

Стохастические системы

© 2020 г. В.М. ВИШНЕВСКИЙ, д-р техн. наук (vishn@ipu.ru)
(Институт проблем управления им. В.А. Трапезникова РАН, Москва),
К.Е. САМУЙЛОВ, д-р техн. наук (ksam@sci.pfu.edu.ru),
Н.В. ЯРКИНА, канд. физ.-мат. наук (natyarkina@sci.pfu.edu.ru)
(Российский университет дружбы народов, Москва)

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ СОТЫ LTE С ТРАФИКОМ МЕЖМАШИНЫХ И ШИРОКОПОЛОСНЫХ КОММУНИКАЦИЙ¹

Представлена система массового обслуживания с эластичными и неэластичными заявками для анализа совместной передачи трафика межмашинных и широкополосных коммуникаций в сети LTE. Поступающие в систему эластичные заявки образуют марковский поток и обслуживаются в соответствии с дисциплиной справедливого разделения процессора на выделяемых блоками ресурсах. Получена система уравнений равновесия для расчета стационарного распределения вероятностей состояний и выражения основных вероятностно-временных характеристик. На примере рассматриваемой системы исследована связь характеристик производительности системы и коэффициента корреляции последовательных интервалов между поступлениями заявок в марковском потоке.

Ключевые слова: разделение ресурсов, марковский поток, межмашинные коммуникации, интернет вещей, MAP, IoT, mMTC, LTE.

DOI: 10.31857/S0005231020040054

1. Введение

Дальнейшее развитие сетей сотовой подвижной связи направлено на решение двух задач: расширение возможностей существующих систем по широкополосной передаче данных (mobile broad band, МВВ) и обеспечение комплексной инфраструктуры для межмашинных коммуникаций (machine-type communications, МТС) как для потребительских, так и для промышленных нужд, призванной стать основой для полноценного развертывания Интернета вещей (Internet of Things, IoT) [1]. Таким образом, в одной сети связи необходимо организовать эффективную передачу трафика двух типов с принципиально различными характеристиками и требованиями к качеству обслуживания. Действительно, если трафик МВВ, как правило, связан с услугами телефонии и мультимедиа и характеризуется продолжительными соединениями и передачей больших объемов данных с малыми задержками, то для трафика МТС характерна передача пакетов данных малого размера от чрезвычайно большого количества устройств.

¹ Публикация подготовлена при финансовой поддержке Минобрнауки России (проект № 2.882.2017/4.6).

Поддержка двух принципиально различных типов трафика, необходимость реализации которой возникла в сетях Long Term Evolution (LTE), по-прежнему ставит задачи оптимизации ресурсов соты, причем как канальных, так и сигнализации. Например, в статье [2] построена вероятностная модель для анализа эффективности механизма доступа устройств МТС в соту LTE. Пропускная способность соты здесь оценивается не с точки зрения канальных ресурсов, а как среднее число успешных попыток получить доступ в сеть за единицу времени. Модель позволяет настроить параметры стандартных процедур радиодоступа LTE таким образом, чтобы максимизировать пропускную способность соты.

В [3] предложена математическая модель для анализа совместной передачи межмашинного и широкополосного трафика в соте сети пятого поколения с поддержкой механизмов нарезки (slicing) сети доступа. Модель построена в виде системы массового обслуживания и учитывает потери запросов как из-за нехватки канальных ресурсов, так и из-за коллизий на этапах установления соединения. Численный анализ модели показывает, что основные потери в соте вызваны именно нехваткой сигнальных ресурсов для обслуживания большого числа устройств МТС.

Одним из возможных способов избежать перегрузок на уровне радиодоступа при обслуживании очень большого числа устройств МТС и обеспечить покрытием труднодоступные места (например, подвальные помещения) является агрегация трафика МТС посредством кластеризации (группировки) устройств, при которой один или несколько узлов сети, называемых агрегаторами, осуществляют передачу данных на базовую станцию от группы устройств МТС [4, 5]. В [6, 7] в роли агрегаторов предложено использовать устройства МТС, первыми установившие соединение с базовой станцией.

В настоящей статье предложена модель для анализа разделения канальных ресурсов соты сети, обслуживающей трафик МВВ и МТС с агрегацией трафика МТС и резервированием ресурсов. Модель построена в виде системы массового обслуживания (СМО) с двумя типами заявок. Заявки, соответствующие запросам от устройств МТС, образуют марковский входящий поток [8, 9] и обслуживаются в соответствии с дисциплиной справедливого разделения процессора на выделяемых блоками ресурсах. Поскольку скорость обслуживания заявок данного типа зависит от числа заявок в системе, называем их эластичными.

На численном примере исследовано поведение системы с входящими потоками некоторых специальных видов и связь корреляции длин последовательных интервалов между поступлениями заявок марковского потока с характеристиками системы. Существенное влияние корреляции во входящем потоке на характеристики функционирования СМО отмечено в [10, 11], где при численном анализе ряда систем замечено ухудшение показателей производительности с ростом коэффициента корреляции (см., например, [11, с. 290]). В настоящей статье показано, что коэффициент корреляции во входящем марковском потоке не обязательно отражает характер распределения поступлений заявок во времени, вызывающий ухудшение показателей производительности СМО.

Статья организована следующим образом. В разделе 2 производится построение СМО, описаны входящие потоки и особенности выделения ресурсов заявкам разных типов. В разделе 3 получена система уравнений равновесия (СУР) для нахождения стационарного распределения вероятностей состояний системы. Эффективный алгоритм для решения СУР представлен в разделе 4. В разделе 5 даны выражения для основных вероятностно-временных характеристик СМО. В разделе 6 представлены результаты численного анализа.

2. Построение системы массового обслуживания

Рассмотрим соту сети LTE, в которой осуществляется предоставление двух типов услуг связи: передача потоков данных МВВ, например телефония, и передача блоков данных от устройств МТС, например счетчиков потребления электроэнергии. Для построения математической модели такой соты сделаем следующие упрощающие предположения. Пусть все оконечные устройства не меняют своего положения относительно базовой станции и имеют одинаковое значение отношения сигнал/шум. Таким образом, все устанавливаемые в соте радиоканалы имеют одинаковые характеристики, и скорость передачи данных определяется лишь количеством выделенных единиц канального ресурса. Под единицей (канального) ресурса будем понимать условную величину, соответствующую минимально допустимой скорости передачи данных (например, в кбит/с) для заданного количества выделенных физических ресурсных блоков.

С ростом предложенной нагрузки планировщик на базовой станции LTE определяет оптимальный размер диапазона выделяемых радиоресурсов, исходя из установленных оператором сети требований к качеству обслуживания. Будем считать, что при поступлении запросов на МВВ-соединения планировщик выделяет канальные ресурсы оконечному устройству на все время сеанса связи (телефонного разговора). Передача же МТС-данных осуществляется через агрегатора следующим образом: при поступлении первого запроса на МТС-передачу агрегатор соединяется с базовой станцией и обслуживание последующих МТС-запросов производится по уже установленному каналу. Если количество одновременно обслуживаемых посредством одного канала МТС-устройств не позволяет соблюсти требования к качеству обслуживания, то агрегатор запрашивает установление еще одного канала с теми же характеристиками. При отсутствии активных соединений с МТС-устройствами канал между базовой станцией и агрегатором разрывается.

Для того чтобы сделать сети LTE эффективным механизмом обслуживания МТС-трафика, необходимы методы распределения радиоресурсов, минимизирующие влияние МТС на качество обслуживания традиционных абонентов, приносящих пока основной доход операторам сетей связи. Поэтому будем считать, что часть ресурсов соты доступна только трафику МВВ и не может использоваться для соединений с агрегатором МТС.

