На правах рукописи

Жукова Ксения Алексеевна

ОЦЕНИВАНИЕ КАЧЕСТВА ОБСЛУЖИВАНИЯ КОММУНИКАЦИОННЫХ СИСТЕМ С ИСПОЛЬЗОВАНИЕМ ТЕОРИИ БОЛЬШИХ УКЛОНЕНИЙ И РЕГЕНЕРАТИВНОГО АНАЛИЗА

05.13.18 — математическое моделирование, численные методы
и комплексы программ

Автореферат

диссертации на соискание учёной степени

кандидата физико-математических наук

Петрозаводск — 2018

Работа выполнена в Институте прикладных математических исследований федерального государственного учреждения науки федерального исследовательского центра «Карельский научный центр Российской академии наук».

Научный руководитель: **Морозов Евсей Викторович**
доктор физико-математических наук, профессор

Официальные оппоненты: **Гайдамака Юлия Васильевна**
доктор физико-математических наук, доцент, ФГАОУ ВО «Российский университет дружбы народов», профессор кафедры прикладной информатики и теории вероятностей

Ушаков Владимир Георгиевич
доктор физико-математических наук, профессор, ФГБОУ ВО «Московский государственный университет имени М.В.Ломоносова», заведующий лабораторией статистического анализа

Ведущая организация: Институт проблем информатики Российской академии наук федерального исследовательского центра «Информатика и управление» Российской академии наук

Защита состоится «21» декабря в 14 часов на заседании диссертационного совета Д 212.190.03 на базе ФГБОУ ВО «Петрозаводский государственный университет» по адресу: 185910, г. Петрозаводск, пр. Ленина, 33.

С диссертацией можно ознакомиться в научной библиотеке Петрозаводского государственного университета и на сайте petrsu.ru.

Автореферат разослан « » __________ 2018 г.

Ученый секретарь
диссертационного совета Воронов Роман Владимирович

Общая характеристика работы

Актуальность темы. Важной и актуальной задачей при разработке и поддержке современных инфокоммуникационных систем для обработки и передачи данных является возможность построения и выбора режимов работы, позволяющих обеспечить выполнение заданных требований качества обслуживания (quality of service, QoS). Эти требования, которые отражают интересы пользователя, должны быть сбалансированы с интересами разработчиков или провайдеров услуг, что приводит к необходимости постановки и решения соответствующих оптимизационных задач. В частности, такими требованиями могут быть ограничения на среднее время ожидания выполнения задания или на вероятность того, что время ожидания задания в очереди превысит заданный порог. Как правило, параметры QoS выбираются в зависимости от особенностей системы и тех задач, которые она решает. Важнейшими параметрами QoS в инфокоммуникационных системах является вероятность превышения большого значения очереди и/или заданное ограничение времени ожидания. Так, например, в высокоответственных системах ключевым требованием является ограничение вероятности превышения процессом нагрузки заданного значения. В связи с постоянным совершенствованием и усложнением современных систем и сетей передачи данных для анализа параметров качества обслуживания необходимо привлекать сложные современные математические методы, в частности, теорию больших уклонений и регенеративный метод. Применение этих методов к анализу инфокоммуникационных систем позволяет не только получить новые теоретические результаты, в частности, касающиеся асимптотики вероятности большого уклонения, но и построить эффективные методы оценивания параметров системы, которые определяют качество обслуживания.

Цель диссертационной работы – разработать модели и методы оценивания качества обслуживания коммуникационных систем на основе методов теории больших уклонений и регенеративного анализа.

Для достижения поставленной цели были решены следующие **задачи**:

1. Исследована асимптотика стационарной вероятности большого значения очереди для широкого класса систем с повторными вызовами.
2. Проведен сравнительный анализ методов оценивания эффективной пропускной способности модели коммуникационного узла с приме-

нением методов теории больших уклонений и регенеративного анализа.

3. Исследованы свойства средней стационарной нагрузки в системах с входным процессом с распределением специального класса.

Методы исследования. В диссертационной работе применяются методы теории больших уклонений и теории регенеративных процессов, метод каплинга, метод группового среднего и регенеративного оценивания, а также методы имитационного моделирования.

Научная новизна работы.

1. Доказана экспоненциальная асимптотика стационарной вероятности большого размера орбиты для широкого класса систем с повторными вызовами.
2. Разработан метод оценивания эффективной пропускной способности узла системы, гарантирующий требуемое качество обслуживания в высокоответственных системах.
3. Доказаны неравенства типа Поллачека-Хинчина, задающие границы характеристики качества обслуживания для систем с с входным процессом с распределением специального класса.

Теоретическая и практическая значимость. В диссертационной работе доказана экспоненциальная асимптотика стационарной вероятности большого размера орбиты для широкого класса систем с повторными вызовами путем анализа эквивалентной модели с очередью. Полученные результаты могут быть использованы для оценивания качества обслуживания широкого класса существующих, а также при разработке новых коммуникационных систем. В частности, выработаны некоторые рекомендации для провайдеров и разработчиков для более точного оценивания эффективной пропускной способности. Разработан комплекс программ имитационного моделирования для оценивания эффективной пропускной способности различными методами и оценивания вероятности большого размера орбиты в системах с постоянной интенсивностью повторных вызовов.

На защиту выносятся следующие основные результаты и положения:

1. Экспоненциальная асимптотика стационарной вероятности растущего значения очереди в односерверной и многосерверной системах с классической и постоянной интенсивностью повторных вызовов.
2. Рандомизированный метод группового среднего для оценивания эффективной пропускной способности, обеспечивающий заданное качество обслуживания системы.
3. Неравенства типа Поллачека-Хинчина для многосерверной системы с входным потоком с распределением специального класса.
4. Комплекс программ имитационного моделирования для оценивания эффективной пропускной способности коммуникационного узла.

Положения, выносимые на защиту, соответствуют пунктам 1, 2, 3, 8 раздела «Области исследований» паспорта специальности 05.13.18 – математическое моделирование, численные методы и комплексы программ.

Апробация работы. Результаты диссертационной работы представлялись на следующий международных конференциях:

1. Всероссийская конференция с международным участием:«Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем 2015» (20 – 24 апреля 2015, Москва, РУДН);
2. 29th European Conference on Modelling and Simulation (26 – 29 мая, 2015, Болгария);
3. Distributed computer and communication networks: control, computation, communications, DCCN-2015 (19 – 22 октября 2015 г., Москва, Россия);
4. The Second International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management (Израиль, 15 – 18 февраля 2016 г.);
5. IX международная конференция «Вероятностные методы в дискретной математике» (30 мая – 3 июня 2016 г., Петрозаводск, Россия);
6. 19th International Conference on distributed computer and communication networks DCCN-2016 (21 – 25 ноября 2016 г., Москва, Россия);

7. Двадцатая международная научная конференция «Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь» (25 – 29 сентября 2017 г., Москва, ИПУ РАН);
8. International Conference on Man–Machine Interactions (3 – 6 октября 2017 г., Краков, Польша);
9. European Conference on Queueing Theory 2018 (2 – 4 июля 2018 г., Израиль);
10. Distributed computer and communication networks: control, computation, communications, DCCN-2018 (17 – 21 сентября 2018 г., Москва, ИПУ РАН).

