

МАТЕМАТИЧЕСКАЯ ТЕОРИЯ ТЕЛЕТРАФИКА

УДК 519.6:621.39

Построение и анализ моделей системы с групповым обслуживанием заявок

К. Е. Самуйлов, Н. В. Першаков, И. А. Гудкова

*Кафедра систем телекоммуникаций
Российский университет дружбы народов
Россия, 117198, Москва, ул. Миклухо-Маклая, 6*

Статья посвящена исследованию системы с групповым обслуживанием со случайной длиной поступающей заявки и таймером, ограничивающим время формирования группы. Для простейшей модели типа $M|M^{[K]}|1|\infty$ излагается метод расчёта квантилей функции распределения времени ожидания начала обслуживания. В статье также представлены результаты имитационного моделирования для системы с таймером и примеры численного анализа.

КЛЮЧЕВЫЕ СЛОВА: система массового обслуживания, групповое обслуживание, таймер формирования группы заявок, время ожидания начала обслуживания.

Введение

Системы массового обслуживания (СМО) с групповым обслуживанием заявок находят применение в таких сферах жизнедеятельности человека, как промышленность, транспорт и связь. Начало исследованиям данного класса СМО в середине 1950-х гг. положили работы зарубежных учёных Н. Бейли и Ф. Даунтона [1, 2]. Бейли впервые рассмотрел систему с пуассоновским входящим потоком и переменным, ограниченным сверху размером группы заявок. Методом производящих функций он получил выражения для среднего значения и дисперсии числа заявок в очереди для цепи Маркова, вложенной по моментам, непосредственно предшествующим моментам окончания обслуживания групп заявок. В качестве дисциплины обслуживания им было рассмотрено распределение χ^2 (хи-квадрат) с чётным числом степеней свободы. Даунтон, в продолжение исследований Бейли, нашёл связь производящей функции с преобразованием Лапласа–Стилтьеса функции распределения (ФР) времени ожидания начала обслуживания, получил выражения для среднего значения и дисперсии, а также дал некоторые предварительные замечания к методу вычисления квантилей.

В настоящее время существует достаточное число работ (см. например, список источников в [3]), посвящённых исследованию СМО с групповым обслуживанием и предназначенных для решения различных прикладных задач. С целью унификации результатов этих исследований и их применения к анализу показателей качества современных протоколов передачи данных мы модифицируем и расширяем классификацию Бапарина–Кендалла [4] для систем с групповым обслуживанием заявок. Анализ простейшей СМО с групповым обслуживанием проведён с использованием матрично-аналитического метода, а для численного анализа системы с таймером применён метод имитационного моделирования.

1. Система массового обслуживания с ограничением на время формирования группы заявок

Рассмотрим СМО типа $GI_{l_\lambda}|GI^{[X,T]}|1|\langle R, L \rangle$, состоящую из одного обслуживающего прибора и накопителя ограниченной ёмкости. На систему поступает рекуррентный стационарный поток заявок с интенсивностью λ , а длина l_λ поступающей заявки является дискретной случайной величиной (СВ) с распределением $\gamma_m = P\{l_\lambda = m\}$, $m = \overline{1, M}$. Обслуживание заявок производится группами, причём ёмкость накопителя рассчитана на R групп, и предполагается, что обслуживаемые прибором заявки сохраняют за собой место в накопителе и одновременно покидают систему в момент окончания обслуживания. Время формирования каждой группы заявок ограничено таймером $0 \leq T \leq \infty$, и, поэтому, длина X группы является дискретной СВ, максимальное значение которой равно L . Обслуживание группы начинается либо в момент освобождения прибора, если группа сформирована, либо в момент завершения формирования группы, если прибор не занят обслуживанием, а время обслуживания группы заявок распределено по произвольному закону с ФР $B(x)$. На рис. 1 изображена диаграмма, поясняющая процедуру формирования группы, которая считается сформированной либо в момент истечения таймера, либо в момент поступления заявки, длина которой превышает доступное ей место в группе, причём при условии занятости прибора, заявки могут поступать в группу и после истечения таймера.

