Эта статья была подготовлена в 1999 году для публикации в Соросовском образовательном журнале (СОЖ), которая по техническим причинам не была осуществлена.

Стохастические системы и сети обслуживания

С.Г. Фосс*

«Очереди есть бедствие
нашей эпохи»

Действительно, кто из нас не сталкивался с очередями? Нам часто приходится тратить время на ожидание: в магазине мы стоим в очереди, в автобусе или автомашине ждем зеленого сигнала светофора или (что хуже) попадаем в дорожные пробки; долго пытаемся дозвониться по телефону или войти в сеть ИНТЕРНЕТ... Каждый легко вспомнит массу примеров потери времени на ожидание — еженедельной, ежедневной, даже ежечасной. Можно произвести приближенный подсчет, сколько в среднем за неделю вы простаиваете во всех очередях, с которыми сталкиваетесь. Затем умножить это время на число недель в году — и ужаснуться!

Из-за чего возникают очереди? Во всех случаях схема, по сути, одна и та же: есть некоторое количество «клиентов», каждый из которых желает «обслужиться» в одном и том же месте, и сделать это одновременно они не могут — их интересы вступают в «конфликт». Для того, чтобы этот конфликт разрешать, формируется некоторое правило (алгоритм, дисциплина), упорядочивающее обслуживания клиентов. В одних случаях (например, в случае со светофором) правила задаются заранее, «третьим лицом»; в других они формируются стихийно.

Математическая теория систем обслуживания — область прикладной математики, использующая методы теории вероятностей и математической статистики. Часто ее называют также теорией массового обслуживания, а в англоязычной литературе — теорией очередей (queueing theory). Стимулом к развитию теории систем обслуживания явилось стремление научиться предсказывать случайно изменяющиеся потребности по результатам наблюдений и на основе этого организовывать обслуживание с приемлемым временем ожидания.

1 Введение

Первые математические работы по системам обслуживания появились в начале двадцатого века. Они были тесно связаны с практическими задачами, касавшимися вопросов обслуживания телефонных линий, определения оптимального количества касс и продавцов в торговых предприятиях, выработки правил расчета запасов в магазинах, достаточных для их бесперебойной работы, и других. Среди этих работ особо важное место занимают исследования датского ученого А.К. Эрланга (1878-1929). Благодаря развитию теории вероятностей, к середине двадцатого столетия теория систем обслуживания получила хороший математический фундамент. Среди фамилий ученых, внесших наибольший вклад в теорию очередей, следует назвать такие, как Ф. Поллачек, А.Я. Хинчин, Б.В. Гнеденко, Ж.Ф.С. Кингман, Р.М. Лойнес, С.М. Росс, В.Л. Смит, Г. Коэн, А.А. Боровков, У. Прабху, П. Франкен, В.А. Малышев.

В последние годы произошел бурный всплеск исследований в области теории систем обслуживания. Научные работы, в которых одновременно встречаются слова «очередь» и «случайность», составляют в мире

*Почтовый адрес: Новосибирск, 630090, проспект Академика Коптюга, 4, Институт математики СО РАН. Электронный адрес: foss@math.nsc.ru.

- среди математических статей за 1980-1995 годы — 13 процентов ;
- среди диссертаций за 1980-1995 годы — 24 процента ;
- среди работ, опубликованных в научных и инженерных журналах и сборниках в областях физики, электроники, вычислительных методов и информационных технологий — 60 процентов (данные по индексу INSPEC, разработанному американским и немецким обществами электронной инженерии).

В практике возникают новые и новые задачи, связанные с очередями и требующие математического решения, что способствует появлению новых и развитию известных направлений исследований. Например, с каждым годом компьютерные системы работают все быстрее и быстрее, но их очереди становятся все длиннее и длиннее. Поэтому стали актуальными проблемы, связанные со «взаимодействием» очередей в течение длительных промежутков времени. А это привело к интенсивному развитию направления, получившего название «теория больших уклонений» ("large deviation theory").

Статьи по математической теории систем обслуживания побликуются в десятках различных журналов. С 1986 года издается специализированный математический журнал "Queueing Systems" («Системы очередей»).

В любой системе обслуживания предполагается наличие объектов двух типов: *обслуживающих устройств* (другие названия: *обслуживающие приборы, серверы, каналы* и т. д.) и *клиентов* (другие названия: *заявки, вызовы, требования* и т. д.), нуждающихся в обслуживании. Мы будем использовать термины *сервер* и *вызов*. Правило или алгоритм взаимодействия вызовов и серверов мы будем называть *дисциплиной обслуживания*, или *дисциплиной*. Отметим, что, вообще говоря, вызову может требоваться несколько обслуживаний на одном или нескольких серверах. Обычно термин «*система обслуживания*» (по-английски: "queueing system") употребляется при рассмотрении относительно простых моделей, в которых каждый вызов может иметь только одно обслуживание на некотором сервере. Если же вызовы должны пройти обслуживания на ряде приборов в соответствии с заданными маршрутами, то принято говорить о «*сети обслуживания*» (по английски: "queueing network"). Другими словами, сеть — это просто сложная система.

Число моделей систем (сетей) обслуживания, используемых на практике и изучающихся в теории, очень и очень велико. Даже для того, чтобы схематично описать основные их типы, требуется не один десяток страниц. Поэтому в данной работе мы рассмотрим только три «характерных» вида систем обслуживания: системы с очередью, системы множественного доступа и системы поллинга. При этом будем предполагать, что эти системы являются

- «открытыми» для вызовов, т. е. вызовы поступают в систему извне (в некотором «входном потоке»), каждому из них требуется конечное число обслуживаний, по окончании последнего из которых вызов навсегда покидает систему;

а дисциплины обслуживания таковы, что

- в любой момент времени каждый сервер может обслуживать не более одного вызова (другими словами, не допускается «параллельного» обслуживания двух и более вызовов одним сервером).

Во всех случаях мы обсудим условия, которые гарантируют стабильную работу системы.

Так как процесс поступления и обслуживания вызовов могут зависеть от множества факторов, носящих случайный характер, то они тоже являются случайными, или стохастическими.

2 Сведения из теории вероятностей

В этом разделе мы приводим основные понятия и утверждения теории вероятностей, использующиеся в последующих параграфах.