Публикации. По материалам и результатам диссертации опубликовано 12 печатных работ, из них 1 статья в российском журнале из списка ВАК [1], 5 статей в журналах и сборниках трудов международных конференций [2] – [6], индексируемые в реферативной базе Scopus, 1 статья в журнале [7] и 5 тезисов докладов [8] – [12], индексируемых РИНЦ. Получено свидетельство о государственной регистрации программы для ЭВМ [13].

Структура и объем диссертации. Диссертация состоит из введения, 4 глав, заключения, списка литературы и приложения. Общий объем диссертации 111 страниц, включая 16 рисунков. Список литературы включает 101 наименование.

Содержание работы

Во введении обоснована актуальность темы диссертационного исследования, сформулированы цель и задачи, аргументирована научная новизна исследований, обозначена практическая и теоретическая значимость работы, а также представлены выносимые на защиту научные положения.

В первой главе кратко изложены ключевые понятия и результаты теории больших уклонений, необходимые для дальнейшего анализа, а также представлен обзор основных исследований по применению принципа больших уклонений при анализе стохастических моделей инфокоммуникационных систем. Теория больших уклонений описывает асимптотическое поведение хвостов распределений случайных величин (с.в.) и является одним из важных методов для анализа стохастических моделей, описывающих процессы в системах передачи данных. Эти системы постоянно усложняются в связи с развитием технологий и исследованиями свойств реального интернет- и инфокоммуникационного трафика. Под вероятностью большого уклонения понимается вероятность превышения некоторым процессом большого значения, например, вероятность большого размера очереди в системе. Важную роль в описании вероятности большого уклонения играет преобразование Лежандра и связанная с ним производящая логарифмическая функция моментов. В главе 1 рассмотрены некоторые их свойства. Также обсуждаются возможности применения результатов теории больших уклонений в теории массового обслуживания и возможности расширения этих результатов на некоторые модели инфокоммуникационных систем. Особый интерес представляют случаи, когда в качестве вероятности большого уклонения рассматривается вероятность превышения процессом нагрузки или процессом общего числа заявок в системе некоторого большого значения. Подобные вероятности лежат в основе некоторых критериев качества обслуживания, поэтому важно иметь возможность оценивать их даже в случаях, когда классические методы неприменимы ввиду сложности моделей реальных систем передачи данных.

Во второй главе представлены теоретические результаты исследования асимптотики вероятности большого размера орбиты в системах с повторными вызовами.

Рассмотрим односерверную систему с орбитой с входным процессом восстановления. Обозначим моменты поступления заявок в систему $t_n,\ n \geqslant 1$, независимые одинакого распределенные (н.о.р.) времена между приходами заявок $\tau_n := t_{n+1} - t_n,\ n \geqslant 1,\ t_1 = 0$, и н.о.р. времена обслуживания $S_n, n \geqslant 1$. В дальнейшем для удобства будем опускать порядковый индекс у с.в. τ_n и S_n, обозначая через τ типичный интервал между приходами и через S – типичное время обслуживания. Пусть входной поток заявок в систему имеет интенсивность $\lambda := 1/\mathsf{E}\tau \in (0,\infty)$, а интенсивность обслуживания при этом $\mu := 1/\mathsf{E}S \in (0,\infty)$.

В рассматриваемых системах поступившая заявка либо мгновенно идет на обслуживание (если сервер свободен в момент ее прихода), либо мгновенно отправляется на виртуальную орбиту с неограниченной вместимостью (если сервер был занят в момент поступления заявки в систему). Находясь на орбите, заявка повторяет попытки занять сервер через промежутки времени, распределенные экспоненциально с некоторым параметром $\gamma > 0$. В системах с классической орбитальной дисциплиной суммарная интенсивность попыток занять сервер зависит (как правило линейно) от числа заявок на орбите, в то время как при постоянной орбитальной скорости эта интенсивность равна γ независимо от размера орбиты (т. е. числа заявок на ней). В этом случае удобно рассматривать орбиту как очередь типа FIFO, в которой только первая в очереди (т. е. самая «старая») орбитальная заявка делает попытки занять сервер. Стоит отметить, что дисциплина FIFO здесь относится именно к заявкам, поступающим на орбиту и уходящим с нее на обслуживание.

Рассмотрим систему с классической интенсивностью повторных вызовов. Основное предположение для анализа заключается в том, что система работает в стационарном режиме. Для односерверной системы с орбитой с классической дисциплиной повторных вызовов известно достаточное условие стационарности

$$\rho := \lambda/\mu < 1. \tag{1}$$

Как видно, условие (1) совпадает с критерием стационарности классической системы $GI/G/1$. Обозначим $Q(t)$ общее число заявок в системе (на сервере и на орбите) в момент времени t^-. Пусть начальное состояние процесса нулевое: $Q(0) = 0, t_1 = 0$. Определим моменты регенерации процесса Q:

$$Z_{n+1} = \min(t_k > Z_n : Q(t_k) = 0),\ n \geqslant 0,\ Z_0 = 0. \tag{2}$$

Стационарность означает, что регенеративный процесс Q является положительно возвратным, т. е. типичный цикл регенерации имеет конечное среднее. Пусть T – типичная длина цикла занятости системы. Рассмотрим вероятность превышения процессом общего числа заявок в системе некоторого (большого) значения $N > 1$ на цикле занятости.

Пусть β — число приходов заявок в течение цикла занятости, а D_k — интервал между $k-1$ и k уходами из системы, $k \geqslant 2$. Обозначим K_N номер первой заявки, которая достигла уровня N на цикле занятости, т. е.:

$$K_N = \min\left\{n \geqslant N : \sum_{k=1}^{n-1} \tau_k < \sum_{k=1}^{n-N+1} D_k\right\}, \tag{3}$$

где мы полагаем $\min\emptyset = \infty$, а значит $K_N := \infty$, для случая, когда на цикле занятости не было превышения уровня N. Если $K_N < \infty$, то превышение на цикле было, и значит $K_N \leqslant \beta$. Обозначим K_0 номер заявки, которая поступает в пустую систему (т. е. видит систему пустой в момент прихода)

$$K_0 = \min\left\{ l > 1 : \sum_{k=1}^{l-1} \tau_k > \sum_{k=1}^{l-1} D_k \right\}. \tag{4}$$

Нетрудно заметить, что порядок поступления заявок в систему не обязательно совпадает с порядком их поступления на обслуживание и ухода из системы. Однако, учитывая тот факт, что времена обслуживания заявок $S_i,\ i \geqslant 1$, являются н.о.р. с.в., а орбиту можно рассматривать как очередь типа FIFO, будем назначать время обслуживания S_i заявки непосредственно при поступлении ее на сервер. Таким образом, мы перенумеровываем заявки, сохраняя соответствие между временами обслуживания $S_i,\ i \geqslant 1$, интервалами между приходами $\tau_i,\ i \geqslant 1$ и интервалами между уходами $D_i,\ i \geqslant 1$. Это позволяет определить интервалы D_k следующим образом:

$$D_i = S_i + I_i, \quad i \geqslant 1, \quad D_1 = S_1,\, I_1 = 0. \tag{5}$$

где I_i — время простоя между $(i-1)$-м уходом заявки с сервера и i-м приходом заявки с орбиты на сервер. При этом искомая вероятность превышения может быть записана как $\mathsf{P}(K_N < K_0)$.