Рассмотрим изображенную на рис. 1 многолинейную СМО, обслуживающую заявки двух типов. Заявки первого типа – эластичные – обслуживаются с переменной скоростью, зависящей от числа заявок данного типа в систе-

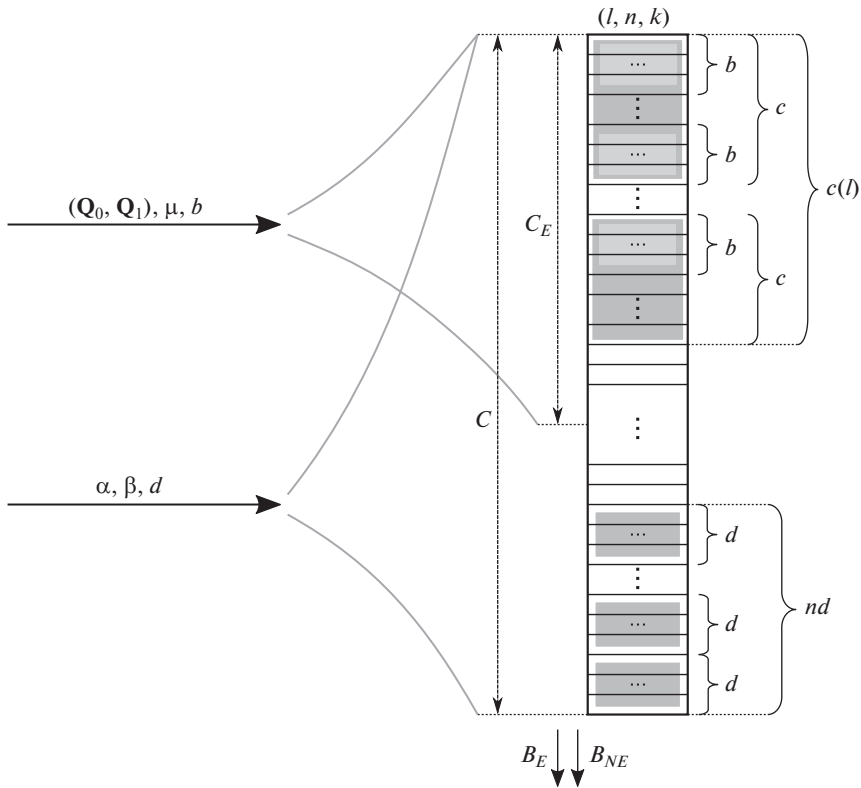


Рис. 1. Система массового обслуживания с эластичными и неэластичными заявками.

ме, и описывают соединения МТС. Заявки второго типа – неэластичные – обслуживаются с постоянной скоростью и соответствуют соединениям МВВ.

Поток эластичных заявок является марковским, или МАР-поток (Markovian Arrival Process, подробнее см., например, [8–10]), и задан двумя квадратными матрицами Q_0 и Q_1 порядка K , $Q_1 \geq 0$. Матрица $Q = Q_0 + Q_1$ представляет собой инфинитезимальную матрицу управляющей марковским потоком цепи Маркова (ЦМ) с непрерывным временем $\{\xi(t), t \geq 0\}$, причем Q_0 содержит интенсивности ее переходов без генерации заявок, тогда как Q_1 описывает переходы с генерацией заявок. Будем считать, что матрица Q_1 отлична от нулевой, а управляющая цепь $\{\xi(t), t \geq 0\}$ является неприводимой. Обозначим через $\mathbf{1}$ вектор-столбец из единиц и через $\mathbf{0}$ – вектор-строку из нулей соответствующей контексту размерности. Вектор-строка \mathbf{q} стационарных вероятностей состояний ЦМ $\{\xi(t), t \geq 0\}$ находится как единственное решение системы линейных алгебраических уравнений $\mathbf{q}Q = \mathbf{0}$, $\mathbf{q}\mathbf{1} = 1$. Средняя интенсивность потока дается выражением

$$(1) \quad \lambda = \mathbf{q}Q_1\mathbf{1},$$

дисперсия длин интервалов между поступлениями равна

$$(2) \quad \sigma^2 = \frac{2}{\lambda} \mathbf{q}(-Q_0)^{-1}\mathbf{1} - \frac{1}{\lambda^2},$$

а коэффициент корреляции длин последовательных интервалов между поступлениями находится по формуле

$$(3) \quad r = \frac{\frac{1}{\lambda} - \mathbf{qQ}_0^{-1}\mathbf{Q}_1\mathbf{Q}_0^{-1}\mathbf{1}}{\frac{1}{\lambda} + 2\mathbf{qQ}_0^{-1}\mathbf{1}}.$$

Длины эластичных заявок (т.е. требуемые длительности обслуживания с единичной скоростью) являются независимыми случайными величинами, распределенными по экспоненциальному закону с параметром μ . Во время своего пребывания в системе каждая эластичная заявка занимает не менее $b > 0$ единиц ресурса.

Заявки второго типа – неэластичные – образуют пуассоновский поток интенсивности α и обслуживаются с постоянной скоростью в течение случайного времени, распределенного по экспоненциальному закону с параметром β . Каждая принятая на обслуживание неэластичная заявка получает ровно d единиц ресурса системы.

Пусть общий объем ресурсов СМО для обслуживания заявок равен C , из которых $C_E \leq C$ единиц могут выделяться заявкам обоих типов, а $C - C_E$ единиц доступны только неэластичным заявкам (часть ресурсов резервируется для трафика МВВ). Для обслуживания эластичных заявок ресурсы выделяются блоками по $c \geq b$ единиц, что соответствует выделению ресурсов агрегатору. Обозначим через

$$M = \lfloor c/b \rfloor = \max \{y \in \mathbb{N} : y \leq c/b\}$$

максимальное число эластичных заявок, которые могут обслуживаться одновременно одним блоком ресурсов размера c . Тогда если в системе находятся l эластичных заявок, то объем занимаемых ими ресурсов равен

$$c(l) = c \cdot \lceil l/M \rceil = c \cdot \min \{y \in \mathbb{N} : y \geq l/M\}.$$

Этот объем ресурсов поровну делится между эластичными заявками, находящимися в СМО, т.е. скорость обслуживания каждой эластичной заявки равна $\frac{c(l)}{l}$.

Далее, если в СМО, в которой обслуживаются l эластичных заявок, поступает $(l + 1)$ -я эластичная заявка, то возможен один из следующих вариантов:

- если в момент поступления заявки $c(l + 1) = c(l)$, то заявка принимается на обслуживание без выделения дополнительных ресурсов, при этом объем ресурсов, занятых эластичными заявками, перераспределяется поровну между $l + 1$ заявками и скорость обслуживания снижается;

- если в момент поступления заявки $c(l) < c(l + 1) \leq C_E$ и в СМО имеются s свободных единиц ресурса, то заявка принимается на обслуживание с выделением ресурсного блока размером s единиц, при этом $c(l + 1)$ единиц ресурса поровну перераспределяются между эластичными заявками в системе и скорость их обслуживания возрастает;

- если в момент поступления заявки $c(l + 1) > c(l)$ и при этом $c(l + 1) > C_E$ и/или в СМО нет свободных s единиц ресурса, то поступившая заявка теряется, не оказывая влияния на дальнейшее функционирование СМО.