Для математически строгого определения понятий теории вероятностей требуется хорошо развитый математический аппарат. Поэтому ниже мы предложим неформальное определение случайного события и дискретной случайной величины, принимающей конечное число значений. Затем определим понятия математического ожидания и дисперсии, обсудим свойства независимости и одинаковой распределенности случайных величин и сформулируем две основные теоремы теории вероятностей.

Если событие A при некотором испытании может как произойти, так и не произойти, то мы назовем это событие *случайным*. Случайному событию приписывается некоторое число между 0 и 1, называющееся его вероятностью и обозначающееся через $\mathbf{P}(A)$. Например, если подбрасывается симметричная монета, то событие $A = \{$ выпадает герб $\}$ является случайным, и ему естественно приписать вероятность $\mathbf{P}(A) = \frac{1}{2}$.

События A и B называются *несовместными*, если они не могут происходить одновременно. Набор событий $A_1, A_2, \ldots, A_k$ образует *полную группу*, если, во-первых, любые два из них несовместны и, во-вторых, одно из этих событий обязательно происходит. Из последнего следует, что $\mathbf{P}(A_1) + \mathbf{P}(A_2) + \ldots + \mathbf{P}(A_k) = 1$.

Например, при бросании игрального кубика могут выпасть числа $1, 2, 3, 4, 5, 6$. Введем шесть событий $A_1 = \{$выпадает единица$\}$, $A_2 = \{$выпадает цифра 2$\}$, …, $A_6 = \{$выпадает цифра 6$\}$. Нетрудно видеть, что эти события образуют полную группу. В силу симметрии кубика, естественно положить все вероятности $\mathbf{P}(A_i)$ равными $\frac{1}{6}$. Можно предложить и другие полные группы событий. Например, взять три события B_1, B_2, B_3, где $B_1 = \{$выпадает единица$\}$, $B_2 = \{$выпадает 2, 3 или 4$\}$, $B_3 = \{$выпадает 5 или 6$\}$. При этом, естественно, $\mathbf{P}(B_1) = \frac{1}{6}$; $\mathbf{P}(B_2) = \frac{3}{6} = \frac{1}{2}$ и $\mathbf{P}(B = 3) = \frac{2}{6} = \frac{1}{3}$.

Пусть даны полная группа событий $A_1, A_2, \ldots, A_k$ с вероятностями $p_i = \mathbf{P}(A_i)$, где $i = 1, 2, \ldots, k$ и набор чисел $x_1, x_2, \ldots, x_k$. *Дискретная случайная* величина X — это величина, принимающая значение x_i, если происходит событие A_i (то есть с вероятностью p_i). Соответствие между x_i и p_i выражается формулой $\mathbf{P}(X = x_i) = p_i$, читаемой как: «вероятность того, что X принимает значение x_i, равна p_i». При этом $p_1 + p_2 + p_3 + \ldots + p_k = 1$. Набор пар $\{(x_i, p_i)\}$ называется *законом распределения*, или просто *распределением* величины X. Для любого числа d можно определить вероятность того, что X примет значение, не меньшее чем d:

$$\mathbf{P}(X \geq d) = \sum_{i:\ x_i \geq d} p_i.$$

Аналогично, для любых чисел $c < d$

$$\mathbf{P}(c \leq X \leq d) = \sum_{i:\ c \leq x_i \leq d} p_i.$$

Математическим ожиданием, или *средним* $M(X)$ случайной величины X называется число

$$M(X) = \sum x_i p_i,$$

а *дисперсией* $D(X)$ — число

$$D(X) = M\left(X - M(X)\right)^2,$$

которое можно вычислить по формуле

$$D(X) = \sum x_i^2 p_i - \left(M(X)\right)^2.$$

Среднее квадратическое отклонение (или — неформально – *средний разброс*) случайной величины X есть квадратный корень из дисперсии: $\sigma = \sqrt{D(X)}$.

Например, если X — значение, выпадающее при случайном бросании кубика, то возможных значений только шесть, от 1 до 6, и все шесть вероятностей p_i равны $\frac{1}{6}$. Значит,

$$M(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \ldots + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3,5,$$

$$D(X) = 1 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + \ldots + 36 \cdot \frac{1}{6} - (M(X))^2 = \frac{35}{12} \approx 2,92$$

и $\sigma \approx 1,71$.

Случайная величина X называется *вырожденной*, если она принимает только одно значение (скажем, x) с вероятностью единица (то есть $\mathbf{P}(X = x) = 1$). Для вырожденной случайной величины $M(X) = x \cdot 1 = x$ и $D(X) = x^2 - x^2 = 0$ (значит, и $\sigma = 0$). Можно доказать, что справедливо и обратное утверждение: если $D(X) = 0$, то случайная величина вырождена (для «невырожденных» случайных величин всегда $D(X) > 0$).

Математические ожидания и дисперсии обладают рядом хороших свойств. Например, если c — некоторое число, то случайная величина $Y = X + c$ принимает значения $x_i + c$ с вероятностями p_i и, следовательно,

$$M(Y) = \sum (x_i + c)p_i = \sum x_i p_i + c \sum p_i = M(X) + c.$$

Поэтому

$$D(Y) = M\left(Y - M(Y)\right)^2 = M\left(X + c - M(X) - c\right)^2 = D(X).$$

Если случайная величина X может принимать только целые неотрицательные значения, то $M(X) = \sum_{l=1}^{L} \mathbf{P}(X \geq l)$. Здесь L — максимальное из возможных значений. Действительно, пусть X принимает значение 0 с вероятностью p_0, значение 1 — с вероятностью p_1, и т. д. Тогда

$$M(X) = p_1 + 2p_2 + 3p_3 + \ldots + Lp_L$$

$$= (p_1 + p_2 + p_3 + \ldots + p_L) + (p_2 + p_3 + \ldots + p_L) + \ldots + p_L,$$

где сумма, стоящяя в первой скобке, совпадает с $\mathbf{P}(X \geq 1)$, во второй скобке — с $\mathbf{P}(X \geq 2)$, и т. д.

Несколько случайных величин называются *одинаково распределенными*, если их распределения совпадают (т. е. они принимают одинаковые значения с одинаковыми вероятностями). Как следствие, одинаково распределенные случайные величины имеют одинаковые средние и дисперсии.

Например, если подбросить кубик несколько раз и обозначить через X_n значение, выпадающее при n-ом подбрасывании, то эти случайные величины будут одинаково распределены.