Заметим, что в частных случаях рассматриваемая СМО была изучена в ряде работ отечественных и зарубежных авторов. Например, в [5] получены вероятностно-временные характеристики (ВВХ) СМО $M_1|G^{[K,0]}|1|\langle \infty, K \rangle \triangleq M|G^{[K,0]}|1|\infty$. В [3] проведено исследование СМО $M_1|G^{[K,\infty]}|1|\langle \infty, K \rangle \triangleq M|G^{[K]}|1|\infty$, кодированной как $M|G^{[K]}|1|FS$ (англ. FS – Fixed Size), в случае, когда размер группы заявок является фиксированным.

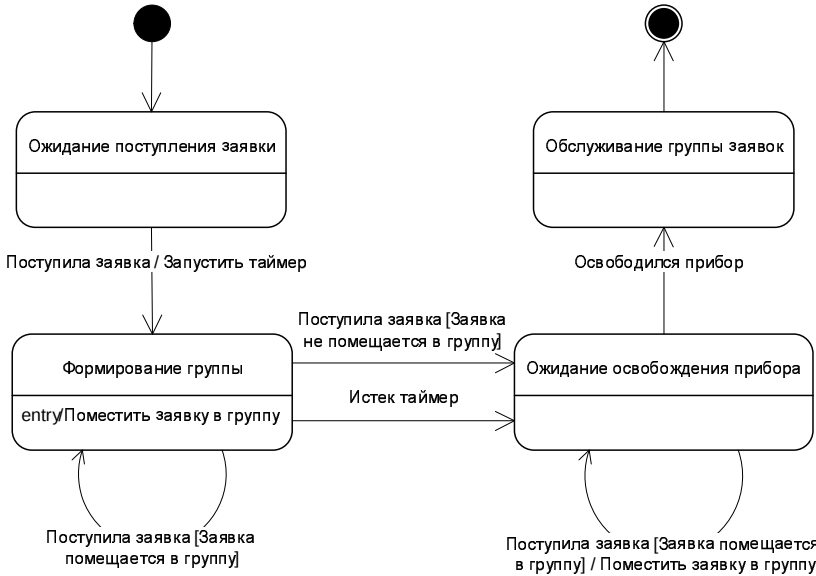


Рис. 1. Диаграмма состояний группы заявок

В следующем разделе статьи для анализа простейшей системы типа $M|M^{[K]}|1|\infty$ применён матрично-аналитический метод [6], а в последнем разделе для численного анализа ВВХ СМО $M_{l_\lambda}|D^{[X,T]}|1|\langle R, L \rangle$ — метод имитационного моделирования [7].

2. Простейшая система с групповым обслуживанием заявок

Рассмотрим СМО типа $M|M^{[K]}|1|\infty$, в которой обслуживание группы заявок распределено экспоненциально с параметром μ . Введём марковский процесс (МП) $X(t)$ как число заявок в очереди в момент времени $t \geq 0$, диаграмма переходов состояний которого имеет вид, показанный на рис. 2. Целью анализа является стационарное распределение $p_j = \lim_{t \rightarrow \infty} P\{X(t) = j\}$, $j \geq 0$ и квантилей w^θ ФР $W(x)$ времени ожидания начала обслуживания.

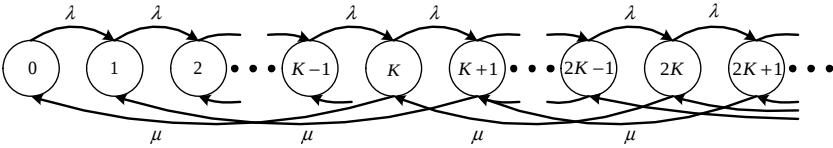


Рис. 2. Диаграмма переходов состояний СМО $M|M^{[K]}|1|\infty$

Предложение 1. Если $K > 1$ и $\rho < K$, то распределение $\{p_j\}$ МП $X(t)$ вычисляется по формулам

$$p_j = \begin{cases} p_{K-1} \frac{\sum_{i=0}^j R^i}{\rho + 1 - R^K}, & j = \overline{0, K-2}, \\ \left[\frac{\frac{\rho}{1-R} + \sum_{i=0}^{K-2} \sum_{k=0}^i R^k}{\rho + 1 - R^K} + 1 \right]^{-1}, & j = K-1, \\ p_{K-1} \frac{1}{\rho + 1 - R^K}, & j = K, \\ p_K R^{j-K}, & j > K, \end{cases} \quad (1)$$