Принятая в СМО эластичная заявка обслуживается с переменной скоростью до тех пор, пока ее остаточная длина не будет равна нулю, после чего покидает систему. В момент ухода эластичной заявки ресурсы перераспределяются между оставшимися $l - 1$ эластичными заявками следующим образом:

- если $c(l - 1) = c(l)$, то ресурсы не высвобождаются и $c(l)$ единиц ресурсов поровну делятся между оставшимися $l - 1$ эластичными заявками;
- если $c(l - 1) < c(l)$, то ресурсный блок объемом c единиц высвобождается и $c(l - 1)$ единиц ресурса системы перераспределяется поровну между оставшимися $l - 1$ эластичными заявками.

Неэластичная заявка принимается на обслуживание, если в момент ее поступления в СМО имеется хотя бы d свободных единиц ресурса. Заявка занимает этот объем ресурсов на время обслуживания и высвобождает в момент выхода из СМО. Если в момент поступления неэластичной заявки объем свободных ресурсов системы меньше d , то заявка теряется, не оказывая влияния на дальнейшее функционирование системы.

3. Стационарное распределение вероятностей состояний

Обозначим через $S = \lfloor C_E/c \rfloor$ максимальное число ресурсных блоков, которые могут быть выделены эластичным заявкам. Тогда максимально возможное число эластичных заявок в СМО равно $L = MS$. Введем еще обозначение для максимального числа неэластичных заявок в системе, в которой уже обслуживаются l эластичных заявок:

$$(4) \quad N(l) = \left\lfloor \frac{C - c(l)}{d} \right\rfloor, \quad 0 \leq l \leq L.$$

Теперь пусть $l(t) \in \{0, \dots, L\}$ и $n(t) \in \{0, \dots, N(0)\}$ — соответственно числа эластичных и неэластичных заявок в СМО, а $k(t) \in \{1, \dots, K\}$ — состояние управляющей марковским потоком ЦМ в момент времени $t \geq 0$. Тогда стохастическое поведение рассматриваемой системы можно описать трехмерной ЦМ с непрерывным временем $\{X(t) = (l(t), n(t), k(t)), t \geq 0\}$ над пространством состояний

$$(5) \quad \mathcal{X} = \{(l, n, k) : 0 \leq l \leq L, n \geq 0, c(l) + nd \leq C, 1 \leq k \leq K\}.$$

Расположим состояния ЦМ $\{X(t), t \geq 0\}$ в лексикографическом порядке и обозначим через \mathbf{A} ее матрицу интенсивностей переходов. Обозначим через \mathbf{I} единичную матрицу порядка K и доопределим функцию $N(l)$, заданную в (4), положив $N(l) = N(L)$ при $l > L$.

Лемма 1. Инфинитезимальная матрица \mathbf{A} имеет блочно-трехдиагональную структуру:

$$(6) \quad \mathbf{A} = \begin{bmatrix} \mathbf{D}_0 & \Lambda_0 & & & \\ \mathbf{M}_1 & \mathbf{D}_1 & \Lambda_1 & & \mathbf{0} \\ & \mathbf{M}_2 & \ddots & \ddots & \\ & & \ddots & \mathbf{D}_{L-1} & \Lambda_{L-1} \\ \mathbf{0} & & & \mathbf{M}_L & \mathbf{D}_L \end{bmatrix}.$$

Блоки, расположенные на пересечении i -й блочной строки и j -го блочного столбца, представляют собой блочные матрицы блочных размеров $(N(i-1)+1) \times (N(j-1)+1)$, составленные из квадратных матриц порядка K . Диагональные блоки матрицы \mathbf{A} имеют вид

$$(7) \quad \mathbf{D}_l = \begin{cases} \mathbf{D}_{l,1} & \text{при } N(l) = N(l+1), \\ [\mathbf{D}_{l,1} \quad \mathbf{D}_{l,2}] & \text{при } N(l) > N(l+1), \end{cases}$$

$$\mathbf{D}_{l,1} = \begin{bmatrix} \mathbf{F}_{l,0} & \alpha \mathbf{I} & & & & \\ \beta \mathbf{I} & \mathbf{F}_{l,1} & \alpha \mathbf{I} & & & \mathbf{0} \\ & 2\beta \mathbf{I} & \ddots & & \ddots & \\ & & \ddots & \mathbf{F}_{l,N(l+1)-1} & & \alpha \mathbf{I} \\ & & & N(l+1)\beta \mathbf{I} & & \mathbf{F}_{l,N(l+1)} \\ \mathbf{0} & & & & & (N(l+1)+1)\beta \mathbf{I} \\ & & & & & \mathbf{0} \end{bmatrix}, \quad 0 \leq l < L,$$

$$\mathbf{D}_{l,2} = \begin{bmatrix} \mathbf{0} & & & & & & \\ & \alpha \mathbf{I} & & & & & \\ \mathbf{G}_{l,N(l+1)+1} & & \alpha \mathbf{I} & & & & \mathbf{0} \\ (N(l+1)+2)\beta \mathbf{I} & & \mathbf{G}_{l,N(l+1)+2} & \alpha \mathbf{I} & & & \\ & & (N(l+1)+3)\beta \mathbf{I} & \ddots & & \ddots & \\ & & & \ddots & \mathbf{G}_{l,N(l)-1} & & \alpha \mathbf{I} \\ & & & \mathbf{0} & & N(l)\beta \mathbf{I} & \mathbf{G}_{l,N(l)} \end{bmatrix}, \quad 0 \leq l < L,$$

$$\mathbf{D}_{L,1} = \begin{bmatrix} \mathbf{G}_{L,0} & \alpha \mathbf{I} & & & & \\ \beta \mathbf{I} & \mathbf{G}_{L,1} & \alpha \mathbf{I} & & & \mathbf{0} \\ & 2\beta \mathbf{I} & \ddots & & \ddots & \\ & & \ddots & \mathbf{G}_{L,N(L)-1} & & \alpha \mathbf{I} \\ \mathbf{0} & & & N(L)\beta \mathbf{I} & & \mathbf{G}_{L,N(L)} \end{bmatrix},$$

где

$$\mathbf{F}_{l,j} = \begin{cases} \mathbf{Q}_0 - (\alpha + j\beta + c(l)\mu)\mathbf{I} & \text{при } j \neq N(l), \\ \mathbf{Q}_0 - (N(l)\beta + c(l)\mu)\mathbf{I} & \text{при } j = N(l) \end{cases}$$

и

$$\mathbf{G}_{l,j} = \mathbf{F}_{l,j} + \mathbf{Q}_1.$$

Под- и наддиагональные блоки матрицы \mathbf{A} являются блочными диагональными прямоугольными матрицами вида

$$(8) \quad \mathbf{\Lambda}_l = \begin{bmatrix} \mathbf{Q}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{Q}_1 \\ \mathbf{0} & \dots & \mathbf{0} \end{bmatrix},$$

$$(9) \quad \mathbf{M}_l = \begin{bmatrix} c(l)\mu\mathbf{I} & & \mathbf{0} & \mathbf{0} \\ & \ddots & & \vdots \\ \mathbf{0} & & c(l)\mu\mathbf{I} & \mathbf{0} \end{bmatrix}.$$

Доказательство. Доказательство основано на анализе переходов ЦМ $\{X(t), t \geq 0\}$ за бесконечно малый интервал времени. Переходы происходят в результате наступления следующих событий:

— Принятие на обслуживание эластичной заявки. Поступление эластичных заявок происходит при переходах управляющей ЦМ $\{\xi(t), t \geq 0\}$, сопровождающихся генерацией заявок. Интенсивности таких переходов содержит матрица \mathbf{Q}_1 . В матрице \mathbf{A} данные интенсивности объединены в блоки $\mathbf{\Lambda}_l$, $l = 0, \dots, L - 1$, имеющие блочные размеры $(N(l) + 1) \times (N(l + 1) + 1)$ и, следовательно, являющиеся квадратными при $N(l + 1) = N(l)$. Ненулевые внедиагональные элементы $\mathbf{\Lambda}_l$ являются интенсивностями переходов из состояния (l, n, k) в состояние $(l + 1, n, k^*)$, где $k \neq k^*$. Диагональные элементы $\mathbf{\Lambda}_l$ соответствуют переходам из (l, n, k) в $(l + 1, n, k)$.