Несколько (скажем, n) случайных величин $X_1, X_2, \ldots, X_n$ называются *независимыми*, если вероятность того, что одновременно первая из них приняла значение c_1, вторая — значение c_2, …, n-ая — значение c_n, совпадает с произведением вероятностей $\mathbf{P}(X_1 = c_1) \cdot$ $\cdot\, \mathbf{P}(X_2 = c_2) \cdot \ldots \cdot \mathbf{P}(X_n = c_n)$ — каковы бы ни были числа $c_1, c_2, \ldots, c_n$. Независимость можно понимать как «отсутствие влияния результатов одних испытаний на другие».

Например, если мы подбросим два одинаковых кубика по разу, то всего различных исходов 36 (каждому варианту выпадения первого кубика может соответствовать 6 вариантов выпадения второго), и все они «равновозможны», т. е. вероятность того, что на первом выпадет число c_1, а на втором – c_2 (где c_1, c_2 — любые целые числа от 1 до 6), равна $\frac{1}{36}$, что совпадает с произведением $\frac{1}{6} \cdot \frac{1}{6}$. Значит, соответствующие случайные величины независимы.

Можно также ввести понятия (дискретной) случайной величины, принимающей бесконечное (счетное) число значений $x_1, x_2, x_3, \ldots$, а также ее математического ожидания

и дисперсии[1]. При этом нужно лишь определить, что означают бесконечные суммы $\sum x_i p_i$ и $\sum x_i^2 p_i$.

Пусть $X_1, X_2, \ldots, X_n$ — набор из независимых, одинаково распределенных случайных величин. Обозначим через a их общее среднее, через σ — средний разброс и через $S_n = X_1 + X_2 + \ldots X_n$– их сумму. Ниже формулируются два основных утверждения теории вероятностей.

Теорема 2.1 (Закон больших чисел, ЗБЧ[2]). *При всех достаточно больших n*

$$\frac{S_n}{n} \approx a.$$

Более точно, для любого как угодно малого числа $\varepsilon > 0$ можно указать достаточно большое число N так, что при каждом целом $n \geq N$ вероятность неравенства $\left\{\left|\frac{S_n}{n} - a\right| > \varepsilon\right\}$ не превосходит ε. Последнее можно записать в виде формулы:

$$\mathbf{P}\left(\left|\frac{S_n}{n} - a\right| > \varepsilon\right) \leq \varepsilon.$$

Например, если будем бросать кубик достаточно много раз и складывать выпадающие значения, то при делении полученной суммы на количество испытаний получим число, очень близкое к $3,5$.

Пусть $\Phi(t)$ — функция вида

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\, dx,$$

где число $e \approx 2,71828$ — основание натуральных логарифмов. Эта функция (называющаяся функцией Лапласа, или функцией распределения нормального закона) играет одну из центральных ролей в теории вероятностей. Она табулирована, ее значения заложены в программы многих калькуляторов и, конечно же, компьютеров. Отметим, что функция $\Phi(t)$

[1] Мы приведем один пример такой случайной величины (который потребуется в следующем параграфе) и посчитаем ее математическое ожидание. Рассмотрим последовательность независимых бросаний монеты, в каждом из которых герб выпадает с вероятностью p, а решетка — с вероятностью $q = 1-p$. Будем считать, что $0 < p < 1$. Пусть Y — случайная величина, означающая номер испытания, при котором герб выпадает в первый раз. Найдем вероятность $p_l = \mathbf{P}(Y = l)$ при всех $l = 1, 2, \ldots$. Ясно, что $p_1 = p$. Далее, событие $\{Y = 2\}$ означает, что одновременно при первом бросании выпадает решетка и при втором — герб. И так как испытания независимы, то $p_2 = qp$. Совершенно аналогично, при каждом $l > 2$ событие $\{Y = l\}$ означает, что при каждом из первых $(l-1)$ бросаний выпадает решетка, а при l-ом бросании — герб. И из независимости испытаний следует, что $p_l = q^{l-1}p$. Значит, случайная величина Y принимает значение $x_1 = 1$ с вероятностью $p_1 = p$, значение $x_2 = 2$ с вероятностью p_2 и так далее. Из школьного курса вы знаете, что сумма геометрической прогрессии $1 + q + q^2 + \ldots$ равна $\frac{1}{1-q} = \frac{1}{p}$. Поэтому

$$p_1 + p_2 + p_3 + \ldots = p(1 + q + q^2 + \ldots) = p \cdot \frac{1}{p} = 1.$$

Так как случайная величина Y принимает лишь неотрицательные целые значения, то естественно определить ее среднее равенством $M(Y) = \sum \mathbf{P}(Y \geq l)$, где сумма берется по всем натуральным l. И так как

$$\mathbf{P}(Y \geq l) = p_l + p_{l+1} + p_{l+2} + \ldots = pq^{l-1}(1 + q + q^2 + \ldots) = pq^{l-1}\frac{1}{1-q} = q^{l-1},$$

то $M(Y) = 1 + q + q^2 + \ldots = \frac{1}{p}$. Если, в частности, $p = \frac{1}{2}$ (монета симметрична), то равенство $M(Y) = 2$ означает, что нам «в среднем» потребуется два бросания до выпадения герба. Если $p = \frac{1}{3}$, то потребуется «в среднем» три бросания, и т. д.

[2] Понятие «закон больших чисел» хорошо известно не только математикам — его можно найти, скажем, и в философском словаре, где оно трактуется так: «с ростом числа испытаний (наблюдений) эмпирическое среднее сближается с математическим средним». Физики говорят проще и выразительнее: «среднее по времени стремится к среднему по пространству».

является монотонно возрастающей, и ее значения при всех t строго положительны и меньше единицы. Она очень медленно растет при $t < -3$ (в частности, $\Phi(-3) \approx 0,00135$), затем достаточно быстро возрастает при $-3 < t < 3$ (в частности, $\Phi(3) = 1 - \Phi(-3) \approx 0,99865$) и потом снова растет очень медленно.

Теорема 2.2 (Центральная предельная теорема, ЦПТ). *Пусть $c < d$ — два произвольных вещественных числа. Предположим, что $\sigma > 0$. Тогда при всех достаточно больших n*

$$\mathbf{P}\left(c \le \frac{S_n - na}{\sigma\sqrt{n}} \le d\right) \approx \Phi(d) - \Phi(c). \tag{2.1}$$

Более точно, для любого как угодно малого числа $\varepsilon > 0$ можно указать достаточно большой номер N так, что при каждом $n \ge N$ абсолютное значение разности левой и правой частей приближенного равенства (2.1) будет не больше, чем ε.

