии, что в начальный момент она была свободна. Это событие может произойти следующим образом. За время Δt не поступит ни одного требования, тогда, очевидно, все обслуживающие аппараты будут свободны. Вероятность эта равна

$$1-W(\Delta t)=1-\lambda\Delta t+o(\Delta t),$$

так как $W(\Delta t)$ есть вероятность поступления хотя бы одного требования за время Δt, а следовательно, $1-W(\Delta t)$ есть вероятность отсутствия требований за это время.

Но возможно, что за время Δt все-таки поступит одно требование, при этом оно будет обслужено и покинет систему. Тогда по истечении промежутка времени Δt все обслуживающие аппараты будут свободны. Пусть требование поступило в начальный момент отрезка Δt, тогда вероятность того, что оно будет обслужено за время, не превосходящее Δt, равна

$$F(\Delta t)=1-e^{-\nu\Delta t}.$$

Разлагая $e^{-\nu\Delta t}$ в ряд по степеням $\nu\Delta t$, получаем

$$F(\Delta t) = 1 - 1 + \nu\Delta t - \frac{(\nu\Delta t)^2}{2!} + \frac{(\nu\Delta t)^3}{3!} - \ldots =$$

$$= \nu\Delta t + o(\Delta t).$$

Сумма

$$-\frac{(\nu\Delta t)^2}{2!} + \frac{(\nu\Delta t)^3}{3!} - \frac{(\nu\Delta t)^4}{4!} + \ldots \leqslant \frac{(\nu\Delta t)^2}{2!} = o(\Delta t),$$

так как это знакопеременный ряд и сумма его по абсолютной величине не превосходит первого члена, т. е. $\frac{(\nu\Delta t)^2}{2!}$, а это величина порядка $(\Delta t)^2$.

Если требование поступило не в начальный момент промежутка Δt, то вероятность того, что оно будет обслужено за оставшееся время, еще меньше. Но вероятность поступления хотя бы одного требования за время Δt равна

$$W(\Delta t) = \lambda\Delta t + o(\Delta t).$$

Таким образом, вероятность того, что за время Δt требование поступит и будет обслужено, не превосходит

$$W(\Delta t)\, F(\Delta t) = [\lambda\Delta t + o(\Delta t)]\,[\nu\Delta t + o(\Delta t)] =$$

$$= \lambda\nu\,(\Delta t)^2 + o(\Delta t),$$

т. е. эта вероятность есть величина бесконечно малая более высокого порядка, чем Δt. Вероятность же того, что за время Δt поступят и будут обслужены больше двух требований, еще меньше. Следовательно, по теореме сложения вероятностей

$$P_{00}(\Delta t) = 1 - \lambda\Delta t + o(\Delta t).$$

Теперь найдем значение $P_{kk}(\Delta t)$. Для вычисления $P_{kk}(\Delta t)$ воспользуемся тем, что

$$\sum_{i=0}^{n} P_{ki}(\Delta t) = 1.$$

Справедливость этого равенства очевидна, так как оно означает, что если в некоторый момент занято k ап-

паратов, то через время Δt система или остается в прежнем состоянии, т. е. будет занято k аппаратов, или перейдет в какое-нибудь другое возможное состояние, т. е. к i занятым аппаратам ($i=0, 1, 2, \ldots, k-1, k+1, \ldots, n$). Из этого равенства вытекает, что

$$P_{kk}(\Delta t) = 1 - P_{k0}(\Delta t) - P_{k1}(\Delta t) - \ldots - P_{k,k-1}(\Delta t) - P_{k,k+1}(\Delta t) - \ldots - P_{kn}(\Delta t). \quad (3.3)$$

Покажем, что все члены правой части этого равенства, кроме $P_{k,k-1}(\Delta t)$ и $P_{k,k+1}(\Delta t)$ есть величины бесконечно малые более высокого порядка, чем Δt. Действительно $P_{k,i}(\Delta t)$ при $|i-k| \geqslant 2$ есть вероятность того, что за время Δt в систему поступит не меньше двух требований (при $i-k \geqslant 2$), а вероятность этого

$$\psi(\Delta t) = o(\Delta t) \quad (\Delta t \to 0),$$

или за это время систему покинут не меньше двух обслуженных требований (при $k-i \geqslant 2$). Вероятность этого последнего события равна

$$[F(\Delta t)]^{k-i} = [1 - e^{-\nu \Delta t}]^{k-i} = [-\nu \Delta t + o(\Delta t)]^{k-i}.$$

Здесь мы воспользовались ранее полученным равенством

$$F(\Delta t) = -\nu \Delta t + o(\Delta t) \quad (\Delta t \to 0).$$

Но так как, $k-i \geqslant 2$, то

$$[F(\Delta t)]^{k-i} \leqslant [-\nu \Delta t + o(\Delta t)]^2 = o(\Delta t).$$

Следовательно, вероятность того, что за время Δt систему покинут не меньше чем два обслуженных требования, есть бесконечно малая величина более высокого порядка, чем Δt. Таким образом, вероятность $P_{ki}(\Delta t)$ при $|i-k| \geqslant 2$ имеет порядок $o(\Delta t)$, т. е.

$$P_{k,i}(\Delta t) = o(\Delta t) \quad |i-k| \geqslant 2.$$

Учитывая это, равенство (3.3) можно переписать в следующем виде:

$$P_{k,k}(\Delta t) = 1 - P_{k,k-1}(\Delta t) - P_{k,k+1}(\Delta t) + o(\Delta t). \quad (3.4)$$

Вычислим величину $P_{k,\,k-1}(\Delta t)$, которая является вероятностью того, что из k занятых аппаратов за время Δt освободится по крайней мере один из них. Вероятность того, что за время Δt занятый аппарат не освободится, равна вероятности того, что время обслуживания превзойдет Δt, а так как вероятность того, что время обслуживания будет меньше Δt, равна $F(\Delta t) = 1 - e^{-\nu t}$, то, следовательно, вероятность того, что оно будет больше Δt, равна

$$1 - F(\Delta t) = e^{-\nu \Delta t}.$$

Если в момент t занято k аппаратов, то вероятность того, что не освободится ни один из них, может быть найдена по теореме умножения вероятностей. Так как $e^{-\nu \Delta t}$ есть вероятность того, что один занятый аппарат за время Δt не освободится, то вероятность того, что все k аппаратов не освободятся, равна

$$[e^{-\nu \Delta t}]^k = e^{-k\nu \Delta t}.$$

Следовательно, вероятность того, что за это время освобождается хотя бы один из k аппаратов, равна

$$1 - e^{-\nu k \Delta t} = \nu k \Delta t + o(\Delta t)^* \quad (\Delta t \to 0).$$

Так как выше было показано, что вероятность освобождения двух и более аппаратов за время Δt имеет порядок $o(\Delta t)$, то, следовательно, вероятность того, что за время Δt освободится точно один аппарат из k, равна

$$P_{k,\,k-1}(\Delta t) = \nu k \Delta t + o(\Delta t) \quad (0 < k \leqslant n).$$

Нами опущено доказательство того, что вероятность поступления и окончания обслуживания одного и того же числа требований за время Δt имеет порядок $o(\Delta t)$, но это можно показать методом, аналогичным предыдущему.

Осталось вычислить вероятность $P_{k,\,k+1}(\Delta t)$. Нетрудно видеть, что эта величина с точностью до $o(\Delta t)$ равна вероятности поступления требования за время Δt, т. е.

$$P_{k,\,k+1}(\Delta t) = \lambda \Delta t + o(\Delta t) \quad (0 \leqslant k \leqslant n - 1).$$

---

* Это выражение получено путем разложения $e^{-k\nu \Delta t}$ в ряд по степеням показателя подобно тому, как раньше было получено выражение для $F(\Delta t) = \nu \Delta t + o(\Delta t)$.

Представляя полученные выражения для $P_{k,k-1}$ и $P_{k,k+1}$ в (3.4), получаем

$$P_{k,k}(\Delta t)=1-\lambda\Delta t-\nu k\Delta t+o(\Delta t)\quad(0<k\leqslant n-1).$$

Из формулы (3.4) можно получить значения $P_{n,n}(\Delta t)$, если учесть, что $P_{n,n+1}(\Delta t)=0$, так как переход от n занятых аппаратов к $n+1$ занятому аппарату невозможен в силу того, что их всего n по условию. Поэтому

$$P_{n,n}(\Delta t)=1-\nu n\Delta t+o(\Delta t).$$

Объединяя все полученные выражения, получаем следующую сводку асимптотических формул для переходных вероятностей:

$$\left.\begin{aligned}
&P_{0,0}(\Delta t)=1-\lambda\Delta t+o(\Delta t)\\
&P_{k,k}(\Delta t)=1-\lambda\Delta t-\nu k\Delta t+o(\Delta t)\\
&\qquad\qquad(1\leqslant k\leqslant n-1)\\
&P_{n,n}(\Delta t)=1-\nu n\Delta t+o(\Delta t)\\
&P_{i,k}(\Delta t)=o(\Delta t)\quad(|i-k|\geqslant 2)\\
&P_{k,k-1}(\Delta t)=\nu k\Delta t+o(\Delta t)\\
&P_{k,k+1}(\Delta t)=\lambda\Delta t+o(\Delta t)
\end{aligned}\right\}. \tag{3.5}$$

Подставляя эти выражения в (3.2) при $k=0$, $1\leqslant k\leqslant n-1$ и $k=n$, получаем следующую группу уравнений:

$$\left.\begin{aligned}
P_0(t+\Delta t)&=P_0(t)(1-\lambda\Delta t)+P_1(t)\,\nu\Delta t+o(\Delta t)\\
P_k(t+\Delta t)&=P_{k-1}(t)\,\lambda\Delta t+P_k(t)(1-\lambda\Delta t-\nu k\Delta t)+\\
&\quad+P_{k+1}(t)(k+1)\Delta t+o(\Delta t)\\
&\qquad\qquad(1\leqslant k\leqslant n-1)\\
P_n(t+\Delta t)&=P_{n-1}(t)\,\lambda\Delta t+\\
&\quad+P_n(t)(1-\nu n\Delta t)+o(\Delta t)
\end{aligned}\right\}. \tag{3.6}$$

В процессе подстановки все члены, содержащие $o(\Delta t)$, сгруппированы и заменены $o(\Delta t)$. Это законно, так как сумма конечного числа бесконечно малых величин одного

порядка есть также бесконечно малая величина того же порядка.

Теперь произведем некоторые преобразования в этой группе уравнений (3.6): перенесем в первом уравнении слагаемое $P_0(t)$, во втором уравнении перенесем $P_k(t)$, а в третьем $P_n(t)$ влево, после чего разделим обе части всех уравнений на Δt. В результате получим

$$\left.\begin{aligned}
&\frac{P_0(t+\Delta t)-P_0(t)}{\Delta t}=-\lambda P_0(t)+\nu P_1(t)+\frac{o(\Delta t)}{\Delta t}\\
&\frac{P_k(t+\Delta t)-P_k(t)}{\Delta t}=\lambda P_{k-1}(t)-\\
&\quad-(\lambda-\nu k)\,P_k(t)+\nu(k+1)\,P_{k+1}(t)+\\
&\quad+\frac{o(\Delta t)}{\Delta t}\quad(1\leqslant k\leqslant n-1)\\
&\frac{P_n(t+\Delta t)-P_n(t)}{\Delta t}=\lambda P_{n-1}(t)-\nu n P_n(t)+\frac{o(\Delta t)}{\Delta t}
\end{aligned}\right\}.$$

Переходя к пределу при $\Delta t\to 0$, замечаем, что пределы левых частей равенств есть производные по времени от $P_i(t)$:

$$\left.\begin{aligned}
&P_0'(t)=-\lambda P_0(t)+\nu P_1(t)\\
&P_k'(t)=\lambda P_{k-1}(t)-(\lambda+\nu k)\,P_k(t)+\\
&\quad+\nu(k+1)\,P_{k+1}(t)\\
&\quad(1\leqslant k\leqslant n-1)\\
&P_n'(t)=\lambda P_{n-1}(t)-\nu n P_n(t)
\end{aligned}\right\}. \qquad (3.7)$$

Эта система $n+1$ линейных однородных дифференциальных уравнений относительно неизвестных функций $P_0(t)$, $P_1(t),\ldots,P_n(t)$ называется системой Эрланга.

Из этой системы можно найти $P_k(t)$ $(k=0,\ 1,\ 2,\ldots,n)$ как функции от параметров λ и ν. Для определения произвольной постоянной, входящей в решение, можно использовать следующее условие:

$$\sum_{k=1}^{n}P_k(t)=1. \qquad (3.8)$$

Однако хотя задача интегрирования такой системы уравнений принципиально разрешима, практическое решение ее связано со значительными вычислительными трудностями. Вот тут-то и приходят на помощь предыдущие рассуждения, доказывающие наличие предельного решения. Отыскать его гораздо проще, чем проинтегрировать эту систему. Покажем, как это может быть сделано. Переходя в (3.7) к пределу и используя наличие

$$\lim_{t\to\infty} P_k(t) = p_k \quad (k=0,\ 1,\ 2,\ \ldots,\ n),$$

получаем

$$\left.\begin{aligned} 0 &= -\lambda p_0 + \nu p_1 \\ 0 &= \lambda p_{k-1} - (\lambda+\nu k)\, p_k + \\ &+ \nu(k+1)\, p_{k+1} \quad (1 \leqslant k \leqslant n-1) \\ 0 &= \lambda p_{n-1} - \nu n p_n \end{aligned}\right\}. \tag{3.9}$$

Пределы левых частей равны нулю, так как если предположить противное, то это означало бы, что $|P_k(t)| \to \infty$ при $t \to \infty$, что невозможно потому, что $|P_k(t)| \leqslant 1$. Система (3.9) является системой линейных однородных алгебраических уравнений относительно неизвестных p_0, p_1, p_2, ..., p_n. Таким образом, если для отыскания $P_k(t)$ $(k=0,\ 1,\ \ldots,\ n)$ нужно интегрировать систему дифференциальных уравнений, то для определения p_k $(k=0,$ $1,\ 2,\ \ldots,\ n)$ нужно всего лишь решить систему алгебраических уравнений. Для этого заменим

$$\lambda p_{k-1} - \nu k p_k = z_k \qquad (1 \leqslant k \leqslant n),$$

тогда $z_1 = \lambda p_0 - \nu p_1$; но из первого уравнения (3.9) следует, что эта величина равна нулю, поэтому $z_1 = 0$.

Преобразуя второе уравнение из (3.9), получаем

$$(\lambda p_{k-1} - \nu k p_k) - [\lambda p_k - \nu(k+1)\, p_{k+1}] = 0,$$

т. е.

$$z_k - z_{k+1} = 0 \qquad (1 \leqslant k \leqslant n-1).$$

Последнее уравнение из (3.9) дает $z_n = 0$.

Таким образом:

$$z = 0,$$

$$z_k - z_{k+1} = 0, \qquad (1 \leqslant k \leqslant n-1),$$

$$z_n = 0.$$

Подставляя $z_1 = 0$ во второе уравнение при $k = 1$, получим, что $z_2 = 0$. Аналогично приходим к выводу, что $z_1 = z_2 = \ldots z_n = 0$. Это означает, что

$$\lambda p_{k-1} = \nu k p_k$$

и

$$p_k = \frac{\lambda}{\nu k} p_{k-1} \qquad (k = 1, 2, \ldots, n).$$

Отсюда получаем, что

$$p_1 = \frac{\lambda}{\nu} p_0; \quad p_2 = \frac{\lambda}{2\nu} p_1 = \left(\frac{\lambda}{\nu}\right)^2 \frac{1}{1 \cdot 2} p_0;$$

$$p_3 = \frac{\lambda}{3\nu} p_2 = \left(\frac{\lambda}{\nu}\right)^3 \frac{1}{1 \cdot 2 \cdot 3} p_0, \ldots$$

Легко показать, что этот закон имеет место при любом k, т. е.

$$p_k = \frac{p_0}{k!}\left(\frac{\lambda}{\nu}\right)^k \qquad (k = 1, 2, \ldots, n). \tag{3.10}$$

Для определения p_0 воспользуемся условием, которое получается из (3.8) при $t \to \infty$:

$$\sum_{m=0}^{n} p_m = 1.$$

Подставляя в это равенство выражения p_m, получаем, что

$$p_0 \sum_{m=0}^{n} \frac{1}{m!}\left(\frac{\lambda}{\nu}\right)^m = 1,$$

$$p_0 = 1 : \sum_{m=0}^{n} \frac{1}{m!}\left(\frac{\lambda}{\nu}\right)^m. \tag{3.11}$$

Таким образом, формулы (3.10) и (3.11) дают решение поставленной задачи. С их помощью можно вывести формулу для вычисления одного из основных критериев функционирования обслуживающей системы — вероятности отказа. Очередное требование не будет принято на обслуживание в том случае, если все аппараты заняты, т. е. если $k=n$. Поэтому вероятность отказа

$$p_n=\frac{\left(\frac{\lambda}{\nu}\right)^n\frac{1}{n!}}{\sum\limits_{m=0}^{n}\frac{1}{m!}\left(\frac{\lambda}{\nu}\right)^m}. \qquad (3.12)$$

Математическое ожидание числа занятых аппаратов равно

$$M=\sum_{k=1}^{n}kp_k=\sum_{k=1}^{n}\frac{1}{(k-1)!}\left(\frac{\lambda}{\nu}\right)^k p_0. \qquad (3.13)$$

Нужно отметить, что несмотря на то, что формулы

$$p_k=\frac{p_0}{k!}\left(\frac{\lambda}{\nu}\right)^k \qquad (k=1,\ 2,\ \ldots,\ n) \qquad (3.10)$$

были выведены в предположении, что время обслуживания требований обслуживающей системой подчинено показательному закону, тем не менее они имеют гораздо более широкое применение. Как показал в своей работе [17] советский ученый Б. А. Севастьянов, эти формулы справедливы и при произвольном законе распределения времени обслуживания.

Выводы. 1. Вероятность того, что в обслуживающей системе находится точно k требований, т. е. занято k обслуживающих аппаратов:

$$p_k=\frac{p_0}{k!}\left(\frac{\lambda}{\nu}\right)^k \quad (k=1,\ 2,\ \ldots,\ n).$$

Эти формулы называются обычно формулами Эрланга. Напомним, что λ — среднее число требований, поступающих за единицу времени; а $\frac{1}{\nu}$ — среднее время обслуживания одного требования.

2. Вероятность того, что все обслуживающие аппараты свободны,

$$p_0 = \frac{1}{\sum_{m=0}^{n} \frac{1}{m!}\left(\frac{\lambda}{\nu}\right)^m}.$$

Здесь n — число обслуживающих аппаратов системы.

3. Вероятность отказа очередному требованию в обслуживании

$$p_n = \frac{\left(\frac{\lambda}{\nu}\right)^n \frac{1}{n!}}{\sum_{m=0}^{n} \frac{1}{m!}\left(\frac{\lambda}{\nu}\right)^m}.$$

4. Среднее число занятых обслуживающих аппаратов

$$M = \sum_{k=1}^{n} \frac{1}{(k-1)!}\left(\frac{\lambda}{\nu}\right)^k p_0.$$

Примеры. Рассмотрим следующий пример. На вокзале, в мастерской бытового обслуживания работают 3 мастера.

Если клиент заходит в мастерскую, когда все мастера заняты обслуживанием ранее прибывших клиентов, то он уходит из мастерской не ожидая обслуживания. Предположим, что среднее число клиентов, обращающихся в мастерскую в течение часа, равно 24 и что сред-

нее время, которое затрачивает мастер на обслуживание одного клиента, равно 5 минутам. Спрашивается, какова вероятность того, что клиент не будет обслужен, и насколько полно загружены мастера работой?

Прежде чем приступить к решению этой задачи, попытаемся качественно обосновать допустимость применения выведенных нами формул. Начнем разбор с потока требований, т. е. с потока клиентов, обращающихся в мастерскую за помощью. Покажем, что этот поток обладает свойствами, которые позволяют считать его простейшим. Очевидно, что вероятность того, что за время $(t, t+h)$ в мастерскую обратятся k клиентов, в первую очередь зависит от h и k. Если вокзал большой или мастерская работает только в часы, когда имеется достаточное число пассажиров, то эта вероятность мало зависит от t. Чем более оживленный вокзал, тем меньше эта зависимость. Таким образом, приходим к выводу, что поток требований обладает свойствами, которые не противоречат утверждению, что он стационарный. Поэтому будем считать, что поток требований на обслуживание, поступающих в мастерскую, — стационарный.

Так как по предположению вокзал достаточно большой и, следовательно, в нем довольно большое количество пассажиров, то вероятность прибытия k новых клиентов за время $(t, t+h)$ мало зависит от того, сколько клиентов обратилось за обслуживанием до момента t. Поэтому будем считать, что данный поток требований обладает свойством отсутствия последействия.

Кроме того, можно предполагать, что пассажиры будут обращаться за помощью более или менее равномерно в течение суток, т. е. вероятность того, что за малый промежуток времени Δt за обслуживанием обратятся не меньше двух пассажиров, есть малая величина. Это позволяет предполагать, что поток будет обладать свойством ординарности. Читатель, конечно, заметил, что во всех этих рассуждениях очень часто встречаются слова «можно предполагать», «будем считать», но на первом этапе такие качественные рассуждения неизбежны, они помогают найти правильное направление дальнейшей работы. После этих качественных рассуждений необходимо их подтвердить количественно, доказать, что это действительно так. Возможно, что некоторые

из них окажутся ложными, и тогда нужно будет искать другие пути решения.

Будем считать, что поток требований является простейшим. Таким образом, для его полной характеристики (§ 1 гл. 2) достаточно задать одну постоянную величину — математическое ожидание числа требований за единицу времени. По условию задачи эта величина известна ($\lambda=24$).

Перейдем теперь к вопросу о законе распределения времени обслуживания. Так как в мастерскую бытового обслуживания, тем более на вокзале, клиенты, как правило, обращаются только с несложными просьбами, удовлетворение которых не требует много времени, то вероятность того, что обслуживание очень затянется, мала, а вероятность того, что оно закончится в первые минуты после начала, велика. Поэтому будем считать, что время обслуживания подчинено показательному закону. Так как по предположению среднее время обслуживания составляет 5 минут, то, принимая за единицу времени час, можно найти параметр показательного закона ν. Математическое ожидание времени обслуживания есть величина, обратная параметру ν (§ 2, гл. 2), поэтому

$$\nu = 1 : \frac{1}{12} = 12, \text{ так как } 5 \textit{ мин} = \frac{1}{12} \textit{ час}.$$

Таким образом, все условия рассмотренной задачи соблюдены, поэтому ответы на поставленные вопросы могут быть получены путем применения выведенных формул *. Так, из (3.12) следует, что вероятность того, что все мастера в момент обращения очередного пассажира будут заняты и он вынужден будет покинуть мастерскую, равна

$$p_3 = \frac{2^3}{3!\left(1+2+\frac{2^2}{2!}+\frac{2^3}{3!}\right)} \approx 0{,}21.$$

---

* Напомним, что требование показательного закона распределения времени обслуживания является слишком жестким, но, как было указано выше, решение не изменится и при произвольном законе распределения времени обслуживания.

Это значит, что из ста пассажиров в среднем 79 будут обслужены, а 21 — нет.

Теперь определим, какова средняя занятость мастера. Математическое ожидание числа занятых мастеров получим из (3.13):

$$M = \sum_{k=1}^{n} k p_k.$$

Величины p_0, p_1, p_2, p_3 приведены в следующей таблице:

Число работающих мастеров, k	$\frac{p_k}{p_0}$	p_k	kp_k
0	1	0,16	0
1	2	0,32	0,32
2	2	0,32	0,64
3	$^4/_3$	0,21	0,63
Сумма	$6^1/_3$	1,01	1,59

При вычислении первым заполняется второй столбец, в котором величины $\frac{p_k}{p_0}$ вычисляются по формулам $\frac{p_k}{p_0} = \frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k = \frac{2^k}{k!}$, так как $\lambda = 24$, $\nu = 12$.

Учитывая, что $\sum_{k=1}^{3} p_k = 1$, и суммируя второй столбец, получаем

$$\sum_{k=0}^{3} \frac{p_k}{p_0} = \frac{1}{p_0} = \frac{19}{3},$$

поэтому $p_0 \approx 0{,}16$. Заметим, что для контроля можно использовать сумму величин третьего столбца, которая должна быть равна единице. У нас получается $\sum_{k=0}^{3} p_k = 1{,}01$, что приемлемо в пределах точности вычислений*.

* На первый взгляд мы пришли к нелепости — вероятность события равна 1,01! Ведь она не может быть больше единицы! Но это только кажущаяся нелепость. Так как вычисления производились с точностью до $5 \cdot 10^{-3}$, то погрешность, равная $1 \cdot 10^{-2}$, допустима, она появилась как результат округлений. С подобными ситуациями часто приходится сталкиваться при вычислениях.

Умножая элементы второго столбца на p_0, получаем p_k ($k=0, 1, 2, 3$), которые записаны в третьем столбце. Умножая элементы третьего столбца на k (первый на 0, второй на 1 и т. д.) и суммируя их, получаем $M=\sum_{k=1}^{3} kp_k$. Таким образом, в среднем будет занято $M=1{,}59$ мастера. Это означает, что в среднем каждый мастер будет занят $M/3=0{,}53$ часть рабочего дня.

Этот результат на первый взгляд является неожиданным. Как же так! В среднем за час обращается за помощью 24 пассажира, на обслуживание одного уходит в среднем 5 минут, значит за час может быть обслужено 36, т. е. в среднем каждый мастер будет загружен $\frac{24}{36}=0{,}36\,(6)$ часть рабочего времени. Однако здесь не учтено, что часть клиентов попадут в тот момент, когда все мастера будут заняты, а так как они спешат на поезд, то им придется покинуть мастерскую. Если число мастеров в мастерской уменьшить до двух и произвести аналогичные вычисления, то получим, что математическое ожидание числа занятых мастеров

$$M_2=\sum_{k=1}^{2} kp_k=1,\ 2.$$

Следовательно, каждый мастер будет занят в среднем 0,6 рабочего дня. При этом вероятность того, что пассажир покинет мастерскую, равна

$$p_2=\frac{2^2}{2(1+2+2)}=0{,}4,$$

т. е. из 100 пассажиров будет обслужено 60.

Если качество работы мастерской оценивать коэффициентом обслуживания, равным отношению числа обслуженных клиентов к общему числу обращавшихся в мастерскую, то уменьшение числа мастеров на одного приводит к уменьшению этого коэффициента от 0,79 до 0,60.

Этот пример носит характер шутки, однако за ним нетрудно увидеть ряд подобных примеров, имеющих серьезное практическое значение.

Рассмотрим еще один пример, который является частным случаем одиннадцатого примера, сформулированного в общем виде в § 3 гл. 1. Пусть телефонная станция может одновременно обслуживать вызовы n абонентов. Поток требований, поступающих на станцию, является простейшим со средним числом вызовов в минуту $\lambda=2$. Продолжительность телефонного разговора является случайной величиной. Пусть вероятность того, что разговор не продлится больше t минут, равна $F(t)=1-e^{-2t}$, т. е. мы предполагаем, что продолжительность разговора подчинена показательному закону с параметром $\nu=2$.

Нужно заметить, что эти предположения о характере потока требований и законе распределения времени обслуживания не противоречат реальным условиям эксплуатации телефонных станций. Необходимо определить, какое число вызовов n должна одновременно обслуживать телефонная станция для того, чтобы вероятность отказа ε не превосходила 0,01. Предполагается, что если в момент поступления вызова все n линий связи заняты, то абонент получает отказ. В этом примере обслуживающим «аппаратом» является одна линия связи. Таким образом, если телефонная станция является автоматической, т. е. единым аппаратом с физической точки зрения, то с точки зрения теории массового обслуживания она состоит из n обслуживающих аппаратов.

Эта задача является частным случаем решенной задачи, поэтому ответ может быть получен с помощью формулы (3.12). Необходимо подобрать такое число n, чтобы вероятность найти все линии связи занятыми не превосходила ε, т. е. $p_n \leqslant \varepsilon$.

Отсюда, подставляя $\lambda=2$ и $\nu=2$ в (3.12), получаем

$$\frac{1}{n!\sum_{m=0}^{n}\frac{1}{m!}} \leqslant 0{,}01.$$

Нетрудно видеть, что при $n=4$ мы имеем $p_n=0{,}015$, а при $n=5$ величина $p_n=0{,}003$. Следовательно, если телефонная станция может одновременно обслуживать 5 разговоров, то вероятность отказа будет меньше необходимой более чем в три раза, а если 4 разговора, то

в 1,5 раза больше необходимой. Теперь определим математическое ожидание числа занятых линий. Результаты сведем в следующую таблицу, которая вычисляется так же, как в предыдущем примере.

Число работающих линий	$\frac{p_k}{p_0}$	p_k	kp_k
0	1,000	0,368	0,000
1	1,000	0,368	0,368
2	0,500	0,184	0,368
3	0,167	0,061	0,183
4	0,041	0,015	0,060
5	0,009	0,003	0,015
Сумма	2,717	0,999	0,994

Здесь

$$\sum_{k=0}^{5} \frac{p_k}{p_0} = \frac{1}{p_0} = 2{,}72, \text{ т. е. } p_0 = 0{,}368.$$

Математическое ожидание числа занятых линий $M = 0{,}994$, т. е. каждая линия в среднем будет занята меньше 0,2 рабочего времени. Из этой таблицы видно, что 0,368 рабочего времени все линий будут свободны от вызовов и 0,368 рабочего времени будет занята только одна линия. Вероятность того, что будет занято не меньше двух линий, равна

$$p_{\geqslant 2} = p_2 + p_3 + p_4 + p_5 = 0{,}263,$$

т. е. только чуть больше четверти рабочего времени будет занято не менее двух линий.

2. ОБСЛУЖИВАНИЕ В СИСТЕМЕ С НЕОГРАНИЧЕННЫМ ЧИСЛОМ АППАРАТОВ

Как указывалось выше, в ряде процессов массового обслуживания приходится сталкиваться с системами, число обслуживающих аппаратов которых неограниченно (примеры 5, 6, 12, 13). Анализу процесса обслуживания в таких системах посвящен этот параграф. Как и раньше, рассмотрим абстрактную обслуживающую си-

стему, выведем основные характеристики ее функционирования, а затем применим полученные результаты к решению конкретных задач.

Постановка задачи. Предположим, что обслуживающая система состоит из неограниченного числа обслуживающих аппаратов. В этом случае понятие потери требования для такой системы теряет смысл, так как в любой момент система, состоящая из неограниченного числа обслуживающих аппаратов, сможет принять на обслуживание очередное требование. Однако сохраняет смысл понятие вероятности состояния. В частности, вопрос о том, какова вероятность $P_k(t)$ того, что в момент времени t занято точно k аппаратов системы, сохраняет свой смысл.

Ответ на этот вопрос может помочь оценить степень использования системы, состоящей из очень большого числа обслуживающих аппаратов, а следовательно, определить ожидаемый износ аппаратов системы и другие экономические показатели. Во всем остальном будем предполагать, что выполнены условия задачи, рассмотренной в предыдущем параграфе, т. е. будем считать, что в систему на обслуживание поступает простейший поток требований с параметром λ и что время обслуживания подчинено показательному закону с параметром ν. Целью решения поставим отыскание вероятности $P_k(t)$ ($k=0, 1, 2, \ldots$) того, что в момент времени t занято k аппаратов системы.

Дифференциальные уравнения решения задачи. Вывод уравнения для определения $P_0(t)$ ничем не будет отличаться от того вывода, который был сделан выше, поэтому уравнение для определения $P_0(t)$ можно взять из формулы (3.7). Подобным образом можно использовать уравнения для определения $P_k(t)$. Отличие будет заключаться только в том, что если в предыдущей задаче при $k=n$ уравнение приобретало иной вид за счет того, что система не могла перейти от n занятых аппаратов к $n+1$, так как их всех было n, т. е. не могла из состояния n перейти в состояние $n+1$, то теперь она может перейти в это состояние, так как по условию $n \leftarrow \infty$ и, следовательно, $P_{n+1}(t) \neq 0$. Поэтому второе уравнение из (3.7) пригодно при любом $k>0$, а уравнение для $P_n'(t)$ из этой системы для данной задачи теряет смысл.