— Окончание обслуживания эластичной заявки вызывает переход из состояния (l, n, k) в состояние $(l - 1, n, k)$. Интенсивности данных переходов $c(l)\mu$ объединены в блоки \mathbf{M}_l , $l = 1, \dots, L$, блочного размера $(N(l) + 1) \times (N(l - 1) + 1)$. Блоки являются диагональными матрицами, квадратными при $N(l) = N(l - 1)$.

— Принятие на обслуживание неэластичной заявки вызывает переход из состояния (l, n, k) в состояние $(l, n + 1, k)$. Интенсивности данных переходов α содержат наддиагональные подблоки $\alpha\mathbf{I}$ блоков \mathbf{D}_l , $l = 0, \dots, L$.

— Окончание обслуживания неэластичной заявки вызывает переход из (l, n, k) в $(l, n - 1, k)$. Интенсивности данных переходов $n\beta$ объединены в поддиагональные подблоки блоков \mathbf{D}_l , $l = 0, \dots, L$.

— Переход ЦМ $\{\xi(t), t \geq 0\}$ без генерации заявки (интенсивности \mathbf{Q}_0) или с генерацией заявки, но в таком состоянии системы, где нет свободных ресурсов для ее принятия на обслуживание (интенсивности \mathbf{Q}), вызывает переход процесса $\{X(t), t \geq 0\}$ из состояния (l, n, k) в (l, n, k^*) , $k^* \neq k$. Данные интенсивности содержатся в диагональных подблоках блоков \mathbf{D}_l , $l = 0, \dots, L$, причем подблоки $\mathbf{F}_{l,j}$ соответствуют переходам $\{\xi(t), t \geq 0\}$ без генерации заявки в системе, где достаточно ресурсов для принятия заявки, тогда как подблоки $\mathbf{G}_{l,j}$ соответствуют любым переходам $\{\xi(t), t \geq 0\}$ в системе, где ресурсов для принятия эластичной заявки недостаточно.

Диагональные элементы матрицы \mathbf{A} отрицательны, и их абсолютные значения соответствуют интенсивностям выхода процесса $\{X(t), t \geq 0\}$ из соответствующего состояния. Поскольку за бесконечно малый промежуток време-

ни не может произойти более одного из перечисленных выше событий, остальные элементы матрицы \mathbf{A} равны нулю. Лемма 1 доказана.

Так как при сделанных предположениях ЦМ $\{X(t), t \geq 0\}$ регулярна и неприводима, а ее пространство состояний \mathcal{X} конечно, то существует стационарное распределение вероятностей состояний данного процесса. Представим его в векторном виде в соответствии с разбиением (6) матрицы интенсивностей переходов \mathbf{A} на блоки: $\mathbf{p} = (\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_L)$, где $\mathbf{p}_l = (\mathbf{p}_{l,0}, \mathbf{p}_{l,1}, \dots, \mathbf{p}_{l,N(l)})$, $l = 0, \dots, L$, и $\mathbf{p}_{l,n} = (p_{l,n,1}, p_{l,n,2}, \dots, p_{l,n,K})$, $n = 0, \dots, N(l)$. Стационарные вероятности \mathbf{p} удовлетворяют системе уравнений равновесия

$$(10) \quad \mathbf{p}\mathbf{A} = \mathbf{0}$$

с условием нормировки $\mathbf{p}\mathbf{1} = 1$. В случаях когда размерность данной системы невелика, ее легко решить стандартными методами с помощью компьютера. В противном случае можно воспользоваться специальным эффективным алгоритмом, приведенным в следующем разделе и представляющим собой частный случай блочного метода исключения Гаусса.

4. Алгоритм вычисления стационарного распределения

Разобьем матрицы $\mathbf{\Lambda}_l$, $l = 0, \dots, L - 1$, в соответствии с разбиением нижележащих блоков \mathbf{D}_{l+1} следующим образом:

$$(11) \quad \mathbf{\Lambda}_l = \begin{cases} \mathbf{\Lambda}_{l,1} & \text{при } N(l+1) = N(l+2), \\ [\mathbf{\Lambda}_{l,1} \quad \mathbf{\Lambda}_{l,2}] & \text{при } N(l+1) > N(l+2), \end{cases}$$

$$\mathbf{\Lambda}_{l,1} = \begin{bmatrix} \mathbf{Q}_1 & & 0 \\ & \ddots & \\ & & \mathbf{Q}_1 \\ & & & 0 \end{bmatrix}, \quad \mathbf{\Lambda}_{l,2} = \begin{bmatrix} 0 \\ \mathbf{Q}_1 \\ \ddots \\ 0 \quad \mathbf{Q}_1 \end{bmatrix},$$

где $\mathbf{\Lambda}_{l,1}$ имеет $N(l+1) + 1$ блочных столбцов.

Лемма 2. Компоненты \mathbf{p}_l , $l = 0, \dots, L$, вектора – решения СУР (10) связаны соотношениями

$$(12) \quad \mathbf{p}_l = \mathbf{p}_0 \mathbf{H}_l, \quad l = 1, \dots, L,$$

где матрицы \mathbf{H}_l , $l = 1, \dots, L$, размеров $K(N(0) + 1) \times K(N(l) + 1)$ вычисляются рекуррентно по формулам:

$$\mathbf{H}_0 = \mathbf{I},$$

$$\mathbf{H}_1 = -\frac{1}{c(1)\mu} \mathbf{D}_{0,1},$$

$$\mathbf{H}_{l+1} = -\frac{1}{c(l+1)\mu} (\mathbf{H}_{l-1} \mathbf{\Lambda}_{l-1,1} + \mathbf{H}_l \mathbf{D}_{l,1}), \quad l = 1, 2, \dots, L - 1.$$

Вектор \mathbf{p}_0 является единственным решением системы линейных уравнений

$$(13) \quad \mathbf{p}_0 \mathbf{Z} = \mathbf{z},$$

в которой вектор-строка \mathbf{z} получается заменой последних K элементов нулевого вектора-строки длины $K(N(0) + 1)$ вектором \mathbf{q} , тогда как матрица \mathbf{Z} является невырожденной квадратной матрицей порядка $K(N(0) + 1)$ и вида $\mathbf{Z} = [\mathbf{Z}_0, \mathbf{Z}_1, \dots, \mathbf{Z}_L]$, где $\mathbf{Z}_0 = \mathbf{D}_{0,2}$, $\mathbf{Z}_l = \mathbf{H}_{l-1} \mathbf{\Lambda}_{l-1,2} + \mathbf{H}_l \mathbf{D}_{l,2}$, $l = 1, \dots, L-1$, а матрица \mathbf{Z}_L получена заменой последних K столбцов $\mathbf{H}_{L-1} \mathbf{\Lambda}_{L-1} + \mathbf{H}_L \mathbf{D}_L$ матрицей $\sum_{l=0}^L \mathbf{H}_l (\mathbf{1} \otimes \mathbf{I})$. (Символ \otimes здесь обозначает произведение Кронекера.)