Часто центральную предельную теорему называют также «принципом инвариантности»[3]. Действительно, каково бы ни было «индивидуальное» распределение случайных величин X_i, с ростом n исчезает влияние этой «индивидуальности» на «поведение» случайной величины $Y_n = \dfrac{S_n - na}{\sigma\sqrt{n}}$ — оно достаточно точно задается функцией Лапласа.

В примере с кубиками, если взять $c = -3$ и $d = 3$, то получим, что при достаточно большом числе n испытаний вероятность того, что отношение $\dfrac{S_n - na}{\sigma\sqrt{n}}$ содержится в интервале $(-3, 3)$. будет не меньше, чем $1 - 2 \cdot 0,00135 = 0,9873$. Если мы разрешим двойное неравенство

$$-3 < \frac{S_n - na}{\sigma\sqrt{n}} < 3,$$

относительно S_n и вспомним найденные ранее значения a и σ, то получим, что S_n лежит в диапазоне

$$3,5n - 5,13\sqrt{n} < S_n < 3,5n + 5,13\sqrt{n}$$

с вероятностью, очень близкой к единице.

3 Системы и сети очередей

В этом параграфе мы обсудим несколько примеров систем и сетей очередей.

Начнем с рассмотрения так называемой **одноканальной системы обслуживания**. В системе находится один прибор (сервер). Вызовы поступают в систему группами случайного объема через единичные интервалы времени (пусть $X_n \ge 0$ означает количество вызовов, поступающих в момент времени $n = 1, 2, \ldots$; все возможные значения X_n являются целыми числами). Все вызовы нумеруются и обслуживаются в порядке поступления (внутри каждой группы вызовы нумеруются в произвольном порядке). Как только сервер заканчивает обслуживание некоторого вызова, он немедленно начинает обслуживать следующий (если вызовы в системе есть), а обслуженный вызов покидает систему. Сервер может простаивать только тогда, когда в системе нет вызовов.

Предположим, простоты ради, что все вызовы имеют одно и то же неслучайное время обслуживания $b > 0$. Будем считать, что случайные величины X_n независимы и одинаково распределены. Следовательно, они имеют одно и то же математическое ожидание (среднее), которое обозначим через $a = \mathbf{E}X_n > 0$. В начальный момент времени $n = 0$ вызовы в системе отсутствуют.

[3] На житейский язык термин «принцип инвариантности» можно перевести как «с чего бы мы ни начали, всегда придем к одному и тому же». Один из известнейших «принципов инвариантности» звучит так: «все дороги ведут в Рим».

Обозначим через W_n количество «работы», скопившейся в системе к моменту времени n (то есть суммарное время, нужное для обслуживания имеющихся в наличии вызовов). Ясно, что $W_1 = bX_1$ и при всех $n = 2, 3, \ldots$ справедливы рекуррентные соотношения

$$W_n = bX_n + \max(0, W_{n-1} - 1),$$

где $\max(x, y)$ означает наибольшее из чисел x, y. Величины W_n являются, вообще говоря, случайными, то есть могут принимать различные значения с положительными вероятностями.

Естественно считать, что система работает «стабильно», если с ростом времени n распределения случайных величин W_n остаются ограниченными — в некотором смысле, но в каком? Можно, к примеру, назвать систему стабильной, если найдется некоторое число K, при котором неравенство $W_n \le K$ выполняется всегда (т. е. при каждом n с вероятностью единица). Но для выполнения этого условия необходимо, чтобы с вероятностью единица имело место неравенство

$$bX_1 \le 1. \tag{3.2}$$

Действительно, если это не так и $\mathbf{P}(bX_1 > 1) > 0$, то найдется целое число $l > 0$ такое, что $bl > 1$ и $\mathbf{P}(X_1 = l) > 0$. Обозначим эту вероятность через p, а разность $bl - 1$ через δ. Тогда при каждом n вероятность одновременного наступления событий $\{X_1 = l\}, \{X_2 = l\}, \ldots, \{X_n = l\}$ равняется произведению вероятностей (в силу независимости случайных величин), т. е. просто p^n. Но при этом выполняется равенство $W_n = n\delta$. Для любого числа K можно выбрать номер n так, чтобы $n\delta > K$. И при таком n вероятность того, что значение W_n превосходит K, является положительной — значит, система не может быть стабильной в предложенном выше смысле.

Отметим, что предположение (3.2) является сильно ограничительным и малореалистичным, так как на практике возможны резкие (хотя и редкие) колебания входного потока во времени (так называемые «пиковые нагрузки»), т. е. величины X_n могут принимать большие значения — но с маленькими вероятностями. Поэтому представляется разумным следующее (более «широкое») определение стабильности.

Система обслуживания является *стабильной*, если для любого (как угодно малого) числа ε найдется такое число K, что при всех n

$$\mathbf{P}(W_n \le K) \ge 1 - \varepsilon,$$

и *нестабильной* — в противном случае. Это определение остается таким же и для любой другой системы обслуживания, если под W_n понимать (более общо) количество работы, скопившейся к моменту n на всех серверах системы.

Оказывается, что для того, чтобы определить, является одноканальная система стабильной или нет, не требуется знания распределения случайных величин X_n — достаточно лишь знать значение их среднего a. Более точно, имеет место следующая теорема

Теорема 3.1. *Одноканальная система обслуживания является стабильной при $ab < 1$ и нестабильной при $ab > 1$. В случае $ab = 1$ система является стабильной тогда и только тогда, когда величины X_n являются вырожденными, т. е. $\mathbf{P}(X_n = a) = 1$.*

Поясним, как доказывается эта теорема. Начнем со случая $ab = 1$. Если все X_n вырождены (и совпадают с a), то $W_n \equiv 1$ при всех n. Поэтому система стабильна. В случае невырожденных случайных величин, заметим, что при каждом n

$$W_{n+1} \ge W_n - 1 \ge bX_n - 1 + W_{n-1} - 1.$$

С использованием индукции, получаем неравенство

$$W_{n+1} \ge (bX_n - 1) + (bX_{n-1} - 1) + \ldots + (bX_1 - 1).$$

Так как $b = \frac{1}{a}$, то правая часть этого неравенства совпадает с $\dfrac{S_n - na}{a}$, где $S_n = X_1 + \ldots + X_n$. Поэтому для любого числа $c > 0$

$$\mathbf{P}(W_{n+1} > c\sqrt{n}) \geq \mathbf{P}(\frac{S_n - na}{a} > c\sqrt{n}) = \mathbf{P}(\frac{S_n - na}{\sigma\sqrt{n}} > \frac{ca}{\sigma}) \approx 1 - \Phi(\frac{ca}{\sigma}).$$

Так как число $q \equiv 1 - \Phi(\frac{ca}{\sigma})$ положительно, то и $\mathbf{P}(W_{n+1} > c\sqrt{n}) \approx q > 0$ при всех достаточно больших n. А число $c\sqrt{n}$ можно сделать как угодно большим. Значит, система не может быть стабильной.

