УДК 519.872

НОВАЯ МОДЕЛЬ РАСПРЕДЕЛЕНИЯ РЕСУРСА МЕЖДУ ЭЛАСТИЧНЫМИ И НЕЭЛАСТИЧНЫМИ ЗАПРОСАМИ

А.Н. Дудин[1)], **С.А. Дудин**[2)]

[1), 2)] *Белорусский государственный университет, пр.Независимости, 4, 220030, г. Минск, Беларусь, dudin@bsu.by, dudins@bsu.by*

Предложена математическая модель распределения ресурсов между потоками эластичного и неэластичного трафика в виде системы массового обслуживания с двумя типами запросов. Запросы первого типа (неэластичный трафик) обслуживаются как в стандартной многолинейной системе массового обслуживания с потерями. Оставшаяся часть пропускной способности системы отдается запросам второго типа (эластичный трафик), которые получают обслуживание в соответствии с дисциплиной разделения процессора.

Ключевые слова: гибридная дисциплина обслуживания; марковский входной поток; эластичный трафик.

Введение

Системы массового обслуживания (СМО) широко используются для оценки и оптимизации производительности систем, в которых некоторые ограниченные ресурсы динамически распределяются между конкурирующими пользователями. В частности, эти системы полезны для анализа систем и сетей телекоммуникаций, производства, транспорта, здравоохранения, экстренной помощи и логистики, см., например, [1-3]. Наибольшее внимание в литературе получили так называемые однолинейные системы, где весь ресурс используется для обслуживания запросов строго по одному. Однако во многих реальных системах ресурс может предоставляться одновременно нескольким пользователям. Популярны два варианта организации одновременного обслуживания.

Первым является разделение ресурса, присущее классическим многолинейным СМО. Весь ресурс системы делится на фиксированное число так называемых приборов. Каждый запрос занимает один прибор и обслуживается им независимо от других запросов. Недостаток такой организации обслуживания состоит в том, что, когда количество запросов в системе невелико, пропускная способность системы используется далеко не полностью.

В качестве альтернативы этому варианту, более эффективо использующей ресурс системы, рассматриваются СМО с дисциплиной обслуживания типа PS (PS – processor sharing). Эта дисциплина предполагает, что все запросы на обслуживании постоянно используют

всю емкость системы. Системы с дисциплиной обслуживания PS широко используются, в частности, в телекоммуникационных сетях. Более подробную информацию о состоянии исследований, связанных с дисциплиной разделения процессора, можно найти, например, в [4-8].

Однако классическая дисциплина PS может применяться не во всех СМО. Два основных недостатка этой дисциплины относятся к ситуациям, когда произвольный запрос имеет ограничение: а) минимальной требуемой пропускной способности и б) максимальной требуемой пропускной способности. Чтобы преодолеть эти недостатки, рассматриваются различные модификации дисциплины PS. Например, ограниченная дисциплина PS предполагает, что количество запросов, которые могут одновременно использовать прибор, ограничено некоторым конечным целым числом. В качестве работ, посвященных моделям с ограниченным совместным использованием процессоров, можно привести работы [9-12]. Для преодоления недостатка б) вводится модификация дисциплины PS, которая предполагает, что запросы имеют предпочтительную пропускную способность, а снижение получаемой пропускной способности происходит только в ситуации, когда сумма этих способностей всех обслуживающих запросов превышает пропускную способность системы.

В данной статье формулируется СМО с гибридной дисциплиной обслуживания. В качестве важного практического примера реальных систем, где можно применить эту дисциплину, можем упомянуть современные сети связи, в которых передаваемая информация неоднородна. В частности, в таких сетях есть пользователи, которым для передачи информации необходима фиксированная полоса пропускания, которую нельзя сократить или расширить. В то время как другие пользователи могут обслуживаться и с переменной интенсивностью. Ниже мы будем называть запросы, генерируемые такими пользователями, как неэластичными (потоковыми) и эластичными, соответственно.

Для моделирования одновременной передачи эластичного и неэластичного трафика в одном потоке и оптимизации планирования ресурсов системы между этими типами трафика в [13] была предложена гибридная дисциплина обслуживания, представляющая собой смесь классической многолинейной и дисциплины обслуживания с разделением процессора. В данной статье мы рассматриваем следующий сценарий. Запросы первого типа требуют постоянной пропускной способности, а запросы второго типа – нет. Часть пропускной способности системы выделяется для поддержания фиксированного количества каналов с постоянной интенсивностью, доступных для обслуживания запросов первого типа, которые обслуживаются как в классической многолинейной СМО.

Оставшаяся часть пропускной способности плюс часть пропускной способности, выделенная для запросов первого типа, но временно не используемая, используется для обслуживания запросов второго типа, которые обслуживаются в соответствии с дисциплиной классического разделения процессора.