Доказательство. Используя блочную структуру матрицы \mathbf{A} и учитывая специальный вид ее поддиагональных блоков \mathbf{M}_l , систему уравнений (10) можно записать в виде:

$$(14a) \quad \mathbf{p}_0 \mathbf{D}_{0,1} + c(1) \mu \mathbf{p}_1 = \mathbf{0},$$

$$(14b) \quad \mathbf{p}_0 \mathbf{D}_{0,2} = \mathbf{0},$$

$$(14c) \quad \mathbf{p}_{l-1} \mathbf{\Lambda}_{l-1,1} + \mathbf{p}_l \mathbf{D}_{l,1} + c(l+1) \mu \mathbf{p}_{l+1} = \mathbf{0}, \quad l = 1, \dots, L-1,$$

$$(14d) \quad \mathbf{p}_{l-1} \mathbf{\Lambda}_{l-1,2} + \mathbf{p}_l \mathbf{D}_{l,2} = \mathbf{0}, \quad l = 1, \dots, L-1,$$

$$(14e) \quad \mathbf{p}_{L-1} \mathbf{\Lambda}_{L-1,2} + \mathbf{p}_L \mathbf{D}_L = \mathbf{0}.$$

Из уравнений (14a) и (14c) данной системы получаются рекуррентные соотношения (12), тогда как уравнения (14b), (14d) и (14e) составляют систему (13) с условием нормировки $\sum_{l=0}^L \sum_{n=0}^{N(l)} p_{l,n,k} = \mathbf{q}_k$, $k = 1, \dots, K$. Лемма 2 доказана.

5. Вероятностно-временные характеристики

Зная стационарное распределение вероятностей состояний системы, легко получить набор важных для приложений вероятностно-временных характеристик. В частности, вероятности потерь эластичных и неэластичных заявок равны соответственно:

$$(15) \quad B_E = \frac{1}{\lambda} \left(\sum_{l=0}^{L-1} \sum_{n=N(l+1)+1}^{N(l)} \mathbf{p}_{l,n} \mathbf{Q}_1 \mathbf{1} + \sum_{n=1}^{N(L)} \mathbf{p}_{L,n} \mathbf{Q}_1 \mathbf{1} \right),$$

$$(16) \quad B_{NE} = \sum_{l=0}^L \mathbf{p}_{l,N(l)} \mathbf{1}.$$

Среднее число занятых единиц ресурса системы можно получить по формуле

$$(17) \quad \bar{c} = \sum_{l=0}^L \sum_{n=0}^{N(l)} (dn + c(l)) \mathbf{p}_{l,n} \mathbf{1}.$$

Средние числа эластичных и неэластичных заявок в СМО даются соответственно выражениями:

$$(18) \quad N_E = \sum_{l=1}^L \sum_{n=0}^{N(l)} l \mathbf{p}_{l,n} \mathbf{1},$$

$$(19) \quad N_{NE} = \sum_{l=0}^L \sum_{n=1}^{N(l)} n \mathbf{p}_{l,n} \mathbf{1}.$$

Для нахождения среднего времени обслуживания эластичной заявки применима формула Литтла:

$$(20) \quad T_E = \frac{N_E}{\lambda(1 - B_E)}.$$

6. Численный анализ

В данном разделе рассмотрим систему со следующими структурными параметрами: $C = 200$, $C_E = 180$, $d = 10$ и $b = 1$. Положим $\alpha = 10^{-3}$, $\beta = 10^{-4}$ и $\mu = 0,05$. В качестве входящего потока эластичных заявок рассмотрим примеры потоков, предложенные С. Чакраварти в [10], а именно: некоррелированные пуассоновский, гиперэкспоненциальный и Эрланга, отрицательно коррелированный марковский (МАР-NC) и положительно коррелированный марковский (МАР-PC) с параметрами, приведенными в таблице, а также параметризованный марковский поток вида

$$(21) \quad \mathbf{Q}_0 = \begin{bmatrix} -a & a & 0 & 0 & 0 \\ 0 & -a & a & 0 & 0 \\ 0 & 0 & -a & a & 0 \\ 0 & 0 & 0 & -a & 0 \\ 0 & 0 & 0 & 0 & -100a \end{bmatrix},$$

$$\mathbf{Q}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ a\pi_1 & 0 & 0 & 0 & a(1 - \pi_1) \\ 100a(1 - \pi_2) & 0 & 0 & 0 & 100a\pi_2 \end{bmatrix}.$$

Поток вида (21) (будем кратко обозначать его МАР-PAR), условно говоря, имеет два режима функционирования: как поток Эрланга порядка четыре и интенсивности $a/4$ и как пуассоновский поток интенсивности $100a$. Параметры π_1 и π_2 определяют переходы между режимами: если поток находится в режиме потока Эрланга, то после поступления заявки с вероятностью π_1 поток останется в данном режиме, а с дополнительной вероятностью перейдет в режим пуассоновского потока. Аналогично π_2 является вероятностью после поступления заявки остаться в режиме пуассоновского потока. Марковский поток МАР-PAR построен таким образом, что, варьируя параметры π_1 и π_2 ,

Входящие потоки эластичных заявок

Обозначение	Q_0	Q_1	r	σ
EXP(1)	$[-1]$	$[1]$	0	1
HYPEREXP	$\begin{bmatrix} -1,9 & 0 \\ 0 & -0,19 \end{bmatrix}$	$\begin{bmatrix} 1,71 & 0,19 \\ 0,171 & 0,019 \end{bmatrix}$	0	2,24472
ERLANG	$\begin{bmatrix} -5 & 5 & 0 & 0 & 0 \\ 0 & -5 & 5 & 0 & 0 \\ 0 & 0 & -5 & 5 & 0 \\ 0 & 0 & 0 & -5 & 5 \\ 0 & 0 & 0 & 0 & -5 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \end{bmatrix}$	0	0,44721
MAP-NC	$\begin{bmatrix} -1,00243 & 1,00243 & 0 \\ 0 & -1,00243 & 0 \\ 0 & 0 & -225,797 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0,01002 & 0 & 0,99241 \\ 223,539 & 0 & 2,258 \end{bmatrix}$	-0,48891	1,40952
MAP-PC	$\begin{bmatrix} -1,00243 & 1,00243 & 0 \\ 0 & -1,00243 & 0 \\ 0 & 0 & -225,797 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \\ 0,99241 & 0 & 0,01002 \\ 2,258 & 0 & 223,539 \end{bmatrix}$	0,48891	1,40952

можно варьировать значения коэффициента корреляции (3) при фиксированной средней интенсивности λ . При этом значение параметра a получается из (1) по формуле

$$(22) \quad a = \lambda \frac{0,01(1 - \pi_1) + 4(1 - \pi_2)}{1 - \pi_1 + 1 - \pi_2}.$$

Средняя интенсивность всех потоков таблицы равна единице; σ обозначает среднеквадратичное отклонение длин интервалов между поступлением заявок.

На основе результатов численного анализа в [10, 11] подчеркивается важная роль корреляции (и особенно положительной корреляции) интервалов между поступлением заявок потока, оцениваемой как коэффициент корреляции длин последовательных интервалов (3). С. Чакраварти в [10] на численных примерах показал существенное различие характеристик потока и производительности СМО MAP|M|5 при негативной и позитивной корреляции во входящем MAP-потоке. В [11, с. 280–290, 393–395] для ряда СМО более сложной структуры замечен эффект ухудшения показателей производительности системы с ростом коэффициента корреляции входящего MAP-потока.

На рис. 2 представлена зависимость вероятности потерь (15) эластичных заявок от параметра s . На рис. 2 видно, что вероятности потерь эластичных заявок действительно оказываются существенно выше для входящих марковских потоков с положительным коэффициентом корреляции $r > 0$ (рис. 2,б), тогда как для потоков с отрицательной корреляцией $r < 0$ (рис. 2,а) значения вероятности потерь ниже и близки к вероятности потерь при простейшем входящем потоке той же интенсивности. Здесь поток MAP-PAR имеет характеристики $r = -0,49804$, $\sigma = 1,1129$ при $\pi = (0,21; 0,01)$ (рис. 2,а) и $r = 0,47697$, $\sigma = 6,36608$ при $\pi = (0,61; 0,99)$ (рис. 2,б).