Рассмотрим случай $ab > 1$. Если случайные величины X_n вырождены, то величины $W_n = ab + (n-1)(ab-1)$ неограниченно растут с ростом n. Если же X_n невырождены, то, повторяя ранее изложенные рассуждения, мы получаем неравенство:

$$W_{n+1} > \frac{S_n - na}{a},$$

из которого следует нестабильность системы.

В случае $ab < 1$ утверждение Теоремы 3.1 доказывается значительно сложнее, и мы его приводить не будем. Опишем лишь один полезный прием, используемый при доказательстве и называемый часто «правилом насыщения» (saturation rule). Предположим, что к некоторому моменту времени n в системе скопилось очень много вызовов (система «насыщена» вызовами). Тогда, начиная с момента n, в течение длительного времени сервер обслуживает только имеющиеся вызовы — независимо от того, что еще в систему поступает. Допустим, что так сервер работает в течение времени T. За это время обслуживается приблизительно $\frac{T}{b}$ вызовов (то есть сервер обслуживает приблизительно $\frac{1}{b}$ вызовов за единицу времени, эту дробь естественно назвать *интенсивностью*, или скоростью обслуживания), а поступает $X_{n+1} + X_{n+2} + \ldots + X_{n+T}$ вызовов. По закону больших чисел,

$$\frac{X_{n+1} + X_{n+2} + \ldots + X_{n+T}}{T} \approx a,$$

то есть «интенсивность» поступления вызовов равна a. И так как $a < \frac{1}{b}$ (в насыщенной системе скорость поступления вызовов меньше скорости обслуживания), то кажется интуитивно очевидным, что величины W_n не могут неограниченно расти, и система стабильна.

И действительно, как показывает более точный и строгий анализ, для одноканальной системы использование правила насыщения оказывается корректным.

Приведем еще два примера систем, где условия стабильности находятся аналогичным образом.

Пример 1. Одноканальная система с двумя типами вызовов. В систему с одним сервером поступают вызовы двух типов: в каждый момент времени n приходят $X_n^{(1)}$ вызовов первого типа и $X_n^{(2)}$ — второго типа. Для каждого $i = 1, 2$, случайные величины $X_1^{(i)}, X_2^{(i)}, \ldots$ независимы и одинаково распределены со средним $a_i = M(X_1^{(i)})$. Время обслуживания каждого вызова первого типа равно b_1, второго типа — b_2. Вызовы обслуживаются на сервере в порядке, задаваемом с помощью некоторого правила (дисциплины) обслуживания. Наиболее известные варианты дисциплин:

- Дисциплина *первый пришел — первый обслуживается* (first come first served, FCFS). При этой дисциплине сначала (в произвольном порядке) обслуживаются вызовы, пришедшие в момент времени $n = 1$, затем вызовы, пришедшие в момент $n = 2$, и т. д.

- *Приоритетная* дисциплина. Вызовы первого типа имеют приоритет перед вызовами второго типа. Это означает, что вызовы разных типов образуют две разные очереди, и в каждый момент, когда сервер заканчивает обслуживание какого-то вызова, на обслуживание направляется вызов первого типа (первый вызов из первой очереди), и только если первая очередь пуста, то разрешается обслуживание вызову второго типа.

Справедлива следующая

Теорема 3.2. *Обозначим $\rho = a_1 b_1 + a_2 b_2$. Каков бы ни был порядок (дисциплина) обслуживания вызовов, одноканальная система с двумя типами вызовов является стабильной при $\rho < 1$ и нестабильной при $\rho > 1$. В случае $\rho = 1$ система является стабильной тогда и только тогда, когда величины $X_n^{(1)}$ и $X_n^{(2)}$ являются вырожденными, т. е. $\mathbf{P}(X_n^{(1)} = a_1) = 1$ и $\mathbf{P}(X_n^{(2)} = a_2) = 1$.*

Доказательства Теорем 3.1 и 3.2 используют одни и те же рассуждения.

Пример 2. Открытая сеть Джексона с двумя серверами и одним типов вызовов. Пусть в сети находятся два сервера, время обслуживания любого вызова на первом неслучайно и равно b_1, а на втором — b_2. В каждый момент времени $n = 1, 2, \ldots$ в сеть извне поступает X_n вызовов, причем случайные величины $X_1, X_2, \ldots$ независимы и одинаково распределены со средним a. Каждый вызов (независимо от других) с вероятностью s_1 встает в очередь к первому серверу и с вероятностью $s_2 = 1 - s_1$ — ко второму. Любой вызов, обслуженный на первом сервере, либо покидает сеть (с вероятностью $r_{1,0}$), либо встает в очередь ко второму (с вероятностью $r_{1,2}$). Здесь $r_{1,0} + r_{1,2} = 1$. Аналогично, после обслуживания на втором сервере вызов с вероятностью $r_{2,0}$ покидает сеть и с вероятностью $r_{2,1}$ переходит в очередь к первому серверу (где $r_{2,0} + r_{2,1} = 1$). Для того, чтобы вызовы имели возможность покинуть сеть, требуется, чтобы сумма $r_{1,0} + r_{2,0}$ была положительной. Посчитаем, какое число раз в среднем любой вызов посетит первый сервер (нам нужно найти $M(\nu_1)$, где ν_1 — случайное число посещений).

В случае, когда вызов сразу поступил в очередь к первому серверу, вероятность q того, что он снова вернется на этот сервер хотя бы раз (другими словами, посетит этот сервер хотя бы два раза), равна вероятности того, что он переходит с первого сервера на второй и затем — со второго на первый. Так как эти события независимы, то $q = r_{1,2} \cdot r_{2,1} < 1$. Аналогично, вероятность того, что вызов вернется на первый сервер хотя бы m раз, если он начал обслуживание с этого сервера, равна q^m. Поэтому среднее число посещений равняется $1 + q + q^2 + \ldots = \frac{1}{1-q}$.

