УДК 519.23

СИСТЕМА ОБСЛУЖИВАНИЯ С БЕСКОНЕЧНЫМ БУФЕРОМ И ДИСЦИПЛИНОЙ ЛИМИТИРОВАННОГО РАЗДЕЛЕНИЯ ПРОЦЕССОРА

А.Н. Дудин, С.А. Дудин, О.С. Дудина

Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь, dudin@bsu.by, dudins@bsu.by, dudina@bsu.by

Мы рассматриваем систему массового обслуживания с ограниченной дисциплиной лимитированного разделения процессора и бесконечным буфером. Входной поток запросов задан марковским входным процессом. Количество одновременно обслуживаемых запросов ограничено. Процесс состояний системы является многомерным марковским процессом с интенсивностями переходов, зависящими от уровня. Получен генератор этого процесса. Найдены основные характеристики производительности системы.

Ключевые слова: Марковский входной поток; лимитированное распределение процессора; бесконечный буфер.

QUEUEING SYSTEM WITH AN INFINITE BUFFER AND LIMITED PROCESSOR SHARING DISCIPLINE

A.N. Dudin, S.A. Dudin, O.S. Dudina

Belarusian State University, 4 Niezalieznasci Avenue, Minsk 220030, Belarus, dudins@bsu.by, dudina@bsu.by
Corresponding author: dudin@bsu.by

We consider a queueing system with limited processor sharing discipline and infinite buffer. The arrival flow is defined by Markov arrival process. The number of customers that can be serviced simultaneously is restricted . The process of the system states is defined as a level-dependent process. The generator of this process is derived. The main performance measures of the system are obtained.

Keywords: Markov arrival flow; processor sharing; infinite buffer.

Введение

Системы массового обслуживания эффективно применяются для моделирования и оптимизации различных производственных, логистических и телекоммуникационных систем и сетей. В некоторых из таких систем запросы обслуживаются по одному в порядке, заданном дисциплиной обслуживания. Однако, зачастую запросы могут обслуживаться в системе одновременно. В этом случае рассматриваются многолинейные системы. То есть пропускная способность системы делится на несколько частей, условно называемыми приборами, и каждый прибор может обслуживать один запрос. Многолинейные системы массового обслуживания являются популярным объектом для исследования. Обзор современного состояния вопроса может быть найден, например, в [1]. Стоит отметить, что многолинейные системы имеют свои недостатки с точки зрения оптимального использования ресурса системы. Например, в ситуации, когда на обслуживания находится один запрос, а приборов много, то основная часть пропускной способности не используется. Как альтернатива многолинейным системам, рассматриваются системы массового обслуживания с дисциплиной разделения процессора. Для обзора работ по системам с разделением процессора, см., например, [2, 3, 4]. Данная дисциплина предполагает, что весь ресурс системы всегда направлен на обслуживание всех имеющихся на обслуживании запросов. То есть, даже когда на обслуживании находится один запрос, ресурс системы используется полностью.

Данная работа посвящена исследованию системы массового обслуживания с дисциплиной ограниченного разделения процессора. В отличие от классических систем, данная модель имеет следующие черты, повышающие ее адекватность современным системам. Во-первых, мы предполагаем, что каждый запрос имеет требуемую скорость обслуживания, которая не может быть превышена. В действительности, если пользователю беспроводной сети связи для работы требуется определенная пропускная способность системы, то совершенно необязательно выделять ему всю пропускную способность системы. Он просто не сможет ее использовать и не будет обслуживаться быстрее. Однако, если запросов на обслуживании становится много, и пропускной способности не хватает на обслуживание всех запросов с требуемой скоростью, то допускается уменьшение средней скорости обслуживания. Во-вторых, мы предполагаем, что число запросов на обслуживании ограничено заданным управляющим параметром. Дело в том, что если не ограничивать доступ в систему, то возможно возникновение ситуации, при

которой число запросов на обслуживании окажется настолько велико, что запросы станут обслуживаться с недопустимо малой скоростью. Кроме того, в данной работе мы предполагает, что входной поток запросов задается марковским входным потоком – MAP (от англ. – Markov arrival process), что позволяет учитывать существенные флуктуации трафика, свойственные современным телекоммуникационным сетям связи. Также, для большей адекватности модели, мы считаем, что запросы, которые не были допущены на обслуживание по приходу в систему, могут ожидать обслуживание в буфере неограниченной емкости.

1. Математическая модель

Мы рассматриваем систему массового обслуживания с бесконечным буфером и дисциплиной обслуживания разделение процессора.