Таким образом, мы приходим к выводу, что для определения вероятностей $P_k(t)$ может быть использована следующая система однородных дифференциальных уравнений:

$$\left.\begin{aligned} P_0'(t) &= -\lambda P_0(t) + \nu P_1(t) \\ P_k'(t) &= \lambda P_{k-1}(t) - (\lambda + \nu k) P_k(t) + \\ &\quad + \nu(k+1) P_{k+1}(t) \\ &\quad (k > 0) \end{aligned}\right\}. \qquad (3.14)$$

Если система (3.7) состоит из конечного числа дифференциальных уравнений и может быть проинтегрирована любым известным методом, то система (3.14) состоит из неограниченного числа уравнений и ее интегрирование требует использования специальных приемов. С этой точки зрения рассматриваемая задача представляет самостоятельный интерес. Использование метода производящих функций показывает, что система (3.14) более простая, чем (3.7). Продемонстрируем на этой системе метод производящих функций и попутно найдем ее решения. Введем вспомогательную функцию

$$\Phi(t, x) = \sum_{k=0}^{\infty} P_k(t) x^k. \qquad (3.15)$$

Эта функция называется *производящей.* Если бы мы сумели найти ее, то задача определения искомых вероятностей $P_k(t)$ свелась бы к разложению производящей функции в ряд по степеням x, так как коэффициент при x^k равен $P_k(t)$. Таким образом, задача интегрирования системы (3.14) сводится к задаче отыскания производящей функции. Ее нужно найти из условия, что $P_k(t)$ $(k=0, 1, 2, \ldots)$ удовлетворяют системе (3.14). При определении сходимости ряда (3.15) оказывается, что он сходится абсолютно при любых t и $|x| \leqslant 1$, так как ряд

$$\sum_{k=0}^{\infty} P_k(t) = 1,$$

т. е. сходится при любом t согласно условию задачи.

Перейдем теперь к отысканию $\Phi(t, x)$; с этой целью продифференцируем по t обе части тождества (3.15):

$$\frac{\partial \Phi(t, x)}{\partial t} = \sum_{k=0}^{\infty} \frac{dP_k(t)}{dt} x^k. \tag{3.16}$$

Покажем, что дифференцирование законно. Для этого нужно доказать, что ряд, образовавшийся в результате дифференцирования, также сходится. В процессе этого доказательства выведем уравнение для $\Phi(t, x)$. Так как значения $P_k(t)$ должны удовлетворять (3.14), то подставим из (3.14) выражения для $P'_k(t)$ в (3.16). Тогда получим

$$\frac{\partial \Phi(t, x)}{\partial t} = -\lambda P_0(t) + \nu P_1(t) + \sum_{k=1}^{\infty} [\lambda P_{k-1}(t) -$$

$$-(\lambda + \nu k) P_k(t) + \nu(k+1) P_{k+1}(t)] x^k =$$

$$= \lambda(x-1) \sum_{k=0}^{\infty} P_k(t) x^k - \nu(x-1) \sum_{k=1}^{\infty} k P_k(t) x^{k-1}.$$

Так как

$$\Phi(t, x) = \sum_{k=0}^{\infty} P_k(t) x^k \quad \text{и} \quad \frac{\partial \Phi(t, x)}{\partial x} = \sum_{k=1}^{\infty} k P_k(t) x^{k-1},$$

(сумма последнего ряда при $x=1$ равна математическому ожиданию числа занятых обслуживающих аппаратов, которое можно считать конечным), то

$$\frac{\partial \Phi(t, x)}{\partial t} = \lambda(x-1) \Phi(t, x) - \nu(x-1) \frac{\partial \Phi(t, x)}{\partial x}. \tag{3.17}$$

Таким образом, показано, что ряд, полученный в результате дифференцирования, сходится, а следовательно, дифференцирование законно. Полученное уравнение является уравнением в частных производных относительно $\Phi(t, x)$. В общем случае интегрирование уравнения в частных производных является нелегкой и пока еще не всегда разрешимой задачей, однако это уравнение может быть решено следующим образом.

Упростим уравнение (3.17), введя новую искомую функцию $F(t, x)$, которая связана с $\Phi(t, x)$, так:

$$\Phi(t, x) = G(t, x) F(t, x),$$

где $G(t, x)$ — известная функция:

$$G(t, x) = e^{\frac{\lambda}{\nu}(x-1)(1-e^{-\nu t})}.$$

Отсюда

$$\frac{\partial \Phi}{\partial t} = G\frac{\partial F}{\partial t} + FG\lambda(x-1)e^{-\nu t},$$

$$\frac{\partial \Phi}{\partial x} = G\frac{\partial F}{\partial x} + FG\frac{\lambda}{\nu}(1-e^{-\nu t}).$$

Перенеся все члены (3.17) влево и подставив эти выражения, получим

$$\frac{\partial \Phi}{\partial t} + (x-1)\nu\frac{\partial \Phi}{\partial x} - \lambda(x-1)\Phi = G\left\{\frac{\partial F}{\partial t} + \right.$$

$$+ F\lambda(x-1)e^{-\nu t} + (x-1)\nu\frac{\partial F}{\partial x} + (x-1)\lambda F(1-e^{-\nu t}) -$$

$$\left. - \lambda(x-1)F\right\} = G\left\{\frac{\partial F}{\partial t} + (x-1)\nu\frac{\partial F}{\partial x}\right\}.$$

Следовательно, уравнение (3.17) равносильно уравнению

$$G\left\{\frac{\partial F}{\partial t} + (x-1)\nu\frac{\partial F}{\partial x}\right\} = 0.$$

Так как $G \neq 0$, то

$$\frac{\partial F}{\partial t} + (x-1)\nu\frac{\partial F}{\partial x} = 0. \qquad (3.18)$$

Рассмотрим вспомогательную функцию $L(t, x)$:

$$(x-1)e^{-\nu t} = L(t, x)$$

и вычислим определитель Остроградского—Якоби

$$\frac{D(F, L)}{D(t, x)} = \begin{vmatrix} \frac{\partial F}{\partial t} & \frac{\partial F}{\partial x} \\ \frac{\partial L}{\partial t} & \frac{\partial L}{\partial x} \end{vmatrix} = \begin{vmatrix} \frac{\partial F}{\partial t} & \frac{\partial F}{\partial x} \\ -\nu(x-1)\,e^{-\nu t} & e^{-\nu t} \end{vmatrix} =$$

$$= e^{-\nu t}\left[\frac{\partial F}{\partial t} + (x-1)\,\nu\,\frac{\partial F}{\partial x}\right].$$

Если функция $F(t, x)$ является решением уравнения (3.18), то

$$\frac{D(F, L)}{D(t, x)} = 0.$$

Равенство нулю определителя Остроградского — Якоби означает, что между искомой функцией $F(t, x)$ и заданной функцией $L(t, x)$ существует функциональная зависимость, т. е. искомая функция

$$F(t, x) = R(L) = R[(x-1)\,e^{-\nu t}],$$

где R—произвольная дифференцируемая по L функция. Таким образом, искомая функция $\Phi(t, x)$ имеет вид

$$\Phi(t, x) = e^{\frac{\lambda}{\nu}(x-1)(1-e^{-\nu t})} R[(x-1)\,e^{-\nu t}]. \qquad (3.19)$$

Но это общее решение, которое определяет семейство производящих функций $\Phi(t, x)$ при произвольных дифференцируемых функциях R. Для того чтобы получить интересующее нас решение, необходимо использовать начальные условия. Естественно считать, что начальное состояние системы известно, т. е. при $t=0$ $P_k(0) = a_k$ $(k=0, 1, 2, \ldots)$, где a_k — известные числа и

$$\sum_{k=0}^{\infty} a_k = 1.$$

С помощью этих начальных данных определим вид функции. Для этого в выражение (3.15) подставим $t = 0$, тогда получим

$$\Phi(0, x) = \sum_{k=0}^{\infty} P_k(0)\,x^k = \sum_{k=0}^{\infty} a_k x^k.$$

Но у нас есть еще уравнение (3.19), описывающее функцию $\Phi(0, x)$. Подставляя в него $t=0$, получаем

$$\Phi(0, x)=R(x-1).$$

Следовательно,

$$R(x-1)=\sum_{k=0}^{\infty} a_k x^k.$$

Обозначим $x-1=z$, тогда

$$R(z)=\sum_{k=0}^{\infty} a_k (1+z)^k.$$

Таким образом, функция R определяется рядом

$$\sum_{k=0}^{\infty} a_k (1+z)^k,$$

где $z=x-1$.

Поэтому

$$R[(x-1)\,e^{-\nu t}]=\sum_{k=0}^{\infty} a_k [1+(x-1)\,e^{-\nu t}]^k.$$

Окончательно производящая функция имеет вид

$$\Phi(t, x)=e^{\frac{\lambda}{\nu}(x-1)(1-e^{-\nu t})}\sum_{k=0}^{\infty} a_k [1+(x-1)\,e^{-\nu t}]^k. \quad (3.20)$$

Этот ряд сходится абсолютно при $|x|\leqslant 1$ и $t>0$, так как при этом

$$|1+(x-1)\,e^{-\nu t}|\leqslant 1.$$

Теперь для получения функций $P_k(t)$ правую часть выражения (3.20) нужно представить в виде разложения по степеням x. Коэффициент при x^k равен $P_k(t)$. Это разложение имеет очень громоздкий вид, поэтому ограничимся выводом выражений $P_k(t)$ для частного, но наиболее интересного случая.

Пусть в начальный момент все обслуживающие аппараты свободны. Это означает, что $P_0(0)=a_0=1$, а $P_k(0)=a_k=0$ $(k=1, 2, \ldots)$. Подставляя эти значения в (3.20), получаем

$$\Phi(t, x)=e^{\frac{\lambda}{\nu}(x-1)(1-e^{-\nu t})}=e^{-\frac{\lambda}{\nu}(1-e^{-\nu t})}e^{\frac{\lambda}{\nu}(1-e^{-\nu t})x}.$$

Разложив второй множитель в ряд по степеням *показателя* e, получим

$$\Phi(t, x)=e^{-\frac{\lambda}{\nu}(1-e^{-\nu t})}\sum_{k=0}^{\infty}\frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k(1-e^{-\nu t})^k x^k.$$

Сравнение коэффициентов с разложением (3.15) показывает, что

$$P_k(t)=\frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k(1-e^{-\nu t})^k e^{-\frac{\lambda}{\nu}(1-e^{-\nu t})} \quad (k=0, 1, 2, \ldots). \tag{3.21}$$

Таким образом, найдены формулы для вычисления вероятности того, что в произвольный момент времени t будет занято k аппаратов при условии, что в начальный момент все они были свободны. В частности, при $k=0$ получаем вероятность того, что все аппараты свободны:

$$P_0(t)=e^{-\frac{\lambda}{\nu}(1-e^{-\nu t})}.$$

В пределе, при $t\to\infty$ получаем, что

$$P_0=\lim_{t\to\infty}P_0(t)=e^{-\frac{\lambda}{\nu}},$$

т. е. вероятность того, что все аппараты будут свободны, тем меньше, чем больше математическое ожидание числа требований в единицу времени и чем больше $\frac{1}{\nu}$ — математическое ожидание времени обслуживания одного требования. Не останавливаясь на подробностях доказательства, заметим, что, используя (3.20), можно пока-

зать, что независимо от начального состояния системы стационарное решение имеет следующий вид:

$$p_k = \frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k e^{-\frac{\lambda}{\nu}} \quad (k = 0,\ 1,\ 2, \ldots), \qquad (3.22)$$

где

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

На примере рассмотрим, какая погрешность допускается при использовании более простых формул стационарного решения (3.22) вместо точных (3.21) для случая, когда в начальный момент времени все обслуживающие аппараты свободны. Определим математическое ожидание числа занятых аппаратов в момент времени t. Оно равно

$$M = \sum_{k=1}^{\infty} kP_k(t).$$

Подставим значения $P_k(t)$ из (3.21):

$$M = e^{-\frac{\lambda}{\nu}(1-e^{-\nu t})} \sum_{k=1}^{\infty} \frac{1}{(k-1)!}\left[\frac{\lambda}{\nu}(1-e^{-\nu t})\right]^k =$$

$$= e^{-\frac{\lambda}{\nu}(1-e^{-\nu t})} \frac{\lambda}{\nu}(1-e^{-\nu t}) \sum_{k=1}^{\infty} \frac{1}{(k-1)!}\left[\frac{\lambda}{\nu}(1-e^{-\nu t})\right]^{k-1} =$$

$$= \frac{\lambda}{\nu}(1-e^{-\nu t}),$$

так как

$$e^{\frac{\lambda}{\nu}(1-e^{-\nu t})} = \sum_{k=0}^{\infty} \frac{1}{k!}\left[\frac{\lambda}{\nu}(1-e^{-\nu t})\right]^k =$$

$$= \sum_{k=1}^{\infty} \frac{1}{(k-1)!}\left[\frac{\lambda}{\nu}(1-e^{-\nu t})\right]^{k-1}.$$

Таким образом, математическое ожидание числа занятых аппаратов в момент времени t равно

$$M = \frac{\lambda}{\nu}(1 - e^{-\nu t}). \qquad (3.23)$$

Теперь вычислим это математическое ожидание, используя стационарное решение (3.22):

$$M' = e^{-\frac{\lambda}{\nu}} \sum_{k=1}^{\infty} \frac{1}{(k-1)!}\left(\frac{\lambda}{\nu}\right)^k = \frac{\lambda}{\nu} e^{-\frac{\lambda}{\nu}} e^{\frac{\lambda}{\nu}} = \frac{\lambda}{\nu}. \qquad (3.24)$$

Здесь использовано преобразование, аналогичное предыдущему. Сравним точное значение математического ожидания M с приближенным M'. Абсолютная погрешность, которая допускается при замене значения M значением M', равна

$$M' - M = \frac{\lambda}{\nu} e^{-\nu t};$$

она тем меньше, чем больше параметр ν, т. е. чем меньше среднее время обслуживания. Естественно, что эта погрешность будет убывать с ростом t.

Выводы. 1. Вероятность того, что в момент времени t занято k обслуживающих аппаратов, при условии, что в начальный момент все они свободны,

$$P_k(t) = \frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k \left(1 - e^{-\nu t}\right)^k e^{-\frac{\lambda}{\nu}(1 - e^{-\nu t})}.$$

$$(k = 0,\ 1,\ 2, \ldots).$$

Напомним, что λ — среднее число требований, поступающих в систему за единицу времени; $\frac{1}{\nu}$ — среднее время обслуживания одного требования.

2. Среднее число аппаратов, занятых в момент времени t, при условии, что в начальный момент все они свободны,

$$M = \frac{\lambda}{\nu}(1 - e^{-\nu t}).$$

3. Вероятность того, что в установившемся процессе будет занято k обслуживающих аппаратов, независимо от начального состояния системы

$$p_k = \frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k e^{-\frac{\lambda}{\nu}} \qquad (k = 0,\ 1,\ 2, \ldots).$$

4. Вероятность того, что в установившемся процессе все обслуживающие аппараты свободны,

$$p_0 = e^{-\frac{\lambda}{\nu}}.$$

Примеры. Рассмотрим один из примеров массового обслуживания, являющийся частным случаем задачи, сформулированной в пятом примере. В процессе эксплуатации автомашины выходят из строя и требуют ремонта. Ремонт является обслуживанием требования, а весь процесс ремонта представляет частный случай массового обслуживания. Общее число всех эксплуатируемых автомашин очень велико, число ремонтируемых также велико, поэтому число возможных состояний систем, состоящих из автомашин, находящихся в данный момент времени в ремонте, может быть очень большим.

Поставим перед собой задачу определить среднее число автомашин, находящихся в ремонте. Пусть среднее число машин, выходящих из строя за месяц, равно 300. Число машин, требующих ремонта, есть случайная функция, зависящая от степени их загрузки, условий эксплуатации, качества повседневного ухода и т. д. Будем считать, что этот поток требований является простейшим. Качественный анализ показывает, что основные свойства, характеризующие простейший поток в рассматриваемом примере, должны иметь место. Есть основания предполагать, что этот поток будет стационарным. Отсутствие последействия также должно иметь место, так как выход из строя части машин, если это не очень большая часть из всех имеющихся, не слишком повлияет на степень загрузки оставшихся машин. Поэтому вероятность выхода автомашин из строя не будет сильно зависеть от того, сколько их вышло из строя до этого. Так как в сутки в среднем из строя вы-

ходит только 10 машин, то, очевидно, будет иметь место и ординарность потока.

Ремонт, как правило, начинается сразу после выхода машины из строя. Правда, на практике могут возникнуть ситуации, когда машина, нуждающаяся в ремонте, простаивает, и иной раз довольно долго, однако мы, исходя из нормальных требований, предъявляемых к системам обслуживания, будем предполагать, что обслуживание (ремонт) машины начинается сразу же, как только машина поступила в мастерскую. Пусть среднее время ремонта одной машины равно 10 суткам. Если принять за единицу времени сутки, то

$$\frac{1}{\nu} = 10 \text{ и } \nu = 0{,}1,$$

а параметр простейшего потока $\lambda = \frac{300}{30} = 10$.

Поставленная задача является задачей Эрланга, поэтому для ее решения могут быть использованы результаты, полученные выше. Так, используя формулу (3.24), найдем, что математическое ожидание числа ремонтируемых машин равно

$$M' = \frac{\lambda}{\nu} = 100.$$

Посмотрим, насколько стационарное решение отличается от нестационарного при условии, что в начальный момент все машины исправны. Уже через месяц после начала эксплуатации математическое ожидание числа неисправных машин, вычисленное по более точной формуле (3.23), равно

$$M = \frac{\lambda}{\nu}(1 - e^{-\nu t}) = 100\,(1 - e^{-3}) \approx 95,$$

так как $t = 30$, а через год

$$M \approx 100\,(1 - e^{-36{,}5}) \approx 100,$$

считая, что в году 365 суток, т. е. практически использование формулы (3.24) вполне допустимо.

Второй пример является более подробным рассмотрением тринадцатого примера из § 3 гл. 1. Как уже было сказано, цифровые электронные машины состоят из

очень большого числа различных элементов. Выход из строя одного элемента арифметического устройства или устройства управления равносилен выходу из строя всей машины, так как до того, как будет заменен неисправный элемент, правильное функционирование машины невозможно.

Выход элемента из строя означает поступление требования на обслуживание машины. Поток этих требований можно считать простейшим. Время обслуживания складывается из времени отыскания неисправности и ее устранения, оно также является случайной величиной, зависящей от опытности инженерно-технического персонала и сложности отыскания неисправного элемента. Время устранения неисправности, как правило, мало, так как ремонт заключается в замене элемента.

Большую часть неисправностей обнаруживают и устраняют быстро, однако встречаются и такие, которые требуют значительных затрат времени. Будем считать, что это время подчинено показательному закону и среднее время обслуживания равно 2 часам, следовательно, $\nu=0{,}5$.

Предположим, что имеются элементы двух типов. Первые дешевые, но менее надежные. Пусть в среднем из общего их числа за 10 часов портится один, т. е. $\lambda_1=0{,}1$. Вторые дороже, но лучше и надежней. В среднем, из общего их числа за 100 часов портится один, т. е. $\lambda_2=0{,}01$.

Далее предположим, что общая стоимость всех элементов первого типа a, общая стоимость всех элементов второго типа b, а потери за час простоя машины равны c. Спрашивается, при каком соотношении a, b и c монтаж машины на более надежных элементах окупится в течение 1000 часов работы машины? Для решения этой задачи можно также использовать полученные выше формулы. Вероятность того, что машина работает, равна вероятности того, что в системе обслуживания нет ни одного требования, т. е. система находится в «нулевом» состоянии. Эта вероятность равна $p_0=e^{-\frac{\lambda}{\nu}}$.

Вероятность того, что имеется хоть один неисправный элемент и машина простаивает, равна

$$1-p_0=1-e^{-\frac{\lambda}{\nu}}.$$

Для элементов первого типа эти величины равны

$$p_0 = e^{-\frac{\lambda_1}{\nu}} = e^{-\frac{0,1}{0,5}} \approx 0,819,$$

$$1 - p_0 \approx 0,181.$$

Для элементов второго типа

$$p_0 = e^{-\frac{\lambda_1}{\nu}} = e^{-\frac{0,01}{0,5}} \approx 0,980,$$

$$1 - p_0 \approx 0,020.$$

Величина ν не меняется, так как время устранения неисправности не зависит от типа элемента. Таким образом, простои машины, смонтированной на элементах первого типа, составляют в среднем 181 час из тысячи, что повлечет потери $181c$ рублей, а простои машины, построенной на элементах второго типа, составят в среднем 20 часов из тысячи, что повлечет потери, равные $20c$ рублей. Следовательно, за 1 000 часов работы машина, смонтированная на элементах второго типа, позволит сэкономить $181c - 20c = 161c$ рублей. Элементы второго типа будут более выгодными и окупят затраты на них за 1000 часов, если

$$b - a \leqslant 161\ c.$$

Так, если стоимость элементов 1-го типа $a = 10\,000$ рублей, а стоимость элементов второго типа $b = 100\,000$ рублей и стоимость работы одного часа машины $c = 1\,000$ рублей, то

$$\frac{b-a}{c} = \frac{90\,000}{1\,000} = 90 < 161,$$

т. е. монтаж на элементах второго типа, при этих условиях, за 1 000 часов работы себя окупит. Дополнительные затраты на элементы составят 90 000 рублей, но экономия за счет более надежной работы машины будет равна 161 000 рублей, т. е. за это время более надежные элементы не только окупят себя, но и обеспечат экономию в 71 000 рублей. При этих условиях даже затраты на элементы второго типа, равные

$$b = 161c + a = 171\,000 \text{ рублей},$$

выгодны, так как при этом за 1 000 часов работы машины они себя окупят. Кроме того, нужно заметить, что при этом надежность работы машины возрастает с 0,819 до 0,980, а это само по себе является величайшим достоинством.

Используя приведенные выше рассуждения, можно без труда найти решение обратной задачи. Так, например, в ряде случаев представляет интерес вопрос о том, какой надежностью должны обладать элементы, из которых монтируется машина, для того, чтобы надежность всей машины имела заданную величину ε. Пусть, например, требуемая степень надежности $\varepsilon=0{,}999$, т. е. вероятность выхода машины из строя не должна превосходить этой величины, а параметр ν имеет прежнее значение. Тогда надежность работы элементов, которая характеризуется величиной λ, т. е. средним числом элементов, выходящих из строя за единицу времени, может быть найдено из уравнения

$$\varepsilon = e^{-\frac{\lambda}{\nu}},$$

т. е.

$$\lambda = -\nu \ln \varepsilon.$$

В нашем случае $\lambda = -0{,}5 \cdot \ln 0{,}999 \approx 0{,}002$.

Таким образом, из общего числа элементов за 1 000 часов работы машины из строя могут выйти не больше чем два, тогда заданная надежность будет обеспечена.

В качестве третьего примера рассмотрим, как может быть решена задача, сформулированная в шестом примере из § 2 гл. 1. На телеграф поступают телеграммы, которые должны быть доставлены адресатам. Момент поступления очередной телеграммы есть случайная величина. Время доставки телеграммы зависит от того, на каком расстоянии от телеграфа живет адресат, на каком этаже расположена его квартира, работает ли в доме лифт, какими видами транспорта сможет воспользоваться почтальон для доставки телеграммы. Предполагается, что почтальон отправляется в путь сразу же по получении телеграммы. Необходимо определить вероятность того, что в процессе доставки будет находиться одновременно k телеграмм. Этот вопрос представляет интерес для определения необходимого

числа почтальонов для доставки телеграмм. Как и прежде, будем считать, что поток требований (т. е. поток телеграмм) является простейшим, с параметром $\lambda=2$. Это, в частности, означает, что вероятность отсутствия телеграммы в течение часа согласно (2.10) равна $V_0=e^{-\lambda}\approx 0{,}135$.

Вероятность поступления за час одной телеграммы равна $V_1=\lambda e^{-\lambda}\approx 0{,}270$, двух телеграмм — равна $V_2=\frac{\lambda^2}{2!}e^{-\lambda}\approx 0{,}270$, трех телеграмм — равна $V_3=\frac{\lambda^3}{3!}e^{-\lambda}\approx 0{,}180$ и т. д.

Пусть время обслуживания, которое равно времени, необходимому на доставку телеграммы и возвращению на телеграф, подчинено показательному закону. В частности, это означает, что вероятность того, что адресат живет очень далеко от телеграфа или очень высоко, мала. Среднее время доставки телеграммы с учетом времени, необходимого на возвращение почтальона на телеграф, равно 2 часам, т. е. $\nu=0{,}5$. Состояния рассматриваемой системы обслуживания определяются числом почтальонов, находящихся в пути, т. е. числом доставляемых телеграмм. Ответ на поставленный вопрос дают формулы (3.22). Вероятность того, что в процессе обслуживания находятся k телеграмм, т. е. в пути находится k почтальонов, равна

$$p_k=\frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k e^{-\frac{\lambda}{\nu}}=\frac{1}{k!}4^k e^{-4}.$$

В следующей таблице приведены значения p_k:

Число k	0	1	2	3	4	5	6	7	8	9
p_k	0,0183	0,0732	0,146	0,195	0,195	0,156	0,104	0,0593	0,0296	0,0132

Из таблицы видно, что вероятность того, что будут загружены сразу девять почтальонов, не больше 1,4%.

Наиболее вероятными являются состояния, при которых заняты одновременно 3—4 почтальона. Вероятность

того, что все почтальоны будут свободны, равна 0,0183, т. е. из 1 000 часов рабочего времени в среднем 18,3 часа почтальоны будут свободны от разноски телеграмм.

Этот пример, конечно, носит условный характер. В нем не учитывается, что почтальон может разносить сразу несколько телеграмм за счет того, что они не сразу отправляются адресату, а некоторое время находятся на телеграфе, и т. п. Но пример можно рассматривать как первое приближение к реальному процессу и при более тщательном качественном и количественном анализе рассматриваемого процесса он может оказать весьма существенную пользу для выявления лучшей организации.

3. ЗАДАЧИ ОБСЛУЖИВАНИЯ В СИСТЕМАХ С ОЖИДАНИЕМ

В этом параграфе мы рассмотрим несколько задач массового обслуживания в системах с ожиданием. Первые две задачи объединяет условие неограниченного ожидания начала обслуживания. Требование, попавшее в систему обслуживания, будет находиться в ней до тех пор, пока его обслуживание не будет закончено.

В первой задаче рассматривается ограниченный поток требований. Это значит, что общее число требований, которые могут одновременно находиться в обслуживающей системе, ограниченно. Во второй задаче рассматривается неограниченный поток требований.

Третья задача отличается от первых двух отсутствием требования неограниченного времени ожидания начала обслуживания. В рассмотренной задаче вновь поступившее требование остается в системе лишь при соблюдении некоторого дополнительного условия. Это условие состоит в том, что общее число требований, ожидающих начала обслуживания, к моменту поступления очередного требования не превосходит определенной величины.

Эти три задачи, естественно, не исчерпывают всего многообразия задач массового обслуживания в системах с ожиданием. Во многих случаях ожидание начала обслуживания не может превзойти определенной величины. В частности, время ожидания начала обслуживания может быть случайной величиной. Так, например, Е. С. Вентцель [8] получены обобщения уравнений и формул Эрланга для случая, когда время ожидания на-

чала обслуживания подчинено показательному закону.

Как и раньше, рассматривать задачи будем в общем виде, применяя в дальнейшем полученные результаты к решению конкретных примеров.

Задача первого типа

Постановка задачи. Пусть обслуживающая система состоит из конечного числа обслуживающих аппаратов. Система относится к числу систем с ожиданием. Каждый аппарат системы может одновременно обслуживать только одно требование. Если в момент поступления очередного требования имеются свободные аппараты, то один из них немедленно приступает к обслуживанию этого требования, если же все аппараты заняты, то требование ждет, пока освободится один из них. Следовательно, если число требований, нуждающихся в обслуживании, превысит количество обслуживающих аппаратов, то образуется очередь. Время обслуживания одного требования есть случайная величина γ, подчиненная показательному закону распределения

$$P\{\gamma < t\} = F(t) = 1 - e^{-\nu t}.$$

Поток поступающих требований ограничен, т. е. одновременно в системе обслуживания не может находиться больше m требований, где m — конечное число. Это дает нам право считать, что требования на обслуживание поступают от m обслуживающих объектов, которые время от времени нуждаются в обслуживании. При этом часть времени они находятся в системе обслуживания, а часть вне ее.

Поток требований обладает следующими свойствами:

1. Вероятность того, что требование поступит на обслуживание за время $(t, t+\Delta t)$, если оно не поступило до момента t, равна $\lambda \Delta t + o(\Delta t)$, где $\lambda > 0$ и не зависит от m, t и числа требований, поступивших до него.

2. Моменты поступления данного требования в непересекающиеся промежутки времени — события независимые.

Эти условия очень близки к тем, которым подчиняется простейший поток требований. Так, в частности, нетрудно показать, что если требование не поступило до момента t_0, то вероятность его непоступления до момента $t_0 + t$

равна $P(t_0, t) = V_0(t) = e^{-\lambda t}$. Действительно, $P(t_0, t+\Delta t)$— вероятность непоступления требования за время $t+\Delta t$ согласно второму условию, по теореме умножения вероятностей, равна произведению вероятности того, что оно не поступит за (t_0, t_0+t), на вероятность того, что оно не поступит за $(t_0+t, t_0+t+\Delta t)$, т. е.

$$P(t_0, t+\Delta t) = P(t_0, t)\, P(t_0+t, \Delta t).$$

Но по первому условию вероятность поступления требования за $(t_0+t, t_0+t+\Delta t)$ равна

$$1 - P(t_0+t, \Delta t) = \lambda \Delta t + o(\Delta t).$$

Поэтому

$$P(t_0, t+\Delta t) = P(t_0, t)(1 - \lambda \Delta t - o(\Delta t)),$$

откуда

$$\frac{P(t_0, t+\Delta t) - P(t_0, t)}{\Delta t} = -\lambda P(t_0, t) + \frac{o(\Delta t)}{\Delta t} P(t_0, t).$$

Переходя к пределу при $\Delta t \to 0$, получаем дифференциальное уравнение

$$\frac{\partial P(t_0, t)}{\partial t} = -\lambda P(t_0, t).$$

Так как $P(t_0, 0) = 1$*, то, интегрируя уравнение, получаем $P(t_0, t) = e^{-\lambda t}$, т. е. $P(t_0, t)$ совпадает с вероятностью отсутствия требований в простейшем потоке $V_0(t) = e^{-\lambda t}$. Здесь параметр λ характеризует частоту возвращения требования в обслуживающую систему. Чем больше λ, тем больше вероятность поступления (или возвращения) требования в обслуживающую систему. Получается, что промежуток времени, в течение которого обслуживаемый объект не требует обслуживания, подчинен показательному закону. В частности, это означает, что среднее время нахождения его вне системы обслуживания равно $\frac{1}{\lambda}$.

В качестве критерия, характеризующего качество функционирования рассматриваемой системы, выберем

* Это условие означает, что если до момента t_0 требование не поступило, то вероятность того, что оно не поступит в момент t_0, равна единице.

отношение средней длины очереди к m — наибольшему числу требований, находящихся одновременно в обслуживающей системе. Это отношение назовем *коэффициентом простоя обслуживаемого объекта.* В качестве другого критерия возьмем *отношение среднего числа незанятых обслуживающих аппаратов к их общему числу.* Назовем это отношение *коэффициентом простоя обслуживающего аппарата.*

Совершенно ясно, что эти два критерия, характеризующие качество функционирования обслуживающей системы, коэффициент простоя обслуживаемого объекта и коэффициент простоя обслуживающего аппарата, при решении многих конкретных задач массового обслуживания имеют важное практическое значение. Если первый критерий характеризует потери времени за счет ожидания начала обслуживания, то второй показывает полноту загрузки обслуживающей системы. Представляют интерес также и такие критерии, как средняя длина очереди, вероятность иметь в очереди больше чем заданное число требований и т. д.

Так как в системе обслуживания одновременно не может находиться больше m требований, то, следовательно, она может находиться в момент времени t не больше чем в $m+1$ различном состоянии. Эти состояния будут определяться числом требований, находящихся на обслуживании и ожидающих очереди. Очевидно, что очередь может возникнуть лишь при условии, что число аппаратов $n<m$. Этот случай представляет наибольший практический интерес, так как нет никакого смысла для обслуживания m требований выделять обслуживающих аппаратов больше чем m, поскольку в этом случае часть из них будет неизбежно простаивать.