Если предположить, что сначала вызов поступает на второй сервер, то вероятность того, что он посетит первый сервер хотя бы раз, равна $r_{2,1}$, хотя бы два раза — $r_{2,1}q$, хотя бы три — $r_{2,1}q^2$ и т. д. Следовательно, в этом случае среднее число посещений первого сервера равно $r_{2,1} + r_{2,1}q + r_{2,1}q^2 + \ldots = \frac{1}{1-q} r_{2,1}$.

И так как первый случай осуществляется с вероятностью s_1, а второй — с вероятностью s_2, то

$$M(\nu_1) = s_1 \frac{1}{1-q} + s_2 \frac{1}{1-q} r_{2,1} = \frac{1}{1-q}(s_1 + s_2 r_{2,1}).$$

Симметрично, среднее число $M(\nu_2)$ посещений одним вызовом второго сервера равно $\frac{1}{1-q}(s_2 + s_1 r_{1,2})$.

Для того, чтобы понять, при каких условиях сеть работает стабильно, воспользуемся снова правилом насыщения. Но теперь нам удобнее рассуждать в терминах не числа вызовов, а количества «работы».

Каждый вызов посетит в среднем первый сервер $M(\nu_1)$ раз. Значит, он в среднем «приносит» на этот сервер количество работы, равное $b_1 M(\nu_1)$. За единицу времени в систему поступает в среднем a вызовов. Значит, все они вместе приносят на первый сервер количество работы $ab_1 M(\nu_1)$. Обозначим это число через ρ_1. Если первый сервер «насыщен», то он постоянно занят (работает с интенсивностью единица). Поэтому для стабильности нужно, чтобы выполнялось условие $\rho_1 \leq 1$. Аналогично, требуется и второе условие: $\rho_2 \equiv ab_2 M(\nu_2) \leq 1$.

Поэтому неудивительно, что имеет место следующая теорема.

Теорема 3.3. *Открытая сеть Джексона с двумя серверами может быть стабильной только в следующих случаях:*

- $\max(\rho_1, \rho_2) < 1$;
- *случайная величина* X_1 *вырождена,* $\max(\rho_1, \rho_2) = 1$ *и выполнено одно из условий: либо* $r_{1,0} = r_{2,0} = 1$*, либо* $r_{1,0} = 1, r_{2,0} = 0$*, либо* $r_{1,0} = 0, r_{2,0} = 1$.

Теорема 3.2 обобщается естественным образом на случай любого числа вызовов, а Теорема 3.3 — любого числа серверов.

До конца 80-х годов у многих ученых, работающих в области теории систем обслуживания, была надежда на то, что «правило насыщения» является универсальным, т. е. с его помощью всегда можно найти верные условия стабильности систем обслуживания. Однако в 1990 году Р.П. Кумар привел относительно простой пример детерминированной (неслучайной) системы, в которой условия стабильности существенно отличаются от получаемых с помощью этого правила. Затем А.Л.Столяр и А.Н.Рыбко (1992) построили пример системы со случайными характеристиками, имеющей те же свойства. Эти и другие примеры привели к развитию так называемой «теории жидкостной аппроксимации», позволяющей анализировать условия стабильности случайных систем обслуживания с помощью вспомогательных динамических «жидкостных» моделей, удовлетворяющих ряду интегро-дифференциальных уравнений.

4 Системы множественного доступа

Теоретическое исследование этих систем началось в семидесятые годы и было вызвано следующими практическими соображениями.

Рассмотрим, к примеру, одноканальную систему с несколькими типами вызовов. Допустим, что любой вызов (любого типа) требует обслуживания в течение единичного времени (скажем, время обслуживания равно 1 секунде). Предположим, что сервер находится на космической станции, все вызовы первого типа скапливаются в очереди где-то в Австралии, второго — в России, третьего — в Бразилии, и т. д. Наземная связь работает намного хуже, чем связь космическая. Упорядочить вызовы в каждой очереди можно, а упорядочивать поступление на сервер вызовов из разных очередей очень сложно, на это может теряться большое количество времени и средств. Поэтому на практике стали применять такой алгоритм:

- заранее каждой очереди указывается некоторое число (i-ой очереди — число p_i, где $0 < p_i < 1$) ;
- в каждую единицу времени, первый вызов из первой очереди посылает сигнал на сервер с вероятностью p_1, первый вызов из второй очереди – с вероятностью p_2, и т. д., причем делают они это независимо один от другого (если некоторая очередь пуста, то из нее сигнала не поступает);
- если число полученных сервером сигналов равно единице (послан только один сигнал), то соответствующий вызов обслуживается и затем (через единицу времени) покидает систему;
- если число сигналов равно нулю, то сервер простаивает в течение единицы времени;
- если число сигналов не меньше двух, то они вступают в «конфликт», и сервер также простаивает в течение единицы времени.

Можно рассматривать и более сложные модели сетей с несколькими серверами, работающими по описанному принципу, и маршрутами передвижения вызовов от одних серверов к другим.

Такого рода системы (и сети) получили название «системы множественного доступа» (“multi-access broadcast channel”). Часто в литературе термин «вызов» заменяют на «сообщение», а «обслуживание вызова» — на «передачу сообщения».

Чтобы понять, какого рода условия требуются для стабильной работы системы, рассмотрим наиболее простой случай **симметричной системы множественного доступа с двумя очередями**.

Предположим, что $p_1 = p_2 = p$. Если в некоторый момент одна очередь пуста, а другая — нет, то вероятность обслуживания за следующую единицу времени равна p. Если же обе очереди непусты, то обслуживание происходит в двух симметричных случаях, когда из одной очереди сигнал поступает, а из другой — нет. Так как сигналы подаются независимо, каждый из случаев осуществляется с вероятностью $p(1-p)$. Следовательно, вероятность обслуживания равна $2p(1-p)$.

Пусть, как и в предыдущем параграфе, $X_n^{(1)}$ означает количество вызовов, поступающих за n-ую единицу времени в первую очередь, и $X_n^{(2)}$ — во вторую. Предположим, что все случайные величины $X_1^{(1)}, X_1^{(2)}, X_2^{(1)}, X_2^{(2)}, \ldots$ независимы и одинаково распределены. Обозначим через a их общее среднее.