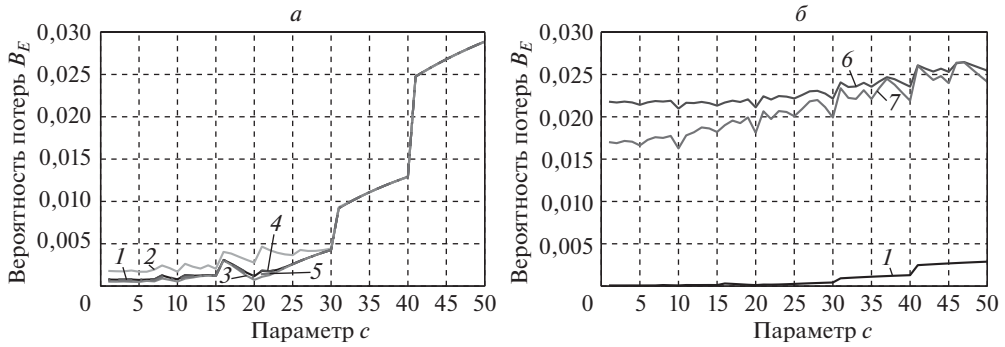


Рис. 2. Вероятность потерь эластичных заявок B_E как функция от c для: 1 – EXP(1), 2 – HYPEREXP, 3 – ERLANG, 4 – MAP-NC, 5 – MAP-PAR при $\pi = (0,21; 0,01)$, 6 – MAP-PC, 7 – MAP-PAR при $\pi = (0,61; 0,99)$.

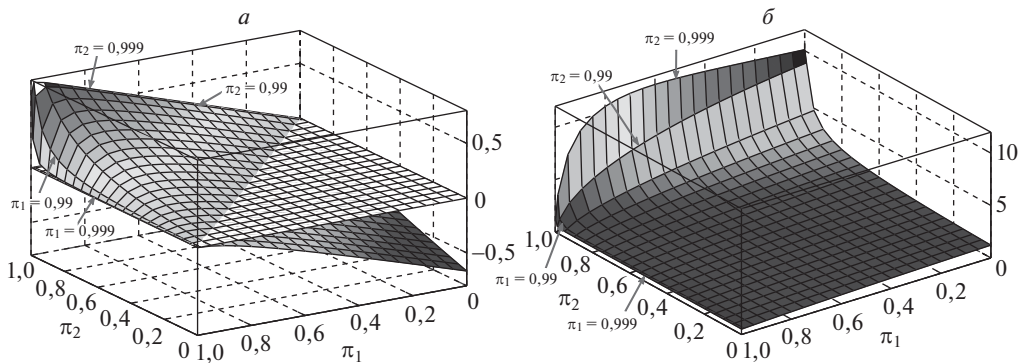


Рис. 3. Характеристики марковского потока MAP-PAR (21) при $\lambda = 1$.

Теперь зафиксируем $c = 15$ и рассмотрим вероятности потерь в СМО с входящим потоком MAP-PAR при различных значениях параметров $0 \leq \pi_1, \pi_2 < 1$. На рис. 3 показаны коэффициент корреляции r (рис. 3,а) и среднеквадратичное отклонение σ (рис. 3,б) для данного потока при $\lambda = 1$. Отметим, что при приближении π_2 к единице имеет место резкий рост дисперсии потока, тогда как коэффициент корреляции плавно возрастает с ростом π_2 , меняя знак при $\pi_2 = 1 - \pi_1$.

На рис. 4 представлена вероятность потерь эластичных заявок B_E как функция от коэффициента корреляции r . Значения обеих характеристик получены для следующей последовательности значений параметра π_2 : 0,1, 0,2, ..., 0,9, 0,91, ..., 0,99, 0,999, 0,99999. Для всех рассмотренных π_1 вероятность потерь резко увеличивается, приближаясь к $\pi_2 = 0,99$, однако падает уже при $\pi_2 = 0,999$. При этом, как видно на рис. 4,а, коэффициент корреляции r монотонно возрастает с ростом π_2 .

Таким образом, как показывает рис. 4, само по себе значение (и даже знак) коэффициента корреляции потока не позволяет для исследуемой СМО предсказать поведение вероятности потерь. В частности, для $\pi_1 = 0$ рост происходит при отрицательных значениях коэффициента корреляции, а при высоких r наблюдаются как высокие так и низкие значения B_E . Вообще го-

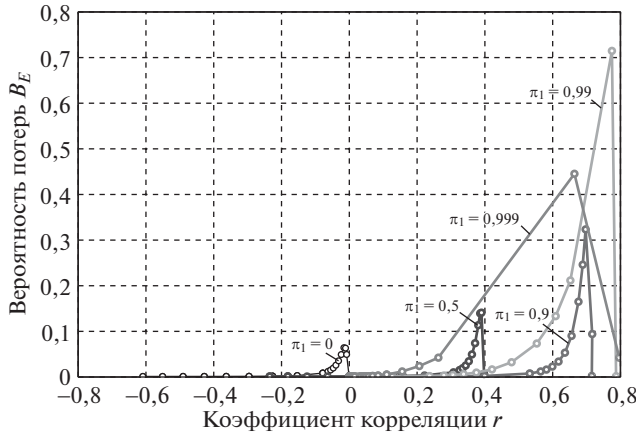


Рис. 4. Вероятность потерь эластичных заявок B_E как функция от коэффициента корреляции r .

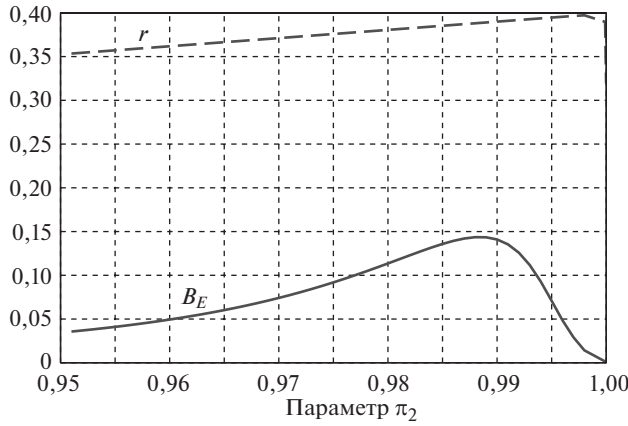


Рис. 5. Зависимость вероятности B_E и коэффициента корреляции r от параметра π_2 при $\pi_1 = 0,5$.

воря, только пользуясь данными графиков рис. 4, можно подобрать такую последовательность пар $(\pi_1; \pi_2)$, на которой коэффициент корреляции будет возрастать, а вероятность потерь – убывать.

На рис. 5 отдельно представлена зависимость вероятности потерь эластичных заявок B_E и коэффициента корреляции r от параметра π_2 при π_2 близких к единице, т.е. в том диапазоне, где вероятность потерь достигает максимума. Здесь рассмотрен случай $\pi_1 = 0,5$. На графике отчетливо видно, что, в частности, на отрезке $[0,99; 0,995]$ вероятность потерь резко убывает, тогда как коэффициент корреляции потока продолжает свой рост.