Пусть логарифмическая функция моментов $\Lambda_S(\theta)$ существует для некоторых $\theta > 0$. Обозначим

$$\hat{\theta} = \max(\theta > 0 : \mathsf{E}e^{\theta S} < \infty) \tag{6}$$

и определим

$$\theta_* = \sup(\theta \in (0, \hat{\theta}) : \Lambda_\tau(-\theta) + \Lambda_S(\theta) \leqslant 0). \tag{7}$$

Заметим, что если $\mathsf{P}(\tau \geqslant S) = 1$, т. е. $S - \tau \leqslant 0$, то по определению (7) имеем $\theta_* = \infty$. При этом процесс Q всегда ограничен сверху значением 1. Так как нас интересует вероятность большого уклонения, мы исключим условие

$\mathsf{P}(\tau \geqslant S) = 1$ из рассмотрения, положив

$$\mathsf{P}(\tau < S) > 0. \tag{8}$$

В рамках условия (8) возможен произвольный рост количества заявок в системе, и можно исследовать асимптотику вероятности превышения $\mathsf{P}(K_N < K_0)$.

Для исследования асимптотики вероятности большого размера орбиты в исходной системе была предложена модель с буфером. Это позволило сравнить системы с орбитой с классическими системами, которые проще анализировать и в которых исследована асимптотика стационарной вероятности большого уклонения. С использованием техники каплинга была построена нижняя граница вероятности $\mathsf{P}(K_N < K_0)$:

Лемма 2.1 (Нижняя граница). *Асимптотика вероятности превышения $\mathsf{P}(K_N < K_0)$ на цикле занятости в односерверной системе с орбитой с классической интенсивностью повторных вызовов удовлетворяет следующему условию:*

$$\liminf_{N\to\infty} \frac{1}{N} \ln \mathsf{P}(K_N < K_0) \geqslant \Lambda_\tau(-\theta_*), \tag{9}$$

где значение $\theta_* > 0$ определяется выражением (7).

Пусть Q, Q_∞ – стационарные процессы общего числа заявок в системе в произвольный момент времени и прямо перед приходом заявки соответственно. С применением результатов теории регенеративных процессов и теоремы Чернова была доказана верхняя граница для вероятности большого уклонения $\mathsf{P}(K_N < K_0)$ на цикле занятости, а также вероятностей $\mathsf{P}(Q > N)$ и $\mathsf{P}(Q_\infty > N)$.

Лемма 2.3 Пусть выполнены условия (8) и

$$\lambda < \frac{\gamma\mu}{\gamma + \mu}. \tag{10}$$

Тогда асимптотика вероятности превышения $\mathsf{P}(K_N < K_0)$ на цикле занятости в односерверной системе с классической интенсивностью повторных вызовов удовлетворяет следующему условию:

$$
\begin{aligned}
\limsup_{N\to\infty}\frac{1}{N}\ln \mathsf{P}(K_N < K_0) &\leqslant \limsup_{N\to\infty}\frac{1}{N}\ln \mathsf{P}(Q(\infty) \geqslant N) &(11)\\
&\leqslant \limsup_{N\to\infty}\frac{1}{N}\ln \mathsf{P}(Q \geqslant N) &(12)\\
&\leqslant \Lambda_\tau(-\theta^*), &(13)
\end{aligned}
$$

где θ^* определяется по формуле

$$
\theta^* = \sup(\theta \in (0,\, \min(\hat{\theta}, \gamma)) : \Lambda_\tau(-\theta) + \Lambda_S(\theta) + \log\frac{\gamma}{\gamma-\theta} \leqslant 0). \qquad (14)
$$

Аналогичное утверждение верно и для нижнего предела (liminf).

Утверждения Лемм 2.1 и 2.3 доказаны и для систем с постоянной интенсивностью повторных вызовов. Объединим полученные результаты:

Теорема 1. *Пусть выполнены условия (8) и (10). Тогда асимптотика вероятности* $\mathsf{P}(K_N < K_0)$ *в односерверной системе с классической или постоянной интенсивностью повторных вызовов удовлетворяет следующему условию:*

$$
\begin{aligned}
\Lambda_\tau(-\theta_*) &\leqslant \limsup_{N\to\infty}\frac{1}{N}\ln \mathsf{P}(K_N < K_0) &(15)\\
&\leqslant \limsup_{N\to\infty}\frac{1}{N}\ln \mathsf{P}(Q(\infty) \geqslant N) &(16)\\
&\leqslant \limsup_{N\to\infty}\frac{1}{N}\ln \mathsf{P}(Q \geqslant N) &(17)\\
&\leqslant \Lambda_\tau(-\theta^*), &(18)
\end{aligned}
$$

где θ_* *и* θ^* *определены в (7) и (14) соответственно. Аналогичное утверждение верно и для нижних пределов.*

Следствие 1. В односерверной системе с постоянной интенсивностью повторных вызовов при $\gamma \to \infty$ $\theta^*(\gamma) \to \theta_*$

$$
\limsup_{N\to\infty}\frac{1}{N}\ln \mathsf{P}(K_N < K_0) = \Lambda_\tau(-\theta_*).
$$

Действительно, при $\gamma \to \infty$ левая часть неравенства под знаком супремума в определении θ^* примет вид

$$\lim_{\gamma\to\infty} \mathsf{E}e^{-\theta(\tau-S)}\frac{\gamma}{\gamma-\theta} = \mathsf{E}e^{-\theta(\tau-S)},$$

что совпадает с выражением под знаком супремума в определении параметра нижней границы θ_* (7).