где R – корень уравнения

$$r^{K+1} - (\rho + 1)r + \rho = 0, \quad |r| < 1. \quad (2)$$

Доказательство. Для доказательства воспользуемся матрично-аналитическим методом [6], который применим для неприводимых МП с инфинитезимальными матрицами вида

$$A = \begin{bmatrix} G_0 & G_{-1} & G_{-2} & G_{-3} & \cdots & G_{-n} & & & \\ G_1 & F_0 & F_{-1} & F_{-2} & \cdots & F_{-n+1} & F_{-n} & & 0 \\ G_2 & F_1 & F_0 & F_{-1} & \cdots & F_{-n+2} & F_{-n+1} & F_{-n} & \\ G_3 & F_2 & F_1 & F_0 & \cdots & F_{-n+3} & F_{-n+2} & F_{-n+1} & F_{-n} \\ \vdots & \ddots & \ddots & \ddots & \ddots & & \ddots & \ddots & \ddots \end{bmatrix}, \quad (3)$$

где G_i и F_i , $i \geq -n$ – прямоугольные матрицы. Нетрудно убедиться, что матрица интенсивностей МП $X(t)$ может быть приведена к виду (3) в случае $n = 1$ путём разбиения её на перечисленные ниже блоки:

$$[G_0]_{ij} = \begin{cases} -\lambda, & i = \overline{1, K}, \quad j = i, \\ \lambda, & i = \overline{1, K-1}, \quad j = i + 1, \\ 0, & \text{иначе;} \end{cases}$$

$$\begin{aligned}
\mathbf{G}_i &= (0^{i-1}, \mu, 0^{K-i}), \quad i = \overline{1, K}; \\
\mathbf{G}_i &= \mathbf{0}, \quad i > K, \quad \dim \mathbf{G}_i = K, \quad \forall i; \\
\mathbf{G}_{-1}^T &= (0^{K-1}, \lambda), \quad \dim \mathbf{G}_{-1}^T = K; \\
\mathbf{F}_0 &= -(\lambda + \mu); \\
\mathbf{F}_K &= \mu; \\
\mathbf{F}_i &= 0, \quad i = \{1, \dots, K-1, K+1, \dots\}; \\
\mathbf{F}_{-1} &= \lambda.
\end{aligned}$$

Вероятности p_j , $j > K$ можно выразить через вероятность p_K , если МП $X(t)$ положительно-возвратный, а матрицы, соответствующие ведущим главным минорам матрицы, которая получается вычёркиванием первой строки и первого столбца блоков матрицы $\mathbf{A}$, не являются единичными. Оба условия выполнены, и, следуя [6], получаем

$$p_j = p_K R^{j-K}, \quad j > K, \quad (4)$$

$$R = \lim_{i \rightarrow \infty} \hat{\Phi}_{i,1},$$

$$\hat{\Phi}_{i,1} = \begin{cases} \frac{\lambda}{\lambda + \mu}, & i = \overline{1, K}, \\ \frac{\lambda}{\lambda + \mu (1 - \hat{\Phi}_{i-1,1}^K)}, & i > K. \end{cases} \quad (5)$$

Переходя к пределу в (5), получим уравнение $\mu r^{K+1} - (\lambda + \mu)r + \lambda = 0$ или $(r^K + r^{K-1} + \dots + r^2 + r - \frac{\lambda}{\mu})(r - 1) = 0$, одним из корней которого и является исследуемая величина R . Уравнение $r^K + r^{K-1} + \dots + r^2 + r - \frac{\lambda}{\mu} = 0$, согласно правилу знаков Декарта, имеет ровно один действительный положительный корень. Если $R \geq 1$, то не выполнено условие нормировки для распределения стационарных вероятностей p_j , и поэтому искомое значение $R \in (0; 1)$ будет единственным действительным положительным корнем уравнения (2). Заметим, что значение $R \in (\frac{\rho}{1+\rho}; \frac{(1+K)\rho}{(1+\rho)K})$ может быть найдено одним из итерационных методов.

