ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И СИСТЕМЫ

УДК 681.2
DOI: 10.17586/0021-3454-2015-58-6-415-420

РАСПРЕДЕЛЕНИЕ ПРИОРИТЕТОВ В СИСТЕМАХ С ВЕРОЯТНОСТНЫМИ ОГРАНИЧЕНИЯМИ

Т. И. Алиев

Университет ИТМО, 197101, Санкт-Петербург, Россия
E-mail: aliev@cs.ifmo.ru

Рассматривается задача назначения приоритетов в одноканальных системах массового обслуживания с неоднородным потоком запросов при наличии ограничений на вероятность превышения допустимого времени пребывания в системе запросов разных классов. Задача решается в классе дисциплин обслуживания со смешанными приоритетами, задаваемых в виде матрицы приоритетов. Элементы матрицы отображают вид (относительный или абсолютный) и уровень приоритетов между запросами разных классов. Для оценки вероятности превышения допустимого времени пребывания запросов в системе используется аппроксимация соответствующих законов распределений по двум первым моментам. Алгоритм назначения приоритетов основан на целенаправленном переборе дисциплин обслуживания. Показателем эффективности дисциплины обслуживания служит производительность системы.

Ключевые слова: *система обслуживания, время пребывания, вероятностные ограничения, производительность, дисциплина обслуживания, смешанные приоритеты.*

Введение. Ограничения на время реакции (задержки запросов разных классов) в информационно-управляющих системах обычно задаются в вероятностном виде, в отличие от информационно-вычислительных систем, в которых ограничения накладываются на средние значения задержки [1]. Аналогично при передаче мультимедийных пакетов в компьютерных сетях ограничения (см. рекомендации ITU-T Y.1541 [2]) накладываются как на средние значения задержек пакетов, так и на вариацию (джиттер) задержки, для расчета которой требуется знание закона распределения задержки. Это существенно усложняет решение задачи выбора приоритетной стратегии управления процессами обработки и передачи данных из-за необходимости проведения математических выкладок на уровне функций распределений. Задача может быть упрощена, если ее решать на уровне двух первых моментов распределений соответствующих случайных величин.

В качестве моделей систем реального времени широко используются системы массового обслуживания с одним обслуживающим устройством и накопителем неограниченной емкости, в которую поступает неоднородный поток запросов, обрабатываемых с заданной производительностью [3]. В этом случае решение задачи распределения приоритетов сводится к выбору дисциплины обслуживания (ДО) запросов, обеспечивающей выполнение заданных ограничений на время пребывания запросов в системе.

Постановка задачи распределения приоритетов. При решении задачи распределения приоритетов в системе с неоднородным потоком запросов в качестве нагрузочных параметров используются: количество классов запросов H, поступающих в систему с интенсивностями $\lambda_1,\ldots,\lambda_H$ и образующих простейшие потоки; ресурсоемкости обработки Θ_h запросов каждого класса (команд или инструкций, выполняемых при обработке запроса соответствующего класса), задаваемые как минимум тремя моментами: средними значениями $\theta_1,\ldots,\theta_H$, коэффициентами вариации $\nu_1,\ldots,\nu_H$ и третьими начальными моментами $\theta_1^{(3)},\ldots,\theta_H^{(3)}$. Время обработки запросов класса h (h-запросов) зависит от производительности V, измеряемой, например, числом команд или инструкций, выполняемых системой за единицу времени, и определяется как $\tau_{b_h} = \Theta_h / V$ $(h = \overline{1,H})$.

Вероятностные ограничения на время пребывания $\tau_{u_1},\ldots,\tau_{u_H}$ запросов в системе задаются в виде:

$$\Pr(\tau_{u_h} > u_h^*) \le \delta_h^*, \qquad (1)$$

где u_h^* — допустимое время пребывания в системе запросов класса h; δ_h^* — допустимая вероятность превышения заданного ограничения u_h^* $(h = \overline{1,H})$.

Время пребывания в системе $\tau_{u_1},\ldots,\tau_{u_H}$ зависит от производительности системы V и ДО, причем наилучшей является дисциплина, обеспечивающая выполнение ограничений (1) при наименьшей производительности системы.