Обозначим через $P_k(t)$ $(k=0, 1, 2, ..., m)$ вероятность того, что в системе обслуживания в момент времени t находится точно k требований, и выведем систему дифференциальных уравнений для $P_k(t)$.

Дифференциальные уравнения решения задачи. Метод, который будет нами использован для вывода системы уравнений, имеет много общего с тем, который был применен в § 1. Воспользуемся формулой (3.2):

$$P_k(t+\Delta t)=\sum_{i=0}^{m} P_i(t)\,P_{ik}(\Delta t)$$

и вычислим вероятности перехода системы из состояния i, когда имеется i требований, в состояние k за время Δt, т. е. $P_{ik}(\Delta t)$. Нетрудно показать, как это было сделано ранее, что вероятность

$$P_{ik}(\Delta t) = o(\Delta t) \text{ при } |i-k| \geqslant 2.$$

Действительно, вероятность поступления даже двух требований за время Δt по теореме умножения вероятностей равна

$$[\lambda \Delta t + o(\Delta t)]^2 = (\lambda \Delta t)^2 + 2\lambda \Delta t + [o(\Delta t)]^2 = o(\Delta t),$$

а вероятность поступления более чем двух требований за это же время тем более является бесконечно малой величиной более высокого порядка, чем Δt.

Вероятность того, что за время Δt систему покинут более двух требований, равна

$$(1 - e^{-\nu \Delta t})^2 = [\nu \Delta t + o(\Delta t)]^2 = o(\Delta t).$$

Тем более бесконечно мала по сравнению с величиной Δt вероятность того, что систему за отрезок времени Δt одновременно покинут и в нее поступят несколько требований.

Таким образом, вероятность перехода рассматриваемой системы из состояния i в состояние k за время Δt для $|i-k| \geqslant 2$ есть бесконечно малая величина более высокого порядка, чем Δt, т. е.

$$P_{i,k}(\Delta t) = o(\Delta t) \text{ при } |i-k| \geqslant 2.$$

Вычислим теперь остальные переходные вероятности. Вероятность того, что в системе через время Δt не будет ни одного требования, если в начальный момент она была свободна, равна вероятности того, что за это время не поступит ни одного из m возможных требований. Так как вероятность поступления одного требования равна $\lambda \Delta t + o(\Delta t)$, а события поступления требований независимы, то, применяя теорему умножения вероятностей, получаем

$$P_{0,0}(\Delta t) = [1 - \lambda \Delta t - o(\Delta t)]^m = 1 - \lambda m \Delta t + o(\Delta t).$$

Вероятность того, что из имеющихся k требований ни одно не покинет систему за время Δt и ни одно новое не поступит за это время, может быть найдена как произведение вероятности того, что из $m-k$ возможных требований ни одно не поступит за время Δt, на вероятность того, что за время Δt не будет окончено обслуживание ни одного из $k (k<n)$ обслуживаемых требований:

$$P_{kk}(\Delta t) = [1 - \lambda \Delta t - o(\Delta t)]^{m-k} [e^{-\nu \Delta t}]^k =$$

$$= 1 - \lambda (m-k) \Delta t - \nu k \Delta t + o(\Delta t) \qquad (0 < k < n).$$

Эта формула верна для случая, когда в системе отсутствует очередь, т. е. $k \leqslant n$. Аналогичная величина для случая $k>n$ отличается тем, что при любом $k>n$ производится обслуживание только n требований, поэтому

$$P_{kk}(\Delta t) = 1 - (m-k) \lambda \Delta t - \nu n \Delta t + o(\Delta t) \qquad (n \leqslant k \leqslant m).$$

Вероятность того, что в системе через время Δt на обслуживании останутся все имеющиеся m требований, легко вычисляется, если учесть, что ни одно новое требование за это время не может поступить, поэтому

$$P_{mm}(\Delta t) = 1 - n \nu \Delta t + o(\Delta t).$$

Вероятность перехода системы от k к $k+1$ требованию, находящемуся в системе обслуживания за время Δt, может быть найдена аналогично. Это событие равно произведению двух событий. Первое — за время Δt в систему поступит одно из $m-k$ возможных требований, вероятность чего равна

$$C^1_{m-k} [\lambda \Delta t + o(\Delta t)] = (m-k) \lambda \Delta t + o(\Delta t),$$

второе — за время Δt не будет закончено обслуживание ни одного из k требований, вероятность чего равна

$$[e^{-\nu \Delta t}]^k = 1 - \nu k \Delta t + \frac{(\nu k \Delta t)^2}{2!} - \ldots = 1 - \nu k \Delta t + o(\Delta t),$$

поэтому

$$P_{k,\ k+1}(\Delta t) = (m-k) \lambda \Delta t + o(\Delta t) \qquad (0 \leqslant k < m).$$

Вероятность того, что из k требований, находящихся на обслуживании в момент t, через Δt останется $k-1$ с точностью до бесконечно малых высшего порядка по сравнению с Δt, равна произведению вероятностей следующих двух событий. Первое — за время Δt будет обслужено только одно из k обслуживаемых требований, вероятность чего равна

$$C_k^1 (1 - e^{-\nu \Delta t}) (e^{-\nu \Delta t})^{k-1} =$$

$$= k [\nu \Delta t + o(\Delta t)] [1 - (k-1) \nu \Delta t + o(\Delta t)] = k \nu \Delta t + o(\Delta t),$$

второе — за это время не поступит ни одно из $m-k$ требований, вероятность чего равна

$$[1 - \lambda \Delta t - o(\Delta t)]^{m-k} = 1 - (m-k) \lambda \Delta t + o(\Delta t),$$

поэтому вероятность того, что через Δt на обслуживании останется $k-1$ из k имеющихся требований, равна

$$P_{k,\, k-1}(\Delta t) = [k \nu \Delta t + o(\Delta t)] [1 - (m-k) \lambda \Delta t + o(\Delta t)] =$$

$$= k \nu \Delta t + o(\Delta t) \qquad (0 < k < n).$$

Эта формула верна для $k<n$, т. е. для случая, когда отсутствует очередь. Если же все аппараты заняты, то вероятность того, что освободится один из них, при $k \geqslant n$ не зависит от k и равна

$$C_n^1 (1 - e^{-\nu \Delta t}) (e^{-\nu \Delta t})^{n-1} = n \nu \Delta t + o(\Delta t),$$

поэтому

$$P_{k,\, k-1} = n \nu \Delta t + o(\Delta t) \quad (n \leqslant k \leqslant m).$$

Теперь, используя полученные значения переходных вероятностей и уравнение (3.2), определим исходную систему уравнений.

Если положим $k=0$, то

$$P_0 (t + \Delta t) = P_0 (t) P_{00} (\Delta t) + P_1 (t) P_{10} (\Delta t) + o(\Delta t).$$

При $0 < k < m$ получим

$$P_k (t + \Delta t) = P_{k-1} (t) P_{k-1,\, k} (\Delta t) + P_k (t) P_{kk} (\Delta t) +$$

$$+ P_{k+1} (t) P_{k+1,k} (\Delta t) + o(\Delta t).$$

При $k=m$ получим

$$P_m(t+\Delta t)=P_{m-1}(t)\,P_{m-1,\,m}(\Delta t)+$$
$$+P_m(t)\,P_{m,\,m}(\Delta t)+o(\Delta t).$$

Подставляя значения $P_{ik}(\Delta t)$ и учитывая, что $P_{kk}(\Delta t)$ и $P_{k+1,\,k}(\Delta t)$ имеют разное значение при $0<k<n$ и $n\leqslant k<m$, получаем

$$\left.\begin{aligned}
&P_0(t+\Delta t)=P_0(t)\,[1-\lambda m\Delta t+o(\Delta t)]+P_1(t)\,[\nu\Delta t+o(\Delta t)]\\
&P_k(t+\Delta t)=P_{k-1}(t)\,[(m-k+1)\,\lambda\Delta t+o(\Delta t)]+\\
&\quad+P_k(t)\,[1-(m-k)\,\lambda\Delta t-\nu k\Delta t+o(\Delta t)]+\\
&\quad+P_{k+1}(t)\,[(k+1)\,\nu\Delta t+o(\Delta t)]\\
&\qquad\qquad(0<k<n)\\
&P_k(t+\Delta t)=P_{k-1}(t)\,[(m-k+1)\,\lambda\Delta t+o(\Delta t)]+\\
&\quad+P_k(t)\,[1-(m-k)\,\lambda\Delta t-\nu n\Delta t+o(\Delta t)]+\\
&\quad+P_{k+1}(t)\,[\nu n\Delta t+o(\Delta t)]\\
&\qquad\qquad(n\leqslant k<m)\\
&P_m(t+\Delta t)=P_{m-1}(t)\,\lambda\Delta t+P_m(t)\,[1-\nu n\Delta t+o(\Delta t)]
\end{aligned}\right\}$$

Преобразовывая эту систему, получаем

$$\left.\begin{aligned}
&\frac{P_0(t+\Delta t)-P_0(t)}{\Delta t}=-\lambda mP_0(t)+\nu P_1(t)+o(\Delta t)\\
&\frac{P_k(t+\Delta t)-P_k(t)}{\Delta t}=(m-k+1)\,\lambda P_{k-1}(t)-[(m-k)\,\lambda+\\
&\quad+\nu k]\,P_k(t)+(k+1)\,\nu P_{k+1}(t)+o(\Delta t)\\
&\qquad\qquad(0<k<n)\\
&\frac{P_k(t+\Delta t)-P_k(t)}{\Delta t}=(m-k+1)\,\lambda P_{k-1}(t)-[(m-k)\,\lambda+\\
&\quad+\nu n]\,P_k(t)+n\nu P_{k+1}(t)+o(\Delta t)\\
&\qquad\qquad(n\leqslant k<m)\\
&\frac{P_m(t+\Delta t)-P_m(t)}{\Delta t}=\lambda P_{m-1}(t)-n\nu P_m(t)+o(\Delta t)
\end{aligned}\right\}$$

Если мы теперь перейдем к пределу при $\Delta t \to 0$, то получим следующую однородную систему обыкновенных линейных дифференциальных уравнений относительно $P_k(t)$ $(0 \leqslant k \leqslant m)$:

$$\left.\begin{aligned}
P_0'(t) &= -\lambda m P_0(t) + \nu P_1(t) \\
P_k'(t) &= (m-k+1)\lambda P_{k-1}(t) - [(m-k)\lambda + k\nu] P_k(t) + \\
&\quad + (k+1)\nu P_{k+1}(t) \qquad (0 < k < n) \\
P_k'(t) &= (m-k+1)\lambda P_{k-1}(t) - [(m-k)\lambda + n\nu] P_k(t) + \\
&\quad + n\nu P_{k+1}(t) \qquad (n \leqslant k < m) \\
P_m'(t) &= \lambda P_{m-1}(t) - n\nu P_m(t)
\end{aligned}\right\}.$$

Эта система состоит из конечного числа уравнений и может быть проинтегрирована одним из известных методов. Не останавливаясь на отыскании этого решения, ограничимся получением стационарного решения системы. Методом, аналогичным изложенному в § 1, можно показать существование стационарного решения, т. е. существование пределов

$$\lim_{t \to \infty} P_k(t) = p_k \qquad (k = 0,\ 1, \ldots, m),$$

не зависящих от начального состояния системы. Найдем это предельное решение. Для этого перейдем к пределу при $t \to \infty$ в предыдущей системе. Легко видеть, что

$$\lim_{t \to \infty} P_k'(t) = 0,$$

так как в противном случае $|P_k(t)| \to \infty$ при $t \to \infty$, а это невозможно по смыслу величин $P_k(t)$. Поэтому пределы левых частей системы равны нулю. Переходя к пределу, получаем

$$\left.\begin{aligned}
0 &= -\lambda m p_0 + \nu p_1 \\
0 &= (m-k+1)\lambda p_{k-1} - [(m-k)\lambda + k\nu] p_k + \\
&\quad + (k+1)\nu p_{k+1} \qquad (0 < k < n) \\
0 &= (m-k+1)\lambda p_{k-1} - [(m-k)\lambda + n\nu] p_k + n\nu p_{k+1} \\
&\qquad (n \leqslant k < m) \\
0 &= \lambda p_{m-1} - n\nu p_m
\end{aligned}\right\}.$$

Для того чтобы решить эту линейную однородную систему алгебраических уравнений, введем новые переменные следующим образом:

$$z_k = (m-k)\lambda p_k - (k+1)\nu p_{k+1} \quad \text{при} \quad 0 \leqslant k < n,$$

$$z_k = (m-k)\lambda p_k - n\nu p_{k+1} \quad \text{при} \quad n \leqslant k \leqslant m.$$

Произведя замену в системе, получим новую систему уравнений относительно z_k.

Первое уравнение дает

$$z_0 = 0.$$

Вторая группа уравнений при подстановке дает

$$z_{k-1} - z_k = 0 \qquad (0 < k < n).$$

Третья группа уравнений при подстановке дает

$$z_{k-1} - z_k = 0 \qquad (n \leqslant k < m).$$

Наконец, последнее уравнение превращается в

$$z_{m-1} = 0.$$

Отсюда следует, что $z_k = 0$ при $k = 0, 1, 2, \ldots, m-1$. Следовательно, возвращаясь к p_k, получаем

$$p_{k+1} = \frac{(m-k)\lambda}{(k+1)\nu} p_k \qquad (0 \leqslant k < n),$$

$$p_{k+1} = \frac{(m-k)\lambda}{n\nu} p_k \qquad (n \leqslant k \leqslant m).$$

Из этих соотношений легко найти выражения для p_k через p_0:

$$p_1 = m \frac{\lambda}{\nu} p_0,$$

$$p_2 = \frac{m-1}{2} \frac{\lambda}{\nu} p_1 = \frac{m(m-1)}{1 \cdot 2} \left(\frac{\lambda}{\nu}\right)^2 p_0 \text{ и т. д.}$$

откуда видно, что

$$p_k = \frac{m!}{k!\,(m-k)!}\left(\frac{\lambda}{\nu}\right)^k p_0 \qquad (1 \leqslant k < n).$$

Для $n \leqslant k \leqslant m$ получаем

$$p_{n+1} = \frac{(m-n)\,\lambda}{n\nu}\,p_n = \frac{m!}{n \cdot n!\,[m-(n+1)]!}\left(\frac{\lambda}{\nu}\right)^{n+1} p_0,$$

$$p_{n+2} = \frac{[m-(n+1)]\,\lambda}{n\nu}\,p_{n+1} = \frac{m!}{n^2 \cdot n!\,[m-(n+2)]!}\left(\frac{\lambda}{\nu}\right)^{n+2} p_0.$$

Легко видеть, что

$$p_k = \frac{m!}{n^{k-n}\,n!\,(m-k)!}\left(\frac{\lambda}{\nu}\right)^k p_0 \qquad (n \leqslant k \leqslant m).$$

Таким образом,

$$p_k = \frac{m!}{k!\,(m-k)!}\left(\frac{\lambda}{\nu}\right)^k p_0 \qquad (1 \leqslant k \leqslant n),$$

$$p_k = \frac{m!}{n^{k-n}\,n!\,(m-k)!}\left(\frac{\lambda}{\nu}\right)^k p_0 \qquad (n < k \leqslant m). \quad (3.25)$$

Для определения величины p_0 необходимо использовать очевидное условие

$$\sum_{k=0}^{m} p_k = 1,$$

откуда

$$p_0 = 1 - \sum_{k=1}^{m} p_k.$$

Другим способом величина p_0 может быть получена путем подстановки в равенство

$$\sum_{k=0}^{m} p_k = 1$$

значений $p_1, p_2, \ldots, p_m$, в которые p_0 входит сомножителем. Подставляя их, получаем следующее уравнение для определения p_0:

$$p_0\left[\sum_{k=0}^{n} \frac{m!}{k!\,(m-k)!}\left(\frac{\lambda}{\nu}\right)^k + \sum_{k=n+1}^{m} \frac{m!}{n^{k-n}\,n!\,(m-k)!}\left(\frac{\lambda}{\nu}\right)^k\right] = 1.$$

Математическое ожидание длины очереди, т. е. среднего числа требований, ожидающих начала обслуживания, равно

$$M_1 = \sum_{k=n+1}^{m} (k-n)\, p_k \tag{3.26}$$

или

$$M_1 = \sum_{k=n+1}^{m} \frac{(k-n)\, m!}{n^{k-n}\, n!\, (m-k)!} \left(\frac{\lambda}{\nu}\right)^k p_0.$$

Следовательно, коэффициент простоя обслуживаемого объекта равен

$$\frac{M_1}{m} = \frac{1}{m} \sum_{k=n+1}^{m} (k-n)\, p_k. \tag{3.27}$$

Математическое ожидание числа требований, находящихся в обслуживающей системе, обслуживаемых и ожидающих обслуживания, равно

$$M_2 = \sum_{k=1}^{n} k p_k$$

или

$$M_2 = \left[\sum_{k=1}^{n} \frac{m!}{(k-1)!\, (m-k)!} \left(\frac{\lambda}{\nu}\right)^k + \right.$$

$$\left. + \sum_{k=n+1}^{m} \frac{k \cdot m!}{n^{k-n}\, n!\, (m-k)!} \left(\frac{\lambda}{\nu}\right)^k \right] p_0.$$

Математическое ожидание числа свободных обслуживающих аппаратов равно

$$M_3 = \sum_{k=0}^{n-1} (n-k)\, p_k = \sum_{k=0}^{n} \frac{(n-k)\, m!}{k!\, (m-k)!} \left(\frac{\lambda}{\nu}\right)^k p_0,$$

поэтому коэффициент простоя обслуживающего аппарата равен

$$\frac{M_3}{n} = \frac{1}{n} \sum_{k=0}^{n-1} (n-k)\, p_k = \sum_{k=0}^{n-1} p_k - \frac{1}{n} \sum_{k=0}^{n-1} k p_k.$$

Вероятность того, что число требований, ожидающих обслуживания, будет больше заданного числа N, равна

$$p_{>N} = \sum_{k=N+1}^{m} p_k$$

или

$$p_{>N} = 1 - \sum_{k=0}^{N} p_k.$$

Выводы. 1. Вероятность того, что занято k обслуживающих аппаратов, при условии, что число требований, находящихся в системе, не превосходит числа обслуживающих аппаратов системы

$$p_k = \frac{m!}{k!\,(m-k)!}\left(\frac{\lambda}{\nu}\right)^k p_0 \qquad (1 \leqslant k \leqslant n).$$

Напомним, что λ — среднее число требований, поступающих в единицу времени, $\frac{1}{\nu}$ — среднее время обслуживания одного требования; m — наибольшее возможное число требований, находящихся в обслуживающей системе одновременно.

2. Вероятность того, что в системе находится k требований, для случая, когда их число больше числа обслуживающих аппаратов,

$$p_k = \frac{m!}{n^{k-n}\,n!\,(m-k)!}\left(\frac{\lambda}{\nu}\right)^k p_0 \qquad (n < k \leqslant m).$$

Напомним, что n — число обслуживающих аппаратов системы;

m, λ, ν — имеют прежний смысл.

3. Вероятность того, что все обслуживающие аппараты свободны,

$$p_0 = \left[\sum_{k=0}^{n} \frac{m!}{k!\,(m-k)!}\left(\frac{\lambda}{\nu}\right)^k + \right.$$

$$\left. + \sum_{k=n+1}^{m} \frac{m!}{n^{k-n}\,n!\,(m-k)!}\left(\frac{\lambda}{\nu}\right)^k\right]^{-1}.$$

4. Среднее число требований, ожидающих начала обслуживания (средняя длина очереди),

$$M_1 = \sum_{k=n+1}^{m} \frac{(k-n)\,m!}{n^{k-n}\,n!\,(m-k)!} \left(\frac{\lambda}{\nu}\right)^k p_0.$$

5. Коэффициент простоя обслуживаемого требования

$$\frac{M_1}{m} = \frac{(m-1)!}{n!} \sum_{k=n+1}^{m} \frac{k-n}{n^{k-n}\,(m-k)!} \left(\frac{\lambda}{\nu}\right)^k p_0.$$

6. Среднее число требований, находящихся в обслуживающей системе,

$$M_2 = \left[\sum_{k=1}^{n} \frac{m!}{(k-1)!\,(m-k)!} \left(\frac{\lambda}{\nu}\right)^k + \right.$$

$$\left. + \sum_{k=n+1}^{m} \frac{k \cdot m!}{n^{k-n}\,n!\,(m-k)!} \left(\frac{\lambda}{\nu}\right)^k \right] p_0.$$

7. Среднее число свободных обслуживающих аппаратов

$$M_3 = \sum_{k=1}^{n-1} (n-k)\,p_k = \sum_{k=0}^{n} \frac{(n-k)\,m!}{k!\,(m-k)!} \left(\frac{\lambda}{\nu}\right)^k p_0.$$

8. Коэффициент простоя обслуживающего аппарата

$$\frac{M_3}{n} = \sum_{k=0}^{n-1} p_k - \frac{1}{n} \sum_{k=0}^{n-1} k p_k.$$

9. Вероятность того, что число требований, ожидающих начала обслуживания, больше некоторого числа N,

$$p_{>N} = \sum_{k=N+1}^{m} p_k = 1 - \sum_{k=0}^{N} p_k, \quad N \geqslant n.$$

Примеры. Рабочий обслуживает группу автоматов, состоящую из m станков. При нормальной работе автомат

не требует вмешательства человека. В среднем автомат останавливается один раз в час. Вероятность того, что он не остановится за время t, убывает с ростом t. Пусть она равна $e^{-\lambda t}$. Следовательно, вероятность того, что время работы станка (γ) до остановки будет меньше t, равна

$$P\{\gamma < t\} = 1 - e^{-\lambda t}.$$

Напомним, что в § 2 гл. 2, было показано, что если случайная величина γ подчиняется закону

$$P\{\gamma < t\} = F(t) = 1 - e^{-\nu t},$$

то математическое ожидание этой случайной величины равно $\frac{1}{\nu}$. Так как в рассматриваемом случае время между остановками автомата подчинено этому закону, то $\frac{1}{\lambda}$ есть среднее время между остановками. Но по условию это время равно 1 часу, поэтому $\lambda = 1$. Далее предположим, что обслуживание одного станка занимает у рабочего в среднем 6 минут и время обслуживания, которое является случайной величиной, подчинено показательному закону с параметром ν. Тогда $\frac{1}{\nu} = 0{,}1$ часа, т. е. $\nu = 10$.

Необходимо определить среднее число автоматов, ожидающих обслуживания, коэффициент простоя автомата, коэффициент простоя рабочего, если рабочий обслуживает 6 станков. Легко заметить, что эта задача является частным случаем задачи, рассмотренной выше. Обслуживающим «аппаратом» здесь является рабочий; так как станки обслуживает один рабочий, то $n = 1$. Закон, которому подчинен поток требований, является частным случаем рассмотренного ($\lambda = 1$). Общее число требований не может превзойти числа станков, т. е. $m = 6$. Система может находиться в 7 различных состояниях: все станки работают; один стоит и обслуживается рабочим, а пять работают; два стоят, из них один обслуживается рабочим, один ждет обслуживания, а четыре работают и т. д., наконец, все станки стоят, один из них обслуживается рабочим, а пять ждут очереди. Таким образом, для ответа на поставленные вопросы можно

воспользоваться ранее выведенными формулами, положив в них $n=1$. В частности, формулы (3.25) имеют вид

$$p_1 = \frac{6!}{1!(6-1)!}(0,1)^1 p_0 = 0,6p_0,$$

$$p_k = \frac{6!}{(6-k)!}(0,1)^k p_0 \qquad (2 \leqslant k \leqslant 6).$$

Напомним, что p_1 — вероятность того, что рабочий занят обслуживанием одного станка, а остальные станки работают, p_k — вероятность того, что рабочий ремонтирует один станок, а $k-1$ стоят.

Остальные формулы преобразуются аналогично. Сведем вычисления в таблицу *.

k	Число автоматов, ожидающих обслуживания $(k-1)$	$\frac{p_k}{p_0}$	p_k	$(k-1)\,p_k$	kp_k
0	0	1,00000	0,4845	0	0
1	0	0,60000	0,2907	0	0,2907
2	1	0,30000	0,1454	0,1454	0,2908
3	2	0,12000	0,0582	0,1164	0,1746
4	3	0,03600	0,0175	0,0525	0,0700
5	4	0,00720	0,0035	0,0140	0,0175
6	5	0,00072	0,0003	0,0015	0,0018

В этой таблице первым вычисляется третий столбец, т. е. отношения $\frac{p_k}{p_0}$ при $k=0, 1, 2, \ldots, 6$, которые получаются из формулы (3.25). Затем, суммируя третий столбец и учитывая, что $\sum_{k=0}^{6} p_k = 1$, получаем

$$\sum_{k=0}^{6} \frac{p_k}{p_0} = \frac{1}{p_0}\sum_{k=0}^{6} p_k = \frac{1}{p_0} = 2,06392,$$

откуда $p_0=0,4845$. Умножая величины третьего столбца на p_0, получаем четвертый столбец. Величина $p_0=0,4845$,

* Пример заимствован у В. Феллера. Как указывает Феллер, этот метод успешно применяется в шведской промышленности.

равная вероятности того, что все автоматы работают, может быть истолкована как вероятность того, что рабочий свободен. Получается, что в рассматриваемом случае рабочий будет свободен около половины всего рабочего времени. Однако это не означает, что «очередь» станков, ожидающих обслуживания, всегда будет отсутствовать. Математическое ожидание числа автоматов, стоящих в очереди, равно

$$M_1 = \sum_{k=2}^{6} (k-1)\, p_k.$$

Суммируя пятый столбец, получим $M_1 = 0{,}3298$, следовательно, в среднем из шести станков 0,33 станка будет простаивать в ожидании, пока освободится рабочий.

Суммируя шестой столбец, получим математическое ожидание числа простаивающих станков (ремонтируемых и ожидающих ремонта):

$$M_2 = \sum_{k=1}^{6} k p_k = 0{,}8454,$$

т. е. в среднем 0,8454 часть рабочего времени один из шести станков не будет выдавать продукцию.

Коэффициент простоя станка равен

$$\frac{M_1}{6} = 0{,}0549,$$

т. е. каждый станок простаивает примерно 0,055 часть рабочего времени в ожидании, пока рабочий освободится. Коэффициент простоя рабочего в данном случае совпадает с p_0, так как $n = 1$, поэтому

$$\frac{M_2}{n} = 0{,}4845.$$

Рассмотрим другой пример. Пусть теперь не один рабочий, а бригада из 3 человек обслуживает 20 станков. В среднем на одного рабочего приходится $6^2/_3$ станка. Все остальные условия предыдущего примера будем считать прежними. Таким образом, теперь одновременно могут обслуживаться три станка. Определим те же

параметры, что и в предыдущей задаче. Результаты вычислений сведем в следующую таблицу:

k	Число обслуживаемых станков	Число станков, ожидающих обслуживания	Число свободных рабочих	p_k	$(k-3)\,p_k$	kp_k
0	0	0	3	0,13626	—	—
1	1	0	2	0,27250	—	0,27250
2	2	0	1	0,25888	—	0,51776
3	3	0	0	0,15533	—	0,46599
4	3	1	0	0,08802	0,08802	0,35208
5	3	2	0	0,04694	0,09388	0,23470
6	3	3	0	0,02347	0,07041	0,14082
7	3	4	0	0,01095	0,04380	0,07665
8	3	5	0	0,00475	0,02375	0,03800
9	3	6	0	0,00190	0,01140	0,01710
10	3	7	0	0,00070	0,00490	0,00700
11	3	8	0	0,00023	0,00184	0,00253
12	3	9	0	0,00007	0,00063	0,00084

При $k>12$ значения $P_k<0{,}5\cdot10^{-5}$, поэтому, ограничившись этой точностью, их можно не учитывать.

Математическое ожидание числа автоматов, ожидающих начала обслуживания, равно

$$M_1=\sum_{k=4}^{20}(k-3)\,p_k=0{,}33863.$$

Коэффициент простоя станка в этом случае равен

$$\frac{M_1}{20}=0{,}01693,$$

т. е. каждый станок в среднем будет простаивать 0,01693 часть рабочего времени. Сравним коэффициенты простоя станка для случая, когда один рабочий обслуживал шесть станков, и для рассматриваемого случая, когда на одного рабочего приходится $6^2/_3$ станка, но обслуживание ведется бригадой. Для первого случая коэффициент простоя был равен 0,0549, а для второго 0,0169, т. е. простой сократился на 0,038 рабочего времени. Таким образом, производительность на один станок увеличилась на 3,8% только за счет более рациональной организации

обслуживания с дополнительной экономией за счет того, что каждый рабочий обслуживает на $^2/_3$ станка больше и без всяких дополнительных материальных затрат.

Определим математическое ожидание числа простаивающих станков (ожидающих обслуживания и обслуживаемых):

$$M_2 = \sum_{k=1}^{20} k p_k = 2{,}12597,$$

т. е. из 20 станков простаивает чуть больше двух. В среднем один станок простаивает, с учетом обслуживания, 0,10630 часть рабочего времени. В том случае, когда обслуживание производил один рабочий, эта величина была равна

$$\frac{0{,}8526}{6} = 0{,}1421.$$

Вычислим коэффициент простоя рабочего, для чего найдем математическое ожидание числа свободных рабочих:

$$M_3 = \sum_{k=0}^{2} (3-k)\, p_k = 3p_0 + 2p_1 + p_2 = 1{,}21266.$$

Коэффициент простоя рабочего равен

$$\frac{M_3}{3} = 0{,}40422,$$

в то время как раньше он был равен 0,4845, т. е. средняя загрузка одного рабочего увеличилась примерно на 8%.

Эти два примера показывают, как применение методов теории массового обслуживания позволяет находить и обосновывать более рациональные способы организации обслуживания, обеспечивающие значительную экономию и повышение производительности труда без дополнительных материальных затрат. Нетрудно видеть, какую роль смогут сыграть эти методы при проектировании полностью автоматизированных обслуживающих систем, в которых устранение типичных неполадок в работе оборудования будет осуществляться автоматически. Предварительные расчеты помогут обеспечить создание экономически обоснованных проектов таких систем.

Рассмотрим еще один пример. Бригада из m однотипных сельскохозяйственных машин (тракторов или комбайнов и т. п.) должна проработать 1000 часов на участке, удаленном от центральных ремонтных мастерских на d *км*. Вероятность того, что одна машина без поломки сможет проработать t часов, равна $e^{-\lambda t}$, где $\lambda=0{,}0006$ (рис. 6).

Предположим, что в данной обстановке центральная ремонтная мастерская может выделить одну полевую

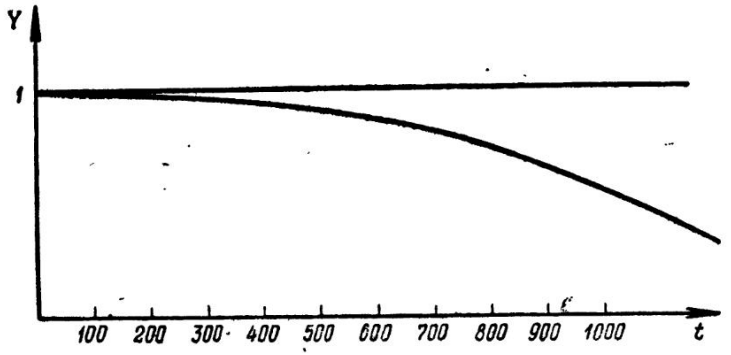

Рис. 6. График функции $Y = e^{-0{,}0006t}$ — вероятности того, что машина не потребует ремонта в течение t часов.

мастерскую, которая будет осуществлять ремонт на месте. При этом полевая мастерская может одновременно ремонтировать не больше одной машины. Время ремонта в полевой мастерской и в центральных мастерских подчинено одному и тому же закону. Пусть время ремонта подчинено показательному закону и среднее время ремонта равно 10 часам, т. е. $\nu=0{,}1$. Ремонт машин может быть организован или на месте, в полевой мастерской, или в центральных ремонтных мастерских. Предположим, что центральные мастерские обладают такой мощностью, которая позволяет им приступать к ремонту очередной неисправной машины в момент ее доставки. Пусть скорость доставки неисправной машины $v=40$ *км/час*.