Основные преимущества нашего исследования по сравнению с [13] заключаются в следующем:

1) Потоки обоих типов запросов в [13] считались стационарными пуассоновскими. Мы предполагаем, что каждый из двух типов запросов поступают в соответствии с гораздо более общими марковскими процессами (*MAP*), см., например, [2].

2) В [13] предполагается ограниченное PS для эластичного трафика. Здесь мы рассматриваем классическую дисциплину PS для эластичного трафика. Успешная компьютерная реализация расчета стационарного распределения количества запросов обоих типов в случае ограниченного PS может быть осуществлена только для относительно небольшого числа одновременно обслуживаемых эластичных пользователей. В некоторых реальных системах это число может достигать многих тысяч. Таким образом, получаемые алгоритмические результаты [13] могут быть сложными для их компьютерной реализации.

3) В [13] предполагается, что запросы абсолютно терпеливы. После принятия в систему любой запрос должен быть обслужен. Это не совсем реально. Например, опыт использования мобильного интернета в выходные дни в зонах отдыха показывает, что из-за реальной бесконечной делимости пропускной способности, разделяемой провайдерами, интенсивность обслуживания становится порой очень маленькой. Соответственно, время пребывания эластичных пользователей в Интернете становится очень продолжительным, и они могут прекратить соединение. Чтобы учесть эту важную особенность, мы предполагаем, что эластичные пользователи нетерпеливы. Каждый такой пользователь может уйти из системы, не получив полного обслуживания, по истечении некоторого времени пребывания в системе.

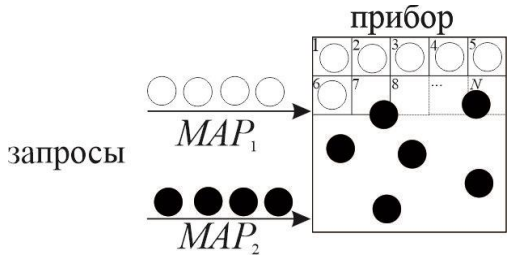
Сформулированная в данном статье модель СМО может быть полезна для выбора оптимальных значений общей пропускной способности системы и ее оптимального распределения между эластичным и неэластичным трафиком.

Следует также отметить, что рассматриваемая СМО имеет гораздо более широкую область потенциального применения, чем только телекоммуникации. Эту систему можно рассматривать как модель любой реальной системы приоритетного обслуживания, в которой привилегиро-

ванные пользователи имеют гарантированный уровень обслуживания и абсолютный приоритет над обычными пользователями.

Математическая модель

Рассмотрим СМО с гибридной дисциплиной обслуживания.



Структура системы

Предположим, что общая пропускная способность системы равна M. Будем считать, что эта величина измеряется в мегабитах в секунду (Мбит/с). Значение параметра M предполагается фиксированным. При этом результаты анализа, проведенного ниже, могут быть использованы, в частности, для правильного выбора этого параметра.

Запросы двух типов поступают в независимых MAP-потоках. Первый поток, обозначаемый MAP_1, задается управляющим процессом $\nu_t, t \geq 0,$ который представляет собой неприводимую цепь Маркова с непрерывным временем, пространством состояний $\{1,2,\ldots,W\}$, и матрицами D_0, D_1. Средняя интенсивность поступления запросов первого типа λ_1 определяются формулой: $\lambda_1 = \theta_1 D_1 \mathbf{e}$, где θ_1 – вектор стационарного распределения управляющего процесса ν_t. Он вычисляется как решение системы $\theta_1(D_0 + D_1) = \mathbf{0}, \theta_1 \mathbf{e} = 1$. Здесь $\mathbf{e}$ – вектор-столбец подходящего размера, состоящий из единиц, а $\mathbf{0}$ – вектор-строка, состоящая из нулей.

Второй поток, обозначаемый MAP_2, задается управляющим процессом $h_t, t \geq 0,$ который представляет собой неприводимую цепь Маркова с непрерывным временем, конечным пространством состояний $\{1,2,\ldots,V\}$, и матрицами H_0, H_1. Обозначим среднюю интенсивность поступления запросов второго типа как λ_2. Она определяются аналогичным образом. Параметр $\lambda = \lambda_1 + \lambda_2$ определяет общую среднюю интенсивность поступления запросов в систему.

Два типа поступающих запросов различаются как средним объемом, так и требованиями к интенсивности обслуживания. Будем считать, что объем одного запроса k – го типа имеет показательное распределение с параметром $1/S_k$, где S_k – средний объём запроса k – го типа (в мегаби-

тах). Для одного запроса первого типа требуется полоса пропускания X Мбит/с. Следовательно, среднее время обслуживания запроса первого типа составляет $b_1^{(1)} = S_1 / X$ секунд, а время обслуживания имеет экспоненциальное распределение с интенсивностью $\alpha = 1/b_1^{(1)}$. Предположим, что в системе одновременно может обслуживаться до N запросов первого типа и выполнено неравенство $XN < M$. Другими словами, для обслуживания запросов первого типа доступно до N каналов (приборов), а для обслуживания запросов второго типа доступна некоторая полоса пропускания. Здесь N – фиксированный параметр управления, который в дальнейшем следует выбирать для обеспечения оптимального качества обслуживания обоих типов запросов.