Задача выбора ДО решается в классе дисциплин со смешанными приоритетами (ДО СП) [4], которые описываются матрицей приоритетов (МП) $Q = [q_{ij}\ (i,j = 1,\ldots,H)]$, где q_{ij} задает приоритет i-запросов по отношению к j-запросам: 0 — нет приоритета, 1 — приоритет относительный (ОП) и 2 — абсолютный (АП).

Таким образом, задача распределения приоритетов в системах с вероятностными ограничениями формулируется следующим образом: найти дисциплину обслуживания запросов в классе ДО СП, обеспечивающую выполнение ограничений (1) при минимальной производительности системы.

Расчетные соотношения. Время пребывания h-запросов в системе τ_{u_h} складывается из времени ожидания начала обслуживания τ_{x_h} и времени нахождения запроса на обработке τ_{z_h}, включающего в себя время ожидания в прерванном состоянии: $\tau_{u_h} = \tau_{x_h} + \tau_{z_h}$. Тогда математическое ожидание u_h и второй начальный момент $u_h^{(2)}$ времени пребывания в системе h-запросов ($h = \overline{1,H}$):

$$u_h = x_h + z_h; \quad u_h^{(2)} = x_h^{(2)} + 2x_h z_h + z_h^{(2)}. \qquad (2)$$

Значения x_h, z_h и $x_h^{(2)}, z_h^{(2)}$ определяются по формулам [5]:

$$\left.\begin{aligned}
x_h &= \frac{\sum_{i=1}^{H} r_4(i,k)\lambda_i b_i^{(2)}}{2(1-R_h^{(2)})(1-R_h^{(3)})}; \qquad z_h = \frac{b_h}{1-R_h^{(1)}};\\
x_h^{(2)} &= \frac{\sum_{i=1}^{H} r_4(i,k)\lambda_i b_i^{(3)}}{3(1-R_h^{(2)})^2(1-R_h^{(3)})} + \frac{\sum_{i=1}^{H} r_3(i,k)\lambda_i b_i^{(2)} \sum_{i=1}^{H} r_4(i,k)\lambda_i b_i^{(2)}}{2(1-R_h^{(2)})^2(1-R_h^{(3)})^2} +\\
&\quad + \frac{\sum_{i=1}^{H} r_2(i,k)\lambda_i b_i^{(2)} \sum_{i=1}^{H} r_4(i,k)\lambda_i b_i^{(2)}}{2(1-R_h^{(2)})^3(1-R_h^{(3)})};\\
z_h^{(2)} &= \frac{b_h^{(2)}}{(1-R_h^{(1)})^2} + \frac{\sum_{i=1}^{H} r_1(i,k)\lambda_i b_i^{(2)}}{(1-R_h^{(1)})^3},
\end{aligned}\right\} \qquad (3)$$

где $b_i^{(n)} = \int_0^\infty \tau^n b_i(\tau)\, d\tau$ — n-й начальный момент времени обслуживания; $b_i(\tau)$ — плотность вероятности распределения времени обслуживания i-запросов ($i = \overline{1,H}$; $n = 1,2,\ldots$); $R_h^{(g)} = \sum_{i=1}^{H} r_g(i,h)\rho_i$ — частичная суммарная загрузка; $r_g(i,h)$ — коэффициент, принимающий значение „0“ или „1“ в зависимости от значений элементов q_{ih} и q_{hi} матрицы приоритетов и позволяющий выделять классы запросов i и h, которые имеют один и тот же вид приоритета (ОП, АП, БП или любое их сочетание): $r_1(i,h) = 0,5\,q_{ih}(q_{ih}-1)$ — принимает значение „1“, если i-запросы имеют АП по отношению к h-запросам; $r_2(i,h) = 0,5\,q_{ih}(3-q_{ih})$ — принимает значение „1“, если i-запросы имеют ОП или АП по отношению к h-запросам; $r_3(i,h) = 1-0,5\,q_{hi}(3-2q_{ih}+q_{hi})$ — принимает значение „1“, если i-запросы имеют БП, ОП или АП по отношению к h-запросам; $r_4(i,h) = 1+0,5\,q_{hi}(1-q_{hi}+q_{ih})$ — принимает значение „0“ только в том случае, если h-запросы имеют АП по отношению к i-запросам.

Синтез дисциплины обслуживания. Решение задачи синтеза ДО сводится к определению значений q_{ij} матрицы приоритетов, при которых выполняются вероятностные ограничения (1), и реализуется путем целенаправленного перебора различных ДО.