Структура системы представлена на рисунке.

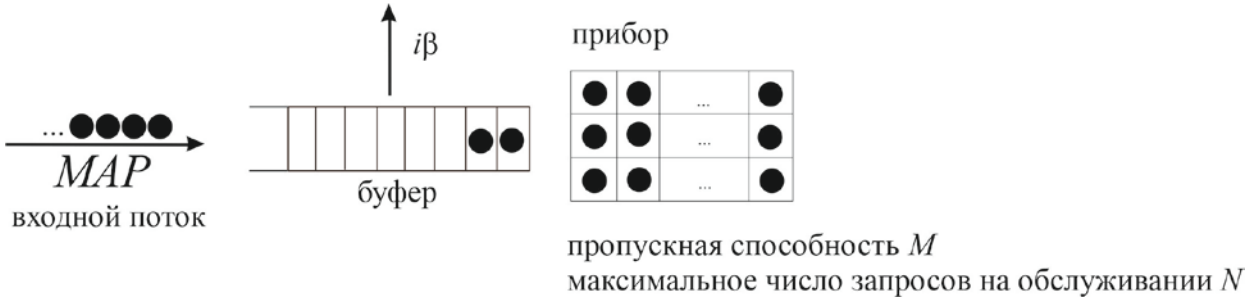

Рисунок – Структура системы

Единственный прибор может обслуживать до $N, N < \infty,$ запросов одновременно. Считаем, что общая пропускная способность прибора равна M мегабит в секунду. Для обслуживания одному запросу требуется средняя скорость X мегабит в секунду. Средний объем одного запроса составляет S мегабит. Такие образом, если запрос обслуживается с требуемой пропускной способностью, то его среднее время обслуживания определяется как $b_1 = S / X$. В данной работе будем предполагать, что время обслуживания одного запроса имеет экспоненциальное распределение. В случае наличия требуемой пропускной способности параметр экспоненциального распределения времени обслуживания задается как $\mu = 1 / b_1$. Если на обслуживании находится такое число запросов i, что $iX \le M$, что все запросы получают требуемую скорость обслуживания и обслуживаются с интенсивностью μ. В противном случае, каждому запросу выделяется пропускная способность $X_i = M / i$ мегабит и интенсивность его обслуживания равняется $\mu_i = X_i / S$.

Логично предположить, что параметр N должен быть выбран таким образом, что $NX > M$. В противном случае, пропускная способность системы не будет эффективно использоваться.

В систему поступает MAP-поток запросов. Данный поток задается управляющим процессом $\nu_t, t \geq 0,$ который является неприводимой цепью Маркова с непрерывным временем и конечным пространством состояний $\{1,\ldots,W\}$, и матрицами D_0 и D_1. Обозначим среднюю интенсивность поступления запросов как λ. Подробное описание MAP-потока, а также формулы для нахождения его характеристик можно найти в [1].

В случае, если в момент прихода запроса число запросов на обслуживании меньше параметра N, то запрос принимается на обслуживание. В противном случае, запрос идет в буфер неограниченной емкости и ожидает, пока освободится место на приборе. Запросы, ожидающие начала обслуживания в буфере, могут проявлять нетерпеливость. Это значит, что каждый запрос может покинуть буфер после экспоненциально распределенного с параметром $\beta, \beta > 0,$ времени.

2. Процесс изменения состояний системы и его анализ

Поведение рассматриваемой системы может быть описано следующей регулярной неприводимой цепью Маркова с непрерывным временем

$$\xi_t = \{i_t, \nu_t\}, t \geq 0,$$

где в момент $t, t \geq 0, i_t$ – число запросов в системе, $i_t \geq 0, \nu_t$ – состояние управляющего процесса MAP, $\nu_t = \overline{1,W}.$

Обозначим через Q генератор цепи Маркова ξ_t. Инфинитезимальный генератор Q цепи Маркова $\xi_t, t \geq 0,$ имеет блочную трехдиагональную структуру

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & O & O & \ldots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & O & O & \ldots \\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & O & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

где ненулевые блоки $Q_{i,j}, |i-j| \leq 1,$ определяются следующим образом:

$$Q_{0,0} = D_0,\ Q_{i,i} = D_0 - i\mu I_W,\ i \le \frac{M}{X},\ Q_{i,i} = D_0 - i\mu_i I_W,\ \frac{M}{X} < i \le N,$$

$$Q_{i,i} = D_0 - (i-N)\beta I_W - N\mu_N I_W,\ Q_{i,i+1} = D_1,\ Q_{i,i-1} = i\mu I_W,\ i \le \frac{M}{X},$$

$$Q_{i,i-1} = i\mu_i I_W,\ \frac{M}{X} < i \le N,\ Q_{i,i-1} = N\mu_N I_W + (i-N)\beta I_W,\ i > N.$$

Здесь O – нулевая матрица, I – единичная матрица соответствующей размерности.