Заметим, что потоки MAP-NC и MAP-PC, предложенные в [10] для иллюстрации роли коэффициента корреляции, аналогично потоку MAP-PAR (21) представимы в виде

$$\mathbf{Q}_0 = \begin{bmatrix} -a & a & 0 \\ 0 & -a & 0 \\ 0 & 0 & -a\gamma \end{bmatrix}, \quad \mathbf{Q}_1 = \begin{bmatrix} 0 & 0 & 0 \\ a\pi_1 & 0 & a(1 - \pi_1) \\ a\gamma(1 - \pi_2) & 0 & a\gamma\pi_2 \end{bmatrix},$$

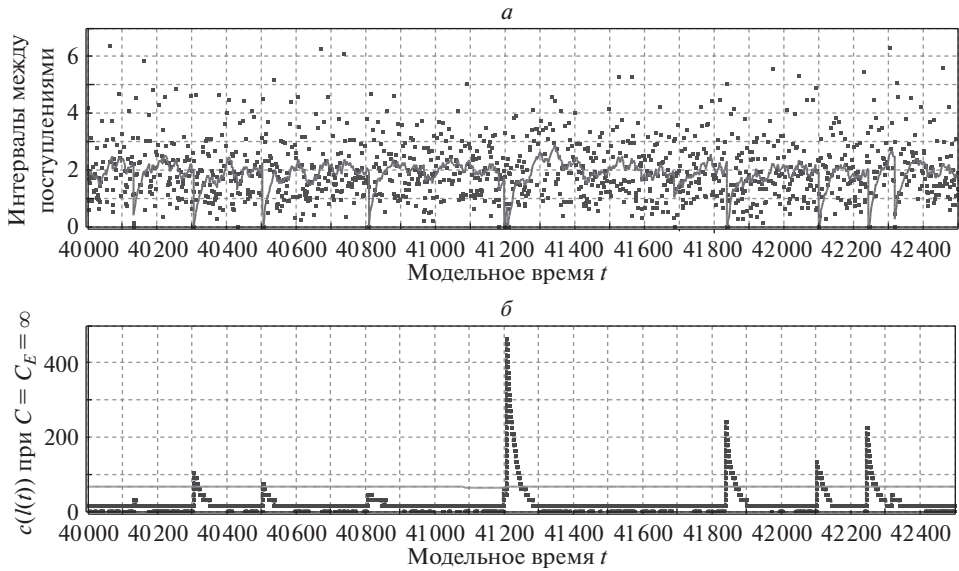


Рис. 6. Нагрузка от эластичного трафика в СМО с MAP-PAR при $(\pi_1, \pi_2) = (0,99; 0,99)$.

при $\gamma = 225,25$, $\pi = (0,01; 0,01)$ в случае MAP-NC, $\pi = (0,99; 0,99)$ в случае MAP-PC. Значение a получается при $\lambda = 1,000212$ по формуле

$$a = \lambda \frac{\gamma^{-1}(1 - \pi_1) + 2(1 - \pi_2)}{1 - \pi_1 + 1 - \pi_2},$$

а зависимости коэффициента корреляции и среднеквадратичного отклонения данного потока от пары параметров (π_1, π_2) качественно повторяют графики рис. 3. Кроме того, выводы в [11] о связи показателей производительности СМО с корреляцией во входящем потоке для MAP-потока сделаны на основе расчетов, выполненных для серии из трех потоков с одинаковой средней интенсивностью и возрастающими коэффициентами корреляции. Эти потоки также имеют по одному относительно большому диагональному элементу в матрице \mathbf{Q}_1 , причем данные элементы, как и корреляция, возрастают от первого потока к третьему.

Чтобы лучше понять природу колебаний нагрузки, приводящих к росту вероятности потерь в системе, и выявить их связь с корреляцией в потоке, обратимся к результатам имитационного моделирования, представленным на рис. 6–8. Результаты получены с помощью специально разработанной в среде OMNeT++ модели рассматриваемой СМО, которая подробно описана в [12]. На графиках показаны фрагменты реализации имитационной модели, а именно: интервалы между поступлениями эластичных заявок (рис. 6,а, 7,а и 8,а; непрерывной линией показано скользящее среднее значение) и соответствующая мгновенная предложенная нагрузка, измеряемая как число занятых эластичными заявками единиц ресурса в системе неограниченной емкости (рис. 6,б, 7,б и 8,б; непрерывной линией показано среднее значение). Здесь, как и ранее, $c = 15$, а в качестве входящего потока эластичных заявок взят MAP-PAR (21) с $\lambda = 1$ и различными значениями параметров (π_1, π_2) .

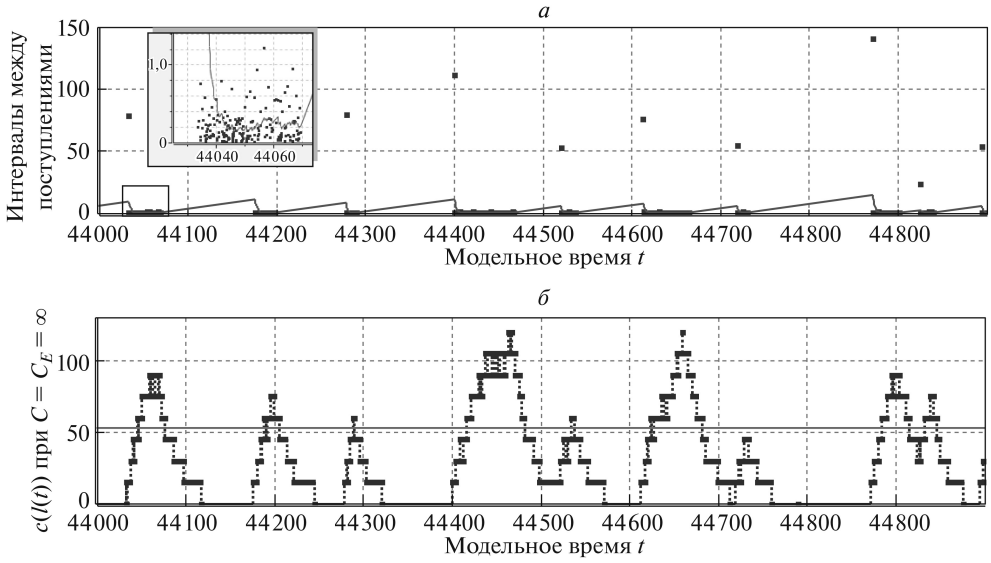


Рис. 7. Нагрузка от эластичного трафика в СМО с MAP-PAR при $(\pi_1, \pi_2) = (0; 0,99)$.

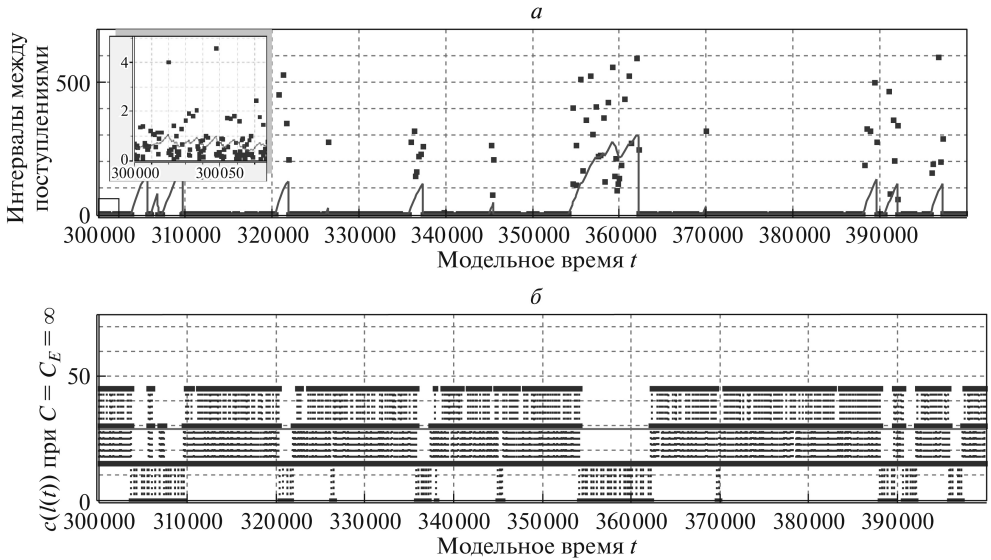


Рис. 8. Нагрузка от эластичного трафика в СМО с MAP-PAR при $(\pi_1, \pi_2) = (0,9; 0,9999)$.