Вероятности p_j , $j \leq K$ находятся из системы уравнений равновесия

$$\begin{cases} (p_0, \dots, p_{K-1}) \mathbf{G}_0 + \sum_{j=K}^{2K-1} p_j \mathbf{G}_{j-K+1} = \mathbf{0}, \\ (p_0, \dots, p_{K-1}) \mathbf{G}_{-1} - (\lambda + \mu)p_K + \mu p_{2K} = 0, \end{cases}$$

и условия нормировки, что совместно с уравнением (4) доказывает сформулированное предположение. $\square$

Для расчёта квантилей w^θ необходимо найти ФР $W(x)$. Эта функция определена ниже в предложении 2, которое мы приводим без доказательства.

Предложение 2. Пусть $a_j = K(b_j + 1) - j - 1$ — число свободных мест для заявок в формирующейся группе, $b_j = \lceil \frac{j}{K} \rceil$ — число сформированных групп в системе при условии, что в момент поступления заявки в системе находилось j

заявок. Если $K > 1$ и $\rho < K$, то ФР $W(x)$ времени ожидания начала обслуживания вычисляется по формулам

$$W(x) = \begin{cases} \sum_{j=0}^{\infty} \left[p_j \int_{0+}^x \frac{\lambda^{a_j} t^{a_j-1}}{(a_j-1)!} e^{-\lambda t} dt \int_{0+}^x \frac{\mu^{b_j} t^{b_j-1}}{(b_j-1)!} e^{-\mu t} dt \right], & a_j > 0, b_j > 0, x > 0, \\ \sum_{j=0}^{\infty} \left[p_j \int_{0+}^x \frac{\mu^{b_j} t^{b_j-1}}{(b_j-1)!} e^{-\mu t} dt \right], & a_j = 0, b_j > 0, x > 0, \\ \sum_{j=0}^{\infty} \left[p_j \int_{0+}^x \frac{\lambda^{a_j} t^{a_j-1}}{(a_j-1)!} e^{-\lambda t} dt \right], & a_j > 0, b_j = 0, x > 0, \\ 1, & a_j = 0, b_j = 0, x > 0, \\ 0, & x \leq 0. \end{cases} \tag{6}$$

Вычисление квантилей w^θ ФР $W(x)$ с использованием формул (6) проводится только численно. Для достижения заданной точности ε необходимо найти значения x_1 и x_2 , $0 < W(x_1) < w^\theta < W(x_2) < 1$, такие что $|W(x_2) - W(x_1)| < \varepsilon$. Пример результатов вычислений ФР $W(x)$ по формуле (6) показан на рис. 3 для случая $K = 5$, $\mu^{-1} = 1,25$ и $\rho \in (0; K)$. Из графика видно, что для малых и предельных значений ρ ФР $W(x)$ ведёт себя как распределение с «тяжёлым хвостом», а в сечениях ФР $W(x)$ горизонтальными плоскостями получаем зависимости квантилей w^θ от значений величины нагрузки ρ .

На рис. 4 для наглядности представлена зависимость среднего времени ожидания начала обслуживания, вычисленная с использованием результатов работы [8]. Поведение графика достаточно очевидное, поскольку, чем меньше нагрузка, тем реже поступают заявки и требуется больше времени на формирование группы, и, следовательно, больше значение среднего времени ожидания. При увеличении нагрузки время ожидания убывает до некоторого минимального значения, а затем начинает возрастать, т. к. группы заявок накапливаются в очереди.