Возникает вопрос, при каком удалении места работы от центральной станции выгоднее организовать ремонт на месте, а при каком доставлять машины в централь-

ные ремонтные мастерские? При ремонте на месте, в том случае, когда потребуется ремонтировать одновременно больше одной машины, остальные, вышедшие из строя, будут ждать своей очереди. За счет этого будут иметь место потери времени на ожидание начала ремонта (начала обслуживания). При ремонте в центральных мастерских потери рабочего времени будут связаны с доставкой машин на центральную станцию и обратно к месту работы. Время, затрачиваемое на ремонт в обоих случаях, считается одинаковым, хотя, вообще говоря, ремонт на центральной станции может быть произведен гораздо быстрее. Естественно, что предпочтение должно быть отдано тому способу, при котором простой машин будет меньше. Таким образом, для ответа на поставленный вопрос необходимо найти среднее время ожидания обслуживания на полевой станции и время, затрачиваемое на доставку машин на центральную станцию и обратно.

Разумеется, если подходить к решению этой задачи только с точки зрения материальных затрат, то ремонт вышедшей из строя техники в полевых условиях является более дешевым, так как к стоимости ремонта в центральных мастерских прибавится еще стоимость доставки неисправных машин к месту ремонта и обратно. Однако при проведении различного рода полевых работ (сев, жатва и т. п.) исключительно важное значение имеет время, в течение которого производятся эти работы, а оно в значительной степени будет зависеть от количества работающих в поле машин или, что то же самое, от времени простоя машин, нуждающихся в ремонте. Поэтому лучшим будет тот способ организации ремонта, для которого это время будет меньше. Решение этой задачи позволяет дать ответ на следующий вопрос. При каком расстоянии выгоднее использовать централизованный ремонт, а при каком местный. На основе качественных рассуждений можно сделать заключение, что если расстояние от места работы до центральных мастерских мало, то выгоднее использовать централизованную организацию ремонта, а если очень велико, то лучше ремонтировать неисправные машины на месте.

Следовательно, существует граница, за которой централизованный ремонт уже не будет давать необходимого эффекта. Найдем эту границу.

Время, затрачиваемое на доставку неисправных машин в центральные мастерские и обратно к месту работы, равно $\frac{2d}{v} = \frac{d}{20}$ часов. Найдем среднее время ожидания начала обслуживания при ремонте в полевой мастерской. Нетрудно видеть, что в такой постановке эта задача является частным случаем рассмотренной выше. Обслуживающей системой здесь является полевая мастерская. Она состоит из одного обслуживающего аппарата, так как по условию может обслуживать только одно требование, т. е. производить ремонт одной машины. До тех пор пока он не будет закончен, все другие машины, потребовавшие ремонта, должны ждать, пока не закончится ремонт первой, после чего можно будет начать ремонт второй и т. д. Наибольшее число требований, находящихся в обслуживающей системе, равно числу обслуживаемых машин m. Поток требований и время обслуживания, по условию, подчинены тем же законам. Следовательно, можно воспользоваться формулами, выведенными выше. Для определения среднего времени ожидания начала обслуживания вычислим математическое ожидание длины очереди

$$M_1 = \sum_{k=1}^{m} (k-1)\, p_k.$$

Пусть число машин $m = 5$. Вычислим

$$p_1 = \frac{5!}{1!4!}\left(\frac{0,0006}{0,1}\right) p_0,$$

$$p_k = \frac{5!}{(5-k)!}\,(0,006)^k\, p_0 \qquad (2 \leqslant k \leqslant 5),$$

где

$$\frac{1}{p_0} = \sum_{k=0}^{5} \frac{p_k}{p_0}.$$

Вычислим значения всех этих величин, задаваясь различным количеством машин, нуждающихся в ремонте, и результаты этих вычислений сведем в следующую таблицу:

k	Число машин, ожидающих обслуживания	$\frac{p_k}{p_0}$	p_k	$(k-1)\,p_k$
0	0	1,000000	0,970183	—
1	0	0,030000	0,029105	—
2	1	0,000720	0,000698	0,000698
3	2	0,000013	0,000012	0,000024
4	3	0,000000	0,000000	0,000000
5	4	0,000000	0,000000	0,000000

Вероятности p_4 и p_5 с точностью до 10^{-7} равны нулю. Математическое ожидание длины очереди $M_1=0{,}000722$. Коэффициент простоя машины равен

$$\frac{M_1}{5}=0{,}000145.$$

Таким образом, в среднем, одна машина будет за 1000 часов простаивать в ожидании начала ремонта 0,145 часа. Даже если не учитывать того, что за это время одна машина может потребовать ремонта больше чем один раз, а считать, что ремонт потребуется только один раз, то и в этом случае в центральную мастерскую невыгодно везти машины на ремонт с расстояния, большего $d=20\cdot 0{,}145=2{,}8$ *км*, так как при этом потери времени на перевозку будут бо́льшие, чем при ожидании обслуживания на месте. Коэффициент простоя мастерской при этом равен $M_3=p_0=0{,}970183$, т. е. загрузка мастерской окажется очень невысокой. Объясняется это высокой надежностью работы машин.

Если же предположить, что $\lambda_1=10\lambda=0{,}006$, при прочих неизменных условиях, то аналогичные расчеты приводят к следующим результатам:

k	Число машин, ожидающих обслуживания	$\frac{p_k}{p_0}$	p_k	$(k-1)\,p_k$
0	0	1,000000	0,721119	—
1	0	0,300000	0,216336	—
2	1	0,072000	0,051921	0,051921
3	2	0,012960	0,009346	0,018692
4	3	0,001555	0,001121	0,003363
5	4	0,000093	0,000067	0,000268

Математическое ожидание длины очереди увеличивается при этом до $M_1 = 0{,}074244$. Коэффициент простоя машины становится равным $\frac{M_1}{5} = 0{,}014848$. Таким образом, за счет того, что обслуживаемые машины менее надежны, увеличивается время ожидания начала обслуживания. На 1000 часов работы машины время простоя составит в среднем около 15 часов. Таким образом, если считать, что за это время ремонт каждой машины нужно будет произвести не больше одного раза, то доставка в центральные мастерские может считаться оправданной с расстояния $d = 15 \cdot 20 = 300$ *км*. Если считать, что за это время ремонт каждой машины придется произвести 3 раза, то уже на расстоянии до 100 *км* от места работы машин до центральной мастерской выгоднее организовать централизованный ремонт. Напомним, что выше было показано, что $\frac{1}{\lambda}$ есть среднее время нахождения машины вне обслуживания; так как $\lambda_1 = 0{,}006$, то $\frac{1}{\lambda_1} = 166$ часов, поэтому за 1 000 часов работы ремонт потребуется в среднем 6 раз. Следовательно, можно ожидать, что на расстояниях до 50 *км* централизованный ремонт себя оправдает.

Таким образом, в этих двух примерах только за счет изменения надежности работы машины, грубо говоря, в 10 раз величина d изменилась в 30—40 раз. Это показывает, сколь неточным и обманчивым может оказаться недостаточно строгий учет всех факторов и к каким излишним материальным затратам он может привести. Все подобные расчеты имеют исключительно важное практическое значение сейчас, когда большое внимание уделяется широкой механизации полевых работ и повышению надежности сельскохозяйственных машин.

Рассмотрим еще один пример. Для охраны района выделено 10 кораблей. Для ремонта кораблей выделено два дока. Каждый док может одновременно принять для ремонта один корабль. В док корабль ставится тогда, когда он не может нести охрану и нуждается в ремонте. Необходимо определить, какое количество кораблей в среднем будет находиться в строю, а какое количество будет ремонтироваться или ожидать ремонта. Кроме того, необходимо найти, какова вероятность иметь не ме-

нее 8 исправных кораблей для охраны района. Пусть вероятность того, что корабль за время t потребует ремонта, равна $1-e^{-0,02t}$ (рис. 7), где t — время в месяцах. Время ремонта есть случайная величина, которая определяется теми неисправностями, которые имеет ремонти-

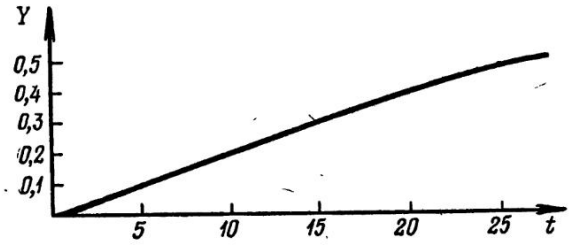

Рис. 7. График функции $Y=1-e^{-0,02t}$ — вероятности выхода корабля из строя.

руемый корабль. Пусть в среднем на ремонт одного корабля затрачивается 2 месяца и время ремонта подчинено показательному закону с параметром ν. Тогда, так как $\frac{1}{\nu}=2$, то $\nu=0,5$.

При такой постановке эта задача может быть решена с помощью рассмотренного метода. Ясно, что все формулы, выведенные выше, пригодны при решении этой задачи; так как система обслуживания ограничена $n=2$ доками, корабли, которые потребуют ремонта в момент, когда оба дока заняты, вынуждены будут ждать освобождения одного из них, т. е. рассматриваемая система является системой обслуживания с ожиданием; общее количество требований, находящихся в системе обслуживания одновременно, также ограничено числом кораблей $m=10$. Для ответа на поставленный вопрос необходимо вычислить математическое ожидание числа кораблей, находящихся в системе обслуживания, т. е. ремонтируемых и ожидающих ремонта. Эта величина равна $M_2=\sum_{k=1}^{10} kp_k$. Результаты вычислений сведем в таблицу:

k	$\frac{p_k}{p_0}$	p_k	kp_k
0	1,000000	0,673242	0
1	0,400000	0,269297	0,269297
2	0,072000	0,048473	0,096946
3	0,011520	0,007756	0,023268
4	0,001613	0,001086	0,004344
5	0,000193	0,000130	0,000650
6	0,000019	0,000013	0,000078
7	0,000002	0,000001	0,000007
8	0,000000	0,000000	0,000000
9	0,000000	0,000000	0,000000
10	0,000000	0,000000	0,000000

Напомним, что из (3.25)

$$p_k = \frac{10!}{k!\,(10-k)!}\left(\frac{0,02}{0,5}\right)^k p_0 \quad (k=1,\ 2),$$

$$p_k = \frac{10!}{2^{k-2}\cdot 2!\,(10-k)!}\left(\frac{0,02}{0,5}\right)^k p_0 \quad (3 \leqslant k \leqslant 10).$$

По этим формулам вычисляется второй столбец таблицы. Сумма всех чисел второго столбца равна $\frac{1}{p_0}$, поэтому, разделив единицу на эту сумму, находим p_0. Третий столбец получается из второго умножением на p_0. Сумма всех чисел четвертого столбца равна математическому ожиданию числа кораблей, находящихся в ремонте. Таким образом, $M_2 = 0,394590$. Следовательно, в среднем не больше одного корабля будет находиться в ремонте, а девять в строю. Вероятность того, что не менее восьми кораблей находится в строю, равна вероятности того, что в ремонте находится не больше двух кораблей:

$$p_{\geqslant 8} = p_0 + p_1 + p_2 = 0,991074.$$

Эти вычисления дают основание предполагать, что двух доков для обслуживания 10 кораблей много. Посмотрим, как изменится картина, если для ремонта

имеется только один док. Положим $n=1$ и произведем аналогичные вычисления. Теперь

$$p_1 = \frac{10!}{1!\,(10-1)!}\left(\frac{0,02}{0,5}\right)p_0,$$

$$p_k = \frac{10!}{(10-k)!}\left(\frac{0,02}{0,5}\right)^k p_0 \quad (2 \leqslant k \leqslant 10).$$

Задаваясь различными значениями k, вычислим значение p_k, а затем, как и в предыдущем случае, найдем значение p_0. Результаты всех вычислений сведем в следующую таблицу:

k	$\frac{p_k}{p_0}$	p_k	kp_k
0	1,000000	0,622351	—
1	0,400000	0,248940	0,248940
2	0,144000	0,089618	0,179236
3	0,046080	0,028678	0,086034
4	0,012902	0,008029	0,032116
5	0,003096	0,001927	0,009635
6	0,000619	0,000385	0,002310
7	0,000099	0,000062	0,000434
8	0,000012	0,000007	0,000056
9	0,000001	0,000001	0,000009
10	0,000000	0,000000	0,000000

Получается, что математическое ожидание числа кораблей, находящихся в ремонте, при этом становится равным

$$M_2 = \sum_{k=1}^{10} kp_k = 0,558770,$$

т. е. увеличивается всего на 0,16418. Это означает, что один док сможет не только успешно обеспечить ремонт этих десяти кораблей, но и будет обладать при этом значительными свободными производственными мощностями. Вероятность же того, что в строю будет находиться не менее восьми кораблей, в этом случае равна

$$p_{\geqslant 8} = p_0 + p_1 + p_2 = 0,961909,$$

т. е. уменьшится по сравнению с 0,991074 незначительно.

Вероятность $p_0=0{,}622351$ говорит о том, что больше 62% рабочего времени док будет пустовать, т. е. нагрузка на него может быть значительно увеличена. Этим примером закончим рассмотрение задачи первого типа.

Задача второго типа

В только что рассмотренной задаче первого типа существенным было то, что число требований, одновременно находящихся в обслуживающей системе, было ограниченным и не превышало некоторого известного или заданного числа. Снимем это ограничение и предположим, что в обслуживающую систему поступает неограниченный входящий поток. Такого рода потоки имеют место там, где число требований практически может быть очень большим, как, например, число заявок на электроэнергию в энергетической системе, число неисправных автомобилей, нуждающихся в ремонте в целом по всей стране, число абонентов, обращающихся на АТС за обслуживанием в большом городе с хорошо развитой телефонной сетью, и т. п.

Постановка задачи. Предположим, что в обслуживающую систему, состоящую из n аппаратов, поступает простейший поток требований с параметром λ. Каждый аппарат одновременно может обслуживать только одно требование. Если в момент поступления очередного требования в системе на обслуживании уже находится не меньше n требований, то это требование «становится» в очередь и ждет начала обслуживания. Время обслуживания одного требования подчинено показательному закону распределения с параметром ν.

При изучении таких систем нас могут интересовать различные показатели эффективности обслуживающей системы, в зависимости от конкретного содержания задачи. Так, например, часто необходимо знать, какова вероятность того, что все обслуживающие аппараты заняты. Для полной характеристики очереди необходимо знать закон распределения β — времени ожидания начала обслуживания. Во всех случаях полезно знать вероятность того, что в системе находится точно k требований (как обслуживаемых, так и ожидающих обслуживания). Они позволят определить такие важные показатели, как математическое ожидание длины очереди, математическое ожидание числа требований, находящихся

в системе обслуживания, математическое ожидание числа свободных обслуживающих аппаратов.

Как и прежде, через $P_k(t)$ обозначим вероятность того, что в момент времени t в системе на обслуживании и в ожиданий его находятся k требований. Если в предыдущей задаче общее число требований, находящихся в обслуживающей системе, не могло превосходить m, то в этой задаче в системе одновременно может находиться любое число требований, так как по условию входящий поток неограничен. При этом не следует представлять себе дело таким образом, что в обслуживающей системе будет постоянно находиться в очереди большое количество требований. Все зависит от соотношения между интенсивностью входящего потока, временем обслуживания одного требования и количеством обслуживающих аппаратов в системе. В дальнейшем увидим, что если $\frac{\lambda}{\nu} \leqslant n$, то очередь не может расти безгранично. Это условие имеет, грубо говоря, следующий смысл: λ — среднее число требований, поступающих за единицу времени, $\frac{1}{\nu}$ — среднее время обслуживания одним аппаратом одного требования, поэтому $\lambda \frac{1}{\nu} = \frac{\lambda}{\nu}$ — среднее число аппаратов, которое необходимо иметь, чтобы обслужить в единицу времени все поступающие требования. Поэтому условие $\frac{\lambda}{\nu} \leqslant n$ означает, что n, число обслуживающих аппаратов, не должно быть меньше среднего числа аппаратов, необходимых для того, чтобы за единицу времени обслужить все поступившие требования.

Решение задачи. Не будем повторять всех рассуждений, приведенных в предыдущей задаче при выводе системы дифференциальных уравнений для $P_k(t)$. Эти рассуждения будут отличаться только тем, что если раньше $P_k(t) \equiv 0$ при $k = m+1, m+2, \ldots$, так как в системе одновременно не могло находиться на обслуживании больше m требований, то теперь в системе одновременно может находиться на обслуживании любое число требований, поэтому $P_k(t) \not\equiv 0$ при любом k. Это означает, что число дифференциальных уравнений системы также неограниченно. В остальном, повторяя рассуждения поч-

ти дословно, придем к следующей системе дифференциальных уравнений:

$$\left.\begin{aligned} P_0'(t) &= -\lambda P_0(t) + \nu P_1(t) \\ P_k'(t) &= \lambda P_{k-1}(t) - (\lambda + k\nu) P_k(t) + (k+1)\nu P_{k+1}(t) \\ & \qquad (1 \leqslant k < n) \\ P_k'(t) &= \lambda P_{k-1}(t) - (\lambda + n\nu) P_k(t) + n\nu P_{k+1}(t) \\ & \qquad (k \geqslant n) \end{aligned}\right\}.$$

Как уже было сказано выше, ограничимся отысканием предельного решения системы, т. е. предположим, что пределы

$$\lim_{t\to\infty} P_k(t) = p_k$$

существуют при любых значениях k. Напомним, что при переходе к пределу в системе дифференциальных уравнений $\lim\limits_{t\to\infty} P_k'(t) = 0$, так как если бы он не был равен нулю, то $P_k(t)$ неограниченно возрастала бы с ростом t, а это невозможно. По смыслу ясно, что $P_k(t) \leqslant 1$, поэтому пределы левых частей при $t \to \infty$ равны нулю, что имеет место при любом k. Поэтому, переходя в системе дифференциальных уравнений к пределу, получаем следующую систему алгебраических уравнений:

$$\left.\begin{aligned} 0 &= -\lambda p_0 + \nu p_1 \\ 0 &= \lambda p_{k-1} - (\lambda + k\nu) p_k + (k+1)\nu p_{k+1} \quad (1 \leqslant k < n) \\ 0 &= \lambda p_{k-1} - (\lambda + n\nu) p_k + n\nu p_{k+1} \quad (k \geqslant n) \end{aligned}\right\}.$$

Так как обслуживающая система всегда находится в одном из возможных состояний, то к этим уравнениям необходимо добавить еще следующее условие:

$$\sum_{k=0}^{\infty} p_k = 1,$$

которое, как и прежде, будет использовано для определения величины p_0. Аналогично тому, как мы это делали

при рассмотрении первой задачи, произведем в этой системе замену неизвестных p_k на z_k:

$$z_k = \lambda p_{k-1} - k\nu p_k \quad (1 \leqslant k \leqslant n),$$

$$z_k = \lambda p_{k-1} - n\nu p_k \quad (k \geqslant n),$$

тогда система алгебраических уравнений преобразуется следующим образом:

$$z_1 = 0,$$

$$z_k - z_{k+1} = 0 \quad (k \geqslant 1).$$

Следовательно, $z_k = 0$ при любом k. Это означает, что

$$\lambda p_{k-1} = k\nu p_k \text{ при } 1 \leqslant k \leqslant n,$$

$$\lambda p_{k-1} = n\nu p_k \text{ при } k \geqslant n.$$

Из этих уравнений получаем, что

$$p_1 = \frac{\lambda}{\nu} p_0; \quad p_2 = \frac{1}{2}\left(\frac{\lambda}{\nu}\right)^2 p_0; \quad p_3 = \frac{1}{2 \cdot 3}\left(\frac{\lambda}{\nu}\right)^3 p_0 \text{ и т. д.,}$$

т. е. при $1 \leqslant k \leqslant n$

$$p_k = \frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k p_0, \tag{3.28}$$

а при $k \geqslant n$

$$p_k = \left(\frac{\lambda}{n\nu}\right)^{k-n} p_n$$

или

$$p_k = \frac{\lambda}{n\nu} p_{k-1} = \frac{1}{n! n^{k-n}}\left(\frac{\lambda}{\nu}\right)^k p_0. \tag{3.29}$$

Для определения p_0 воспользуемся условием

$$\sum_{k=0}^{\infty} p_k = 1,$$

откуда, подставляя полученные значения p_k, имеем

$$p_0\left[\sum_{k=0}^{n-1} \frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k + \frac{1}{n!}\left(\frac{\lambda}{\nu}\right)^n \sum_{k=n}^{\infty}\left(\frac{\lambda}{n\nu}\right)^{k-n}\right] = 1. \tag{3.30}$$

Сумма, стоящая в скобках, сходится к конечному пределу только при условии сходимости ряда

$$\sum_{k=n}^{\infty}\left(\frac{\lambda}{n\nu}\right)^{k-n}=\sum_{k=0}^{\infty}\left(\frac{\lambda}{n\nu}\right)^{k},$$

но этот ряд является суммой членов геометрической прогрессии со знаменателем $\frac{\lambda}{n\nu}$. Как известно, эта сумма существует при условии, что $\frac{\lambda}{n\nu}<1$. Она может быть при этом вычислена по известной формуле суммы членов бесконечно убывающей геометрической прогрессии

$$\sum_{k=0}^{\infty}\left(\frac{\lambda}{n\nu}\right)^{k}=\frac{1}{1-\frac{\lambda}{n\nu}}=\frac{n\nu}{n\nu-\lambda}.$$

Подставляя это выражение в квадратную скобку формулы (3.30), получаем, что

$$p_0\left[\sum_{k=0}^{n-1}\frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^{k}+\frac{\nu}{(n-1)!\,(n\nu-\lambda)}\left(\frac{\lambda}{\nu}\right)^{n}\right]=1, \quad (3.31)$$

откуда получаем, что

$$p_0=\frac{1}{\sum_{k=0}^{n-1}\frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^{k}+\frac{\nu}{(n-1)!\,(n\nu-\lambda)}\left(\frac{\lambda}{\nu}\right)^{n}}. \quad (3.32)$$

Если величина $\frac{\lambda}{n\nu}\geqslant 1$, то сумма ряда будет неограниченно возрастать и, как нетрудно видеть из (3.30), величина p_0 равна нулю. Действительно p_0 есть частное от деления единицы на бесконечно большую величину. Поэтому из (3.28) и (3.29) следует, что $p_k=0$ при любом k, т. е. вероятность того, что в обслуживающей системе нет ни одного требования или находится любое конечное число требований, при $t\to\infty$ равна нулю. Это означает, что число требований неограниченно возра-

стает, т. е. очередь неограниченно велика и система не справляется с обслуживанием. Объединяя (3.28), (3.29) и (3.32), получаем вероятности p_k для любых значений k, т. е. вероятности того, что в системе находится k требований.

Определим вероятность того, что все обслуживающие аппараты заняты. Очевидно, что это событие может иметь место тогда, когда в системе находятся n, $n+1$... требований одновременно. Эти события независимы, поэтому вероятность того, что все обслуживающие аппараты заняты, может быть найдена как сумма вероятностей p_n, p_{n+1}...

Обозначим через П вероятность того, что все обслуживающие аппараты заняты, тогда

$$\Pi = \sum_{k=n}^{\infty} p_k = \frac{n^n}{n!} p_0 \sum_{k=n}^{\infty} \left(\frac{\lambda}{n\nu}\right)^k.$$

Такая сумма, при $\frac{\lambda}{n\nu} < 1$, была найдена выше, поэтому, суммируя геометрическую прогрессию, получаем

$$\Pi = \frac{\nu \cdot p_0}{(n-1)!\,(n\nu - \lambda)} \left(\frac{\lambda}{\nu}\right)^n \text{ при } \frac{\lambda}{\nu} < n. \qquad (3.33)$$

Отсюда, в частности, следует, что

$$p_n = \left(1 - \frac{\lambda}{n\nu}\right) \Pi.$$

Теперь определим закон распределения длительности ожидания начала обслуживания. Время ожидания начала обслуживания является в общем случае случайной величиной, которая зависит от того, как много требований находится в системе обслуживания в данный момент и как скоро закончится их обслуживание. Определим распределение времени ожидания начала обслуживания в установившемся процессе, т. е. будем рассматривать только ранее полученные вероятности p_k.

Если β, время ожидания начала обслуживания, — случайная величина, то необходимо найти $P\{\beta > t\}$, т. е.

вероятность того, что время ожидания начала обслуживания больше произвольного отрезка времени t. Обозначим через $P_k\{\beta > t\}$ условную вероятность того, что время ожидания $\beta > t$, при условии, что в момент поступления требования в системе находилось k требований. Вероятности того, что в обслуживающей системе находится k требований, нам известны: это величина p_k. Поэтому, применяя формулу полной вероятности*, получаем

$$P\{\beta > t\} = \sum_{k=0}^{\infty} p_k \cdot P_k\{\beta > t\}.$$

Если в момент поступления очередного требования в системе будет находиться меньше n требований, то обслуживание очередного требования начинается немедленно, поэтому при $k < n$

$$P\{\beta > t\} = 0$$

для любого $t > 0$. Отсюда вытекает, что

$$P\{\beta > t\} = \sum_{k=n}^{\infty} p_k P_k\{\beta > t\}. \qquad (3.34)$$

Определим условные вероятности $P_k\{\beta > t\}$ при $k \geqslant n$.

---

* Формула полной вероятности имеет следующий смысл. Если событие A может произойти вместе с одним из событий $H_1, H_2, \ldots, H_n$, образующих полную группу несовместных событий (несовместность заключается в невозможности одновременного появления двух разных событий H, а полнота — в том, что одно из событий H_i должно произойти обязательно), то вероятность появления события A равна

$$P(A) = \sum_{i=1}^{n} P(H_i) P(A/H_i).$$

Здесь $P(H_i)$ — вероятность появления события H_i, а $P(A/H_i)$ — условная вероятность появления события A при условии, что событие H_i произошло. Эта формула является следствием теорем сложения и умножения вероятностей. В нашем случае событие H_i заключается в наличии i требований в обслуживающей системе, вероятность чего равна p_i, а условная вероятность того, что время обслуживания $\beta > t$ равна $P_i\{\beta > t\}$. Группа событий — полная, так как в системе или нет требований ($i = 0$), или есть какое-то число требований ($i = 1, 2, 3 \ldots$).

Пусть в момент поступления очередного требования в системе находится k требований. При этом $k>n$, тогда n из них обслуживаются, а $k-n$ ожидает своей очереди. Пусть обслуживание производится в порядке очереди, тогда вновь поступившее требование поступит на обслуживание только после того, как будет закончено обслуживание $(k-n+1)$-го требования. Обозначим через $q_s(t)$ вероятность того, что за время t закончится обслуживание ровно s требований. Тогда условная вероятность $P_k\{\beta>t\}$ по теореме сложения вероятностей равна

$$P_k\{\beta>t\}=\sum_{s=0}^{k-n} q_s(t) \quad (k\geqslant n),$$

так как время ожидания будет больше t в том случае, если за это время будет закончено обслуживание не больше $k-n$ требований.

В связи с тем, что время обслуживания подчинено показательному закону и не зависит от того, сколько требований находится в данный момент в очереди, то вероятность того, что за время t не освободится ни один обслуживающий аппарат, по теореме умножения вероятностей равна

$$q_0(t)=(e^{-\nu t})^n=e^{-\nu nt}. \tag{3.35}$$

Здесь $e^{-\nu t}$ есть вероятность того, что обслуживание не закончит один аппарат, а $q_0(t)$ — вероятность того, что обслуживание не закончат все n аппаратов.

Поток обслуженных требований является простейшим (см. § 1 гл. 2). Действительно, он обладает всеми тремя свойствами, характеризующими простейший поток: стационарностью, отсутствием последействия и ординарностью. Стационарность его вытекает из свойства показательного времени обслуживания — закон распределения времени обслуживания не зависит от того, сколько времени оно уже длилось (§ 2 гл. 2). Отсутствие последействия в этом потоке очевидно, так как число ранее обслуженных требований не может повлиять на последующий ход обслуживания. Что же касается ординарности этого потока, то она может быть легко установлена

следующим образом. Вероятность того, что будет закончено обслуживание двух требований за время Δt, равна

$$(1-e^{-\nu\Delta t})^2 \approx [\nu\Delta t + o(\Delta t)]^2 = o(\Delta t) \quad (\Delta t \to 0),$$

т. е. есть бесконечно малая величина по сравнению с Δt.

Сравнение выражений (3.55) и (2.7) показывает, что параметр этого потока обслуженных требований равен $\lambda_1 = n\nu$, поэтому вероятность того, что за время t будет закончено обслуживание точно s требований, можно вычислить по формуле (2.10), подставив в нее $\lambda_1 = n\nu$. Следовательно,

$$q_s(t) = \frac{(\nu nt)^s}{s!} \cdot e^{-\nu nt}$$

есть вероятность появления точно s требований в выходящем потоке. Иными словами, вероятность $q_s(t)$ есть характеристика продвижения очереди на s единиц за время t. Возвращаясь к условной вероятности $P_k\{\beta > t\}$, можно теперь написать, что

$$P_k\{\beta > t\} = \sum_{s=0}^{k-n} q_s(t) = \sum_{s=0}^{k-n} \frac{(\nu nt)^s}{s!} \cdot e^{-\nu nt}.$$

Подставляя выражение $P_k\{\beta > t\}$ в (3.34), получаем

$$P\{\beta > t\} = \sum_{k=n}^{\infty} p_k \sum_{s=0}^{k-n} \frac{(\nu nt)^s}{s!} e^{-\nu nt}.$$

Подставим сюда значение p_k из (3.29):

$$P\{\beta > t\} = p_n e^{-\nu nt} \sum_{k=n}^{\infty} \left(\frac{\lambda}{n\nu}\right)^{k-n} \sum_{s=0}^{k-n} \frac{(\nu nt)^s}{s!}.$$

Изменим порядок суммирования, тогда

$$P\{\beta > t\} = p_n e^{-\nu nt} \sum_{s=0}^{\infty} \frac{(\nu nt)^s}{s!} \sum_{k=n+s}^{\infty} \left(\frac{\lambda}{n\nu}\right)^{k-n} =$$

$$= p_n e^{-\nu nt} \sum_{s=0}^{\infty} \frac{(n\lambda t)^s}{n^s s!} \sum_{k=n+s}^{\infty} \left(\frac{\lambda}{n\nu}\right)^{k-n-s}.$$

Последняя сумма есть не что иное, как сумма членов геометрической прогрессии со знаменателем $\frac{\lambda}{n\nu}$, поэтому

$$P\{\beta > t\} = p_n e^{-\nu n t} \frac{n\nu}{n\nu - \lambda} \sum_{s=0}^{\infty} \frac{(\lambda t)^s}{s!}.$$

Сумма $\sum_{s=0}^{\infty} \frac{(\lambda t)^s}{s!}$ представляет разложение $e^{\lambda t}$ по степеням показателя, поэтому

$$P\{\beta > t\} = p_n e^{-\nu n t} \frac{n\nu}{n\nu - \lambda} e^{\lambda t} = \frac{n\nu p_n}{n\nu - \lambda} e^{-(n\nu - \lambda)t}.$$

Подставляя выражение

$$p_n = \frac{1}{n!} \left(\frac{\lambda}{\nu}\right)^n p_0 = \left(1 - \frac{\lambda}{n\nu}\right) \Pi,$$

получаем

$$P\{\beta > t\} = \Pi e^{-(n\nu - \lambda)t} \quad (t \geqslant 0). \tag{3.36}$$

Для отрицательных значений t вероятность того, что время ожидания больше t, очевидно, равна единице (рис. 8).

Следовательно, при $t = 0$ у функции $P\{\beta > t\}$ имеется разрыв типа скачка. Скачок функции равен $1 - \Pi$, так как при $t = +0$ $P\{\beta > +0\} = \Pi$, а при $t = -0$ $P\{\beta > -0\} = 1$, т. е. величина этого скачка равна вероятности того, что по крайней мере один обслуживающий аппарат свободен. Соответственно вероятность того, что время ожидания не превзойдет t, равна

$$P\{\beta \leqslant t\} = 1 - \Pi e^{-(n\nu - \lambda)t} \quad (t \geqslant 0),$$

$$P\{\beta \leqslant t\} = 0 \quad (t < 0).$$

Последнее вытекает из того, что время ожидания начала обслуживания или ноль, или положительная величина и неравенство $\beta \leqslant t$ при $t < 0$ не может иметь места. Эта функция также имеет разрыв при $t = 0$ (рис. 9).

Знание закона распределения времени начала обслуживания позволяет ответить на ряд важных вопросов.