Мы предполагаем, что в системе нет буфера. Если количество запросов первого типа в обслуживании на момент поступления запроса первого типа меньше параметра N, то этот запрос принимается к обслуживанию. В противном случае он теряется. Запросы второго типа не имеют жестких требований к необходимой им пропускной способности. Мы предполагаем, что количество запросов второго типа, которые могут быть обслужены одновременно, не ограничено. Эти запросы обслуживаются в соответствии с дисциплиной разделения процессора, и для их обслуживания выделяется вся полоса пропускания системы, не используемая для обслуживания запросов первого типа.

Например, если есть один запрос второго типа и $n, 0 \le n \le N,$ запросов первого типа, то под этот запрос второго типа выделяется вся доступная полоса пропускания системы размером $M - nX$ Мбит/с. Это означает, что среднее время обслуживания такого запроса составляет $b_1^{(2)}(n) = S_2 / (M - nX)$ секунд, а интенсивность его обслуживания $\mu_n = 1/b_1^{(2)}(n)$. Если n приборов занято обслуживанием запросов первого типа и имеются $i, i > 0,$ запросов второго типа на обслуживании, то эти запросы второго типа делят поровну между собой всю оставшуюся пропускную способность и каждый из них обслуживается с интенсивностью $\mu_n / i.$

Запросы второго типа могут быть нетерпеливыми во время обслуживания. Это означает, что каждый запрос может покинуть систему, не получив полного обслуживания, по истечении экспоненциально распределенного времени с параметром $\gamma, \gamma \ge 0.$

Как уже говорилось выше, число N и пропускная способность прибора M являются управляющими параметрами. Целью исследования является получение возможности определения оптимального значения этих

параметров, при котором выбранный стоимостной критерий, описывающий эффективность системы, принимает оптимальное значение.

Анализ сформулированной модели проведен в работе [14].

Библиографические ссылки

1. *Chakravarthy S. R.* Introduction to Matrix-Analytic Methods in Queues 2: Analytical and Simulation Approach-Queues and Simulation. John Wiley & Sons, 2022.

2. *Dudin A. N., Klimenok V. I., Vishnevsky V. M.* The theory of queuing systems with correlated flows. Cham: Springer, 2020. T. 430.

3. *Naumov V. et al.* Matrix and analytical methods for performance analysis of telecommunication systems. Springer Nature, 2022.

4. *Ahmadi M. et al.* Processor sharing queues with impatient customers and state-dependent rates //IEEE/ACM Transactions on Networking. 2021. T. 29. № 6. C. 2467-2477.

5. *Ghosh S., Banik A. D.* An algorithmic analysis of the BMAP/MSP/1 generalized processor-sharing queue // Computers & Operations Research. 2017. T. 79. C. 1-11.

6. *Guillemin F., Quintuna Rodriguez V. K., Simonian A.* Sojourn time in a processor sharing queue with batch arrivals // Stochastic Models. 2018. T. 34. № 3. C. 322-361.

7. *Kim C. et al.* Mathematical models for the operation of a cell with bandwidth sharing and moving users // IEEE Transactions on Wireless Communications. 2019. T. 19. № 2. C. 744-755.

8. *Yashkov S. F., Yashkova A. S.* Processor sharing: A survey of the mathematical theory // Automation and Remote Control. 2007. T. 68. C. 1662-1731.

9. *D'Apice C. et al.* Priority queueing system with many types of requests and restricted processor sharing // Journal of Ambient Intelligence and Humanized Computing. 2023. T. 14. № 9. C. 12651-12662.

10. Dudin A. N. et al. Analysis of queueing model with processor sharing discipline and customers impatience // Operations Research Perspectives. 2018. T. 5. C. 245-255.

11. *Nunez-Queija R.* Sojourn times in non-homogeneous QBD processes with processor sharing // Stochastic models. 2001. T. 17. № 1. C. 61-92.

12. *Telek M., Van Houdt B.* Response time distribution of a class of limited processor sharing queues // ACM SIGMETRICS Performance Evaluation Review. 2018. T. 45. № 3. C. 143-155.

13. *Nunez-Queija R., Berg J. L., Mandjes M. R. H.* Performance evaluation of strategies for integration of elastic and stream traffic. CWI (Centre for Mathematics and Computer Science), 1999.

14. *Дудин А. Н., Дудина О. С.* Анализ новой дисциплины разделения ресурса между эластичными и неэластичными запросами // Сборник трудов Международной научной конференции «Теория вероятностей, математическая статистика и приложения» (Беларусь, Минск, 22–24 апреля 2024 г.). 2024. Р. 1-6.