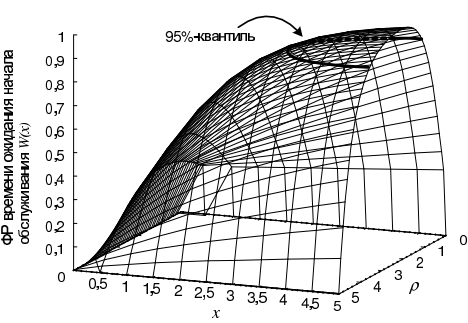


Рис. 3. ФР времени ожидания начала обслуживания и её 95% квантиль

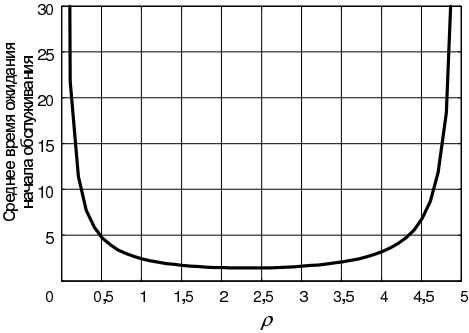


Рис. 4. Среднее время ожидания начала обслуживания

3. Численный анализ модели с таймером

Рассмотрим СМО типа $M_{l_\lambda}|D^{[X,T]}|1|\langle R, L \rangle$, для численного анализа ВВХ которой была разработана имитационная модель [7], а для вычисления квантили $w^{0,95}$ использовался метод, предложенный в [9]. Модель была разработана для анализа показателей качества обслуживания протокола управления потоковой передачей (англ. SCTP — Stream Control Transmission Protocol) [10]. Процесс передачи

по протоколу состоит в следующем: поступающие сообщения инкапсулируются в порции данных (заявки), которые объединяются в пакеты (группы заявок). Пакеты передаются по каналу с пропускной способностью 2 Мбит/с, а длины порций данных заданы распределением, указанным в табл. 1, где единицей измерения длины заявки является один байт. Значение таймера T было выбрано

Таблица 1

Распределение длин порций данных (байт)

l_λ	60	36	32
γ_m	0,23	0,41	0,36

равным 12 мс, что соответствует реальным параметрам оборудования компании Cisco Systems, поддерживающего протокол SCTP, а максимальная длина пакета $L = 1468$ байт, исходя из ограничений на максимальную длину кадра Ethernet.

Как видно из рис. 5, график среднего времени пребывания заявки в системе (задержки передачи сообщения в канал) имеет три экстремальные точки, определяющих четыре промежутка монотонности. Для их объяснения воспользуемся зависимостью среднего времени ожидания начала обслуживания (задержки в очереди на передачу) от нагрузки. При сравнительно низкой нагрузке (на рисунке эта область значений обозначена римской цифрой I) интенсивность поступления заявок мала, время формирования группы определяет таймер и, следовательно, величина задержки близка к значению таймера. С ростом нагрузки время задержки быстро убывает до того момента, когда среднее число заявок в группе, находящейся на обслуживании, становится более одной. При последующем увеличении интенсивности поступления заявок (область II) время формирования группы по-прежнему превышает среднее время обслуживания заявки, которое ещё не достигло максимального значения. Этим обуславливается дальнейшее убывание величины задержки, что в сравнении с предыдущим интервалом происходит медленнее, поскольку число заявок в группе увеличивается, приводя к росту времени их обслуживания. После того как группа достигает максимальной длины $L = 1468$ (область III), чем больше становится нагрузка, тем меньше на формирование группы влияет таймер, но, как и на только что рассмотренном интервале, время формирования группы превышает время обслуживания, что приводит к дальнейшему более быстрому уменьшению задержки. И, наконец, когда среднее время обслуживания превысит среднее время формирования группы (область IV), заявки начинают накапливаться в накопителе, способствуя увеличению среднего времени ожидания начала обслуживания.

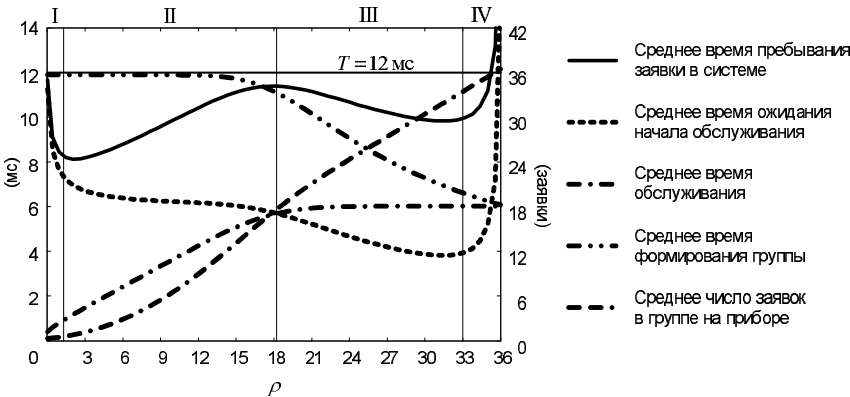


Рис. 5. Средние значения вероятностных характеристик СМО $M_{l_\lambda} | D^{[X,12]} | 1 | \langle \infty, 1468 \rangle$