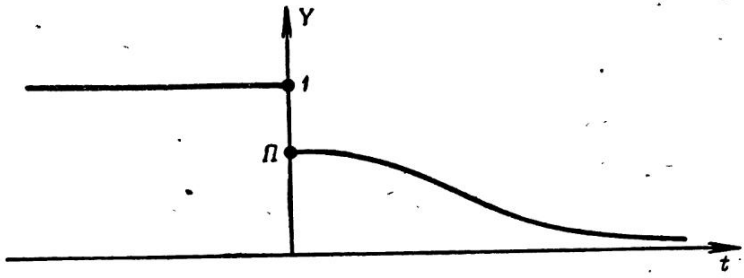

Рис. 8. График закона распределения времени ожидания начала обслуживания $Y = P\{\beta > t\}$.

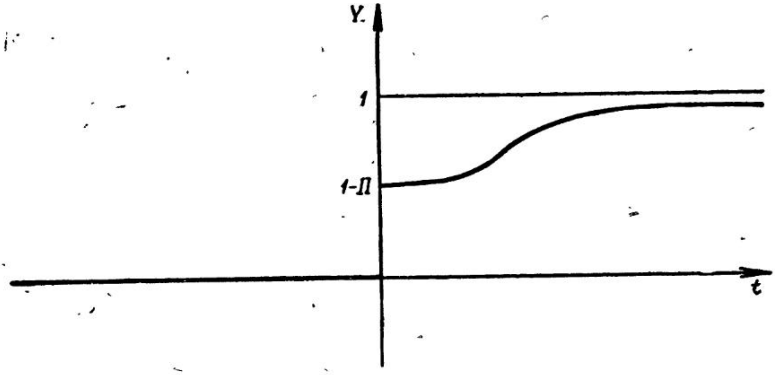

Рис. 9. График закона распределения $Y = P\{\beta \leqslant t\}$.

Так, например, можно определить среднее время ожидания начала обслуживания T_n при наличии n обслуживающих аппаратов, которое равно

$$T_n = M[\beta] = -\int_0^\infty t dP\{\beta > t\} =$$

$$= \Pi \int_0^\infty t(n\nu - \lambda) e^{-(n\nu-\lambda)t} dt.$$

Последний интеграл вычисляется по частям. В результате получаем

$$T_n = \Pi\left\{-te^{-(n\nu-\lambda)t}\Big|_0^\infty + \int_0^\infty e^{-(n\nu-\lambda)t}\,dt\right\} =$$

$$= \Pi\left\{0 - \frac{e^{-(n\nu-\lambda)t}}{n\nu-\lambda}\Big|_0^\infty\right\} = \frac{\Pi}{n\nu-\lambda}.$$

Таким образом, получаем, что среднее время ожидания начала обслуживания при n обслуживающих аппаратах равно

$$T_n = \frac{\Pi}{n\nu-\lambda}, \qquad (3.37)$$

т. е. оно пропорционально вероятности того, что все обслуживающие аппараты заняты.

Величины p_k позволяют вычислить среднее число требований, ожидающих начала обслуживания, т. е. математическое ожидание длины очереди M_1. Как и прежде [см. формулу (3.26)],

$$M_1 = \sum_{k=n}^{\infty}(k-n)\,p_k.$$

Отличие только в том, что если в (3.26) сумма состояла из конечного числа членов, то здесь число членов неограниченно. Подставив в это выражение значение p_k из (3.29), получим

$$M_1 = \sum_{k=n}^{\infty}(k-n)\,\frac{p_0}{n!n^{k-n}}\left(\frac{\lambda}{\nu}\right)^k = \frac{p_0}{n!}\left(\frac{\lambda}{\nu}\right)^n \sum_{k=0}^{\infty} k\left(\frac{\lambda}{n\nu}\right)^k =$$

$$= p_n \sum_{k=0}^{\infty} k\left(\frac{\lambda}{n\nu}\right)^k,$$

так как

$$p_n = \frac{p_0}{n!}\left(\frac{\lambda}{\nu}\right)^n.$$

Можно показать, что ряд $\sum_{k=0}^{\infty} k\left(\frac{\lambda}{n\nu}\right)^k$ сходится при $\frac{\lambda}{n\nu}<1$*, а это условие, как было показано выше, означает, что очередь не растет неограниченно.

Таким образом, среднее число требований, ожидающих в очереди начала обслуживания, равно

$$M_1 = p_n \sum_{k=0}^{\infty} k\left(\frac{\lambda}{n\nu}\right)^k.$$

Найдем более удобное выражение для M_1, для чего обозначим $\frac{\lambda}{n\nu}=a$ и вычислим сумму ряда

$$S=\sum_{k=0}^{\infty} ka^k.$$

Для этого вычислим частную сумму S_n и перейдем к пределу при $n\to\infty$, тогда

$$S=\lim_{n\to\infty} S_n.$$

Частная сумма S_n может быть преобразована следующим образом:

$$S_n=\sum_{k=0}^{n} ka^k = a+2a^2+3a^3+\ldots+na^n =$$

$$=(a+a^2+a^3+\ldots+a^n)+(a^2+a^3+\ldots+a^n)+$$

$$+(a^3+a^4+\ldots+a^n)+\ldots+a^n,$$

* Показать сходимость этого ряда можно с помощью признака Даламбера: ряд $\sum_{k=0}^{\infty} a_k$ сходится, если $\lim_{k\to\infty}\frac{a_{k+1}}{a_k}=\rho<1$. В нашем случае $a_k=k\left(\frac{\lambda}{n\nu}\right)^k$, поэтому $\lim_{k\to\infty}\frac{a_{k+1}}{a_k}=\frac{\lambda}{n\nu}$, а это отношение по условию меньше единицы, таким образом ряд сходится.

причем общее число таких слагаемых равно n. Каждая сумма, стоящая в круглых скобках, является суммой членов геометрической прогрессии со знаменателем a. Они отличаются друг от друга только первыми членами и числом членов. Применяя формулу суммы членов геометрической прогрессии к каждой скобке, получаем

$$S_n = \frac{a - a^{n+1}}{1-a} + \frac{a^2 - a^{n+1}}{1-a} + \ldots + \frac{a^n - a^{n+1}}{1-a} =$$

$$= \frac{1}{1-a}\left[(a + a^2 + \ldots + a^n) - na^{n+1}\right].$$

Суммируя еще раз выражение, стоящее в круглых скобках, которое также является суммой членов геометрической прогрессии, получаем

$$S_n = \frac{1}{1-a}\left(\frac{a - a^{n+1}}{1-a} - na^{n+1}\right). \qquad (3.38)$$

Теперь для получения S перейдем к пределу при $n \to \infty$, при этом заметим, что так как $a < 1$, то

$$\lim_{n\to\infty} a^{n+1} = 0 \text{ и } \lim_{n\to\infty} na^{n+1} = 0^*.$$

Тогда

$$S = \lim_{n\to\infty} S_n = \frac{a}{(1-a)^2}.$$

Возвращаясь к M_1 и подставляя $a = \frac{\lambda}{n\nu}$, получаем

$$M_1 = p_n \frac{\lambda}{n\nu\left(1 - \frac{\lambda}{n\nu}\right)^2}. \qquad (3.39)$$

---

* Последнее может быть показано с помощью правила Лопиталя. Действительно, $\lim\limits_{n\to\infty} na^{n+1}$ есть неопределенность типа $\infty \cdot 0$, но $\frac{n}{a^{-(n+1)}}$ при $n \to \infty$ представляет неопределенность типа $\frac{\infty}{\infty}$, к которой можно применить правило Лопиталя. Продифференцировав числитель и знаменатель дроби, получим

$$\lim_{n\to\infty} \frac{n}{a^{-(n+1)}} = \lim_{n\to\infty} \frac{1}{a^{-(n+1)} \ln a} = \lim_{n\to\infty} \frac{a^{n+1}}{\ln a} = 0.$$

Среднее число требований, находящихся в системе обслуживания (как обслуживаемых, так и ожидающих обслуживания), равно

$$M_2 = \sum_{k=1}^{\infty} k p_k .$$

Подставляя p_k из (3.28) и (3.29), получаем

$$M_2 = \sum_{k=1}^{n-1} \frac{k}{k!} \left(\frac{\lambda}{\nu}\right)^k p_0 + \sum_{k=n}^{\infty} k \left(\frac{\lambda}{n\nu}\right)^{k-n} p_n =$$

$$= p_0 \sum_{k=1}^{n-1} \frac{1}{(k-1)!} \left(\frac{\lambda}{\nu}\right)^k + p_n \sum_{k=0}^{\infty} (n+k) \left(\frac{\lambda}{n\nu}\right)^k =$$

$$= p_0 \sum_{k=1}^{n-1} \frac{1}{(k-1)!} \left(\frac{\lambda}{\nu}\right)^k + p_n \left[n \sum_{k=0}^{\infty} \left(\frac{\lambda}{n\nu}\right)^k + \sum_{k=0}^{\infty} k \left(\frac{\lambda}{n\nu}\right)^k \right].$$

Суммы, стоящие в квадратных скобках, могут быть определены следующим образом: первая может быть вычислена как сумма членов геометрической прогрессии, а вторая вычислена выше и равна $\frac{M_1}{p_n}$. Таким образом, среднее число требований, находящихся в системе обслуживания, равно

$$M_2 = M_1 + p_0 \sum_{k=1}^{n-1} \frac{1}{(k-1)!} \left(\frac{\lambda}{\nu}\right)^k + \frac{n p_n}{1 - \frac{\lambda}{n\nu}}. \qquad (3.40)$$

Среднее число аппаратов, свободных от обслуживания, равно

$$M_3 = \sum_{k=0}^{n-1} (n-k) p_k .$$

Подставляя значение p_k из (3.28), получаем

$$M_3 = \sum_{k=0}^{n-1} \frac{n-k}{k!} \left(\frac{\lambda}{\nu}\right)^k p_0 . \qquad (3.41)$$

Выводы. 1. Вероятность того, что занято точно k обслуживающих аппаратов, при условии, что общее число требований, находящихся на обслуживании, не превосходит числа обслуживающих аппаратов,

$$p_k = \frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k p_0 \qquad (1 \leqslant k \leqslant n).$$

Напомним, что λ — среднее число требований, поступающих в обслуживающую систему за единицу времени;

$\frac{1}{\nu}$ — среднее время обслуживания одного требования;

p_0 — вероятность того, что все обслуживающие аппараты свободны;

n — число обслуживающих аппаратов системы.

2. Вероятность того, что в системе находится k требований, в случае, когда их число больше числа обслуживающих аппаратов,

$$p_k = \frac{1}{n!n^{k-n}}\left(\frac{\lambda}{\nu}\right)^k p_0 \qquad (k \geqslant n).$$

3. Вероятность того, что все обслуживающие аппараты свободны,

$$p_0 = \frac{1}{\sum\limits_{k=0}^{n-1}\frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k + \frac{\nu}{(n-1)!(n\nu-\lambda)}\left(\frac{\lambda}{\nu}\right)^n} \qquad \left(\frac{\lambda}{n\nu} < 1\right).$$

4. Вероятность того, что все обслуживающие аппараты заняты,

$$\Pi = \frac{\nu p_0}{(n-1)!(n\nu-\lambda)}\left(\frac{\lambda}{\nu}\right)^n \qquad \left(\frac{\lambda}{n\nu} < 1\right).$$

5. Вероятность того, что время ожидания начала обслуживания, т. е. время пребывания в очереди, больше t,

$$P\{\beta > t\} = \Pi e^{-(n\nu-\lambda)t} \qquad (t \geqslant 0).$$

6. Вероятность того, что время ожидания начала обслуживания (β) меньше t,

$$P\{\beta < t\} = 1 - \Pi e^{-(n\nu-\lambda)t} \qquad (t \geqslant 0).$$

7. Среднее время ожидания начала обслуживания при n обслуживающих аппаратах

$$T_n = \frac{\Pi}{n\nu - \lambda} \qquad \left(\frac{\lambda}{n\nu} < 1\right).$$

8. Средняя длина очереди (среднее число требований, ожидающих начала обслуживания)

$$M_1 = \frac{p_n \lambda}{n\nu\left(1 - \frac{\lambda}{n\nu}\right)^2}.$$

9. Среднее число требований, находящихся в системе обслуживания,

$$M_2 = M_1 + \frac{np_n}{1 - \frac{\lambda}{n\nu}} + p_0 \sum_{k=1}^{n-1} \frac{1}{(k-1)!}\left(\frac{\lambda}{\nu}\right)^k.$$

10. Среднее число свободных обслуживающих аппаратов

$$M_3 = \sum_{k=0}^{n-1} \frac{n-k}{k!}\left(\frac{\lambda}{\nu}\right)^k p_0.$$

Примеры. Пусть мастерская, ремонтирующая телевизоры, имеет трех мастеров. Количество телевизоров, находящихся в эксплуатации, весьма велико. Как известно, в настоящее время надежность их заставляет желать много лучшего. Поэтому число телевизоров, поступивших на ремонт, может быть очень большим. Практически это число можно считать неограниченным. Правда, вероятность того, что в мастерской будет одновременно находиться очень много телевизоров, мала, но потенциальная возможность того, что они поступят в мастерскую, имеется.

Моменты выхода телевизоров из строя — случайны, поэтому поток заказов в мастерскую является случай-

ным. Будем считать этот поток простейшим. Это предположение недалеко от истины. Интуитивно ясно, что этот поток должен быть стационарным, ординарным и без последействия. Пусть среднее число телевизоров, поступающих на ремонт за день, равно $\lambda=8$. Тогда вероятность того, что за t дней поступит на ремонт точно k телевизоров, будет равна

$$V_k(t)=\frac{(8t)^k}{k!}\cdot e^{-8t}.$$

Видно, что с ростом k эта вероятность при $k>8t$ быстро убывает, так как при этом знаменатель $(k!)$ начинает расти быстрее, чем числитель $(8t)^k$. Время ремонта одного телевизора есть случайная величина, так как это время зависит от характера неисправностей, опытности мастера и т. д. Пусть три мастера за день могут отремонтировать шесть телевизоров и время ремонта подчинено показательному закону. Тогда среднее время ремонта одного телевизора равно $^1/_2$ рабочего дня, следовательно, параметр показательного закона $\nu=2$.

Возникает вопрос о том, какое время каждый владелец неисправного телевизора будет ждать окончания ремонта. Это основной вопрос, интересующий владельца. С точки зрения ателье, представляет интерес вопрос о том, какое среднее количество телевизоров будет находиться на ремонте. А с точки зрения треста бытовых предприятий необходимо знать, насколько хорошо мастерская справляется с ремонтом всего потока телевизоров, какое количество их в среднем будет ждать начала ремонта в мастерской и т. д. Ответы на все эти вопросы можно легко найти, если воспользоваться полученными выше результатами. Ясно, что в такой постановке этот пример является частным случаем рассмотренной задачи. Система обслуживания состоит из трех мастеров, работающих в мастерской. Требование на обслуживание не покинет системы до тех пор, пока не будет закончено обслуживание, так как неисправный телевизор не нужен владельцу. Конечно, можно предположить, что, кроме этой мастерской, есть другие, куда владелец телевизора может обратиться, если эта мастерская сильно загружена работой. Но в этом случае можно рассматривать систему обслуживания, состоящую

из всех мастерских, и тогда неисправный телевизор не сможет покинуть ее до окончания ремонта.

Таким образом, данную систему обслуживания мы можем рассматривать как систему без потерь. Прежде чем приступить к вычислению всех интересующих нас показателей, рассмотрим, способна ли эта мастерская справиться с ремонтом всех телевизоров. Напомним, что нами выше был установлен следующий факт: если $\frac{\lambda}{\nu} \geqslant n$, то очередь неограниченно возрастает. Это означает, что мастерская в этом случае не сможет обеспечить ремонта всех телевизоров. Проверим выполнение этого условия для нашего примера. Так как $\lambda = 8$, $\nu = 2$, то $\frac{\lambda}{\nu} = \frac{8}{2} = 4 > n = 3$.

Следовательно, со временем количество неисправных телевизоров будет неограниченно возрастать и мастерская не сможет обеспечить их ремонта.

Этим самым фактически получен ответ на все поставленные выше вопросы. Владельцам неисправных телевизоров придется долго ждать окончания их ремонта. Мастерская будет постоянно загружена работой и, следовательно, мастера никогда не будут простаивать. Обслуживание населения будет организовано очень плохо. Следовательно, его нужно улучшить, т. е. увеличить число мастеров в мастерской. Очевидно, что если их меньше четырех, то картина не изменится, поэтому их число должно быть увеличено по меньшей мере до пяти человек.

Рассмотрим, как будет выглядеть процесс обслуживания в том случае, когда в мастерской работает пять мастеров. Начнем с вычисления вероятности того, что в момент поступления очередного телевизора на ремонт все мастера заняты. Эту величину можно найти по формуле (3.33):

$$\Pi = \frac{2}{4!(2 \cdot 5 - 8)} \cdot \left(\frac{8}{2}\right)^5 \cdot p_0 = \frac{4^5}{4!} p_0,$$

так как

$$n = 5; \ \lambda = 8; \ \nu = 2.$$

С помощью (3.32) вычислим p_0:

$$p_0 = \frac{1}{\sum_{k=0}^{5} \frac{1}{k!} 4^k + \frac{2 \cdot 4^6}{5! \cdot 2}} = 0{,}01299.$$

Следовательно, вероятность того, что в момент поступления очередного телевизора все мастера свободны, равна $p_0 = 0{,}01299$, а вероятность того, что все они заняты, равна

$$\Pi = \frac{4^5}{4!} p_0 \approx 0{,}55424,$$

т. е. заняты все мастера только около половины всего рабочего времени. Закон распределения времени ожидания имеет следующий вид (3.36):

$$P\{\beta > t\} = 0{,}55424 e^{-2t},$$

где β — время ожидания начала ремонта телевизора. Так, например, вероятность того, что в течение рабочего дня ремонт телевизора, поступившего утром, не будет начат, равна

$$P\{\beta > 1\} = 0{,}55424 e^{-2} \approx 0{,}0750,$$

т. е. в среднем не больше 8 телевизоров из 100 будут ждать начала ремонта больше одного рабочего дня. На рис. 10 изображен график этого закона распределения,

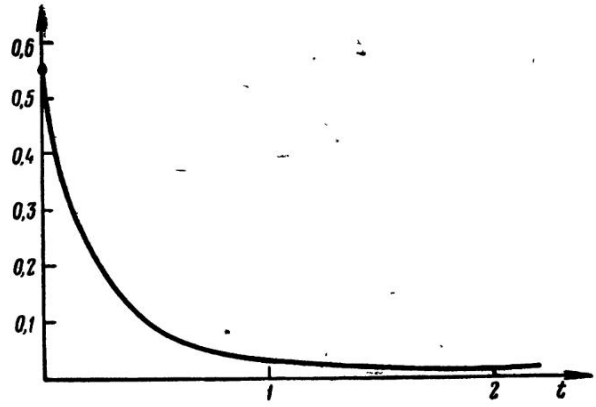

Рис. 10. График распределения времени ожидания начала обслуживания $Y = P\{\beta > t\} = 0{,}55424 \cdot e^{-2t}$ (t — в рабочих днях).

показывающий вероятность того, что время ожидания начала обслуживания будет больше t. Среднее время ожидания начала ремонта согласно (3.37) равно

$$T \approx \frac{0{,}55424}{5 \cdot 2 - 8} = 0{,}27712 \text{ рабочего дня.}$$

Вероятность того, что время ожидания начала ремонта превзойдет среднее время ожидания, равна

$$P\{\beta > T\} \approx 0{,}55424 \cdot e^{-0{,}55424} \approx 0{,}3184.$$

Следовательно, с вероятностью 0,6816 ремонт будет начат через 0,27712 часть рабочего дня. При 7-часовом рабочем дне это составляет чуть меньше двух часов. Точнее, вероятность того, что время ожидания начала ремонта будет не больше 2,0 часа, не меньше 0,68, т. е. из 100 телевизоров к ремонту 68 из них приступят раньше, чем через 2 часа после их поступления в мастерскую. Найдем математическое ожидание длины очереди, т. е. среднее число телевизоров, ожидающих начала ремонта. Согласно (3.39) это число равно

$$M_1 = p_5 \frac{0{,}8}{\left(1 - \frac{8}{10}\right)^2} = p_5 \frac{0{,}8}{0{,}04} = 20 p_5.$$

Величина p_5 есть вероятность того, что на ремонте находится точно пять телевизоров. Найдем ее по формуле (3.28):

$$p_5 = \frac{4^5}{5!} p_0 \approx 0{,}1108.$$

Следовательно, среднее число телевизоров, ожидающих начала ремонта, равно $M_1 = 2{,}216$.

Определим еще среднее число мастеров, свободных от работы. Эту величину можно найти с помощью (3.41):

$$M_3 = \sum_{k=0}^{4} \frac{5-k}{k!} 4^k p_0 \approx 1{,}0000.$$

Следовательно, в среднем каждый мастер свободен 0,2 часть рабочего времени, т. е. около 1,4 часов при 7-часовом рабочем дне.

Все эти вычисления показывают, что пять мастеров хорошо обеспечат ремонт телевизоров, при этом загрузка их будет довольно высокой, но и заказчики при этом не будут терпеть больших неудобств.

Рассмотрим другой пример. В процессе проектирования завода необходимо определить, каким количеством испытательных стендов, предназначенных для контроля качества готовой продукции, он должен обладать. Количество стендов должно быть таким, чтобы время, затрачиваемое готовым изделием на ожидание начала контроля, было мало. Время, которое затрачивается на контроль качества одного изделия, вообще говоря, есть случайная величина. Если завод производит сложные изделия, то в процессе контроля возникнет необходимость, в зависимости от того, как протекает испытание, или увеличить или сократить время на контроль.

Предположим, что поток изделий, поступающих на контроль, является простейшим. Свойства простейшего потока (стационарность, ординарность и отсутствие последействия) не противоречат тем свойствам, которыми обладает реальный поток, состоящий из готовых изделий. Пусть среднее число изделий, изготовляемых за смену, λ, равно 5. Это не означает, что на контроле будет находиться не больше 5 изделий. Так как процесс производства непрерывен, то при малом числе испытательных стендов на контроле может скопиться неограниченное количество изделий, поэтому эту задачу можно отнести к рассмотренному выше типу задач с неограниченным входящим потоком.

Предположим, что время, затрачиваемое на контроль одного изделия, подчинено показательному закону и среднее время контроля одного изделия равно 2 сменам, т. е. $\nu=0{,}5$. Тогда, если ограничиться расчетом, в среднем можно грубо определить необходимое число испытательных стендов. Для этого можно воспользоваться условием, обеспечивающим отсутствие неограниченного возрастания длины очереди $\frac{\lambda}{\nu}<n$. Фактически это условие для данного случая означает, что число испытательных стендов n должно быть не меньше, чем произведение числа изделий за единицу времени на среднее время контроля одного из них, т. е. $\frac{\lambda}{\nu}=5\cdot 2=10$. Следова-

тельно, необходимо иметь больше десяти испытательных стендов.

Однако это число испытательных стендов еще не гарантирует того, что система контроля хорошо справится с возложенными на нее задачами. Может оказаться при этом, что время ожидания начала обслуживания будет очень велико и готовая продукция будет простаивать в ожидании начала контроля. Кроме того, необходимо добиться такого положения, при котором все испытательные стенды достаточно хорошо загружены работой, но при этом имеется резерв, который позволяет производить их наладку и ремонт. Пусть, например, необходимо иметь полную загрузку всех стендов не выше 90%. Это означает, что вероятность того, что все стенды заняты, не должна быть больше 0,9. Будем считать, что она равна 0,9, тогда, используя ранее введенное обозначение, заметим, что $\Pi=0{,}9$. Если попытаться добиться этого вычислениями в среднем, то можно рассуждать так: для того, чтобы проконтролировать все изделия в среднем, достаточно 10 стендов, значит, грубо говоря, если число их увеличить на один, то будет примерно около 10% резерва. Дальнейшие наши рассуждения покажут опасность и несостоятельность таких выводов. Ведь нормальная загрузка стендов не является для нас самоцелью. Кроме нормальной загрузки стендов, необходимо еще и обеспечить «быстрый» контроль готовой продукции. Пусть необходимо добиться такого положения, при котором вероятность того, что готовое изделие ждет начала контроля больше смены, мала, например 0,01. В наших обозначениях это означает, что

$$P\{\beta > 1\} = 0{,}01.$$

Эти два требования вступают в противоречие. С одной стороны, ожидание начала контроля не должно быть большим, а с другой стороны, загрузка стендов не должна быть больше заданной. Разрешение этого противоречия может быть найдено за счет подбора соответствующего числа стендов. Для этой цели воспользуемся ранее выведенным соотношением (3.36) между Π, $P\{\beta > t\}$, n — числом обслуживающих стендов, λ и $\bar{\nu}$. По предположению, число готовых изделий в смену в среднем равно 5, т. е. $\lambda=5$, контроль одного изделия занимает, в среднем, 2 смены, т. е. $\nu=0{,}5$. В уравнении

(3.36) остается одна неизвестная величина n, которую мы и найдем решив это уравнение. Подставим в него все известные величины и получим

$$0{,}01 = 0{,}9e^{-(0{,}5n-5)\cdot 1}.$$

Здесь $t=1$, так как 0,01 есть вероятность того, что за 1 смену контроль готового изделия не будет начат, а время мы измеряем в данном примере в сменах. Решая это уравнение, получаем

$$-0{,}5n + 5 = \ln \frac{1}{90},$$

откуда

$$n = 2(5 + \ln 90) \approx 2 \cdot 9{,}5 = 19.$$

Таким образом, получаем, что для удовлетворения этих двух требований необходимо иметь не менее 19 испытательных стендов. Если бы мы ограничились расчетами в среднем, то ясно, что запроектированные 11 стендов стали бы «узким» местом в производственном процессе и не обеспечили бы своевременного контроля готовой продукции.

Посмотрим, какой вид имеет закон распределения времени ожидания начала обслуживания при $n=19$. Для этого воспользуемся формулой (3.36), откуда

$$P\{\beta > t\} = 0{,}9e^{-4{,}5t}.$$

Результаты вычислений сведем в следующую таблицу:

t в сменах	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
$P\{\beta > t\}$	0,56	0,37	0,23	0,16	0,09	0,06	0,04	0,02	0,015	0,001

Таким образом, получается, что при 19 стендах вероятность того, что очередному готовому изделию приходится ждать начала контроля больше 0,5 смены, равна 0,009, т. е. с вероятностью 0,91 контроль изделия будет начат в течение полсмены после его изготовления и с вероятностью 0,44 почти немедленно, т. е. не больше чем через 1/10 смены. В течение смены после изготовления контроль будет начат почти наверняка (с вероятностью 0,99).

Предположим, что вероятность полной загрузки стендов не может превосходить $\Pi = 0{,}9$ из-за необходимости производить их систематическую проверку и ремонт. Обратимся к рис. 11, на котором представлены законы распределения времени ожидания начала обслуживания при $n = 11$, 15 и 19. Для $n = 11$

$$P\{\beta > t\} = 0{,}9e^{-0{,}5t}.$$

Для $n = 15$

$$P\{\beta > t\} = 0{,}9e^{-2{,}5t}.$$

Таким образом, если имеется только 11 испытательных стендов, то вероятность того, что время ожидания

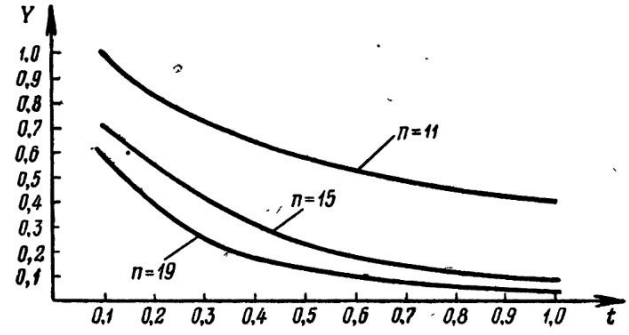

Рис. 11. Графики распределения времени ожидания начала обслуживания $Y = P\{\beta > t\}$ для $n = 11$, $n = 15$ и $n = 19$.

начала обслуживания не превзойдет одной смены, равна 0,45, т. е. в среднем меньше половины готовых изделий за смену попадут на испытательные стенды. Для случая, когда число стендов равно 15, эта вероятность равна 0,93, т. е. картина резко улучшается. В среднем из 100 только 7 изделий будут ждать больше смены, пока освободится испытательный стенд.

Рассмотренный пример показывает, каким образом можно использовать методы теории массового обслуживания при промышленном проектировании в том случае, когда в производственном потоке имеются переходы от одного вида обработки продукции (обслуживания) к дру-

гому и как теория может помочь избежать «узких» мест в некоторых производственных процессах.

Рассмотрим еще один пример. В информационную логическую машину поступает непрерывный поток сообщений. Конструкция машины такова, что она может одновременно обрабатывать только одно сообщение. Если в момент поступления очередного сообщения машина занята обработкой ранее поступившего сообщения, то новое сообщение записывается в буферную память и ждет, пока будет закончена обработка предыдущего. Все последующие сообщения также записываются в буферную память и ждут своей очереди. При этом информация, которая содержится в каждом сообщении, теряет свою ценность через 2 минуты после его получения.

Таким образом, если за две минуты сообщение не поступит в машину, оно может быть отброшено. Возникает вопрос о том, какова вероятность того, что при известной скорости обработки сообщения и определенном их потоке поступившее сообщение не будет своевременно обработано и, следовательно, потеряно. Эта задача является задачей рассмотренного выше типа. Обслуживающим аппаратом в этом случае будет информационная логическая машина. Вся обслуживающая система состоит из одного аппарата, так как машина по условию одновременно может обслуживать только одно требование — обрабатывать одно сообщение. Количество сообщений, записанных в буферной памяти и ожидающих обработки, практически неограниченно. Предположим, что среднее число сообщений в минуту равно 10, а весь поток сообщений является простейшим, т. е. вероятность поступления точно k сообщений за время t равна

$$V_k(t) = \frac{(10t)^k}{k!} e^{-10t} \quad (k = 0, 1, 2, \ldots),$$

где $\lambda = 10$ — математическое ожидание числа сообщений в минуту, а t — время в минутах. Время обработки одного сообщения зависит от ряда факторов, таких, как число знаков в нем, способа его обработки, который может зависеть от характера сообщения, и т. д.

Следовательно, оно является случайной величиной. Предположим, что в среднем за минуту машина обрабатывает 20 сообщений и время обработки как случайная

величина подчинено показательному закону. Эти предположения, как мы уже неоднократно напоминали, нужны нам только для того, чтобы проиллюстрировать на конкретном численном примере способ применения теории. В частности последнее предположение на практике не всегда может быть оправдано. В конкретных случаях нужно изучить фактические законы, которым подчиняются как поток требований, так и время обслуживания и уже для них разрабатывать или выбирать один из существующих методов решения. Возвращаясь к примеру и предполагая, что показательный закон распределения обслуживания требований (обработки поступающих сообщений) имеет место, напомним, что

$$F(t) = 1 - e^{-20t}$$

есть вероятность окончания обработки одного сообщения за время t минут. Здесь $\nu = 20$, так как среднее время обработки одного сообщения $\frac{1}{\nu} = \frac{1}{20}$ минуты.

После всех этих рассуждений для того, чтобы ответить на поставленный вопрос, а именно, определить вероятность того, что сообщение, записанное в буферной памяти, потеряет свою ценность прежде, чем поступит в машину, можно воспользоваться законом распределения времени ожидания. Очевидно, что искомая вероятность p равна вероятности того, что время ожидания начала обслуживания больше двух минут, т. е.

$$p = P\{\beta > 2\}.$$

Воспользуемся формулой (3.36):

$$P\{\beta > t\} = \Pi e^{-(n\nu - \lambda)t} \quad (t \geqslant 0).$$

Чтобы вычислить $P\{\beta > 2\}$, необходимо найти только одну величину Π, так как все остальные величины уже известны: $\lambda = 10$; $\nu = 20$, $n = 1$. Напомним, что Π—это вероятность того, что обслуживающие аппараты заняты. В рассматриваемом примере это вероятность того, что машина занята обработкой сообщения. Эту величину вычислим с помощью формулы (3.33):

$$\Pi = \frac{\nu p_0}{(n-1)!(\nu n - \lambda)} \cdot \left(\frac{\lambda}{\nu}\right)^n,$$

которая справедлива при $\frac{\lambda}{\nu} < n$. В нашем примере это условие выполнено, так как $\frac{\lambda}{\nu} = \frac{10}{20} < n = 1$. Подставляя все величины в эту формулу, получаем

$$\Pi = \frac{20 p_0}{0!(20-10)} \left(\frac{10}{20}\right)^1 = p_0.$$

Совпадение Π с p_0 является следствием соотношения ν, λ и n. Таким образом, для определения Π достаточно вычислить p_0. Это сделать совсем просто, если вспомнить, что, с другой стороны,

$$\Pi = \sum_{k=1}^{\infty} p_k.$$

Так как $\sum_{k=0}^{\infty} p_k = 1$, то отсюда следует, что

$$p_0 = 1 - \sum_{k=1}^{\infty} p_k = 1 - \Pi, \text{ но } \Pi = p_0,$$

поэтому $p_0 = 1 - p_0$, т. е. $p_0 = 0{,}5$. Эта величина p_0 может быть вычислена и по формуле (3.32).