Аналогичные результаты были получены и для стационарной вероятности большого уклонения в многосерверной системе $GI/GI/m$ с постоянной интенсивностью повторных вызовов γ, входным процессом восстановления с н.о.р. промежутками между приходами заявок $\{\tau_i,\ i \geqslant 1\}$ и н.о.р. временами обслуживания $\{S_i, i \geqslant 1\}$. Пусть $\Lambda_S(\theta)$ существует для некоторых $\theta > 0$. Обозначим

$$\hat{\theta} = \max(\theta > 0 : \mathsf{E}e^{\theta S} < \infty) \tag{19}$$

и определим

$$\theta_* = \sup(\theta \in (0, \hat{\theta}) : \Lambda_\tau(-m\theta) + \Lambda_S(\theta) \leqslant 0). \tag{20}$$

Будем полагать, что выполнено следующее условие

$$\mathsf{P}(S > m\tau) > 0. \tag{21}$$

В противном случае $\mathsf{P}(S \leqslant m\tau) = 1$, откуда следует $\theta_* = \infty$, см. аргументацию перед формулой (8). Заметим, что условие (21) выполнено тогда и только тогда когда

$$\mathsf{P}(S > \tau_1 + \cdots + \tau_m) > 0, \tag{22}$$

что, в свою очередь, гарантирует возможность произвольно большого роста числа заявок в системе.

Доказаны следующие результаты:

Лемма 2.4 В многосерверной системе $GI/GI/m$ с постоянной интенсивностью повторных вызовов, в которой выполнены условия (21) и $\mathsf{P}(\tau > S) > 0$,

$$\limsup_{N\to\infty} \frac{1}{N} \ln \mathsf{P}(K_N < K_0) \ \geqslant\ \Lambda_\tau(-m\theta_*). \tag{23}$$

Определение 2.1. С.в. принадлежит классу «Новое лучше старого» (НЛС), если ее функция распределения F для любых $x, y \geqslant 0$ удовлетворяет следующему условию

$$\bar{F}(x+y) \leqslant \bar{F}(y)\bar{F}(x), \tag{24}$$

где $\bar{F} = 1 - F$ – хвост распределения.

Лемма 2.5 *Рассмотрим многосерверную систему $GI/GI/m$ с постоянной интенсивностью повторных вызовов и промежутками между приходами, принадлежащими классу распределений НЛС. Если выполнено условие (21), то*

$$\begin{aligned}
\limsup_{N\to\infty} \frac{1}{N} \ln \mathsf{P}(K_N \leqslant K_0) &\leqslant \limsup_{N\to\infty} \frac{1}{N} \ln \mathsf{P}(Q(\infty) \geqslant N) && (25)\\
&\leqslant \limsup_{N\to\infty} \frac{1}{N} \ln \mathsf{P}(Q \geqslant N) && (26)\\
&\leqslant \Lambda_\tau(-\theta^*), && (27)
\end{aligned}$$

где $\theta^* = \min(\hat{\theta}, \gamma)$.

Также в главе 2 приведены результаты численных экспериментов, которые позволяют проанализировать точность полученных выше границ и некоторые их свойства на примере односерверной системы с постоянной интенсивностью повторных вызовов. Приведем результаты имитационных экспериментов системы с повторными вызовами типа $M/Weibull/1$ с интенсивностью входного потока λ и временами обслуживания S с функцией распределения Вейбулла

$$f(x) = \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} e^{-(x/b)^a},\ x \geqslant 0. \tag{28}$$

Известно, что

$$\mu = \frac{1}{\mathsf{E}S} = \frac{1}{b\Gamma(1+\frac{1}{a})}, \tag{29}$$

где Γ – гамма-функция. Так как τ распределены экспоненциально, то

$$\Lambda_\tau(-\theta) = \ln \mathsf{E}e^{\theta\tau} = \ln \frac{\lambda}{\lambda+\theta}. \tag{30}$$

Для функции распределения Вейбулла (28) в общем случае затруднительно, а иногда и невозможно рассчитать аналитически логарифмическую функцию моментов $\Lambda_S(\theta)$. Тем не менее, в некоторым специальных случаях это

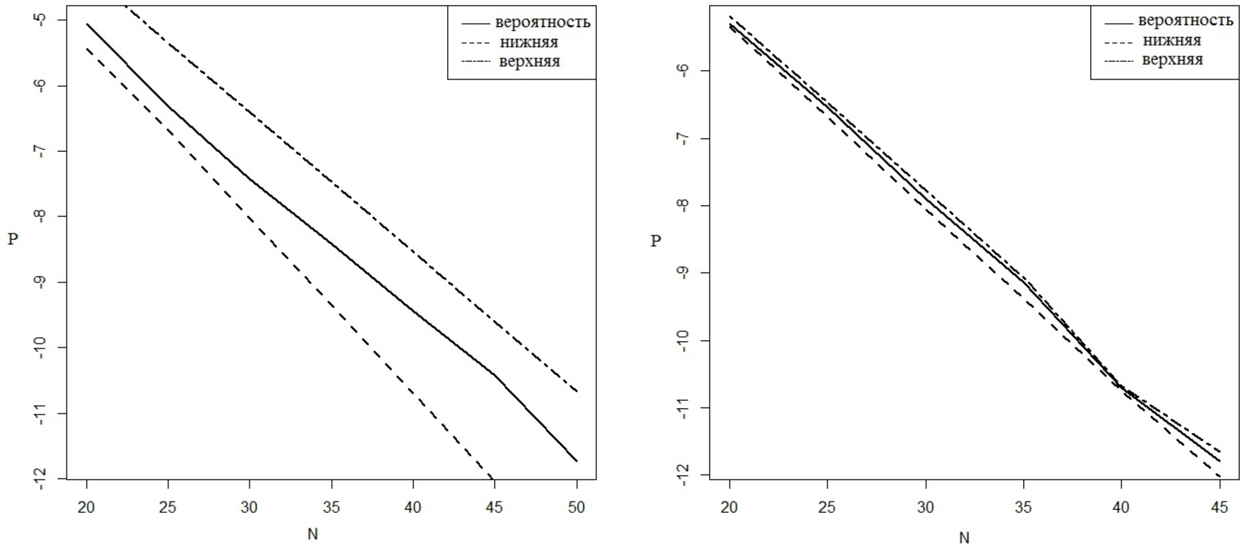

Рис. 1 — Оценка вероятности большого уклонения и ее теоретические границы для $\gamma = 30$ (слева) и $\gamma = 200$ (справа), распределение Вейбулла для времен обслуживания, логарифмическая шкала.

удается сделать. В данном эксперименте положим параметры распределения Вейбулла, равные $a = 2$ и $b = 1$. Тогда

$$\Lambda_S(\theta) = \ln \mathsf{E} e^{\theta S} = \ln \int\limits_0^{\infty} 2x e^{-x^2+\theta x} dx = \ln \left[1 + \frac{\theta}{2} \sqrt{\pi} e^{\frac{\theta^2}{4}} (1 - \Phi(-\theta/2)) \right], \quad (31)$$

где $\Phi(x)$ – интеграл вероятности

$$\Phi(x) = \frac{2}{\sqrt{\pi}} \int\limits_0^x e^{-t^2} dt.$$

Это позволяет вычислить значения θ_* и θ^*, а также численно рассчитать верхнюю и нижнюю границы, см. Рис. 1, 2.