Используя правило насыщения, можно предложить достаточные условия стабильности системы.

Теорема 4.1. *Симметричная система множественного доступа с двумя очередями стабильна, если выполняется одно из двух:*

- $p \geq \frac{1}{2}$ *и* $2p(1-p) > 2a$ *;*
- $2a < p < \frac{1}{2}$.

Действительно, рассмотрим первый случай (во втором рассуждения аналогичны). Если система насыщена, то либо обе очереди непусты (и обслуживание происходит с вероятностью $2p(1-p)$), либо одна пуста, а другая нет (и обслуживание происходит с вероятностью $p \geq 2p(1-p)$). Значит, в течение длительного времени T интенсивность обслуживания не меньше чем $2p(1-p)$, а интенсивность поступления вызовов строго меньше этого числа.

Ясно, почему сформулированные в Теореме 4.1 условия являются только достаточными для стабильности. Предположим, что система насыщена, и нам удалось посчитать, какую часть времени в течение длительного времени T обе очереди непусты (скажем, эта доля равна c). Тогда «средняя» интенсивность обслуживания за время T равняется $c \cdot 2p(1-p) + (1-c) \cdot p$, и «верное» условие стабильности есть $2a < c \cdot 2p(1-p) + (1-c) \cdot p$. Другими словами, система может работать стабильно и при других условиях. Оказывается, что, вообще говоря, определить число c можно, но для этого требуется хорошо развитая техника «теории случайных блужданий», являющейся разделом теории вероятностей.

Таким же образом решается задача нахождения условий стабильности для несимметричных систем с двумя очередями. Чем больше число очередей, тем сложнее решать эту задачу. К настоящему времени она полностью решена только для случаев двух, трех и четырех очередей.

В практике часто применяется и другой тип систем множественного доступа, который мы условно назовем **системы без очередей**. Предположим, что каждый поступающий вызов имеет связь только с обслуживающим сервером — поэтому никакие два из них не могут «договориться», кому из них идти на обслуживание первым. Следовательно, видится единственный способ возможного «разрешения конфликтов» — это предлагать каждому ожидающему вызову в каждый момент времени с некоторой вероятностью либо посылать сигнал на сервер, либо не посылать.

Пусть, например, в систему за единицу времени поступает X_n вызовов (где, как и ранее, случайные величины $X_1, X_2, \ldots$ являются независимыми и одинаково распределенными, $M(X_1) = a > 0$), и в каждый момент времени n каждый из имеющихся вызовов посылает

сигнал на станцию с вероятностью p_n (независимо от всех других). Будем рассматривать случай невырожденных случайных величин X_n, причем $\mathbf{P}(X_1 > 1) > 0$. Набор вероятностей $\{p_n\}$ естественно назвать «дисциплиной», или «алгоритмом» обслуживания.

Самый простой вариант — когда вероятности p_n от n не зависят и равны одному и тому же числу $p < 1$. Но в этом случае справедлива

Теорема 4.2. *При любом выборе числа p система без очередей нестабильна!*

Следовательно, надо брать числа p_n различными. Сформулируем следующий вопрос: можно ли задать заранее такую последовательность чисел $p_1, p_2, \ldots$, при которой система будет работать стабильно? «Заранее» означает, что при выборе этих чисел не может использоваться никакая информация о состоянии системы. Делалось много попыток ответить на этот вопрос; опубликован ряд статей, где показывается, что для различных классов последовательностей ответ всегда является отрицательным. Однако до сих пор не все случаи изучены, и вопрос остается открытым.

Допустим теперь, что в каждый момент времени известно общее число вызовов в системе (пусть Q_n — число вызовов в момент n), и можно использовать знание чисел $Q_1, Q_2, \ldots, Q_n$ для определения вероятности p_n. Тогда имеет место следующая

Теорема 4.3. *Если $a < e^{-1}$, то существует алгоритм, при котором система стабильна. Если $a \geq e^{-1}$, то таких алгоритмов нет.*

Здесь e — основание натуральных логарифмов и $e^{-1} = \frac{1}{e}$.

Доказательство этого утверждения также проводится с использованием правила насыщения. Предложим алгоритм вида $p_n = 1$, если $Q_n = 0$ и $p_n = \dfrac{1}{Q_n}$, если $Q_n > 0$. Пусть в момент n в системе находится $Q_n = L$ вызовов, где L очень велико. Тогда $p_n = \frac{1}{L}$. Обслуживание будет происходить, если на станцию поступит ровно один сигнал (т. е. один вызов сигнал пошлет, а остальные $L - 1$ — нет). Это может происходить в одном из L случаев, каждый из которых имеет вероятность $p_n(1 - p_n)^{L-1}$. Следовательно, вероятность обслуживания равна

$$Lp_n(1 - p_n)^{L-1} = (1 - \frac{1}{L})^{L-1},$$

и, как доказывается в курсе высшей математики, это число сближается с e^{-1} с ростом L. Поэтому при больших L в течение длительного времени интенсивность обслуживания приблизительно равна e^{-1}, что должно быть больше интенсивности входа a.

Отметим, что $e^{-1} \approx 0,37$. Как следует из Теоремы 4.3, даже при оптимальном алгоритме сервер вынужден простаивать приблизительно 63 процента времени. Различными авторами рассматривались алгоритмы из более широкого класса (в частности, допускалась возможность вызову каждый раз при определении вероятности подачи сигнала учитывать, сколько раз он пытался подать сигнал до этого), что позволило поднять границу области стабильности с e^{-1} до $\frac{1}{2}$.

5 Системы поллинга

Вернемся к системам множественного доступа с конечным числом очередей. Один из вариантов повышения эффективности работы таких систем состоит в использовании так называемого режима разделения времени. Предполагается, что сервер имеет возможность работать с очередями последовательно, в некотором порядке. Такие системы и называют **системами поллинга**[4]. В системах поллинга, в отличие от одноканальных систем

[4] В английском языке слово «поллинг» (polling) используется в различных смыслах; мы будем иметь в виду один из основных, означающий в переводе «упорядоченный опрос». Термин «система поллинга» уже укоренился в русскоязычной математической литературе, хотя желающие могут называть их и «системами упорядоченного опроса»

с несколькими типами вызовов, сервер при переходе от одной очереди к другой тратит некоторое время на «переключение».