Таким образом, вероятность того, что машина занята обработкой сообщений, равна $\Pi = 0{,}5$. Возвращаясь к основному вопросу примера и подставляя в выражение $P\{\beta > 2\}$ найденное значение Π, получаем

$$p = P\{\beta > 2\} = 0{,}5e^{-(20-10)\cdot 2} = 0{,}5e^{-20} \approx 10^{-9}.$$

Следовательно, практически вероятность того, что сообщение обесценится за счет загрузки машины, ничтожно мала. Машина фактически сможет обработать в приемлемые сроки все поступающие сообщения. Кроме этого, проделанные вычисления позволяют сделать еще некоторые интересные выводы. Так, например, представляет определенный интерес вопрос о том, какую часть времени машина будет свободна от обработки сообщений и, следовательно, может быть использована для других целей. Так как $p_0 = 0{,}5$, то это означает, что половину времени машина свободна от обработки со-

общений. Это очень важный вывод, так как он означает, что возможности машины таковы, что на нее можно возложить или решение более сложной задачи, которая потребует обработки значительно большего количества сообщений или решение сразу двух задач той же сложности. Если же мы занимаемся проектированием вычислительной машины для решения только одной конкретной задачи, то, следовательно, требование к быстродействию машины можно снизить почти в два раза, что до некоторой степени упрощает задачу, стоящую перед конструкторами.

При проектировании информационной логической машины большой интерес будет представлять также вопрос о том, какого объема должна быть буферная память. Ответить на него можно в том случае, когда определены условия, при которых допустимо превышение длины очереди над объемом памяти. Так как выше было определено, что вероятность обесценивания сообщения порядка 10^{-9}, то из этого условия можно получить и соответствующие требования к буферной памяти.

Очевидно, что нет смысла иметь память, обеспечивающую сохранение информации с вероятностью, большей $1-10^{-9}$. Поэтому потребуем, чтобы буферная память с вероятностью $1-10^{-9}$ сохраняла все поступившие сообщения. Расчеты в среднем здесь могут привести к существенным ошибкам. Так, из (3.39) следует, что математическое ожидание длины очереди равно

$$M_1 = p_1 \frac{1}{2\left(1-\frac{1}{2}\right)^2} = 2p_1;$$

так как из (3.28)

$$p_1 = \frac{1}{2} p_0 = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4},$$

то $M_1 = \frac{1}{2}$. Получается, что очередь в среднем не превосходит одного сообщения. Но это не означает, что буферная память из одной ячейки достаточна для хранения сообщений. Чтобы определить ее объем, нужно найти такое число N, для которого вероятность того, что

очередь будет больше N, была не больше 10^{-9}. Эт вероятность равна

$$p_{>N} = \sum_{k=N+1}^{\infty} p_k.$$

Из (3.29)

$$p_k = \left(\frac{1}{2}\right)^k p_0 = \left(\frac{1}{2}\right)^{k+1}, \text{ так как } p_0 = 0{,}5,$$

поэтому

$$p_{>N} = \sum_{k=N+1}^{\infty} \left(\frac{1}{2}\right)^{k+1} = \frac{\left(\frac{1}{2}\right)^{N+2}}{1-\frac{1}{2}} = \left(\frac{1}{2}\right)^{N+1},$$

так как сумма является суммой членов геометрической прогрессии с знаменателем $q = \frac{1}{2}$.

Приравняв $p_{>N} = \left(\frac{1}{2}\right)^{N+1}$ к 10^{-9}, найдем N:

$$(0{,}5)^{N+1} = 10^{-9}.$$

Прологарифмируем это равенство и получим

$$N+1 = \frac{-9}{-\lg 2} \approx \frac{9}{0{,}3010} \approx 30,$$

$$N \approx 29.$$

Таким образом, вероятность того, что в системе будет находиться больше 29 сообщений, равна 10^{-9} и, следовательно, достаточно иметь буферную память на 29 сообщений. Этот пример хорошо иллюстрирует, к чему могут привести рассуждения и расчеты в среднем. Ошибка при таких расчетах может быть весьма значительной.

Кроме того, он показывает, насколько широки области применения теории массового обслуживания.

Задача третьего типа

В задачах первого и второго типов, которые уже были рассмотрены в этом параграфе, существенное значение имело условие неограниченного времени ожидания начала обслуживания. Требование, поступившее

в систему обслуживания, могло его покинуть только тогда, когда обслуживание его было полностью закончено. Это условие является основным отличием задач этого параграфа от задач предыдущего, где требование могло покинуть систему в двух случаях. Во-первых, оно могло покинуть систему после того, как обслуживание его было полностью закончено, во-вторых, оно обязательно покидает систему обслуживания, если в момент его поступления все обслуживающие аппараты заняты обслуживанием ранее поступивших требований.

Очевидно, что кроме этих двух типов задач, практический интерес представляют и задачи промежуточного типа. Это задачи, в которых имеются дополнительные условия, наличие или отсутствие которых определяет, остается ли очередное требование в системе или нет. Так, например, к этому типу относятся задачи, в которых очередное требование остается в системе обслуживания при условии, что общее число требований, уже находящихся в системе, не превосходит определенной величины. Это условие, очевидно, равносильно тому, что очередь из требований, ожидающих обслуживания, не больше определенной длины. Если же длина очереди превосходит эту величину, то требование покидает систему не обслуженным. Эта задача является одной из задач третьей группы. Рассмотрением ее мы и ограничимся.

Постановка задачи. Пусть, как и прежде, имеется некоторая обслуживающая система, состоящая из однотипных обслуживающих аппаратов, каждый из которых может одновременно обслуживать только одно требование. Если в момент поступления требования имеется хотя бы один свободный обслуживающий аппарат, то он немедленно приступает к обслуживанию этого требования. Если же все обслуживающие аппараты заняты, то очередное требование становится в очередь при условии, что в ней стоит меньше m требований. Если в очереди стоит m ранее поступивших требований, то очередное требование покидает систему обслуживания или, что то же самое, обслуживающая система отказывает требованию в обслуживании, если в ней находится $l=n+m$ требований. Из этих l требований n обслуживаются, а m ожидают своей очереди. Очевидно, что при $m=\infty$ эта задача совпадает с рассмотренной выше,

т. е. превращается в задачу с неограниченным ожиданием начала обслуживания, а при $m=0$ совпадает с задачами первой группы, т. е. превращается в задачу обслуживания с потерями.

Будем предполагать, что в систему поступает простейший поток с параметром λ. Время обслуживания одного требования подчинено показательному закону с параметром ν. При изучении такого процесса обслуживания нас в первую очередь интересует такой показатель, как вероятность отказа, т. е. вероятность того, что в системе на обслуживании находится $l=n+m$ требований. Этот показатель определяет, насколько вероятно, что новое требование вообще не будет принято на обслуживание. Кроме того, для требований, принятых на обслуживание, важно знать, какое время они будут находиться в ожидании начала обслуживания. Ясно, что время ожидания начала обслуживания является случайной величиной. Если обозначить ее, как и прежде, через β, то полезно знать закон распределения β, так как он дает полную характеристику времени ожидания начала обслуживания. Кроме того, представляют интерес такие показатели процесса обслуживания, как среднее число требований, ожидающих начала обслуживания, среднее число свободных обслуживающих аппаратов, среднее число требований, находящихся в системе.

Как и в предыдущих случаях, ограничимся отысканием этих показателей для предельного решения. Как и прежде, через $P_k(t)$ обозначим вероятность того, что в момент времени t в системе находится точно k требований. Очевидно, что для рассматриваемой задачи отличны от нуля только $P_k(t)$ при $k=0, 1, 2, \ldots, l$. Для значений $k>l$ $P_k(t)=0$, так как требования, поступившие в момент, когда в системе находится l ранее поступивших требований, покидают систему и она остается в прежнем состоянии. Следовательно, вероятность того, что система находится в состояниях $l+1$, $l+2, \ldots$, т. е. в ней находится точно $l+1$, $l+2, \ldots$ требований, равна нулю.

Получается, что система может находиться не больше, чем в $l+1$ разных состояниях. Будем считать, что функции $P_k(t)$ имеют предел при $t \to \infty$, т. е.

$$\lim_{t\to\infty} P_k(t) = p_k \quad (k=0, 1, 2, \ldots, l).$$

Решение задачи начнем, как и прежде, с отыскания величин p_k, которые определяют вероятность того, что в установившемся процессе система будет находиться в состоянии k.

Решение задачи. Для отыскания p_k применим хорошо известный способ. Составим систему дифференциальных уравнений, связывающих $P_k(t)$, перейдем в ней к пределу при $t \to \infty$ и получим систему алгебраических уравнений, откуда определим величины p_k. Для составления системы дифференциальных уравнений нет необходимости повторять все ранее приведенные рассуждения. Можно воспользоваться результатами, полученными выше. До тех пор, пока число требований, находящихся в системе, меньше n, процесс обслуживания не будет ничем отличаться от того процесса, который имел место в предыдущей задаче. Точно так же характер процесса сохранится и для числа требований $n \leqslant k \leqslant l-1$, так как при этом каждое очередное требование становится в очередь и ждет начала обслуживания.

То, что требования, поступившие после того, как в очередь станут m требований, будут потеряны, не может оказать влияние на течение процесса обслуживания. Поэтому для $k=0$, $1 \leqslant k < n$ и $n \leqslant k < l-1$ можно воспользоваться готовыми дифференциальными уравнениями для $P_k(t)$ из предыдущей задачи. Течение процесса изменится с того момента, когда в системе будет находиться l требований, поэтому для $P_l(t)$ уравнение будет иметь другой вид. Не повторяя всех рассуждений, которые были неоднократно проведены в предыдущих задачах, заметим, что вывод последнего уравнения имеет много общего с выводом последнего уравнения системы (3.7) § 1 гл. 3. Приведем систему дифференциальных уравнений решения поставленной задачи:

$$
\left.
\begin{aligned}
&P_0'(t) = -\lambda P_0(t) + \nu P_1(t),\\
&P_k'(t) = \lambda P_{k-1}(t) - (\lambda + k\nu) P_k(t) + (k+1)\nu P_{k+1}(t)\\
&\qquad\qquad (1 \leqslant k < n),\\
&P_k'(t) = \lambda P_{k-1}(t) - (\lambda + n\nu) P_k(t) + n\nu P_{k+1}(t),\\
&\qquad\qquad (n \leqslant k \leqslant l-1)\\
&P_l'(t) = \lambda P_{l-1}(t) - n\nu P_l(t).
\end{aligned}
\right\}
$$

Эта система из $l+1$ линейных дифференциальных уравнений может быть проинтегрирована и из нее могут быть получены значения $P_k(t)$. Однако, как мы условились выше, целью наших вычислений являются предельные значения p_k, поэтому перейдем к пределу при $t \to \infty$. Как и раньше, ясно, что пределы левых частей должны быть равны нулю. В противном случае $P_k(t)$ неограниченно возрастало бы с ростом t. Но это невозможно, так как $P_k(t) \leqslant 1$, поэтому

$$\lim_{t\to\infty} P_k'(t) = 0 \quad (k = 0, 1, \ldots, l).$$

Следовательно, система при $t \to \infty$ преобразуется к виду

$$\left.\begin{aligned}
0 &= -\lambda p_0 + \nu p_1 \\
0 &= \lambda p_{k-1} - (\lambda + k\nu) p_k + (k+1)\nu p_{k+1} \quad (1 \leqslant k < n) \\
0 &= \lambda p_{k-1} - (\lambda + n\nu) p_k + n\nu p_{k+1} \quad (n \leqslant k \leqslant l-1) \\
0 &= \lambda p_{l-1} - n\nu p_l
\end{aligned}\right\}.$$

Не имеет смысла повторять все рассуждения, которые уже проводились при решении аналогичной системы в предыдущей задаче, каждый читатель сможет их повторить самостоятельно. Поэтому мы ограничимся только тем, что приведем окончательные результаты:

$$p_k = \frac{p_0}{k!}\left(\frac{\lambda}{\nu}\right)^k \text{ при } 1 \leqslant k \leqslant n, \tag{3.42}$$

$$p_k = \frac{p_0}{n!\,n^{k-n}}\left(\frac{\lambda}{\nu}\right)^k \text{ при } n \leqslant k \leqslant l. \tag{3.43}$$

Вероятность отсутствия требований в системе, т. е. вероятность того, что система свободна (p_0), может быть найдена из условия

$$\sum_{k=0}^{l} p_k = 1.$$

Подставляя в это уравнение значение p_k из (3.42) при $1 \leqslant k < n$ и из (3.43) при $n \leqslant k \leqslant l$, получим

$$p_0\left[\sum_{k=0}^{n-1}\frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k+\frac{1}{n!}\left(\frac{\lambda}{\nu}\right)^n\sum_{k=n}^{l}\left(\frac{\lambda}{n\nu}\right)^{k-n}\right]=1.$$

Вторая сумма может быть найдена как сумма $l-n$ членов геометрической прогрессии со знаменателем $\frac{\lambda}{n\nu}$, поэтому

$$p_0\left[\sum_{k=0}^{n-1}\frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k+\frac{1}{n!}\left(\frac{\lambda}{\nu}\right)^n\frac{1-\left(\frac{\lambda}{n\nu}\right)^{m+1}}{1-\frac{\lambda}{n\nu}}\right]=1. \quad (3.44)$$

Из этого уравнения можно найти p_0.

Выражения, полученные для p_k, позволяют вычислить вероятность того, что очередное требование получит отказ в обслуживании. Так как по условию очередному требованию будет отказано в обслуживании тогда, когда в очереди уже стоит m ранее поступивших требований, то, следовательно, вероятность отказа равна вероятности того, что в системе находится $l=n+m$ требований. Поэтому вероятность отказа равна

$$p_l=\frac{p_0}{n!n^{l-n}}\left(\frac{\lambda}{\nu}\right)^l. \quad (3.45)$$

Определим вероятность того, что все обслуживающие аппараты заняты. Это событие может произойти одним из $m+1$ несовместимых способов: или очередь отсутствует, или в очереди стоит одно требование, или в очереди стоят два требования и т. д., или в очереди стоят m требований. Вероятности этих событий соответственно равны p_n, $p_{n+1}, \ldots, p_{n+m}$. Поэтому вероятность того, что все обслуживающие аппараты заняты, равна по теореме сложения вероятностей, сумме вероятностей p_n, $p_{n+1}, \ldots, p_{n+m}$, т. е.

$$\Pi=\sum_{k=n}^{n+m}p_k,$$

где через П, как и прежде, обозначена вероятность того, что в момент поступления очередного требования все

обслуживающие аппараты будут заняты. Преобразуем это выражение, подставив значение p_k из (3.43):

$$\Pi = \sum_{k=n}^{n+m} \frac{p_0}{n! n^{k-n}} \left(\frac{\lambda}{\nu}\right)^k = \frac{p_0}{n!} \left(\frac{\lambda}{\nu}\right)^n \sum_{k=n}^{n+m} \left(\frac{\lambda}{n\nu}\right)^{k-n}.$$

Сумму найдем как сумму членов геометрической прогрессии со знаменателем $\frac{\lambda}{n\nu}$. Так как

$$\frac{p_0}{n!} \left(\frac{\lambda}{\nu}\right)^n = p_n,$$

то

$$\Pi = p_n \frac{1 - \left(\frac{\lambda}{n\nu}\right)^{m+1}}{1 - \frac{\lambda}{n\nu}}. \qquad (3.46)$$

Отсюда p_n можно выразить через Π:

$$p_n = \Pi \frac{1 - \frac{\lambda}{n\nu}}{1 - \left(\frac{\lambda}{n\nu}\right)^{m+1}}.$$

Теперь определим закон распределения времени ожидания начала обслуживания. Время ожидания начала обслуживания является случайной величиной, так как оно зависит от того, сколько длится обслуживание каждого требования, а это величина случайная, зависящая от количества требований, стоящих в очереди, и многих других условий.

Как и прежде, обозначим через β время ожидания начала обслуживания, тогда нашей целью является отыскание функции $P\{\beta > t\}$, т. е. вероятности, с которой время стояния в очереди β превосходит время t. Для определения этой функций применим тот же метод, который был использован в предыдущей задаче при решении аналогичного вопроса. В связи с тем, что рассуждения будут мало отличаться от приведенных там, изложим их менее подробно, ссылаясь на результаты, полученные в преды-

дущей задаче. Рассуждая как раньше, придем к тому, что $P\{\beta > t\}$ можно вычислить как сумму произведений вероятностей p_k на условные вероятности $P_k\{\beta > t\}$, т. е.

$$P\{\beta > t\} = \sum_{k=n}^{l-1} p_k P_k\{\beta > t\}. \qquad (3.47)$$

Напомним, что эта формула была получена на основании формулы полной вероятности (см. сноску на стр. 180). Напомним также, что $P_k\{\beta > t\}$ есть вероятность того, что время ожидания превзойдет время t, при условии, что в момент поступления требования в системе уже находилось k требований. Суммирование производится от n, так как если в системе обслуживания находится требований меньше, чем n, то обслуживание очередного требования будет начато немедленно, т. е. время ожидания равно нулю, поэтому

$$P_k\{\beta > t\} = 0 \quad (k = 0, 1, 2, \ldots, n-1).$$

Суммирование производится до $l-1$, так как если в момент поступления очередного требования в системе находится l требований, то она отказывает в обслуживании всем последующим требованиям, т. е.

$$P_l\{\beta > t\} = 0.$$

Как и прежде, обозначим через $q_s(t)$ вероятность того, что за время t будет закончено обслуживание точно s требований. Тогда, не повторяя рассуждений, приведенных в предыдущей задаче, напомним, что выходящий поток, состоящий из обслуженных требований, является простейшим и среднее число требований, обслуживание которых заканчивается за единицу времени, равно νn. Поэтому вероятность того, что за время t будет закончено обслуживание точно s требований, равна

$$q_s(t) = \frac{(\nu n t)^s}{s!} e^{-\nu n t}.$$

Если в системе в момент поступления очередного требования находится уже k требований и их число больше числа обслуживающих аппаратов n, то в очере-

ди ждут начала обслуживания $k—n$ требований. Для того чтобы наступила очередь обслуживания этого очередного требования, необходимо, чтобы было начато обслуживание всех требований, поступивших раньше, и освободился бы еще один обслуживающий аппарат. Для этого систему должны покинуть $k—n+1$ обслуженное требование. Если этого не произойдет за время t, то время ожидания начала обслуживания β превзойдет величину t. Это равносильно тому, что за время t произойдет одно из следующих событий: или не будет закончено обслуживание ни одного требования, или будет закончено обслуживание только одного требования, или будет закончено обслуживание двух требований, или будет закончено обслуживание трех требований и т. д., или, наконец, будет закончено обслуживание $k—n$ требований. Во всех этих случаях время ожидания начала обслуживания того требования, которое поступило в систему, когда в ней уже было k требований, превзойдет t. Вероятности этих событий соответственно равны

$$q_0(t),\ q_1(t),\ q_2(t),\ \ldots,\ q_{k-n}(t).$$

Все эти события несовместные, так как не может быть, чтобы за это время закончилось обслуживание ровно одного и ровно двух требований одновременно и т. д. С другой стороны, они исчерпывают все возможности, при которых время ожидания превзойдет t. Поэтому, применяя теорему сложения вероятностей, найдем вероятность того, что время ожидания β превосходит t:

$$P_k\{\beta > t\} = \sum_{s=0}^{k-n} q_s(t).$$

Подставляя в это равенство значения $q_s(t)$ $(s=0,\ 1,\ \ldots,\ k-n)$, получим

$$P_k\{\beta > t\} = \sum_{s=0}^{k-n} \frac{(\nu n t)^s}{s!} e^{-\nu n t}.$$

Подставляя это значение $P_k\{\beta > t\}$ в выражение (3.47), получим выражение для закона распределения β:

$$P\{\beta > t\} = \sum_{k=n}^{l-1} p_k \sum_{s=0}^{k-n} \frac{(\nu t)^s}{s!} e^{-\nu t}.$$

Преобразуем это выражение, для чего изменим порядок суммирования и заменим при этом $l - n$ на m:

$$P\{\beta > t\} = e^{-\nu t} \sum_{s=0}^{m-1} \frac{(\nu t)^s}{s!} \sum_{k=s}^{m-1} p_{n+k}.$$

Если теперь в это выражение подставить значение p_{n+k} из (3.43), то получим

$$P\{\beta > t\} = e^{-\nu t} \sum_{s=0}^{m-1} \frac{(\nu t)^s}{s!} \sum_{k=s}^{m-1} \frac{p_0}{n! n^k} \left(\frac{\lambda}{\nu}\right)^{k+n} =$$

$$= e^{-\nu t} \frac{p_0}{n!} \left(\frac{\lambda}{\nu}\right)^n \sum_{s=0}^{m-1} \frac{(\nu t)^s}{s!} \sum_{k=s}^{m-1} \left(\frac{\lambda}{n\nu}\right)^k.$$

Заметим, что $\frac{p_0}{n!}\left(\frac{\lambda}{\nu}\right)^n = p_n$ и просуммируем вторую сумму как сумму членов геометрической прогрессии со знаменателем $\frac{\lambda}{n\nu}$, тогда

$$P\{\beta > t\} = p_n e^{-\nu t} \sum_{s=0}^{m-1} \frac{(\nu t)^s}{s!} \frac{\left(\frac{\lambda}{n\nu}\right)^s - \left(\frac{\lambda}{n\nu}\right)^m}{1 - \frac{\lambda}{n\nu}}.$$

Подставив значение p_n, найденное из (3.46), получим

$$P\{\beta > t\} = \frac{\Pi e^{-\nu t}}{1 - \left(\frac{\lambda}{n\nu}\right)^{m+1}} \sum_{s=0}^{m-1} \frac{(\nu t)^s}{s!} \left[\left(\frac{\lambda}{n\nu}\right)^s - \left(\frac{\lambda}{n\nu}\right)^m\right]. \quad (3.48)$$

Нужно заметить, что если для предыдущей задачи выражение (3.36) имело смысл только для случая, когда

$\frac{\lambda}{n\nu}<1$, то в этой задаче отношение $\frac{\lambda}{n\nu}$ может быть любым.

Все выведенные формулы являются верными независимо от значения этого отношения. Это объясняется тем, что количество возможных состояний системы обслуживания конечно. Поэтому, например при определении $P\{\beta>t\}$, все рассмотренные суммы состояли из конечного числа членов, следовательно, вопрос о сходимости соответствующих рядов не возникал. Напомним, что условие $\frac{\lambda}{n\nu}<1$ в предыдущей задаче обеспечивало ограниченность очереди. В рассматриваемой задаче очередь ограничена величиной m по условию. Как и прежде, для отрицательных значений t вероятность того, что время ожидания β больше t, равна нулю, т. е.

$$P\{\beta>t\}=0 \text{ при } t<0.$$

Следовательно, $t=0$ является точкой разрыва функции

$$P\{\beta>t\}.$$

При приближении к точке t справа имеем

$$P\{\beta>+0\}=\Pi\frac{1-\left(\frac{\lambda}{n\nu}\right)^m}{1-\left(\frac{\lambda}{n\nu}\right)^{m+1}},$$

а при приближении к этой точке слева имеем

$$P\{\beta>-0\}=0,$$

поэтому в точке $t=0$ имеется скачок функции, равный $P\{\beta>+0\}$.

Найдем M_1 — среднее число требований, ожидающих начала обслуживания. Для этого в выражение

$$M_1=\sum_{k=n}^{l}(k-n)\,p_k$$

подставим значения p_k из (3.43), в результате чего получим

$$M_1=\sum_{k=n}(k-n)\frac{p_0}{n!n^{k-n}}\left(\frac{\lambda}{\nu}\right)^k.$$

Вынеся за знак суммы множители, не зависящие от индекса суммирования k, получим

$$M_1 = \frac{p_0}{n!}\left(\frac{\lambda}{\nu}\right)^n \sum_{k=n}^{l} (k-n)\left(\frac{\lambda}{n\nu}\right)^{k-n}.$$

Заменив индекс суммирования на $k-n$ и заметив, что $l-n=m$ и $p_n = \frac{p_0}{n!}\left(\frac{\lambda}{\nu}\right)^n$, получим

$$M_1 = p_n \sum_{k=0}^{m} k\left(\frac{\lambda}{n\nu}\right)^k.$$

Эта сумма ничем не отличается от S_n, которая была вычислена в предыдущей задаче. Поэтому, применив формулу (3.38), получим

$$M_1 = p_n \frac{1}{1-\frac{\lambda}{n\nu}}\left[\frac{\frac{\lambda}{n\nu} - \left(\frac{\lambda}{n\nu}\right)^{m+1}}{1-\frac{\lambda}{n\nu}} - m\left(\frac{\lambda}{n\nu}\right)^{m+1}\right],$$

откуда

$$M_1 = \frac{p_n}{\left(1-\frac{\lambda}{n\nu}\right)^2}\left[\frac{\lambda}{n\nu} - (m+1)\left(\frac{\lambda}{n\nu}\right)^{m+1} + m\left(\frac{\lambda}{n\nu}\right)^{m+2}\right]. \tag{3.49}$$

Среднее число требований, находящихся в обслуживающей системе, равно

$$M_2 = \sum_{k=1}^{l} k p_k.$$

Подставляя в это выражение значения p_k из (3.42) при $k<n$ и из (3.43) при $k \geqslant n$, получим

$$M_2 = \sum_{k=1}^{n-1} \frac{k}{k!}\left(\frac{\lambda}{\nu}\right)^k p_0 + \sum_{k=n}^{l} \frac{k}{n!n^{k-n}}\left(\frac{\lambda}{\nu}\right)^k p_0.$$

Изменяя индекс суммирования во второй сумме на $k-n$, получим

$$M_2 = p_0 \sum_{k=1}^{n-1} \frac{1}{(k-1)!}\left(\frac{\lambda}{\nu}\right)^k + p_n \sum_{k=0}^{m} (n+k)\left(\frac{\lambda}{n\nu}\right)^k =$$

$$= p_0 \sum_{k=1}^{n-1} \frac{1}{(k-1)!}\left(\frac{\lambda}{\nu}\right)^k + np_n \sum_{k=0}^{m} \left(\frac{\lambda}{n\nu}\right)^k + p_n \sum_{k=0}^{m} k\left(\frac{\lambda}{n\nu}\right)^k.$$

Нетрудно заметить, что в этом выражении третья сумма равна M_1, а вторая может быть найдена как сумма членов геометрической прогрессии со знаменателем $\frac{\lambda}{n\nu}$.

Поэтому

$$M_2 = M_1 + \frac{1-\left(\frac{\lambda}{n\nu}\right)^{m+1}}{1-\frac{\lambda}{n\nu}} np_n + p_0 \sum_{k=1}^{n-1} \frac{1}{(k-1)!}\left(\frac{\lambda}{\nu}\right)^k, \tag{3.50}$$

Заметим, что при $m=\infty$ формулы (3.49) и (3.50) превращаются соответственно в формулы (3.39) и (3.40) так как в этом случае эти две задачи тождественны.

Среднее число аппаратов, свободных от обслуживания, равно

$$M_3 = \sum_{k=0}^{n-1} (n-k)\, p_k.$$

Подставляя p_k из (3.42), получим

$$M_3 = \sum_{k=0}^{n-1} \frac{n-k}{k!}\left(\frac{\lambda}{\nu}\right)^k p_0. \tag{3.51}$$

Эта формула совпадает с (3.41), так как для любого числа требований, которое меньше числа обслуживающих аппаратов, эти процессы обслуживания идентичны.

Выводы. 1. Вероятность того, что занято точно k обслуживающих аппаратов, при условии, что общее

число требований, находящихся на обслуживании, не превосходит числа обслуживающих аппаратов

$$p_k = \frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k p_0 \quad (1 \leqslant k \leqslant n).$$

Напомним, что λ — среднее число требований, поступающих в систему за единицу времени;

$\frac{1}{\nu}$ — среднее время обслуживания одним аппаратом одного требования;

p_0 — вероятность того, что все обслуживающие аппараты свободны.

2. Вероятность того, что в системе находится точно k требований в случае, когда их число не меньше числа обслуживающих аппаратов, или, что то же самое, вероятность того, что в очереди стоит точно $k—n$ требований,

$$p_k = \frac{p_0}{n! n^{k-n}}\left(\frac{\lambda}{\nu}\right)^k \quad (n \leqslant k \leqslant l),$$

Напомним, что n — общее число обслуживающих аппаратов, а l — наибольшее возможное число требований, находящихся в системе одновременно.

3. Вероятность того, что все обслуживающие аппараты свободны,

$$p_0 = \frac{1}{\sum\limits_{k=0}^{n-1} \frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k + \frac{1}{n!\left(1-\frac{\lambda}{n\nu}\right)}\left(\frac{\lambda}{\nu}\right)^n \left[1-\left(\frac{\lambda}{n\nu}\right)^{m+1}\right]}.$$

Здесь m — наибольшая допустимая длина очереди, а n — число обслуживающих аппаратов.

4. Вероятность того, что поступившее требование получит отказ, т. е. не будет принято на обслуживание,

$$p_l = \frac{p_0}{n!\, n^{l-n}}\left(\frac{\lambda}{\nu}\right)^l.$$

5. Вероятность того, что все обслуживающие аппараты будут заняты,

$$\Pi = p_n \frac{1-\left(\frac{\lambda}{n\nu}\right)^{m+1}}{1-\frac{\lambda}{n\nu}}.$$

6. Вероятность того, что время ожидания начала обслуживания β будет больше t (закон распределения времени ожидания начала обслуживания)

$$P\{\beta > t\} = \frac{\Pi e^{-\nu n t}}{1-\left(\frac{\lambda}{n\nu}\right)^{m+1}} \sum_{s=0}^{m-1} \frac{(\nu n t)^s}{s!} \left[\left(\frac{\lambda}{n\nu}\right)^s - \left(\frac{\lambda}{n\nu}\right)^m\right].$$

7. Средняя длина очереди (среднее число требований, ожидающих начала обслуживания)

$$M_1 = \frac{p_n}{\left(1-\frac{\lambda}{n\nu}\right)^2}\left[\frac{\lambda}{n\nu} - (m+1)\left(\frac{\lambda}{n\nu}\right)^{m+1} + m\left(\frac{\lambda}{n\nu}\right)^{m+2}\right].$$

8. Среднее число требований, находящихся в системе обслуживания,

$$M_2 = M_1 + \frac{1-\left(\frac{\lambda}{n\nu}\right)^{m+1}}{1-\frac{\lambda}{n\nu}} n p_n + p_0 \sum_{k=1}^{n-1} \frac{1}{(k-1)!}\left(\frac{\lambda}{\nu}\right)^k.$$

9. Среднее число свободных обслуживающих аппаратов

$$M_3 = \sum_{k=0}^{n-1} \frac{n-k}{k!}\left(\frac{\lambda}{\nu}\right)^k p_0.$$

Пример. Рассмотрим пример, иллюстрирующий изученный процесс обслуживания. Автотранспортная контора принимает заявки на срочную доставку грузов. Так как контора обладает вполне определенным количеством автомашин и возможности этих машин с точки зрения

перевозки определенного количества грузов ограниченны, то она не может принимать от заказчиков любое число заявок на доставку этих грузов. Если количество заявок превзошло некоторую определенную величину, то все последующие заказчики получают отказ в обслуживании до тех пор, пока не уменьшится очередь.

Предположим, что поток заявок является простейшим и что в час в среднем поступает одна заявка ($\lambda=1$), т. е. вероятность поступления точно k заявок за время t равна

$$V_k(t)=\frac{t^k}{k!}\,e^{-t}\quad(k=0,1,\ 2,\ \ldots).$$

Время доставки груза, т. е. время обслуживания, зависит от того, где находится груз, куда необходимо его доставить, какой это груз, от времени суток, качества дороги и т. д. Будем предполагать, что время обслуживания подчинено показательному закону и среднее время, которое затрачивается на удовлетворение одной заявки, равно 1 часу, т. е. параметр показательного закона $\nu=1$.