Рассмотрим два сценария. Сначала будем моделировать работу системы с интенсивностью входного потока $\lambda = 0.95$, времена обслуживания, распределенные по Вейбуллу с параметрами $a = 2$, $b = 1$, и интенсивностями повторных вызовов $\gamma = 30, 200$. Как и ранее, мы оцениваем вероятность достижения общим числом заявок уровня N на цикле занятости. Используя результат, представленный в Теореме 1, рассчитаем нижнюю и верхнюю гра-

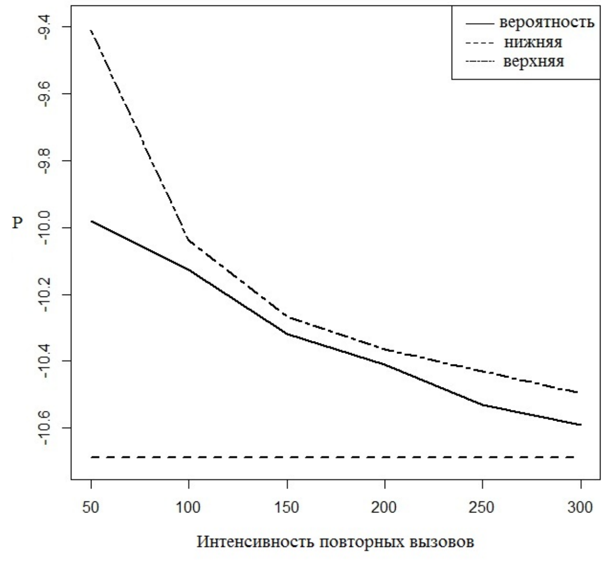

Рис. 2 — Зависимость оценки вероятности большого уклонения и ее теоретические границы от интенсивности повторных вызовов γ для $N = 40$, логарифмическая шкала. Нижняя граница $\Lambda_\tau(\theta_*)N = \Lambda_\tau(0.291) \cdot 30 = -10.688$ не зависит от γ.

ницы и сравним оценку вероятности с этими границами. Из (29) имеем

$$\mu = \frac{1}{\Gamma(1 + \frac{1}{2})} \approx 1.129.$$

Легко проверить, что выбранные в эксперименте значения параметров λ, μ и γ удовлетворяют условию стационарности (10), а также условию (8):

$$\mathsf{P}(\tau > S) = \int_0^\infty \lambda e^{-x^2 - \lambda x} dx = \lambda \frac{\sqrt{\pi}}{2} e^{\frac{\lambda^2}{4}} (1 - \Phi(\lambda/2)) > 0.$$

Следовательно, условия Теоремы 1 выполнены. Чтобы получить значение θ_*, решим численно уравнение (см. (7), (30), (31))

$$\log \frac{0.95}{0.95 + \theta} + \log \left[1 + \frac{\theta}{2} \sqrt{\pi} e^{\frac{\theta^2}{4}} (1 - \Phi(-\theta/2)) \right] = 0,$$

и выберем решение $\theta_* = 0.291$. Отметим, что $\hat{\theta} = \infty$ для времен обслуживания S, распределенных по Вейбуллу, а значит условие $\theta_* < \hat{\theta}$ автоматически выполняется. Для расчета θ^* необходимо решить уравнение (см. (14), (30), (31))

$$\log\frac{0.95}{0.95+\theta}+\log\frac{\gamma}{\gamma-\theta}+\log\left[1+\frac{\theta}{2}\sqrt{\pi}e^{\frac{\theta^2}{4}}(1-\Phi(-\theta/2))\right]=0,$$

и выбрать решение, удовлетворяющее $\theta^*\equiv\theta^*(\gamma)\in(0,\gamma)$. В этом эксперименте $\theta^*(30)=0.226$ и $\theta^*(200)=0.281$.

Результаты первого сценария эксперимента представлены на Рис. 1. Видно, что как и в предыдущем случае, оценка вероятности находится между теоретическими границами, и для большего значения γ границы лежат ближе друг у другу.

В рамках второго сценария зафиксируем уровень $N=40$ и рассчитаем оценку вероями превышения общего числа зачвок в системе этого уровня N на цикле занятости при увеличивающемся значении интенсивности повторных вызовов γ. Рис. 2 демонстрирует поведение границ с ростом γ при не меняющимся уровне N. И снова наблюдается ожидаемый эффект: расстояние между границами уменьшается с увеличением значения γ. Причина этого явления остается той же: доминирующая система с ростом γ сходится к классической системе с буфером.

В третьей главе сформулирована задача вычисления эффективной пропускной способности (ЭПС) узла инфокоммуникационной сети на основе результатов теории больших уклонений, найдено аналитическое решение (для простого случая н.о.р. с.в.), исследованы методы оценивания ЭПС для более сложных моделей, изучены свойства оценок ЭПС на основе результатов имитационного моделирования и предложен рандомизированный метод группового среднего для оценивания ЭПС.

В главе 3 приведен расчет ЭПС узла инфокоммуникационной системы с применением принципа больших уклонений на примере односерверной классической модели типа $GI/G/1$.

Рассмотрим стационарную систему $GI/G/1$ с процессом нагрузки $W(t)\Rightarrow W$ (символом $\Rightarrow$ обозначим сходимость по распределению) и постоянной мощностью (скоростью) обслуживающего устройства C (в единицу времени). Обозначим $v_n,\ n\geqslant 1$ нагрузку, поступающую в систему на интервале времени $[n,n+1)$. Обозначим $A(t)$ число заявок, поступивших в систему на интервале времени $[0,t)$, и введем в рассмотрение общую нагрузку

$V(t) = \sum_{i=1}^{A(t)} v_i$, поступившую в систему на промежутке времени $[0,t)$. Для процесса нагрузки известно следующее выражение:

$$W(t) = V(t) - Ct - \inf_{u \leqslant t}(V(u) - Cu) = \sup_{u \leqslant t}(V(u) - Cu).$$

Стационарный процесс нагрузки W удовлетворяет принципу больших уклонений

$$\lim_{N\to\infty} \frac{1}{N} \log \mathsf{P}(W > N) = -\theta^*, \tag{32}$$

где параметр $\theta^* = \sup_{\theta>0}(\Lambda(\theta) < 0)$, а логарифмическая производящая функция моментов

$$\Lambda(\theta) = \lim_{t\to\infty} \frac{1}{t} \log \mathsf{E} e^{\theta(V(t)-Ct)} = \Lambda_V(\theta) - C\theta.$$

Очевидно, что

$$\Lambda_V(\theta) = \lim_{t\to\infty} \frac{1}{t} \log \mathsf{E} e^{\theta \sum_{i=1}^{A(t)} v_i}. \tag{33}$$