Доказательство теоремы проводится посредством тщательного анализа всевозможных переходов цепи Маркова ξ_t и последующей группировкой интенсивностей в матрицы-блоки генератора.

Исследуемая цепь Маркова ξ_t принадлежит к классу ассимптотически квазитеплицевых цепей Маркова, см. [5]. Воспользовавшись результатами из работы [5] можно формально доказать тот факт, что поскольку запросы, находящиеся в буфере проявляют нетерпеливость, то стационарное распределение системы существуют для всех значений параметров системы.

Обозначим через $\pi(i,\nu), i \ge 0, \nu_t = \overline{1,W}$, стационарные вероятности состояний цепи ξ_t. Сформируем из этих вероятностей векторы строки

$$\pi_i = (\pi(i,1),\ldots,\pi(i,W)), i \ge 0.$$

Широко известно, что вектора стационарных вероятностей π_i могут быть найдены как решение системы уравнений равновесия

$$(\pi_0, \pi_1, \ldots)\mathrm{Q} = 0, (\pi_0, \pi_1, \ldots)\mathrm{e} = 1,$$

где **e** – вектор-столбец, состоящий из единиц, и **0** – вектор-строка, состоящая из нулей.

Поскольку в рассматриваемом случае генератор Q имеет бесконечный размер, а его элементы зависят от номера строки, то решить данную систему стандартными методами не представляется возможным. Для нахождения векторов стационарных вероятностей $\pi_i, i \ge 0$, рекомендуется использовать эффективный алгоритм, разработанный в работе [6].

3. Характеристики производительности системы

Среднее число запросов на обслуживании $N_{serv} = \sum_{i=0}^{\infty} \min\{i, N\}\pi_i$e.

Среднее число запросов в буфере $N_{buffer} = \sum_{i=N+1}^{\infty} (i-N)\pi_i$e.

Среднее число запросов в системе $L = \sum_{i=0}^{\infty} i\pi_i$e $= N_{serv} + N_{buffer}$.

Вероятность того, что прибор простаивает в произвольный момент времени $P_{idle} = \pi_0$e.

Интенсивность потока обслуженных запросов

$$\lambda_{out} = \sum_{i=0}^{\infty} (\delta_{i \le \frac{M}{X}} i\mu\pi_i e + \delta_{i > \frac{M}{X}} min\{i, N\}\mu_{min\{i,N\}}\pi_i e), \text{ где } \delta_a$$

$$= \begin{cases} 1, & \text{если } a \text{ верно,} \\ 0, & \text{в противном случае.} \end{cases}$$

Вероятность того, что произвольный запрос будет потерян

$$P_{loss} = \frac{1}{\lambda} \sum_{i=N+1}^{\infty} (i-N)\beta\pi_i = 1 - \frac{\lambda_{out}}{\lambda}.$$

Вероятность того, что произвольный момент времени запросы получают урезанную скорость обслуживания

$$P_{sharing} = \sum_{i=\lceil \frac{M}{X} \rceil}^{\infty} \pi_i \text{e},$$

где $\lceil a \rceil$ определяет минимальное натуральное число большее, чем a.

Библиографические ссылки

1. Dudin A., Klimenok V. I., Vishnevsky V. M. The Theory of Queuing Systems with Correlated Flows. Cham : Springer, 2020. 430 p.
2. Yashkov S.F., Yashkova A.S. Processor sharing: A survey of the mathematical theory // Automation and Remote Control. 2007. № 68(9). P. 1662–1731.
3. Altman E., Avrachenkov K., Ayesta U. A survey on discriminatory processor sharing // Queueing systems. 2006. № 53(1). P. 53–63.

4. Kim C., Dudin S.A., Dudina O.S., Dudin A.N. Mathematical models for the operation of a cell with bandwidth sharing and moving users // IEEE Transactions on Wireless Communications. 2019. № 19(2). P. 744–755.
5. Klimenok V.I., Dudin A.N. Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory // Queueing System. 2006. № 54. P. 245–259.
6. Dudin S., Dudina O. Retrial multi-server queuing system with PHF service time distribution as a model of a channel with unreliable transmission of information // Applied Mathematical Modelling. 2019. № 65. P. 676–695.