Заключение

Современные прикладные задачи, и в частности, задачи, связанные с анализом показателей качества функционирования протоколов сетей связи, требуют уточнения имеющихся и разработки новых методов исследований. Так, оказывается, что процесс передачи данных по протоколу SCTP — новому решению, предназначенному для обеспечения работы приложений с высокими требованиями к показателям качества обслуживания — может быть описан в виде СМО с ограничением на время формирования группы заявок.

В статье представлена часть результатов исследований системы с групповым обслуживанием с учётом случайной длины заявки и таймера, ограничивающего время формирования группы. Для простейшей модели получен метод расчёта квантилей ФР времени ожидания начала обслуживания и приведены результаты имитационного моделирования для системы с таймером.

В заключении авторы выражают искреннюю признательность профессору кафедры систем телекоммуникаций РУДН Наумову В. А. за помощь и ценные советы в процессе исследований.

Литература

1. *Bailey N.* On Queueing Processes with Bulk Service // *Journal of the Royal Statistical Society, Ser. B.* — Vol. 16, No 1. — 1954. — Pp. 80–87.
2. *Downton F.* Waiting Times in Bulk Service Queues // *Journal of the Royal Statistical Society, Ser. B.* — Vol. 17, No 2. — 1955. — Pp. 256–261.
3. *Tadj L.* Explicit Solution of a Quorum Queueing System // *Stochastic Analysis and Applications.* — Vol. 21. — 2003. — Pp. 703–717.
4. *Башарин Г. П., Харкевич А. Д., Шнепс М. А.* Массовое обслуживание в телефонии. — М.: Наука, 1968.
5. *Саати Т. Л.* Элементы теории массового обслуживания и ее приложения. — М.: Советское Радио, 1971.
6. *Naoumov V. A.* Matrix-Analytic Solution for a Class of Markov Processes with Toeplitz-like Generator Matrices – *Advances in Matrix-Analytic Methods for Stochastic Models* // *Notable Publications.* — New Jersey, 1998.
7. *Гудкова И. А., Першаков Н. В., Самуйлов К. Е.* Имитационная модель для анализа показателей качества функционирования протокола SCTP // Труды 62-ой Научной сессии, посвященной Дню радио. — Москва: РНТОРЭС, 2007. — С. 231–233.
8. *Самуйлов К. Е., Чукарин А. В., Першаков Н. В.* Разработка модели функционирования протокола управления потоковой передачей // Вестник РУДН, серия «Прикладная и компьютерная математика». — Т. 4, № 1. — 2005. — С. 40–47.
9. *Jain R., Chlamtac I.* The P2 Algorithm for Dynamic Calculation of Quantiles and Histograms without Storing Observations // *Communications of the ACM.* — Vol. 28. — 1985. — Pp. 1076–1085.
10. *Гайдамака Ю. В., Першаков Н. В., Чукарин А. В.* Модель протокола SCTP и ее применение к анализу характеристик сигнального трафика в сетях сотовой подвижной связи // *Электросвязь.* — № 8. — 2007.

UDC 519.6:621.39

Batch Service Queue Model Design and Analysis

K. E. Samouylov, N. V. Pershakov, I. A. Gudkova

*Telecommunications Systems Department
Peoples' Friendship University of Russia
6, Miklukho-Maklaya str., Moscow, 117198, Russia*

In this paper we introduce an extended Basharin–Kendall queuing notation with queuing system $GI_{\lambda}|GI^{[X,T]}|1|\langle R, L \rangle$ in which jobs are served in batches, each job having a random length and the batch construction time limited to a maximum. The simplest fixed-size batch queue model $M|M^{[K]}|1|\infty$ is studied. Expressions are obtained for the distribution function of the waiting time and a brief discussion is included of its quantiles computation. We also develop a simulation model and present an example of its application to estimate quality of service in telecommunication networks.