Если задано фиксированное значение нагрузки N, которое можно интерпретировать как размер конечного буфера, то значение θ^* (32) можно использовать для оценивания минимального размера буфера, необходимого для сохранения заданной вероятности переполнения такого буфера

$$P_N := \mathsf{P}(W > N) \approx e^{-\theta^* N} < \Gamma. \tag{34}$$

Если задана функция $\Lambda_V(\theta)$, можно рассчитать параметр θ^* как функцию мощности обслуживанищего устройства C:

$$\theta^*(C) = \sup_{\theta>0}(\Lambda_V(\theta) < C\theta).$$

Другими словами, параметр θ^* является решением уравнения

$$\frac{\Lambda_V(\theta)}{\theta} = C.$$

Если уровень нагрузки N известен, а мощность C можно менять, но можно найти такое минимальное значение мощности (эффективную пропускную способность) $C = C_\Gamma$, которое гарантирует выполнение требования QoS (34):

$$C_\Gamma = \min_C(e^{-\theta^*(C)N} \leqslant \Gamma).$$

Предположим, что существует единственное решение C_Γ уравнения

$$e^{-\theta^*(C)N} = \Gamma.$$

Тогда ЭПС можно рассчитать как

$$C := C_\Gamma = \frac{\Lambda_V(\theta^*)}{\theta^*}, \tag{35}$$

где параметр

$$\theta^* = -\frac{\log \Gamma}{N}. \tag{36}$$

Основная сложность применения формулы (35) заключается в трудности расчета логарифмической производящей функции моментов $\Lambda_V(\theta^*)$ (33). Рассчитать ее аналитически удается лишь в некоторых случаях, когда, например, последовательность поступающей нагрузки состоит из н.о.р. с.в. v_n. Также аналитический расчет возможен для некоторых конкретных распределений с.в. v_n. Сложность расчета Λ_V возникает в случае, когда с.в. $\{v_n\}$ являются зависимыми. Для этого приходится привлекать специальные статистические методы и имитационное моделирование для оценивания логарифмической функции моментов Λ_V и собственно ЭПС. Несмещенной и сильно состоятельной оценкой логарифмической функции моментов Λ_V является выборочное среднее, что означает, что с вероятностью 1

$$\ln \frac{1}{k} \sum_{i=1}^{k} e^{\theta^* v_i} \to \Lambda_V(\theta^*), \quad k \to \infty. \tag{37}$$

Наиболее распространенной техникой построения оценки функции Λ_V является метод группового среднего (batch-mean method), когда последовательность $\{v_i\}$ разбивается на блоки одинаковой длины B. Рассмотрим стандартный вариант разбиения с блоками фиксированной длины B, определяемыми следующим образом:

$$\hat{X}_j = \sum_{i=(j-1)B+1}^{jB} v_i, \quad j \geqslant 1.$$

Основное предположение состоит в том, что при большом B эти блоки можно считать (приближенно) н.о.р. Считая, что общее число наблюдений $n = kB$

кратно величине блока, получим

$$\frac{1}{n}\ln \mathsf{E}e^{\theta\sum_{i=1}^{n}v_i} = \frac{\ln \mathsf{E}e^{\theta\hat{X}}}{B} := \Lambda_V(\theta, B), \tag{38}$$

где $\hat{X}$ означает типичный блок. Очевидно, при фиксированном размере блока B, выборочная оценка

$$\hat{\Lambda}_k(\theta, B) := \frac{1}{B}\ln\frac{1}{k}\sum_{i=1}^{k}e^{\theta\hat{X}_i}$$

является сильно состоятельной оценкой функции $\Lambda_V(\theta, B)$, т. е. при $k \to \infty$

$$\hat{\Lambda}_k(\theta, B) \to \Lambda_V(\theta, B) \quad \text{с в. 1.}$$

Соответствующая оценка ЭПС определяется тогда следующим образом:

$$\hat{C}_k(\theta^*, B) = \frac{\hat{\Lambda}_k(\theta^*, B)}{\theta^*}, \quad \theta^* = -\ln\Gamma/b. \tag{39}$$

Основной проблемой при таком подходе является выбор размера блока B: небольшой размер блока может не соответствовать предположению о независимости, а слишком большая величина неэффективна с точки зрения затрачиваемого на моделирование времени. Более того, результаты моделирования работы тандемной сети и результаты оценивания ЭПС показали, что оценка ЭПС, рассчитанная по методу группового среднего, дает заниженные значения («недооценивание»), что приводит к невыполнению требования надежности (требования QoS): для всех наблюдений разность $\Gamma - \hat{\Gamma}_{BM}$ отрицательна.

Решить проблему недооценивания позволяет использование регенеративного оценивания. Для входной последовательности $\{v_i\}$ с моментами регенерации β_k и периодами регенерации $\alpha_k = \beta_{k+1} - \beta_k$ можно строить оценку ЭПС путем группировки данных по циклам регенерации. Это приводит к регенеративным блокам вида

$$\hat{X}_k := \sum_{i=\beta_k}^{\beta_{k+1}-1} v_i, \quad k \geqslant 0, \quad \beta_0 = 0.$$

Эти блоки, в отличие от блоков, формирующихся при использовании метода группового среднего, действительно являются н.о.р. Обозначим

$$\Lambda_R(\theta^*) := \frac{1}{\mathsf{E}\alpha} \ln \mathsf{E} e^{\theta^* \hat{X}}. \tag{40}$$

Рассмотренные выше аргументы приводят к такой регенеративной оценке, которая строится по k регенеративным блокам,

$$\hat{\Lambda}(\theta^*) := \frac{k}{\beta_k} \ln \frac{1}{k} \sum_{i=1}^{k} e^{\theta^* \hat{X}_i}.$$

Так как $\beta_k/k \to \mathsf{E}\alpha$, то с в. 1

$$\hat{\Lambda}(\theta^*) \to \Lambda_R(\theta^*),\ k \to \infty,$$

и в предельной функции присутствует средний размер блока $\mathsf{E}\alpha$. Таким образом, при данной аппроксимации ЭПС удовлетворяет соотношению

$$C = \frac{1}{\theta^* \mathsf{E}\alpha} \ln \mathsf{E} e^{\theta^* \hat{X}},$$

а соответствующую регенеративную оценку ЭПС, построенную по k циклам регенерации, естественно определить как

$$\hat{C}_k := \frac{\hat{\Lambda}_R(\theta^*)}{\theta^*}. \tag{41}$$

Как показали результаты экспериментов, оценка (41) завышает (переоценивает) значение ЭПС на величину Δ. В высокоответственных системах последнее предпочтительнее, однако избыточная мощность Δ снижает другие показатели эффективности системы. Исследование свойств переоценивания Δ проводилось на основе моделирования «искусственных» циклов, не связанных с состоянием системы. Иными словами, были рассматрены группы данных $\{v_i\}$ случайной длины β (с заданным распределением), не зависящей от $\{v_i\}$. Такие искусственные циклы назовем рандомизированными группами, поскольку они, вообще говоря, уже не связаны с регенеративной структурой процессов в системе. Как выяснилось, рандомизация играет значительную роль в переоценивании. Проведенный анализ позволил предложить следующий подход к построению оценки ЭПС в высокоответственных системах. Поскольку оценка группового среднего недооценивает ЭПС, а рандомизация размера группы переоценивает ее (вне зависимости от наличия регенеративной структуры), то естественно модификацировать стандартный метод группового среднего путем рандомизации величины группы. Именно, выбираем произвольно размер группы B. Затем рандомизируем его, т. е. рассматриваем в качестве β с.в., равномерно распределенную в интервале $[(1-p)B, (1+p)B]$, где величина p изменяется в диапазоне [0, 0.15]. (Данный диапазон соответствует тому, что величина переоценивания не превышает 15% величины ЭПС. Это значение выбиралось исходя из результатов экспериментов.) В Табл. 1 приведена зависимость Δ от оценки дисперсии $Var\hat{\beta}$ длины блока. При значении $p = 0.12$ переоценивание практически равно нулю, указывая «точное» значение ЭПС.