Предположим далее, что контора имеет в своем распоряжении 5 машин, которые работают круглосуточно. Будем считать, что если число принятых и ожидающих удовлетворения заявок стало равным 10, то контора прекращает прием заявок до тех пор, пока не будет обслужена хотя бы одна очередная заявка, т. е. доставка груза, принадлежащего одному заказчику, и, следовательно, до тех пор, пока очередь не уменьшится. Заметим, что число 10 — это максимальная длина очереди, в него не входят 5 заявок, для удовлетворения которых уже выделены машины.

Необходимо определить, какова вероятность того, что все машины заняты, среднюю длину очереди и другие показатели работы конторы. Ясно, что в такой постановке этот пример является частным случаем рассмотренной задачи. Требованием на обслуживание является заявка на доставку груза. Обслуживающим аппаратом является автомашина. Обслуживание заключается в доставке груза. Число обслуживающих аппаратов системы, которой является контора по доставке

грузов, равно пяти. Максимальная длина очереди $m=10$. Наибольшее число заявок, обслуживаемых и ожидающих обслуживания, равно

$$l=m+n=15.$$

Воспользуемся формулами, приведенными выше, и вычислим все интересующие нас показатели. Определим вероятность того, что все машины заняты. Она равна

$$\Pi=p_5\frac{1-(0,2)^{11}}{1-0,2}=\frac{5}{4}[1-(0,2)^{11}]\,p_5,$$

так как

$$\lambda=\nu=1,\quad n=5.$$

Определим величину p_5:

$$p_5=\frac{1}{5!}\left(\frac{1}{1}\right)^5 p_0=\frac{p_0}{5!}.$$

Для определения p_5 необходимо найти p_0:

$$p_0=\frac{1}{\sum\limits_{k=0}^{4}\frac{1}{k!}+\frac{1}{5!0,8}[1-(0,2)^{11}]}=0,5818.$$

Следовательно, вероятность того, что все машины будут свободны, равна 0,5818, т. е. простаивать они будут больше половины рабочего времени. Это означает, что или можно уменьшить их число, или увеличить количество заявок.

Следовательно,

$$p_5=\frac{p_0}{5!}=0,0048.$$

Вероятность того, что все машины заняты, равна

$$\Pi=\frac{5}{4}[1-(0,2)^{11}]\frac{0,5818}{5!}=0,0061,$$

т. е. вероятность полной загруженности очень мала. Эти величины достаточно хорошо характеризуют загрузку

конторы. Качество обслуживания определим средней длиной очереди

$$M_1 = \frac{p_5}{(0,8)^2}[0,2 - 11\,(0,2)^{11} + 10\,(0,2)^{12}] = 0,0015,$$

т. е. практически очередь будет отсутствовать. Следовательно, для выбранных нами значений $n=5$, $\lambda=1$, $\nu=1$ и $m=10$ заказчик почти никогда не получит отказа в обслуживании, но и загрузка машин будет очень маленькой.

На практике, конечно, дело обстоит далеко не так, как в разобранном иллюстративном примере. Однако читатель может самостоятельно, задаваясь фактическими значениями величин n, λ, ν и m, произвести все необходимые расчеты для любой задачи массового обслуживания данного типа.

4. ОБСЛУЖИВАНИЕ В УПОРЯДОЧЕННЫХ СИСТЕМАХ (ЗАДАЧА ПАЛЬМА)

Во всех задачах, которые до сих пор рассматривались, обслуживающие аппараты системы были равноправны: каждое требование на обслуживание, поступившее в систему, могло быть принято любым из свободных аппаратов. Однако не во всех системах обслуживания имеет место такая организация. В ряде систем обслуживание организовано иначе. Обслуживающие аппараты не являются равноправными с точки зрения поступления требования на них. Наиболее простой организацией этого типа является такая, в которой все обслуживающие аппараты пронумерованы. Такая система называется, как уже указывалось выше, *упорядоченной*.

Поступающие требования распределяются между аппаратами согласно их номерам. При этом первым загружается обслуживающий аппарат с номером один; если в момент поступления требования он занят, то загружается аппарат с номером два; если и он занят, то загружается аппарат с номером три, и т. д. В общем случае требование поступает на тот свободный обслуживающий аппарат, который имеет наименьший порядковый номер. Наибольший интерес при изучении таких обслу-

живающих систем представляет вопрос о том, насколько полно будет загружен работой каждый последующий аппарат системы. Если считать, что обслуживание каждого требования будет закончено тем аппаратом, который его начал, то общая схема такого процесса обслуживания будет иметь вид, изображенный на рис. 12.

Степень загрузки каждого обслуживающего аппарата может характеризоваться вероятностью того, что

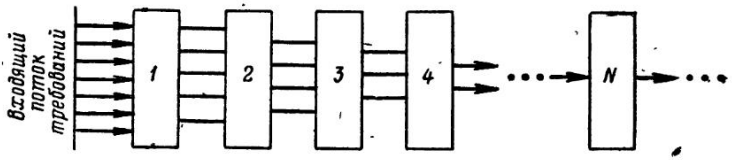

Рис. 12. Схема упорядоченной системы обслуживающих аппаратов
1, 2, 3, 4, . . . , N — обслуживающие аппараты.

требование, поступившее на этот аппарат, найдет его занятым и, следовательно, будет передано следующему аппарату. Простым обобщением этой организации является такая, когда обслуживающие аппараты объединены в группы и все группы пронумерованы. Внутри каждой группы все обслуживающие аппараты равноправны, но все аппараты первой группы имеют преимущество перед аппаратами второй группы и т. д. Во вторую группу требование поступит лишь тогда, когда все аппараты первой группы заняты. В третью группу требование поступит тогда, когда заняты все аппараты как первой, так и второй группы, и т. д. Очевидно, что здесь мы имеем дело с аналогичной задачей, только упорядоченными являются не сами аппараты, а их группы.

Постановка задачи. Имеется упорядоченная система обслуживающих аппаратов, число которых может быть как конечным, так и неограниченным. Обозначим обслуживающие аппараты A_1, A_2, A_3, ..., A_k, ..., где A_k — обслуживающий аппарат с номером k (где $k=1, 2,...$). Очередное требование поступит на обслуживание к аппарату A_k только тогда, когда все предыдущие аппараты A_1, A_2, ..., A_{k-1} заняты обслуживанием ранее поступивших требований. Как и прежде, будем предполагать, что в систему на обслуживание поступает про-

стейший поток требований с параметром λ, т. е. вероятность поступления точно k требований за время t равна

$$V_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad (k = 0, 1, 2, \ldots).$$

Нужно заметить, что это поток требований, поступающих на первый обслуживающий аппарат A_1. Казалось бы, что если на аппарат A_1 поступает простейший поток, то и на аппараты A_2 и A_3 и т. д. будет также поступать простейший поток. Однако это не так. Без доказательства * укажем, что имеет место следующая теорема:

Если на аппарат A_1 поступает простейший поток требований, то на любой аппарат A_k $(k>1)$ поступает стационарный, ординарный поток с ограниченным последействием.

Напомним, что каждый поток такого типа, как было показано в § 1 гл. 2, полностью определяется заданием функции $\varphi_0(t)$, которая определяет вероятность того, что за время t не поступит ни одного требования, при условии, что в начальный момент требование поступило. Каждому обслуживающему аппарату A_k $(k>1)$ будет соответствовать своя функция $\varphi_0(t)$, определяющая поток требований, поступающий на него.

Далее предположим, что среднее время обслуживания одного требования одним аппаратом равно $\frac{1}{\nu}$. При этом закон распределения времени обслуживания может быть любым, лишь бы только среднее время обслуживания было равно $\frac{1}{\nu}$. Пусть перед нами стоит задача — определить степень загрузки обслуживающей системы, которую будем характеризовать вероятностью того, что одновременно заняты n обслуживающих аппаратов.

Решение. Определить вероятность того, что занят первый аппарат A_1, можно без труда, пользуясь ранее полученными результатами. Эта величина может быть получена как частный случай формулы (3.12) § 1 гл. 3. Действительно, рассматриваемый случай является системой обслуживания, состоящей из одного аппарата.

* Читатель при желании может найти доказательство этой теоремы, например, в книге А. Я. Хинчина «Математические методы теории массового обслуживания». Там же читатель найдет полное решение задачи Пальма.

Поэтому, положив в (3.12) $n=1$, получим, что вероятность того, что занят аппарат A_1, которую обозначим через E_1, равна

$$E_1=\frac{\frac{\lambda}{\nu}}{1+\frac{\lambda}{\nu}}=\frac{\lambda}{\nu+\lambda}. \qquad (3.52)$$

Эта величина характеризует долю времени, в течение которой аппарат A_1 работает.

Перейдем к определению вероятности того, что одновременно заняты два аппарата. Для того чтобы определить эту величину, рассмотрим обслуживающую систему, состоящую из двух аппаратов, A_1 и A_2. Система (A_1, A_2) может быть рассмотрена как самостоятельная обслуживающая система. При определении вероятности того, что оба аппарата заняты, тот факт, что на аппарат A_2 требование попадет лишь тогда, когда аппарат A_1 занят, не имеет никакого значения. Следовательно, эта вероятность может быть также вычислена по формуле (3.12) § 1 гл. 3, в которой нужно положить $n=2$. В результате получим, что вероятность того, что заняты два аппарата системы, равна

$$E_2=\frac{\frac{1}{2!}\left(\frac{\lambda}{\nu}\right)^2}{1+\frac{\lambda}{\nu}+\frac{1}{2!}\left(\frac{\lambda}{\nu}\right)^2}=\frac{\lambda^2}{2\nu^2+2\lambda\nu+\lambda^2}=\frac{\lambda^2}{\nu^2+(\lambda+\nu)^2}.$$

Эта величина характеризует долю времени, когда одновременно загружены как аппарат A_1, так и аппарат A_2. Но, кроме этого, величина E_2 есть не что иное, как вероятность того, что очередное требование поступит на аппарат A_3.

В общем случае, рассуждая аналогично, придем к выводу, что вероятность того, что заняты n обслуживающих аппаратов, может быть вычислена с помощью (3.12) § 1 гл. 3 и равна

$$E_n=\frac{\frac{1}{n!}\left(\frac{\lambda}{\nu}\right)^n}{\sum\limits_{k=0}^{n}\frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k} \quad (n=1, 2, \ldots). \qquad (3.53)$$

С одной стороны, эта величина характеризует ту часть времени, на протяжении которой заняты все обслуживающие аппараты. С другой стороны, она равна вероятности того, что очередное требование поступит на A_{n+1} аппарат, если он есть. Независимо от величины отношения $\left(\frac{\lambda}{\nu}\right)$ величина E_n с возрастанием n убывает (рис. 13).

Величины E_n $(n=1, 2,...)$ характеризуют потерю очередного требования на аппаратах $(A_1, A_2,...,A_n)$ в силу

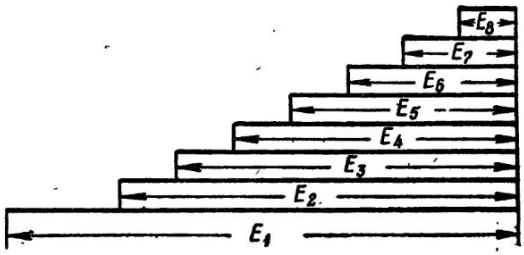

Рис. 13. Характер изменения доли времени одновременной загрузки n аппаратов упорядоченной системы обслуживания $(n=1, 2, 3, \ldots, 8)$.

их занятости. Однако они не дают полной характеристики того, что аппарат A_n занят, так как может оказаться, что аппарат A_n занят, в то время как один из аппаратов $A_1, A_2,....,A_{n-1}$ уже освободился. Характеристикой того, что произвольный аппарат A_n занят, является вероятность потери требования, поступившего на этот аппарат. Требование, поступившее на аппарат A_n $(n=1, 2,...)$, может быть потеряно, т. е. не принято на обслуживание этим аппаратом, только в том случае, если он занят.

Обозначим через Π_n $(n=1, 2,...)$ вероятность того, что аппарат A_n занят. Ясно, что Π_n одновременно является и вероятностью того, что требованию, поступившему на этот аппарат, будет отказано в обслуживании, т. е. оно будет потеряно. Еще раз подчеркнем, что Π_n не совпадает с E_n, так как последнее есть вероятность того, что заняты одновременно аппараты $A_1, A_2, ..., A_n$, в то время как Π_n есть вероятность того, что занят только аппарат A_n.

Найдем величину Π_n. Если заняты аппараты A_1, A_2, ..., A_n, то это означает, что: 1) заняты аппараты A_1, A_2, ..., A_{n-1} и 2) занят аппарат A_n.

Но вероятность первого события есть E_{n-1}, а вероятность второго Π_n, поэтому, по теореме умножения вероятностей, E_n — вероятность того, что заняты A_1, A_2,..., A_n, равна произведению вероятностей E_{n-1} и Π_n, т. е.

$$E_n = E_{n-1}\Pi_n.$$

Так как величины E_n и E_{n-1} нам известны, то можно выразить через них

$$\Pi_n = \frac{E_n}{E_{n-1}}. \tag{3.54}$$

Из (3.53) известно, что

$$E_n = \frac{\frac{1}{n!}\left(\frac{\lambda}{\nu}\right)^n}{\sum_{k=0}^{n}\frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k}.$$

Используя это выражение для E_n, преобразуем его следующим образом. Обозначим

$$\sum_{k=0}^{n}\frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k = S_n \qquad (n>1),$$

тогда

$$E_n = \frac{1}{S_n}\frac{1}{n!}\left(\frac{\lambda}{\nu}\right)^n \text{ и } E_{n-1} = \frac{1}{S_{n-1}}\cdot\frac{1}{(n-1)!}\left(\frac{\lambda}{\nu}\right)^{n-1}.$$

Из (3.54)

$$\frac{1}{\Pi_n} = \frac{E_{n-1}}{E_n}.$$

Подставляя сюда выражения E_n и E_{n-1}, получаем

$$\frac{1}{\Pi_n} = \frac{S_n}{S_{n-1}}\cdot\frac{n!}{(n-1)!}\cdot\frac{\left(\frac{\lambda}{\nu}\right)^{n-1}}{\left(\frac{\lambda}{\nu}\right)^n} = \frac{S_{n-1}+\frac{1}{n!}\left(\frac{\lambda}{\nu}\right)^n}{S_{n-1}}\cdot\frac{n\nu}{\lambda},$$

так как

$$S_n = S_{n-1} + \frac{1}{n!}\left(\frac{\lambda}{\nu}\right)^n.$$

Преобразуя это выражение, получаем

$$\frac{1}{\Pi_n}=\frac{n\nu}{\lambda}\left[1+\frac{\frac{1}{n!}\left(\frac{\lambda}{\nu}\right)^n}{S_{n-1}}\right]=\frac{\nu}{\lambda}\left[n+\frac{\lambda}{\nu}\frac{\frac{1}{(n-1)!}\left(\frac{\lambda}{\nu}\right)^{n-1}}{S_{n-1}}\right].$$

Но так как из (3.53)

$$E_{n-1}=\frac{\frac{1}{(n-1)!}\left(\frac{\lambda}{\nu}\right)^{n-1}}{S_{n-1}},$$

то в предыдущем выражении вместо $\frac{\lambda}{\nu}\frac{\frac{1}{(n-1)!}\left(\frac{\lambda}{\nu}\right)^{n-1}}{S_{n-1}}$ можно записать $\frac{\lambda}{\nu}E_{n-1}$; тогда

$$\frac{1}{\Pi_n}=\frac{\nu}{\lambda}\left[n+\frac{\lambda}{\nu}E_{n-1}\right]$$

и отсюда

$$\Pi=\frac{\lambda}{n\nu+\lambda E_{n-1}} \qquad (n>1). \qquad (3.55)$$

Таким образом, зная E_{n-1}, т. е. вероятность того, что первые $n-1$ аппаратов заняты, мы можем по (3.55) найти вероятность того, что занят аппарат Π_n для любого $n>1$. А для $n=1$ эта вероятность, как отмечалось выше, совпадает с E_1, т. е. $\Pi_1=E_1$.

Выводы. 1. Вероятность того, что одновременно будут заняты n обслуживающих аппаратов, характеризующая долю времени, когда они загружены одновременно, равна

$$E_n=\frac{\frac{1}{n!}\left(\frac{\lambda}{\nu}\right)^n}{\sum_{k=0}^{n}\frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k} \qquad (n=1, 2, \ldots)$$

независимо от того, имеются ли, кроме этих n, еще обслуживающие аппараты или нет.

Напомним, что λ — среднее число требований, поступающих в единицу времени; $\frac{1}{\nu}$ — среднее время обслуживания одного требования.

2. Вероятность того, что очередное требование, поступившее на n-й аппарат, найдет его занятым, т. е. условная вероятность потери требования n-м аппаратом при условии, что оно на него поступило, равна

$$\Pi_n = \frac{\lambda}{n\nu + \lambda E_{n-1}} \qquad (n = 2, 3, \ldots),$$

$$\Pi_1 = E_1.$$

Напомним, что E_{n-1}, вероятность того, что заняты аппараты $(A_1, A_2, \ldots, A_{n-1})$, вычисляется по формулам, приведенным выше.

Пример. Из различных цехов в упаковочный цех поступает готовая продукция. Упаковочные автоматы установлены последовательно и пронумерованы в том порядке, в каком установлены. Первым загружается работой автомат под номером 1. Если он занят, то к обслуживанию очередного требования приступает следующий по номеру автомат, и т. д. Предположим, что среднее время упаковки одного изделия равно 3,24 *сек*; т. е. $\nu = \frac{1}{3{,}24}$. Далее предположим, что поток готовых изделий может быть описан как простейший поток требований, т. е. вероятность поступления ровно k готовых изделий за время t равна

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

где λ — среднее число готовых изделий, поступающих за час, равно 1000. Необходимо определить степень загрузки упаковочной системы. Мы сознательно не указали, какое количество упаковочных аппаратов установлено. Будем считать, что их n штук. Очевидно, что в такой постановке этот пример является частным случаем задачи, рассмотренной выше.

Начнем с определения вероятностей того, что одновременно заняты 1, 2, 3,..., n аппаратов, т. е. вычислим $E_1, E_2, E_3, \ldots, E_n$.

Результаты вычислений сведем в таблицу:

n	1	2	3	4	5	6	7
E_n	0,47368	0,17570	0,05007	0,01115	0,00200	0,00030	0,00000

При вычислениях нужно помнить, что среднее время обслуживания задано в секундах, а среднее число готовых изделий составляет 1000 за час, поэтому, вычисляя

$$E_n = \frac{\frac{1}{n!}\left(\frac{\lambda}{\nu}\right)^n}{\sum_{k=0}^{n}\frac{1}{k!}\left(\frac{\lambda}{\nu}\right)^k} \qquad (n = 1, 2, \ldots),$$

необходимо подставлять в это выражение $\frac{\lambda}{\nu} = \frac{1\,000 \cdot 3 \cdot 24}{3\,600} = 0{,}9$, где $\lambda = \frac{1\,000}{3\,600}$ — среднее число готовых изделий, поступающих за секунду.

Таким образом, видим, что если имеется 6 упаковочных автоматов, то вероятность того, что все они будут загружены одновременно, равна 0,00030, а вероятность того, что одновременно будут загружены 7 автоматов, с точностью до $0{,}5 \cdot 10^{-6}$ равна нулю. Поэтому ясно, что нет смысла устанавливать больше 6 упаковочных автоматов. Возможно, что и 5-й автомат является лишним, так как из 1000 часов работы примерно 2 часа одновременно будут загружены все 5 автоматов. Правда, если 5-й автомат убран, то за это время около 2000 изделий не будет упаковано и их придется вернуть на упаковочную линию.

Решение подобных задач для производственников представляет определенный практический интерес, так как оно позволяет не только выявить нужное количество обслуживающих аппаратов (упаковочных автоматов и т. п.), но и степень загрузки каждого из них как в единицу времени, так и за определенный срок. Это, в свою очередь, позволит до некоторой степени точно предвидеть сроки выхода автоматов из строя и, следовательно, заблаговременно принимать меры по их своевременному ремонту.

Простое обобщение предыдущей задачи. В рассмотренной задаче обслуживающая система состояла из ряда последовательно расположенных обслуживающих аппаратов. Простым обобщением такой системы обслуживания является система, в которой аппараты объединены в группы таким образом, что все аппараты одной

группы равноправны, но аппараты различных групп не равноправны. Группы образуют упорядоченную последовательность, т. е. пронумерованы (или расположены) таким образом, что требование поступает сначала в первую группу.

Если все аппараты этой группы заняты, то требование поступает в следующую группу, если и в ней все аппараты заняты, то в третью, и т. д., т. е. очередное требование всегда поступает на обслуживание в группу с наименьшим номером, где имеется хоть один свобод-

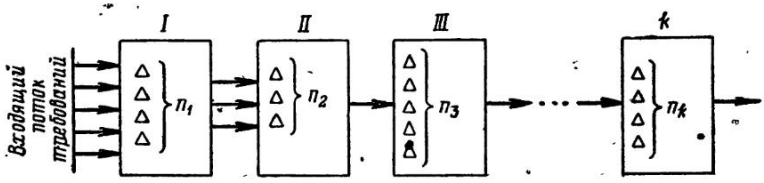

Рис. 14. Схема упорядоченной системы групп обслуживающих аппаратов.

I, II, III, . . . , k — группы обслуживающих аппаратов; Δ — обслуживающие аппараты.

ный обслуживающий аппарат. Предположим, что в эту систему обслуживания поступает простейший поток требований с параметром λ. Еще раз напомним, что для простейшего потока λ есть среднее число требований, поступающих в единицу времени. Далее предположим, что среднее время обслуживания равно $\frac{1}{\nu}$.

Пусть число обслуживающих аппаратов первой группы равно n_1, число обслуживающих аппаратов второй группы n_2, число обслуживающих аппаратов k-й группы n_k и т. д. (рис. 14). Ограничимся определением вероятности того, что одновременно будет занято k групп. Очевидно, это равносильно тому, что очередное требование поступит на обслуживание в $(k+1)$-ю группу и останется в ней при условии, что не все обслуживающие аппараты в ней заняты. Обозначим через E_1 вероятность того, что все обслуживающие аппараты первой группы заняты, через E_2 вероятность того, что заняты как все обслуживающие аппараты первой группы, так и все обслуживающие аппараты второй группы, и т. д. Для получения ответа на поставленный вопрос можно воспользоваться результатами, полученными в § 1 гл. 3. Вероят-

ность того, что все обслуживающие аппараты первой группы заняты, может быть найдена с помощью (3.12). Она равна

$$E_1 = \frac{\frac{1}{n_1!}\left(\frac{\lambda}{\nu}\right)^{n_1}}{\sum_{m=0}^{n_1}\frac{1}{m!}\left(\frac{\lambda}{\nu}\right)^{m}}.$$

Рассуждая аналогичным образом, придем к выводу, что вероятность того, что одновременно заняты 1-я, 2-я,..., k-я группы обслуживающих аппаратов, равна

$$E_k = \frac{\frac{1}{s_k}\left(\frac{\lambda}{\nu}\right)^{s_k}}{\sum_{m=0}^{s_k}\frac{1}{m!}\left(\frac{\lambda}{\nu}\right)^{m}} \quad (k=1, 2\ldots), \qquad (3.56)$$

где $s_k = n_1 + n_2 + n_3 + \ldots + n_k$, т. е. числу обслуживающих аппаратов первых k групп.

Еще раз напомним, что E_k есть вероятность того, что первых k групп недостаточно, чтобы справиться с обслуживанием потока требований и, следовательно, требование попадет в $(k+1)$-ю группу, если она имеется, или будет потеряно, если ее нет.

Рассмотрим отвлеченный пример из области военного дела. Пусть имеется три зоны противоракетной обороны какого-то важного объекта. В первой зоне находятся n_1 противоракетных установок, во второй зоне — n_2 и в третьей зоне — n_3.

Относительно противоракетных установок предполагается, что все они одинаковые и что каждая в данный момент времени может обстреливать не больше одной ракеты. Очевидно, время обстрела не зависит от числа летящих ракет и времени обстрела ракет, летевших до этого. Обстрел внутри данной зоны производится той установкой, которая свободна в данный момент. Пусть поток летящих ракет характеризуется законом Пуассона с параметром потока λ, а время обслуживания, т. е. время, необходимое на обстрел одной ракеты, подчинено

показательному закону и математическое ожидание времени обслуживания равно $\frac{1}{\gamma}$.

Для простоты будем считать, что обстрел ракеты (ее обслуживание) равносилен ее обитию. Летящая ракета должна последовательно преодолеть все три зоны для того, чтобы достигнуть объекта, по которому наносится удар. Очевидно, что те ракеты, которые не были обстреляны в первой зоне, попадут во вторую, а те, которые не будут обстреляны во второй, попадут в третью. Большой интерес представляет вопрос о том, в какой степени система противоракетной обороны способна отразить налет ракет.

Чтобы решить такую задачу, необходимо выбрать критерий для оценки способности системы противоракетной обороны отразить налет. Возьмем в качестве такого критерия вероятность прорыва ракетой каждой зоны.

Эти вероятности вычисляются по формулам (3.56). Вероятности прорыва ракетой первой, второй и третьей зоны соответственно равны E_1, E_2 и E_3.

Для примера пусть в первой зоне находятся три противоракетные установки, во второй две, а в третьей зоне одна, величина $\lambda=4$, а $\frac{1}{\nu}=\frac{1}{2}$. В этом случае вероятность прорыва ракеты через первую зону равна

$$E_1=\frac{\frac{2^3}{3!}}{1+2+\frac{2^2}{2!}+\frac{2^3}{3!}}=0{,}21053.$$

Таким образом, больше чем одну пятую времени все установки первой зоны будут заняты стрельбой и больше чем одна пятая всех ракет преодолеют первую зону, а следовательно, попадут во вторую. Вероятность того, что ракета преодолеет и вторую зону, равна

$$E_2=\frac{\frac{2^5}{5!}}{\sum\limits_{k=0}^{5}\frac{2^k}{k!}}=0{,}03670$$

и вероятность того, что ракета преодолеет все три зоны, равна

$$E_3 = \frac{\frac{2^6}{6!}}{\sum_{k=0}^{6} \frac{2^k}{k!}} = 0{,}00121,$$

т. е. можно ожидать, что через третью зону прорвется меньше двух ракет из тысячи.

Может быть поставлена и иная задача. Например, сколько нужно иметь в зоне противоракетных установок для того, чтобы вероятность прорыва ракеты была меньше заданной величины ε. При известных λ и ν это число s можно найти, решая неравенство

$$\frac{\left(\frac{\lambda}{\nu}\right)^s \frac{1}{s!}}{\sum_{m=0}^{s} \frac{1}{m!}\left(\frac{\lambda}{\nu}\right)^s} < \varepsilon.$$

ЗАКЛЮЧЕНИЕ

Итак, рассмотрение основных типов задач массового обслуживания закончено. Следуя истории развития массового обслуживания, мы сделали основной упор на изучение процессов обслуживания с простейшим входящим потоком и показательным законом распределения времени обслуживания. Во многих задачах массового обслуживания, встречающихся на практике, входящий поток действительно является простейшим и время обслуживания подчинено закону, близкому к показательному. В этом случае могут быть использованы рассмотренные аналитические методы. Однако, как указывалось выше, входящий поток требований может оказаться не только не простейшим, а даже и не стационарным, закон распределения времени обслуживания может быть любым или организация обслуживания может носить сложный многофазовый характер, или поток требований может оказаться неоднородным.

Как быть во всех этих случаях?

Как решать возникающие при этом задачи?

Как уже отмечалось выше, теория массового обслуживания, зародившись относительно недавно, продолжает развиваться. Еще очень многие процессы массового обслуживания не исследованы и соответствующие задачи не решены. Кроме приведенного метода составления и решения системы дифференциальных уравнений, при решении некоторых задач помощь оказывает аппарат интегральных уравнений. Но нужно еще раз подчеркнуть, что аналитические методы решения задач массового обслуживания еще недостаточно сильны для того, чтобы удовлетворить все возникающие потребности.

Однако положение оказывается не настолько безнадежным, как это может показаться на первый взгляд. Имеется возможность уже в настоящее время проана-

лизировать почти любой процесс массового обслуживания, вычислить все необходимые характеристики его независимо от сложности описания входящего потока и закона распределения времени обслуживания (при сложной структуре организации системы). Эту возможность представляет современная электронная вычислительная техника.

Как отмечалось во введении, она, с одной стороны, является мощным стимулом к развитию аппарата теории массового обслуживания, а с другой стороны, дает оружие для практического решения задач массового обслуживания даже в том случае, когда аналитические методы их решения неизвестны. Коротко укажем, каким образом это может быть сделано.

Выше уже отмечалось, что характеристики многих процессов массового обслуживания могут быть получены путем его моделирования на электронных цифровых вычислительных машинах с использованием метода статистических испытаний (метод Монте-Карло). Коротко рассмотрим, как это может быть сделано. При этом будем пользоваться результатами Н. П. Бусленко [1, 2].

Решение задач массового обслуживания методом статистических испытаний, реализуемым на электронных цифровых вычислительных машинах, сводится к построению алгоритма, моделирующего процесс функционирования системы при обслуживании потока требований. Многократная реализация процесса обслуживания с помощью машины при фиксированных условиях задачи с последующей статистической обработкой результатов, полученных при всех реализациях, позволяет найти основные характеристики процесса обслуживания. Очевидно, для того, чтобы иметь возможность это сделать, необходимо соответствующим образом формализовать реальные процессы функционирования системы, для которых строятся моделирующие алгоритмы.

Процесс построения моделирующего алгоритма может быть разбит на ряд этапов в соответствии с теми основными этапами, по которым протекает процесс обслуживания в реальных системах.

Первым этапом анализа любого процесса массового обслуживания является изучение входящего потока, поэтому в первую очередь рассмотрим соотношения, при помощи которых реализуются потоки, формируемые при

моделировании процессов обслуживания. Для того чтобы описать поток однородных требований, достаточно задать закон распределения моментов времени t_1, t_2,..., t_m..., в которые данные требования поступают. Для моделирования удобнее вместо t_1, t_2,..., t_m,... рассматривать случайные величины z_1, z_2,..., z_m,..., определенные следующим образом:

$$t_1 = z_1,$$
$$t_2 = z_1 + z_2,$$
$$\ldots\ldots\ldots$$
$$t_k = z_1 + z_2 + \ldots + z_k.$$

Случайная величина z_i является длиной интервала времени между последовательными моментами t_i и t_{i-1} ($t_0 = 0$; $i = 1, 2 \ldots$).

Покажем, каким образом могут быть получены величины z_i для стационарных ординарных потоков с ограниченным последействием. Так как для потоков с ограниченным последействием величины $z_1, z_2, \ldots, z_n$ — независимы, то совместная функция плотности

$$f(z_1, z_2, \ldots, z_n) = f_1(z_1) \cdot f_2(z_2) \ldots f_n(z_n).$$

Но из формулы (2.12) следует, что

$$f_2(z_2) = f_3(z_3) = \ldots = f_n(z_n) = f(z).$$

Из формулы (2.11) легко получить, что

$$f_1(z) = \lambda \left[1 - \int_0^z f(u)\, du\right].$$

Таким образом, для потоков с ограниченным последействием получение их реализаций на электронных цифровых вычислительных машинах сводится к получению последовательности случайных чисел z_i с законом распределения $f(z)$.

Как правило, на машинах вырабатывается последовательность равномерно распределенных на интервале (0,1) чисел R_i. Это осуществляется либо по специальным алгоритмам, дающим последовательности псевдо-

случайных чисел, либо с помощью специальных приставок — генераторов случайных чисел. Эти последовательности могут быть преобразованы в последовательности с заданным распределением *. Так, случайные числа z_i, определяемые из соотношения

$$\int_{-\infty}^{z_i} f(u)\,du = R_i,$$

имеют закон распределения $f(z)$.

Без доказательства укажем, что из этой формулы следует, что для простейшего потока требований случайные числа z_i могут быть получены следующим образом:

$$z_i = -\frac{1}{\lambda}\ln(1-R_i),$$

где λ — параметр простейшего потока, а R_i — последовательность случайных величин, равномерно распределенных на интервале (0,1).

Некоторые приближенные способы получения случайных чисел z_i читатель найдет в работе [2].