Рассмотрим систему поллинга с двумя очередями (системы с большим числом очередей рассматриваются аналогично). Вспомним описание системы с двумя типами вызовов (см. Пример 1 параграфа 3) и предположим, что вызовы каждого типа образуют свою очередь. Пусть сервер обходит очереди по некоторому циклическому правилу (например, правило $\{1, 2, 1, 1, 2\}$ означает, что в течение каждого цикла сервер сначала посещает первую очередь, затем вторую, затем дважды первую и, наконец, вторую. На этом очередной цикл заканчивается и начинается следующий). Предположим, что за один цикл сервер посещает первую очередь c_1 раз, а вторую — c_2 раз. За каждый цикл сервер тратит суммарно на переключение время V (будем считать его неслучайным). Зададим два целых положительных числа F_1 и F_2 и определим следующее правило обслуживания: при каждом очередном посещении первой очереди, если сервер застает там x вызовов, то он обслуживает $\min(x, F_1)$ из них (т. е. если $x \le F_1$, то обслуживается x вызовов, иначе — F_1 вызовов). После этого сервер покидает эту очередь и переключается на следующую очередь из его маршрута. Аналогично, при приходе во вторую очередь он каждый раз обслуживает все находившиеся там вызовы, но не больше, чем F_2.

Предположим, что распределения случайных величин $X_n^{(i)}$ невырождены. Справедлива

Теорема 5.1. *Условие*

$$a_1 b_1 + a_2 b_2 + V \max\left(\frac{a_1}{F_1 c_1}, \frac{a_2}{F_2 c_2}\right) < 1. \tag{5.3}$$

является необходимым и достаточным для стабильности системы поллинга с двумя очередями.

В частности, если $V = 0$, то мы получаем первую часть утверждения Теоремы 3.2.

Поясним смысл условия (5.3). Для этого применим несколько видоизмененный вариант правила насыщения. Возьмем число n достаточно большим и рассмотрим вспомогательную систему, в которую в момент времени 0 поступает $S_n^{(1)} = X_1^{(1)} + X_2^{(1)} + \ldots + X_n^{(1)}$ вызовов первого типа и $S_n^{(2)} = X_1^{(2)} + X_2^{(2)} + \ldots + X_n^{(2)}$ — второго, а после этого никаких новых поступлений вызовов нет. Посчитаем приблизительно, за какое время сервер обслужит все пришедшие вызовы. Если это время окажется меньшим, чем n, то система должна работать стабильно (ведь n — это время, за которое все эти вызовы поступают в реальную систему!).

За один цикл обслуживается $F_1 c_1$ вызовов первого типа (если такое число вызовов в системе есть). Поэтому для обслуживания всех $S_n^{(1)}$ вызовов первого типа потребуется приблизительно $\dfrac{S_n^{(1)}}{F_1 c_1}$ циклов (точнее, если эта дробь не является целым числом, то нужно взять ближайшее целое, большее ее). Аналогично, для обслуживания всех вызовов второго типа нужно приблизительно $\dfrac{S_n^{(2)}}{F_2 c_2}$ циклов. Следовательно, для обслуживания всех вызовов, пришедших во вспомогательную систему, требуется $M \approx \max\left(\dfrac{S_n^{(1)}}{F_1 c_1}, \dfrac{S_n^{(2)}}{F_2 c_2}\right)$ циклов. За эти M циклов сервер потратит $S_n^{(1)} b_1$ единиц времени на обслуживание вызовов первого типа и $S_n^{(2)} b_2$ — второго. Так как в каждом цикле он тратит V единиц времени на переключение, то M циклов будут длиться время $S_n^{(1)} b_1 + S_n^{(2)} b_2 + VM$. Решим приблизительно неравенство

$$S_n^{(1)} b_1 + S_n^{(2)} b_2 + V \max\left(\frac{S_n^{(1)}}{F_1 c_1}, \frac{S_n^{(2)}}{F_2 c_2}\right) < n. \tag{5.4}$$

Разделим обе части неравенства на n и заметим, что (при больших n)

$$\frac{S_n^{(1)}}{n} \approx a_1 \quad \text{и} \quad \frac{S_n^{(2)}}{n} \approx a_2.$$

Значит, неравенство (5.4) эквивалентно неравенству (5.3), что и требовалось показать.

Литература

1. Саати Т.Л. Элементы теории массового обслуживания. М.; Советское радио, 1971.
2. Боровков А.А. Вероятностные процессы в теории массового обслуживания. М.; Наука, 1972.
3. Гнеденко Б.В., Коваленко И.Н. Введение в теорию массового обслуживания, 2-е изд. М.; Наука, 1987.
4. Цыбаков Б.С., Михайлов В.А. Случайный множественный доступ пакетов. Алгоритм дробления. // Проблемы передачи информации, 1980, Т. 16, вып. 4, с. 65-79.
5. Рыбко А.Н., Столяр А.Л. Об эргодичности случайных процессов, описывающих функционирование открытых сетей массового обслуживания.// Проблемы передачи информации, 1992, Т. 28, вып. 2, с. 3-26.
6. Malyshev V.A. Networks and dynamical systems.// Advances in Applied Probability, 1993, Vol. 25, p. 140-175.
7. Baccelli F., Foss S. On the saturation rule for the stability of queues. // Journal of Applied Probability, 1995, Vol. 32, p. 494-507.

6 Приложение.

- Фосс С.Г. Стохастические системы и сети обслуживания;
  Foss S.G. Stochastic Queueing Systems and Networks.
- Название вуза: Новосибирский государственный университет
- Аннотация.

  Вводятся некоторые понятия теории вероятностей и теории систем обслуживания, обсуждаются на примерах условия стабильной работы систем и сетей обслуживания, предлагаются некоторые методы нахождения этих условий.
- Список литературы — см. выше.
- Краткие сведения об авторе.
  Фосс Сергей Георгиевич, доктор физико-математических наук, профессор Новосибирского государственного университета, ведущий научный сотрудник Института математики Сибирского отделения Российской Академии Наук. Область научных интересов: теория вероятностей и теория систем обслуживания, вопросы эргодичности и устойчивости марковских цепей и процессов, стохастически рекурсивных последовательностей, сетей обслуживания и (теле)коммуникаций. Член редколлегий международных журналов “Queueing Systems” и “Markov Processes and Related Fields”. Автор 60 научных статей и 2 учебных пособий.