Таблица 1

Рандомизация длины циклов β, $\Gamma = 10^{-5}$

p	$Var\beta$	$Var\hat{\beta}$	$\hat{C}$	$\hat{\Gamma}$	Δ	$\hat{\Gamma}_{\Delta}$
0	0	0	2.7155	$9.84 \cdot 10^{-3}$	-10%	$3.65 \cdot 10^{-5}$
0.03	0.0300	0.0289	2.8612	$2.75 \cdot 10^{-4}$	-0.08	$1.04 \cdot 10^{-5}$
0.05	0.0833	0.0804	2.9777	$6.03 \cdot 10^{-4}$	-0.06	$2.73 \cdot 10^{-5}$
0.07	0.1633	0.1711	3.0682	$5.28 \cdot 10^{-4}$	-0.05	$4.38 \cdot 10^{-5}$
0.10	0.3333	0.3195	3.2431	$9.36 \cdot 10^{-5}$	-0.03	$3.66 \cdot 10^{-5}$
0.12	0.4800	0.4682	3.4509	$3.89 \cdot 10^{-5}$	0	$1.05 \cdot 10^{-5}$
0.15	0.7500	0.7743	3.5461	$8.02 \cdot 10^{-6}$	0.06	$2.90 \cdot 10^{-5}$

Эксперименты по исследованию особенностей и свойств оценок показали, что рандомизация размера группы (блока) в методе группового среднего может быть эффективным способом оценки ЭПС в высокоответственных инфокоммуникационных системах. Для проведения этих экспериментов был реализован программный комплекс на языке R. Получено свидетельство о государственной регистрации программы для ЭВМ [13].

В четвертой главе приведено доказательство свойства PASTA для процесса нагрузки и процесса размера очереди односерверной классической системы с пуассоновским входным потоком с использованием регенеративного подхода. Этот анализ расширен на односерверные и многосерверные системы с распределением входного потока из класса «новое лучше старого» (НЛС) и «новое хуже старого» (НХС), показана связь между средней стационарной нагрузкой в произвольном времени и по моментам прихода заявок. Эта связь не только позволяет оценивать с помощью моделирования границы средней стационарной нагрузки как параметра качества обслуживания, но и используется в анализе вероятности большого уклонения в многосерверной системе с орбитой и входных потоком с распределением из класса НЛС. Более того, доказаны неравенства типа Поллачека-Хинчина для систем с НЛС (НХС)-входным потоком.

Теорема 3. *Пусть с.в. τ имеют распределение класса НЛС (НХС). Тогда выполнены неравенства типа Поллачека-Хинчина*

$$\mathsf{E}W_\infty \leqslant \mathsf{E}W(\infty) \leqslant \frac{\lambda \mathsf{E}S^2}{2(1-\rho)}$$

$$\left(\mathsf{E}W_\infty \geqslant \mathsf{E}W(\infty) \geqslant \frac{\lambda \mathsf{E}S^2}{2(1-\rho)}\right).$$

Теоретические результаты продемонстрированы с помощью имитационного моделирования. Проведен анализ точности полученных границ для стационарного процесса нагрузки в односерверной модели. В ходе экспериментов для входной последовательности с.в. τ использовалось распределение Вейбулла с плотностью

$$f(x) = \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} e^{-(x/b)^a},\ x \geqslant 0.$$

Удобство использования такого распределения мотивировано тем, что оно относиться к классу распределений НЛС при $a > 1$, к классу распределений НХС для $a < 1$, и является экспоненциальным с параметром $1/b$, если $a = 1$. В главе представлены результаты моделирования работы классической односерверной системы $Weibull/M/1$ с интенсивностью времен обслуживания $\mu = 1$. При этом были рассмотрены два сценария: случай мало загруженной системы, когда коэффициент загрузки $\rho := 1/\mathsf{E}\tau = 0.1$, и сильно загруженной системы с $\rho = 0.9$. Параметр $a \in [0.5, 2]$ меняется с шагом 0.02. При этом при фиксированной загрузке ρ параметр распределения Вейбулла b можно рассчитать из выражения

$$\frac{1}{\rho} = \mathsf{E}\tau = b\Gamma(1 + \frac{1}{a}),$$

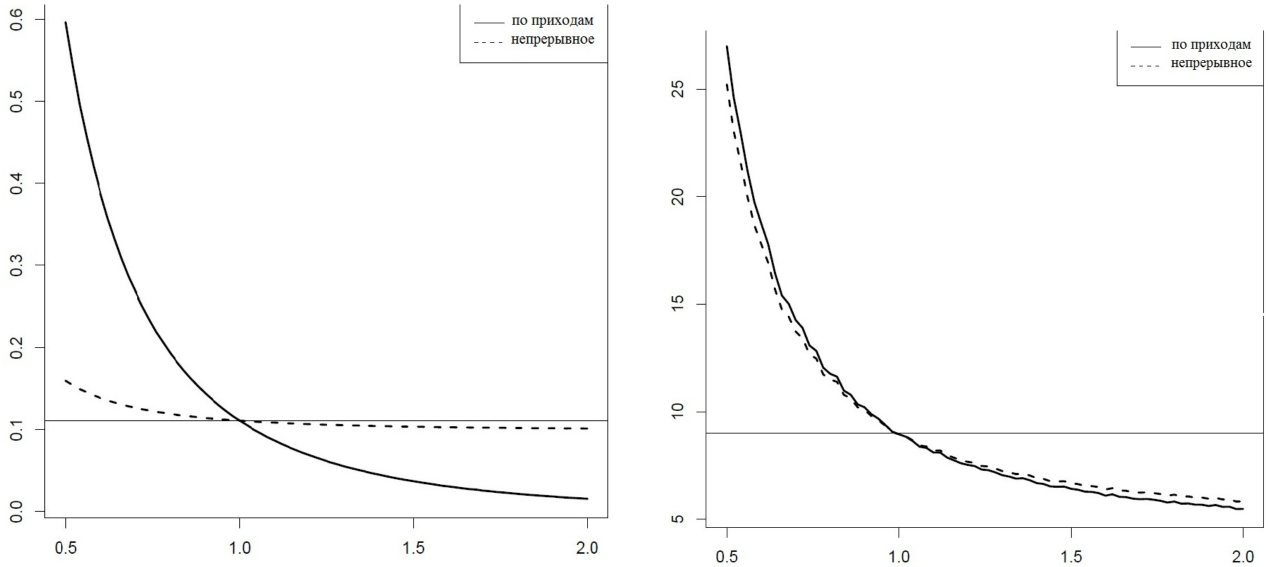

Рис. 3 — Значение Поллачека-Хинчина (прямая) и оценки средней стационарной нагрузки $\hat{W}(\infty)$, $\hat{W}_\infty$ в $Weibull/M/1$: мало загруженная система $\rho = 0.1$ (слева) и сильно загруженная система $\rho = 0.9$ (справа).