Для потоков с равномерным распределением интервалов z_i (ординарных, стационарных с ограниченным последействием) и функцией плотности

$$f(z) = \frac{1}{b} \quad (0 < z < b)$$

значения случайной величины z_1 можно получить из

$$z_{1j} = b(1 - \sqrt{1-R_j}),$$

а случайные величины $z_i\,(i \neq 1)$ из

$$z_{ij} = bR_{ij},$$

где R_j, как и прежде, случайные числа с равномерным распределением в интервале (0,1).

* Как это делается, читатель может найти в книге А. И. Китова и Н. А. Криницкого «Электронные цифровые машины и программирование». Физматгиз, 1959.

Случайные потоки такого типа легко реализуются на электронных цифровых вычислительных машинах. Они часто используются при решении практических задач.

Не останавливаясь на потоках других видов, которые читатель найдет при желании в [1, 2], перейдем к рассмотрению структуры алгоритма, моделирующего процесс обслуживания требований.

Рассмотрим однофазовую систему массового обслуживания, имеющую n обслуживающих аппаратов, в которую в случайные моменты времени t_i поступают требования.

Если в момент времени t_i есть свободные обслуживающие аппараты, то требование занимает один из них на время $\tau_з$. В противном случае оно находится в системе до момента t_n, ожидая начала обслуживания. Обозначим число аппаратов, освободившихся до t_n, через m. Если $m=0$, то требование покидает систему.

Допустим, что аппараты в процессе работы могут выходить из строя, тогда заявка, обслуживаемая ими, получает отказ, а аппарат на время $\tau_{рем}$ становится на ремонт. Заметим, что t_n, $\tau_з$ и $\tau_{рем}$ могут быть случайными величинами с различными законами распределения.

Приведем укрупненную операторную схему, описывающую N-кратное моделирование процесса функционирования системы в интервале времени (0,1).

Каждый оператор этой схемы, в свою очередь, состоит из ряда более простых операторов, однако такая укрупненная схема позволяет проследить логику процесса моделирования.

$$A_1\ P_2\ A_3\ A_4\ P_5\ P_6\ A_7\ A_8\ P_9\ A_{10}\ A_{11}\ \Phi_{12};$$
$$P_{13}\ \Phi_{14}\ A_{15};\ A_{16}.\ P_{17}\ P_{18}\ A_{19};\ A_{20}\ \Pi_{21}\ Я_{22}.$$

Здесь

A_1 — определение момента t_i поступления очередного требования в систему;

P_2 — проверка выполнения неравенства $t_i < T$. Если неравенство не выполнено, то очередная реализация окончена и управление должно быть передано оператору P_{13};

A_3 — определение времени τ_3, в течение которого аппарат должен обслуживать поступившее требование;

A_4 — определение момента освобождения аппарата, который будет занят обслуживанием этого требования;

P_5 — проверка наличия свободных обслуживающих аппаратов $n_{св} > 0$. Если неравенство не выполняется, то свободных аппаратов нет и требование должно стать в очередь. Управление в этом случае передается оператору A_{16};

P_6 — проверка неравенства $n_{св} > 1$. Если имеется только один свободный аппарат, то управление передается оператору A_8;

A_7 — составление перечня свободных аппаратов и выработка условий для реализации правил распределения свободных аппаратов;

A_8 — выбор одного из свободных аппаратов в соответствии с правилами их распределения;

P_9 — проверка условий исправности работы обслуживающего аппарата. Если аппарат исправен и может обслужить требование, то переход к Φ_{12};

A_{10} — определение времени ремонта обслуживающего аппарата, вышедшего из строя;

A_{11} — счетчик числа отказов в обслуживании. Подсчитываются требования, получившие отказ как в связи с истечением времени ожидания, так и в связи с поломкой обслуживающего аппарата;

Φ_{12} — оператор подготовки алгоритма к моделированию процесса обслуживания следующей заявки, управление передается к началу процесса оператору A_1;

P_{13} — проверка неравенства $N_т < N$, где $N_т$ — текущее число реализаций процесса, а N — заданное число их. Если число осуществленных реализаций еще меньше N, то процесс моделирования продолжается. В противном случае управление передается оператору A_{20};

Φ_{14} — оператор подготовки алгоритма к моделированию процесса обслуживания новой реализации потока требований;

A_{15} — счетчик числа реализаций ($N_т$). Управление передается к началу процесса моделирования оператору A_1;

A_{16} — определение момента потери требования $t_п$, поступившего в момент, когда все обслуживающие аппараты заняты;

P_{17} — проверка возможности обслужить требование, поступившее в момент, когда все аппараты заняты. Проверяется неравенство $m > 0$. Если оно не выполняется, то управление передается оператору A_{11};

P_{18} — проверка условия $m > 1$. Если условие не выполнено, то управление передается оператору A_8;

A_{19} — составление перечня аппаратов, которые могут обслужить данное требование, и выработка условий для реализации правил распределения этих аппаратов;

A_{20} — статистическая обработка результатов моделирования;

$П_{21}$ — выдача результатов;

$Я_{22}$ — окончание процесса моделирования.

Таким образом, в электронной цифровой вычислительной машине может быть промоделирован процесс обслуживания.

Кроме моделирования, машина одновременно выполняет роль наблюдателя и анализатора процесса обслуживания, обрабатывает результаты наблюдений — выдает их в обобщенном виде. Так, в результате моделирования можно получить суммарное число отказов m, число отказов, происшедших по причине занятости аппаратов $m_з$, число отказов, происшедших по причине недостаточной надежности обслуживающих аппаратов. Эти величины позволяют найти такие показатели качества обслуживания, как вероятность отказа

$$p \approx \frac{m}{M},$$

где M — число требований, поступивших в систему, вероятности отказа по причине занятости аппаратов и их ненадежности

$$p_з \approx \frac{m_з}{M} \text{ и } p_н \approx \frac{m_н}{M} \text{ и др.}$$

Если входящий поток не является стационарным, то, очевидно, эти оценки непригодны.

В этом случае можно получить характеристики качества обслуживания, фиксируя реализации, в которых имели место 0, 1, 2, ...,k отказов. Если число таких реализаций соответственно $m_0, m_1, \ldots, m_k$, а общее количество реализаций N, то вероятность отказа

$$p \approx 1 - \frac{m_0}{N}.$$

Закон распределения числа отказов может быть получен из оценки вероятностей p_k:

$$p_k \approx \frac{m_k}{N}.$$

Метод статистического моделирования позволяет решать и более сложные задачи массового обслуживания. Так, например, могут решаться задачи, в которых поток заявок состоит не из однородных событий, процесс обслуживания многофазовый и т. д. Ясно, что изложенный подход позволяет учесть такие особенности. Могут быть учтены и возможные отклонения течения процесса от нормального (помехи).

Ограничения в использовании метода статистического моделирования для решения задач массового обслуживания связаны с невозможностью учесть очень большое количество состояний системы обслуживания, так как объем памяти электронных цифровых вычислительных машин ограничен, и с практической невозможностью получить любую точность результатов, так как быстродействие машин также ограниченно.

К числу достоинств этого метода, по сравнению с аналитическим, нужно отнести сравнительную простоту исследования переходных режимов в процессе обслуживания.

Все это, правда, не означает, что нужно отказаться от аналитических методов анализа процессов массового обслуживания. Наоборот, моделирование с помощью электронных вычислительных машин может подсказать пути отыскания прямых методов решения задач.

Теоретическое исследование, в отличие от математического эксперимента, дает возможность исследовать

вопрос в общем виде, может указать перспективу при отыскании различных, в том числе и оптимальных режимов. При этом затраты труда на отыскание оптимального режима могут оказаться гораздо меньшими, чем при использовании методов статистического моделирования.

Широкое применение находят методы теории массового обслуживания в вопросах оценки надежности различных систем. Повышение надежности обеспечивает огромный экономический эффект: уменьшаются затраты на обслуживание системы, уменьшается количество необходимого оборудования, что обеспечивает экономию капитальных вложений как на создание этого оборудования, так и на постройку помещений для его размещения.

В настоящее время, когда автоматизация решительно вторгается во все области производственной и научной деятельности, значение надежности неизмеримо возрастает. Выход из строя одного элемента автоматизированной линии, как правило, приводит к простою всей линии, а следовательно, к значительным материальным потерям. Поэтому обеспечение надежности работы всех элементов автоматизированных линий является чрезвычайно важной задачей.

Особое значение имеет обеспечение надежности работы автоматизированных систем управления. Выход из строя такой системы срывает работу всего оборудования, которым эта система управляет.

Практический эксперимент по определению степени надежности автоматизированных систем на опытных или промышленных образцах таких систем, как уже отмечалось выше, может оказаться и дорогостоящим, и длительным. Поэтому неоценимой является возможность определить степень надежности системы до ее создания, в процессе проектирования. Ясно, что методы теории массового обслуживания в значительной степени обеспечивают такую возможность. Знание эксплуатационной надежности составных элементов автоматизированных систем или отдельных деталей, блоков, устройств этих элементов позволяет, с использованием методов теории массового обслуживания, оценить ожидаемую надежность элементов и всей системы в целом. Естественно, что надежность отдельной детали, блока, устройства необходимо определять экспериментально. Такая экспе-

риментальная проверка, как правило, проводится и не требует таких затрат и такого времени, как проверка всей системы.

Надежность каждого устройства (детали, блока) может количественно характеризоваться, например, числом отказов (выходов из строя). С точки зрения теории массового обслуживания общее описание надежности может быть дано потоком «требований». «Требованием», в смысле надежности, является выход из строя (поломка). При этом описании может быть учтено влияние резерва для восстановления (замены) устройств, вышедших из строя, на характер потока.

Основной из используемых показателей надежности — вероятность безотказной работы устройства в течение установленного времени — является частной характеристикой потока требований, вероятностью отсутствия требований.

Различные элементы системы будут иметь разную надежность. Поэтому задача определения надежности всей системы будет связана с необходимостью учета неоднородных потоков требований.

В тех случаях, когда число различных устройств системы велико, суммарный поток сбоев (под сбоем понимается выход из строя) по своему характеру близок к простейшему. Это свойство в ряде случаев значительно упрощает расчеты.

Если характеристики надежности всех составных блоков (деталей) известны и задана организация ремонта, которая характеризуется возможностями немедленного восстановления блока, вышедшего из строя (заменой), и распределением времени ремонта, то задача определения надежности всей системы в целом может быть сведена к определению характера выходящего потока. В частности, беря за показатель надежности вероятность безотказной работы, работа всей системы в целом будет характеризоваться вероятностью отсутствия необслуженных требований (не устраненных без прекращения функционирования системы) в выходящем потоке.

Последние годы большое внимание уделяется вопросам проектирования и создания надежных систем из менее надежных элементов. Наиболее остро эта проблема стоит при создании радиоэлектронной аппаратуры или систем на базе элементов электронной вычис-

лительной техники. Основной путь решения этой проблемы — создание структур более устойчивых, чем составные элементы системы. Так, например, дублирование работы отдельных элементов системы, очевидно, повышает надежность всей системы. Правда, при этом увеличивается ее стоимость. Поэтому необходимо искать такое соотношение между затратами на дублирование (резервирование) и надежностью, которое обеспечивало бы наибольший эффект.

Задача определения надежности комплекса из n однотипных, дублирующих друг друга устройств в предположении, что комплекс выходит из строя тогда, когда вышли из строя все n устройств, при определенных ограничениях на поток отказов и время ремонта сводится к задаче Эрланга с ограниченным потоком требований. Ограничения на поток отказов (входящий поток) и время ремонта (время обслуживания) состоят в том, что входящий поток должен быть простейшим, а время обслуживания должно подчиняться показательному закону. Если, например, комплекс состоит из электронных вычислительных машин, то для них согласно экспериментальным данным отказы образуют поток, близкий к простейшему.

Остановимся еще на одном примере использования методов теории массового обслуживания, близком к вопросам теории надежности. Большое значение имеет вопрос обеспечения надежности работы цифровых вычислительных машин, используемых как для решения задач, так и в процессе переработки информации при управлении. Одним из основных методов повышения надежности результатов вычислений является метод двойного счета. Потери машинного времени из-за случайных сбоев существенно зависят от величины интервала времени двойного счета. Неправильный выбор величины этого интервала может привести к очень большим потерям машинного времени. Анализ потока сбоев позволяет количественно обосновать выбор интервала двойного счета.

Интересную задачу по применению методов теории массового обслуживания рассмотрел М. В. Медвидь [30, 31]. В автоматизированных производственных линиях большой интерес представляет задача автоматического питания станков деталями (штучными полуфаб-

рикатами). Часто используется бункерная система питания. Из бункера детали поступают в загрузочное устройство, в котором они ориентируются соответствующим образом и откуда поступают в накопитель. Загрузочное устройство имеет ряд карманов, в которые деталь поступает с некоторой вероятностью в зависимости от ориентации детали в момент прохождения очередного кармана.

Возникает вопрос, какой вместимости необходимо сделать накопитель, чтобы станок без перерывов был обеспечен деталями с достаточно большой вероятностью. Это типичная задача массового обслуживания. Накопитель играет роль обслуживающей системы (обслуживание заключается в приеме очередной детали из загрузочного устройства). Если накопитель не заполнен, то очередная деталь поступает в него, а если заполнен, то очередная деталь «получает отказ», остается в кармане заполнителя. Структура потока требований отличается от рассмотренных выше. Здесь требования могут поступать в определенные моменты времени, когда очередной карман подходит к накопителю, и поступают в накопитель с определенной вероятностью.

Значительный интерес представляет круг задач, в которых обслуживание производится с учетом преимущества одних требований перед другими. Так, например, требования, поступающие в систему, могут различаться по времени, необходимому для их обслуживания. В этом случае может оказаться целесообразной следующая система преимуществ. Обслуживание требования после того, как оно начато, продолжается непрерывно, пока не будет закончено. Из очереди на обслуживание поступает первым требование, для которого длина времени обслуживания минимальна. Такая система преимуществ может обеспечить уменьшение длины очереди. Очевидно, что на практике возможна и такая организация системы преимуществ, при которой требования, поступающие в систему и обладающие специальными признаками, обслуживаются вне очереди. При поступлении такого требования аппарат, обслуживающий требование, не имеющее специального признака, переключается на поступившее требование, а ранее обслуживаемое требование становится в очередь или теряется.

Такая организация может иметь место при передаче

сообщений, когда чрезвычайное сообщение передается вне очереди, а остальные или ждут освобождения одного из каналов связи, или обесцениваются. Аналогичное положение может иметь место при ведении боевых действий. С появлением более важной цели огонь переносится (силы перенацеливаются) на нее. Подобное положение имеет место на аэродроме, когда посадка, в первую очередь, предоставляется самолету с минимальным запасом горючего или имеющему повреждения. Рассмотрение этих задач осталось за пределами книги, но их полезность и важность очевидна.

Нами были рассмотрены примеры только из некоторых областей человеческой деятельности, в которых может оказать помощь теория массового обслуживания, однако проще попытаться отыскать такие области, где она не может быть использована, чем попытаться перечислить те, где она может быть полезной. Несомненно, что у теории большое будущее и она будет весьма интенсивно развиваться.

Если авторам удалось заинтересовать читателя, возбудить у него желание продолжить знакомство с методами теории массового обслуживания, то они могут считать свою задачу выполненной.

Лемма 1. Если функция $f(x) \geqslant 0$ не убывает на отрезке $0 < x < a$ и $f(x+y) \leqslant f(x)+f(y)$ при x, y, $x+y \in (0, a)$, то

$$\lim_{x \to 0} \frac{f(x)}{x} = k,$$

где k — конечное число, отличное от нуля, или $\frac{f(x)}{x} \to +\infty$ при $x \to 0$. Предел равен нулю только при $f(a) = 0$.

Доказательство. Разделим отрезок $(0, x)$ на m равных частей, тогда из

$$f(x+y) \leqslant f(x)+f(y) \text{ при } x, y \text{ и } x+y \in (0, a)$$

вытекает, что

$$f(x) \leqslant m f\left(\frac{x}{m}\right) \text{ при } 0 < x \leqslant a \qquad (1)$$

и любом натуральном m. Положим $x = a$, тогда

$$f(a) \leqslant m f\left(\frac{a}{m}\right).$$

Разделим обе части этого неравенства на a:

$$\frac{f\left(\frac{a}{m}\right)}{\frac{a}{m}} \geqslant \frac{f(a)}{a}. \qquad (2)$$

Отсюда следует, что с возрастанием m отношение $\frac{f\left(\frac{a}{m}\right)}{\frac{a}{m}}$ возрастает; так как при этом $\frac{a}{m} \to 0$, то, следова-

тельно, отношение $\frac{f(x)}{x}$ при $x \to 0$ может стремиться к нулю лишь при условии, что $f(a)=0$. Так как по условию теоремы $f(x)$ не убывает, то это значит $f(x) \equiv 0$ на отрезке $(0, a)$. Пусть $\alpha = \lim\limits_{x \to 0} \frac{f(x)}{x}$ и $\alpha < +\infty$, т. е. точная верхняя грань отношения $\frac{f(x)}{x}$ ограничена. Тогда для всякого положительного наперед заданного ε можно подобрать такое $c > 0$, что

$$\frac{f(c)}{c} > \alpha - \varepsilon \qquad (0 < c \leqslant a).$$

Пусть $0 < x < c$, тогда определим натуральное число m из следующего неравенства:

$$\frac{c}{m} \leqslant x \leqslant \frac{c}{m-1}.$$

Воспользуемся монотонностью $f(x)$. Имеет место следующее неравенство:

$$f(x) \geqslant f\left(\frac{c}{m}\right), \text{ так как } x \geqslant \frac{c}{m}.$$

Неравенство усилится, если левую часть его разделить на x, а правую на $\frac{c}{m-1}$, так как $x \leqslant \frac{c}{m-1}$;

$$\frac{f(x)}{x} \geqslant \frac{f\left(\frac{c}{m}\right)}{\frac{c}{m-1}}.$$

Преобразуем правую часть этого неравенства следующим образом:

$$\frac{f\left(\frac{c}{m}\right)}{\frac{c}{m-1}} = \frac{m-1}{m} \cdot \frac{f\left(\frac{c}{m}\right)}{\frac{c}{m}} \geqslant \frac{m-1}{m} \cdot \frac{f(c)}{c},$$

что следует из (2). Поэтому

$$\frac{f(x)}{x} \geqslant \frac{m-1}{m}\frac{f(c)}{c} > \left(1-\frac{1}{m}\right)(\alpha-\varepsilon).$$

Так как при $m \to \infty$ $x \to 0$, то

$$\lim_{x\to 0}\frac{f(x)}{x} \geqslant \lim_{m\to\infty}\left(1-\frac{1}{m}\right)(\alpha-\varepsilon) = \alpha-\varepsilon.$$

Но $\alpha = \lim\limits_{x\to 0}\frac{f(x)}{x}$, поэтому в силу того, что величину ε мы выбрали произвольно,

$$\lim_{x\to 0}\frac{f(x)}{x} = \alpha < +\infty,$$

т. е. первая часть теоремы доказана.

Предположим, что

$$\lim_{x\to 0}\frac{f(x)}{x} = \alpha \to +\infty.$$

Выберем, как и в предыдущем случае, произвольно большое $A > 0$ и подберем такое c, чтобы

$$\frac{f(c)}{c} > A.$$

Тогда, выбирая m аналогично предыдущему, получаем

$$\frac{f(x)}{x} \geqslant \frac{m-1}{m}A = \left(1-\frac{1}{m}\right)A,$$

поэтому

$$\frac{f(x)}{x} \to \infty \quad \text{при } x \to 0.$$

Таким образом, теорема доказана полностью.

Лемма 2. Для стационарного ординарного потока с ограниченным последействием при любом $m > 1$

$$\frac{\phi_{m+1}(u)}{\phi_m(u)} \to 0 \text{ при } u \to 0,$$

где

$$\psi_m(u) = \sum_{k=m}^{\infty} V_k(u).$$

Доказательство. Обозначим через t_k момент поступления k-го требования и $z_k = t_k - t_{k-1}$ $(k = 1, 2, \ldots)$ промежуток времени между моментами поступления $(k-1)$-го и k-го требований.

Если время поступления $(m+1)$-го требования $t_{m+1} < u$, то это означает, что обязательно $t_m < u$ и $z_{m+1} < u$, т. е. m-е требование поступило раньше, а время от момента его поступления до момента поступления $(m+1)$-го требования и подавно меньше u. Поэтому вероятность неравенства $t_{m+1} < u$ не превосходит вероятности одновременного выполнения двух неравенств $t_m < u$ и $z_{m+1} < u$, т. е.

$$\psi_{m+1}(u) = P\{t_{m+1} < u\} \leqslant P\{t_m < u;\ z_{m+1} < u\}.$$

Но z_{m+1} и t_m независимы в силу отсутствия последействия, поэтому

$$\psi_{m+1}(u) \leqslant P\{t_m < u\}\ P\{z_{m+1} < u\}$$

или

$$\psi_{m+1}(u) \leqslant \psi_m(u)\quad F_{m+1}(u),$$

где $F_{m+1}(u)$ — закон распределения z_{m+1}.

Таким образом, для доказательства теоремы достаточно показать, что $F_{m+1}(u) \to 0$ при $u \to 0$. Если поток требований не „пуст", то можно найти такое большое $a > 0$, что $\psi_m(a) > 0$. Пусть $x > 0$ сколь угодно мало и n таково, что $(n-1)x < a < nx$. Ясно, что такое n всегда можно подобрать при любом фиксированном x, сколь мало бы оно ни было.

Разобьем отрезок $[0, nx]$ на участки длиной x каждый. Если $t_m < nx$ и $z_{m+1} < x$, то моменты t_m и t_{m+1} или лежат на одном участке длины x или на двух смежных. Поэтому хоть на одном отрезке длиной $2x$ будут находиться моменты поступления по меньшей мере двух требований. Но может оказаться, что есть другие участки,

на которых лежат моменты поступления не менее двух требований. Пусть C — событие, заключающееся в том, что $t_m < nx$ и $z_{m+1} < x$, а A_k — событие, состоящее в том, что на отрезке $[(k-1)x, (k+1)x]$ поступило хотя бы два требования. Тогда

$$P\{C\} = P\{A_1\} + P\{A_2\} + \ldots + P\{A_n\}.$$

Но в силу стационарности потока

$$P\{A_1\} = P\{A_2\} = \ldots = P\{A_n\} = \psi_2(2x),$$

поэтому

$$P\{t_m < nx;\ z_{m+1} < x\} \leqslant n\psi_2(2x).$$

Но

$$P\{t_m < nx;\ z_{m+1} < x\} = \psi_m(nx) \cdot F_{m+1}(x),$$

следовательно,

$$F_{m+1}(x) \leqslant \frac{n\psi_2(2x)}{\psi_m(nx)} = \frac{2nx}{\psi_m(nx)} \cdot \frac{\psi_2(2x)}{2x}.$$

Так как

$$2nx < 2(a+x) \text{ и } \psi_m(nx) > \psi_m(a),$$

то

$$F_{m+1}(x) < \frac{2(a+x)}{\psi_m(a)} \cdot \frac{\psi_2(2x)}{2x}.$$

Но $\psi_m(a) > 0$, а $\frac{\psi_2(2x)}{2x}$ в силу ординарности потока при $x \to 0$ также стремится к нулю. Поэтому

$$F_{m+1}(x) \to 0 \text{ при } x \to 0,$$

что доказывает лемму 2.

Теорема. Если $\lim\limits_{t\to\infty} P_{ik}(t) = p_k$ $(0 \leqslant i,\ k \leqslant n)$ существует и не зависит от i, то это необходимое и достаточное условие для того, чтобы независимо от начальных данных

$$\lim_{t\to\infty} P_k(t) = p_k \qquad (k = 0,\ 1,\ 2, \ldots,\ n).$$

Доказательство.

Достаточность. Пусть

$$\lim_{t\to\infty} P_{ik}(t) = p_k \qquad (0 \leqslant i;\ k \leqslant n),$$

тогда, так как

$$P_k(t) = \sum_{i=0}^{n} P_i(0)\, P_{ik}(t),$$

то

$$\lim_{t\to\infty} P_k(t) = \sum_{i=0}^{n} P_i(0) \lim_{t\to\infty} P_{ik}(t) = p_k \sum_{i=0}^{n} P_i(0).$$

Но

$$\sum_{i=0}^{n} P_i(0) = 1,$$

поэтому

$$\lim_{t\to\infty} P_k(t) = p_k,$$

т. е. достаточность доказана.

Необходимость. Пусть $\lim\limits_{t\to\infty} P_k(t) = p_k$, где p_k не зависит от начальных данных. Выберем $P_i(0) = 1$ и $P_k(0) = 0$ при $i \neq k$. Тогда

$$P_k(t) = \sum_{i=0}^{n} P_i(0)\, P_{ik}(t) = P_{ik}(t),$$

поэтому

$$\lim_{t\to\infty} P_{ik}(t) = \lim_{t\to\infty} P_k(t) = p_k,$$

что и требовалось доказать.

ЛИТЕРАТУРА

1. Бусленко Н. П. Решение задач теории массового обслуживания методом моделирования на электронных цифровых вычислительных машинах. В сб. «Проблемы передачи информации». Изд-во АН СССР, 1961, вып. 9.

2. Бусленко Н. П., Шрейдер Ю. А. Метод статистических испытаний. Физматиздат, 1961.

3. Гнеденко Б. В. Курс теории вероятностей. Изд. 2-е. ГТТИ, 1954.

4. Гнеденко Б. В. Несколько замечаний к двум работам Д. И. Баррера. Buletinul Institutului Politechnic din Jasi, T. V (IX) fasc. 1 — 2, 1959, p. 111 — 118.

5. Гнеденко Б. В. Про одне узагальнення формул Ерланга. ДАН УРСР, 1959, № 4.

6. Гнеденко Б. В. К теории предельных теорем для сумм независимых случайных величин. Изв. АН СССР, серия матем., 1939, № 2, стр. 181—232.

7. Гнеденко Б. В. О некоторых задачах теории массового обслуживания. Труды Всесоюзного совещания по теории вероятностей и математической статистике, 1958, Ереван, АН Арм. ССР, 1960.

8. Вентцель Е. С. Обобщение уравнений и формул Эрланга на случай системы массового обслуживания смешанного типа с ограниченным временем ожидания. «Морской сборник», 1961, № 1.

9. Дуб Дж. Л. Вероятностные процессы. Пер. с англ. под ред. А. М. Яглом. Изд-во иностранной литературы, 1956.

10. Зитэк Ф. Заметка к одной теореме Королюка. «Чехословацкий математический журнал», 1957, т. 7 (82), стр. 318—319.

11. Зитэк Ф. К теории ординарных потоков. «Чехословацкий математический журнал», т. 8 (83), 1958.

12. Коваленко Н. Н. Исследование многолинейной системы обслуживания с очередью и ограниченным временем пребывания в системе. «Украинский математический журнал», 1960, т. 12, № 3.

13. Колмогоров А. Н. Sur la probleme d'attente. Математический сборник, 1931, № 38, вып. 1—2, стр. 47—50.

14. Мудров В. И. К вопросу об определении вероятности отказа в однолинейных системах массового обслуживания смешанного типа Сб. «Проблемы кибернетики», вып. 4. Физматиздат, 1960.

15. Ососков Г. А. Одна предельная теорема для потоков

однородных событий. Сб. «Теория вероятностей и ее применения», т. 1, вып. 2, 1956, стр. 274—282.

16. Мудров В. И. Очередь с «нетерпеливыми» клиентами и переменным временем обслуживания, линейно зависящим от времени пребывания клиента в очереди. Сб. «Проблемы кибернетики», вып. 5, Физматиздат, 1960.

17. Севастьянов Б. А. Эргодическая теорема для марковских процессов и ее приложение к телефонным системам с отказами. Сб. «Теория вероятностей и ее применения», 1957, т. 2, вып. 1, стр. 106—116.

18. Феллер В. Введение в теорию вероятностей и ее применения. Пер. с англ. Изд-во иностранной литературы, 1953.

19. Фрай Т. Теория вероятностей для инженеров. Гостехиздат, 1934.

20. Хинчин А. Я. Математические методы теории массового обслуживания. «Труды Матем. ин-та им. В. А. Стеклова», т. 49. Изд. АН СССР, 1955.

21. Хинчин А. Я. Потоки случайных событий без последействия. Сб. «Теория вероятностей и ее применения, 1956, т. 1, вып. 1, стр. 3—18.

22. Хинчин А. Я. О пуассоновских потоках случайных событий. Сб. «Теория вероятностей и ее применения», 1956, т. 1, вып. 3, стр. 320—327.

23. Barrer D. Y. Quening with impatient customers and indiferenf clerks. Operations Research. 1957, v. 5, № ,5 p. 644—649.

24. Barrer D. Y., Quening with impatient customers and ordered service. Oper. Research. 1957, v. 5, № 5, p. 650—656.

25. Kendall D. J. Stochastic processes occurring in the theory of quenes and their analysis by the method of the imbedded Markov chain. The Annals of Mathematical Statistical, 1953, v. 24, № 3.

26. Marcinkiewicz J. Sur les founctions independentes. Fundamental Mathematica, 1938, v. 30.

27. Morse Ph. M. Quenes, investories and maintenance J. Wiley, 1958.

28. Saaty Th. L. Resume of the formulas in quening theory. Oper. Res. 1957, v. 5. № 2.

29. Heathcote C. R. The time-dependent problem foraguеue with preemptive priorities. Oper. Res. 1959, v. 7, № 5.

30. Медвидь М. В. К теории автоматического ориентирования деталей в бункерных устройствах. Научные записки Львовского политехнического института. Автоматизация в машиностроении, 1958, вып. XV.

Медвидь М. В. К расчету основных размеров бункерных загрузочных устройств. Там же.

31. Медвидь М. В. Теоретические основы проектирования вибрационных бункерных загрузочных устройств. Научные записки Львовского политехнического института. Автоматизация в машиностроении. 1959, вып. II.

ОГЛАВЛЕНИЕ

В. Я. Розенберг, А. И. Прохоров

ЧТО ТАКОЕ ТЕОРИЯ МАССОВОГО ОБСЛУЖИВАНИЯ

Редактор *И. М. Волкова*
Техн. редактор *А. А. Свешникова*
Обложка *А. А. Свешникова*

Сдано в набор 23.IX.1961 г.
Подписано к печати 31.I.1962 г.
Формат 84×108/32. Печ. л. 13,12
Уч.-изд. л. 12,73 Тираж 11 000 экз.
Г-84517 Цена в переплете № 5—74 к.
Заказ 579

Типография Госэнергоиздата.
Москва, Шлюзовая наб., 10.

ЗАМЕЧЕННЫЕ ОПЕЧАТКИ

Стр.	Строка	Напечатано	Должно быть
70	5 снизу	$k = 0, 1, 2, \ldots$	$i = 0, 1, 2, \ldots$
70	6 снизу	V_k	V_i
70	7 снизу	V_k	V_i
72	16 снизу	$W(t)$	$W(\Delta t)$
74	12 снизу	$\ldots\left[-\lambda V_0(t) - V_0(t)\frac{o(\Delta t)}{\Delta t}\right]$	$\ldots\left[-\lambda V_0(t) + V_0(t)\frac{o(\Delta t)}{\Delta t}\right]$
87	12 сверху	$m > 1$	$m > 0$
116	14 снизу	эта	этого
120	7 снизу	$+ P_{k+1}(t)(k+1)\Delta t + o(\Delta t)$	$+ P_{k+1}(t)\nu(k+1)\Delta t + o(\Delta t)$
121	10 сверху	$-(\lambda - \nu k) P_k(t) + \nu(k+1) P_{k+1}(t) +$	$-(\lambda + \nu k) P_k(t) + \nu(k+1) P_{k+1}(t) +$
123	2 сверху	$z = 0$	$z_1 = 0$
129	13 сверху	$\frac{24}{36} = 0{,}36\,(6)$	$\frac{24}{36} = 0{,}\,6\,(6)$
144	5 сверху	$p_0 = e^{-\frac{\lambda_1}{\nu}} \ldots$	$p_0 = e^{-\frac{\lambda_2}{\nu}}$
145	19 сверху	$s = e^{-\frac{}{\nu}}$	$s = e^{-\frac{\lambda}{\nu}}$
148	16 снизу	обслуживающих	обслуживаемых
158	13 сверху	$M_2 = \sum_{k=1}^{n} k p_k$	$M_1 = \sum_{k=1}^{m} k p_k$
159	11—12 сверху	—среднее число требований, поступающих в единицу времени,	характеризует частоту возвращения требования в систему,
248	2 снизу	$m > 1$	$m > 0$

Цена 74 коп.