При решении задачи синтеза ДО необходимо определить вероятность превышения заданного ограничения на время задержки:

$$\Pr(\tau_{u_h} > u_h^*) = 1 - U_h(u_h^*),$$

где $U_h(\tau)$ — функция распределения времени пребывания (задержки) h-запросов в системе ($h = \overline{1,H}$).

Таким образом, необходимо знать закон распределения задержки, который, в свою очередь, зависит от выбранной дисциплины обслуживания запросов.

Для нахождения закона распределения задержки аппроксимируем распределение времени пребывания h-запросов в системе по двум моментам u_h и $u_h^{(2)}$, рассчитываемым по

формулам (2), (3). В зависимости от значения коэффициента вариации $\alpha_h = \sqrt{u_h^{(2)} - u_h^2}\,/u_h$ в качестве аппроксимирующих законов распределений могут использоваться [6]: распределение Эрланга, если $0 \le \alpha_h \le 1$, или гиперэкспоненциальное распределение, если $\alpha_h > 1$. В соответствии с методикой аппроксимации по двум моментам [6] рассчитываются параметры аппроксимирующих распределений: порядок $k_h = \lceil 1/\alpha_h \rceil$ распределения Эрланга, если $0 \le \alpha_h \le 1$, или вероятности $q_h \le 2/(1+\alpha_h^2)$ и математические ожидания $t_{h_1} = [1+\sqrt{0{,}5(1-q_h)(\alpha_h^2-1)/q_h}\,]u_h$ и $t_{h_2} = [1+\sqrt{0{,}5\,q_h(\alpha_h^2-1)/(1-q_h)}\,]u_h$ экспоненциальных фаз двухфазного гиперэкспоненциального распределения, если $\alpha_h > 1$. Соответствующие функции распределения закона Эрланга и гиперэкспоненциального закона при рассчитанных значениях параметров примут вид:

$$U_h(\tau) = 1 - \exp\left(-\frac{k_h\tau}{u_h}\right)\sum_{i=1}^{k_h-1}\frac{(k_h\tau)^i}{i!\,u_h^i},$$

$$U_h(\tau) = 1 - q_h\exp\left(-\frac{\tau}{t_{h_1}}\right) - (1-q_h)\exp\left(-\frac{\tau}{t_{h_2}}\right).$$

Отсюда вероятность превышения заданного ограничения на время задержки запросов класса $h = \overline{1,H}$ составляет:

$$\left.\begin{aligned}
&\Pr(\tau_{u_h} > u_h^*) = \exp\left(-\frac{k_h u_h^*}{u_h}\right)\sum_{i=1}^{k_h-1}\frac{(k_h u_h^*)^i}{i!\,u_h^i}, \quad \text{если } 0 \le \alpha_h \le 1;\\
&\Pr(\tau_{u_h} > u_h^*) = q_h\exp\left(-\frac{u_h^*}{t_{h_1}}\right) + (1-q_h)\exp\left(-\frac{u_h^*}{t_{h_2}}\right), \quad \text{если } \alpha_h > 1,
\end{aligned}\right\} \tag{4}$$

и ограничения (1) примут вид:

$$\left.\begin{aligned}
&\exp\left(-\frac{k_h u_h^*}{u_h}\right)\sum_{i=1}^{k_h-1}\frac{(k_h u_h^*)^i}{i!\,u_h^i} \le \delta_h^*, \quad \text{если } 0 \le \alpha_h \le 1;\\
&q_h\exp\left(-\frac{u_h^*}{t_{h_1}}\right) + (1-q_h)\exp\left(-\frac{u_h^*}{t_{h_2}}\right) \le \delta_h^*, \quad \text{если } \alpha_h > 1.
\end{aligned}\right\} \tag{5}$$

Поскольку число различных ДО СП даже при небольшом числе классов запросов оказывается значительным (например, при $H = 5$ более 4,5 тыс. корректных ДО, а при 10 — более 100 млн) [6], последовательный перебор всех возможных ДО приводит к значительным затратам времени. Для уменьшения этого времени используются эвристические методы.