На рис. 6 представлен фрагмент реализации модели при $(\pi_1, \pi_2) = (0,99; 0,99)$, $a = 2,005$. Поток имеет характеристики $r = 0,6522367$ и $\sigma = 1,2196604$. В точках, где скользящее среднее опускается к нулю, заявки поступают сериями с очень коротким интервалом. Например, около $t = 41838$ в систему поступает серия из 251 заявки за период $\Delta t \approx 1,3$. Это объясняется относительно большим значением нижнего диагонального элемента матри-

цы Q_1 , который равен $Q_{K,K}^1 = 100a\pi_2 = 198,495$. Такой характер поступления заявок, с одной стороны, порождает существенные всплески нагрузки, а с другой – отражается на коэффициенте корреляции. При относительно большой стационарной вероятности соответствующего состояния управляющей марковским потоком цепи q_K (здесь $q_K = 0,0024938$) данные всплески нагрузки происходят достаточно часто, чтобы также отразиться на стационарной вероятности потерь, которая в системе ограниченной емкости с данным потоком равна $B_E = 0,211593$.

На рис. 7 показан фрагмент реализации модели при $(\pi_1, \pi_2) = (0; 0,99)$, $a = 0,049505$. Здесь также имеют место всплески нагрузки, однако связаны они с наличием одиночных длительных интервалов между поступлениями (вызванных прохождением управляющей потоком ЦМ через первые четыре состояния) и компенсирующими эти интервалы периодами более концентрированного поступления заявок (скользящее среднее опускается к 0,2). В данном случае $Q_{K,K}^1 = 4,9009901$ и $q_K = 0,2$. Поскольку длительные интервалы – одиночные, корреляция в потоке слабо отрицательная: $r = -0,0079719$; среднеквадратичное отклонение велико и составляет $\sigma = 8,9376059$. Стационарная вероятность потерь в системе ограниченной емкости с данным потоком равна $B_E = 0,049382$.

Наконец, на рис. 8 показан фрагмент реализации модели при $(\pi_1, \pi_2) = (0,9; 0,9999)$, $a = 0,013986$, иллюстрирующий поток с высокой корреляцией ($r = 0,7154592$), но отсутствием всплесков нагрузки, приводящих к потерям. Здесь длительные периоды нагрузки средней интенсивности (скользящее среднее около 0,75) сменяются периодами очень слабой нагрузки. В результате вероятность потерь в системе ограниченной емкости с данным потоком сопоставима с показателем в системе с пуассоновским потоком и равна $B_E = 0,002305$ (сравним $B_E^{EXP(1)} = 0,001228$). Большое значение коэффициента корреляции вызвано не сериями малых интервалов между поступлениями, как на рис. 6, а сериями больших интервалов. Среднеквадратичное отклонение составляет $\sigma = 10,107635$, $Q_{K,K}^1 = 1,3984615$, $q_K = 0,7142857$.

Добавим, что приведенный в разделе 4 статьи алгоритм дает существенный выигрыш во времени вычисления и размерности задачи, однако в ряде интересных для приложений диапазонов нагрузочных параметров система (13) оказывается плохо обусловленной. Поэтому часть значений вероятностно-временных характеристик на рис. 2, 4 и 5 были получены путем численного решения непосредственно системы (10).

7. Заключение

В статье представлена система массового обслуживания с неэластичными и эластичными заявками для анализа совместной передачи трафика МВВ и МТС с агрегацией последнего и резервированием ресурсов. На численном примере исследована связь коэффициента корреляции последовательных интервалов между поступлением заявок в марковском потоке с вероятностью потерь в СМО. Показано, что всплески предложенной нагрузки, приводящие к потерям, могут иметь место как при положительной, так и при отрицательной корреляции, и что возможно отсутствие всплесков нагрузки при высокой

корреляции во входящем потоке. Таким образом, в рассматриваемой СМО коэффициент корреляции в марковском входящем потоке не может являться надежным индикатором уровня производительности системы.

СПИСОК ЛИТЕРАТУРЫ

1. *Ericsson* 5G systems: Enabling the transformation of industry and society. Ericsson white paper, 2017. <https://www.ericsson.com/en/white-papers/5g-systems-enabling-the-transformation-of-industry-and-society>
2. *Zhan W., Dai L.* Massive Random Access of Machine-to-Machine Communications in LTE Networks: Modeling and Throughput Optimization // *IEEE Trans. Wireless Communications*. 2018. V. 17. No. 4. P. 2771–2785.
3. *Mancuso V., Castagno P., Sereno M., Marsan M.A.* Slicing Cell Resources: The Case of HTC and MTC Coexistence // *IEEE INFOCOM 2019 – IEEE Conf. on Computer Communications*. Paris, France. 2019. P. 667–675.
4. *Dawy Z., Saad W., Ghosh A., Andrews J.G., Yaacoub E.* Towards Massive Machine Type Cellular Communications // *IEEE Wireless Commun.* 2017. V. 24. No. 1. P. 120–128.
5. *Kim D.M., Sørensen R., Mahmood K., Østerbø O., Zanella A., Popovski P.* Data Aggregation and Packet Bundling of Uplink Small Packets for Monitoring Applications // *IEEE Network*. 2017. V. 31. No. 6. P. 32–38.
6. *Lin C.-Y., Kao H.-W., Tsai M.-H., Chang H.-L.* Gateway-Assisted Two-Stage Radio Access for Machine Type Communication in LTE-Advanced Network // *Comput. Commun.* 2017. No. 105. P. 79–88.
7. *Gharbieh M., Bader A., ElSawy H., Alouini M.-S., Adinoyi A.* The Advents of Device-to-Device Relaying for Massively Loaded 5G Networks // *GLOBECOM 2017 – 2017 IEEE Global Communications Conf.* 2017. P. 1–7.
8. *Lucantoni D.M.* New Results on the Single Server Queue with a Batch Markovian Arrival Process // *Comm. Statist. Stoch. Models*. 1991. V. 7. No. 1. P. 1–46.
9. *Наумов В.А.* Марковские модели потоков требований / Сб. Системы массового обслуживания и информатика. М.: Изд-во УДН, 1987. С. 67–73.
10. *Chakravarthy S.R.* Markovian arrival processes / *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, Inc., 2010.
11. *Вишневецкий В.М., Дудин А.Н., Клименок В.И.* Стохастические системы с коррелированными потоками. Теория и применение в телекоммуникационных сетях. М.: Техносфера, 2018.
12. *Yarkina N., Samouylov K., Vishnevskiy V.* Analysis of Resource Sharing Between MBB and MTC Sessions with Data Aggregation Using Matrix-Analytic Methods and Simulation // *21st Int. Conf. DCCN 2018*. Moscow, Russia, September 17–21, 2018.

Статья представлена к публикации членом редколлегии А.И. Кибзуном.

Поступила в редакцию 22.07.2019

После доработки 13.10.2019

Принята к публикации 28.11.2019