где Γ – гамма-функция. Требовалось оценить близость значений средней стационарной нагрузки в непрерывном времени $\mathsf{E}W(\infty)$, средней стационарной нагрузки по моментам прихода заявок в систему $\mathsf{E}W_\infty$ и средней нагрузки, рассчитанной по формуле Поллачека-Хинчина. Результаты моделирования представлены на Рис. 4 для слабо и сильно загруженной системы. На графиках $\hat{W}(\infty), \hat{W}_\infty$ обозначают оценки значений $\mathsf{E}W(\infty)$ и $\mathsf{E}W_\infty$ соответственно. Как видно, средняя нагрузка по моментам прихода заявок очень близка к значению Поллачека-Хинчина для любого значения параметра a (распределения Вейбулла) в случае мало загруженной системы (Рис. 3 слева). Это говорит о том, что в системах с небольшими значениями коэффициента загрузки и входным процессом из класса распределений НЛС (НХС) формулу Поллачека-Хинчина можно использовать для оценивания верхней (нижней) границы средней нагрузки, которая характеризует фактическую задержку заявок в системе.

В случае сильно загруженной системы (Рис. 3 справа) разница в значениях Поллачека-Хинчина и построенных оценок значительна, однако обе оценки (для средней нагрузки в непрерывном времени и по моментам прихода заявок) очень близки, что говорит о том, что вместо двух оценок, можно строить лишь (более экономичную) оценку $\hat{W}_\infty$.

В заключении сформулированы основные результаты, полученные в рамках научно-квалификационной работы, и возможности их применения.

Заключение

В диссертации представлен обзор основных результатов теории больших уклонений и регенеративного анализа, а также возможности их использования для оценивания качества обслуживания вероятностных моделей коммуникационных систем. Доказана экспоненциальная асимптотика стационарной вероятности растущего значения очереди в односерверной и многосерверной системах с постоянной и классической интенсивностью повторных вызовов.

На основе теории больших уклонений и имитационного моделирования изучены свойства методов оценивания ЭПС и проведен их сравнительный анализ. Разработан рандомизированный метод группового среднего для оценивания ЭПС, обеспечивающий заданное качество обслуживания системы. Полученные результаты могут быть использованы для оценивания качества обслуживания широкого класса коммуникационных систем. Даны некоторые рекомендации по возможному использованию методов оценивания ЭПС.

Доказаны неравенства типа Поллачека-Хинчина, задающие границы качество обслуживания для систем с входным процессом с распределением специального класса.

Разработан комплекс программ имитационного моделирования для оценивания ЭПС рассмотренными методами.

Список публикаций по теме работы

В изданиях из списка ВАК:

1. Калинина К. А., Морозов Е. В. О рандомизации в методе группового среднего при оценивании эффективной пропускной способности высокоответственных систем/ К. А. Калинина, Е. В. Морозов // Системы и средства информатики. – 2017. – Т. 27. – № 4. – С. 36–52.

В изданиях, индексируемых в реферативной базе Scopus:

2. Kalinina K. A. Inequalities for workload process in queues with NBU/NWU input / K. Kalinina, E. Morozov, A. Rumyantsev // Advances in Intelligent Systems and Computing: Springer. – 2018. – V. 659. – P. 535-544.

3. Kalinina K. Evaluating a single-server queue with asynchronous speed scaling / A. Rumyantsev, P. Zueva, K. Kalinina, A. Golovin // Lecture Notes in Computer Science: Springer. – 2018. V. 10740. – P. 157-172.

4. Kalinina K. On the accuracy of the effective bandwidth regenerative estimation / A. Borodina, K. Kalinina., E. Morozov // Proceedings of International Congress on Ultra Modern Telecommunications and Control systems and Workshops. – 2014. – P. 652-656.

5. Kalinina K. Effective bandwidth estimation in highly reliable regenerative networks / K. Kalinina K., E. Morozov, V. Rykov // Proceedings of the Second International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management. – 2016. – P. 323-327.

6. Kalinina K. On the effective bandwidth estimation in communication network / E. Morozov, K. Kalinina // Proceedings of the 29th European Conference on Modelling and Simulation. – 2015. – P. 423-429.

В других изданиях:

7. Калинина К.А. Стохастическое моделирование вычислительного кластера с гистерезисным управлением скоростью обслуживания / А. С. Румянцев, К. А. Калинина, Т. Е. Морозова // Труды КарНЦ РАН. Сер. Математическое моделирование и информационные технологии. – 2017. – №8. – С. 76-85.

8. Калинина К. Об эффективной пропускной способности узлов коммуникационной сети // Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем: материалы Всероссийской конференции с международным участием.— Москва: РУДН. – 2015. – С. 24-26.

9. Kalinina K. Effective bandwidth estimation in the regenerative networks / K. Kalinina, E. Morozov // Proceedings of the 18th international scientific conference DCCN-2015, Moscow, ICS RAN. – 2015. – P. 331-333.

10. Kalinina K. Effective bandwidth estimation in highly critical systems / K. Kalinina // Proceedings of the 19th International Conference on distributed computer and communication networks. – 2016. – P. 215-218.

11. Калинина К. Влияние случайного суммирования на регенеративную оценку эффективной пропускной способности узла сети / К. Калинина // Расширенные тезисы IX международной Петрозаводской конференции «Вероятностные методы в дискретной математике», Петрозаводск, Россия. – 2016. – С. 29-31.

12. Калинина К. О рандомизации в методах оценивания ЭПС в высокоответственных системах / К. Калинина // Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2017): материалы Двадцатой междунар. науч. конфер. Москва: М.: Техносфера. – 2017. – С. 407–410.

Свидетельство о регистрации программы:

13. Калинина К. А. Программа для ЭВМ «Регенеративное оценивание эффективной пропускной способности узла инфо-коммуникационной сети». — Свидетельство о государственной регистрации программы для ЭВМ №2016612795 от 10.03.2016.