Алгоритм распределения приоритетов основан на целенаправленном переборе дисциплин в классе ДО СП. Если ограничения (5) при некоторой ДО не выполняются хотя бы для одного класса запросов, необходимо с заданным шагом увеличивать производительность V системы до тех пор, пока для всех классов не будут выполнены ограничения. При найденном значении производительности ДО для каждого класса запросов в процессе аппроксимации по двум моментам определяются параметры аппроксимирующих распределений и по формулам (4) рассчитывается значение вероятности $\delta_h = \Pr(\tau_{u_h} > u_h^*)$ превышения заданных ограниче-

ний u_h^* для всех классов запросов. Показателем, определяющим необходимость изменения приоритета h-запросов, служит относительное отклонение ζ_h вероятности δ_h превышения заданного ограничения u_h^*, рассчитанного для рассматриваемого варианта назначения приоритетов, от заданного ограничения δ_h^*: $\zeta_h = (\delta_h^* - \delta_h)/\delta_h^*$ $(h = \overline{1, H})$. Необходимо увеличить приоритет класса, у которого ζ_h минимально, за счет уменьшения приоритета класса с максимальным ζ_h. Необходимо стремиться к тому, чтобы значения ζ_h у всех классов были примерно одинаковы.

Заключение. Предлагаемый подход к распределению приоритетов в системах с неоднородной нагрузкой при наличии вероятностных ограничений на среднее время пребывания в системе запросов разных классов позволяет решить задачу выбора наилучшей дисциплины обслуживания в классе дисциплин со смешанными приоритетами при минимальной производительности системы и тем самым обеспечить минимальную стоимость системы.

СПИСОК ЛИТЕРАТУРЫ

1. *Алиев Т. И.* Проектирование систем с приоритетами // Изв. вузов. Приборостроение. 2014. Т. 57, № 4. С. 30—35.
2. Рекомендация МСЭ-Т Y.1541 (02/2006 г.). Требования к сетевым показателям качества для служб, основанных на протоколе IP.
3. *Муравьева-Витковская Л. А.* Обеспечение качества обслуживания в мультисервисных компьютерных сетях за счет приоритетного управления // Изв. вузов. Приборостроение. 2012. Т. 55, № 10. С. 64—68.
4. *Алиев Т. И.* Основы моделирования дискретных систем. СПб: СПбГУ ИТМО, 2009. 363 с.
5. *Алиев Т. И.* Дисциплины обслуживания на основе матрицы приоритетов // Научно-технический вестник информационных технологий, механики и оптики. 2014. № 6 (88). С. 91—97.
6. *Алиев Т. И.* Аппроксимация вероятностных распределений в моделях массового обслуживания // Научно-технический вестник информационных технологий, механики и оптики. 2013. № 2 (84). С. 88—93.

Сведения об авторе

Тауфик Измайлович Алиев — д-р техн. наук, профессор; Университет ИТМО, кафедра вычислительной техники; E-mail: aliev@cs.ifmo.ru

Рекомендована кафедрой вычислительной техники

Поступила в редакцию 06.02.15 г.

Ссылка для цитирования: *Алиев Т. И.* Распределение приоритетов в системах с вероятностными ограничениями // Изв. вузов. Приборостроение. 2015. Т. 58, № 6. С. 415—420.

PRIORITY DISTRIBUTION IN QUEUEING SYSTEM WITH PROBABILITY CONSTRAINTS

T. I. Aliev
ITMO University, 197101, Saint Petersburg, Russia
E-mail: aliev@cs.ifmo.ru

The problem of prioritization in single-channel queueing system with non-uniform flow of demands is considered for the case of certain constraints on probability of permissible service time excess for demands of various classes. A solution to the problem is obtained in the class of queueing disciplines with mixed priorities specified in the form of priority matrix. The matrix elements represent the type (relative or absolute) and the level of priorities for demands of various classes. The algorithm of prioritization is based on purposeful search of queueing disciplines, the system productivity serves as the discipline effectiveness indicator.

Keywords: queueing system, time in system, probability constraints, productivity, queueing discipline, mixed priorities.

Data on author

Taufik I. Aliev — Dr. Sci., Professor; ITMO University; Department of Computer Science,
E-mail: aliev@cs.ifmo.ru

Reference for citation: *Aliev T. I.* Priority distribution in queueing system with probability constraints // Izvestiya Vysshikh Uchebnykh Zavedeniy. Priborostroenie. 2015. Vol. 58, N 6. P. 415—420 (in Russian).

DOI: 10.17586/0021-3454-2015-58-6-415-420