КОМПЬЮТЕРНЫЕ СИСТЕМЫ И СЕТИ

УДК 681.2

Т. И. Алиев

ПРОЕКТИРОВАНИЕ СИСТЕМ С ПРИОРИТЕТАМИ

Рассматриваются особенности проектирования систем с приоритетами при наличии ограничений на среднее время пребывания запросов в них. В процессе проектирования определяются дисциплина обслуживания запросов в классе дисциплин со смешанными приоритетами и производительность устройства, обеспечивающие минимальную стоимость системы.

Ключевые слова: *система с приоритетами, производительность, дисциплина обслуживания, смешанные приоритеты, время пребывания, стоимость системы.*

Введение. Качество функционирования вычислительных систем и компьютерных сетей, задаваемое в виде ограничений на время реакции (задержки запросов), превышение которых недопустимо или крайне нежелательно, обеспечивается за счет применения приоритетных стратегий управления процессами обработки и передачи данных. Например, в информационно-управляющих системах, находящихся в контуре систем автоматического управления технологическим оборудованием или подвижными объектами, превышение ограничений на время реакции может привести к резкому снижению эффективности функционирования системы или вообще к выходу ее из строя. В маршрутизаторах и коммутаторах компьютерной сети при передаче мультимедийных пакетов применение приоритетных стратегий управления трафиком направлено на обеспечение допустимых задержек, значения которых приведены в рекомендациях ITU-T Y.1541 [1]. Поскольку указанные ограничения в таких системах могут составлять доли секунд и даже миллисекунд, то одна из особенностей таких систем заключается в отсутствии обмена с внешней памятью в процессе управления. Таким образом, внешняя память не влияет на эффективность функционирования системы в целом и, в частности, на время реакции, являющееся основной характеристикой функционирования системы. При исследовании таких систем применяются модели в виде системы массового обслуживания с одним обслуживающим устройством и неоднородным потоком запросов, обрабатываемых с заданной производительностью. При этом емкость накопителей предполагается неограниченной, что справедливо для реальных систем, в которых вероятность потери запросов из-за ограниченной емкости накопителей не превышает 10^{-3} [2]. В этом случае задача проектирования системы сводится к синтезу стратегии управления потоком поступающих в систему запросов, задаваемой в виде некоторой дисциплины обслуживания (ДО), и определению производительности системы, обеспечивающих заданные ограничения на время пребывания запросов в системе.

Постановка задачи проектирования. В качестве исходных данных для решения задачи проектирования системы с неоднородным потоком запросов и приоритетным обслуживанием используется следующая совокупность величин: число классов запросов H, поступаю-

щих в систему; интенсивности $\lambda_1,\ldots,\lambda_H$ потоков запросов, которые будем полагать простейшими; средние ресурсоемкости $\theta_1,\ldots,\theta_H$ обработки запросов, задаваемые в виде среднего числа команд (инструкций), выполняемых при обработке запроса соответствующего класса; коэффициенты вариации $\nu_1,\ldots,\nu_H$ ресурсоемкости обработки запросов; ограничения $u_1^*,\ldots,u_H^*$ на время пребывания в системе запросов:

$$u_k \le u_k^* \quad (k=\overline{1,H}), \tag{1}$$

где u_k — среднее значение задержки (времени пребывания в системе) запросов класса k (k-запросов), зависящее от производительности V системы и ДО.

В качестве критерия эффективности рассмотрим стоимость системы: $S = S_1 + S_2$, где $S_1 = \gamma V^{\chi}$ — стоимость устройства (процессора), связанная зависимостью с его производительностью V через коэффициенты пропорциональности γ и нелинейности χ; $S_2 = s_0 E$ — стоимость памяти, предназначенной для хранения поступающих в систему запросов (s_0 — стоимость единицы памяти, например, байта; E — емкость памяти). Емкость E определяется максимальным числом k-запросов $\tilde{m}_k$, которые могут одновременно находиться в системе: $E = \sum_{k=1}^{H} d_k \tilde{m}_k$, где d_k — объем памяти, занимаемый одним запросом класса k. Максимальное число запросов $\tilde{m}_k$ может быть представлено как $\tilde{m}_k = f_k m_k$, где $m_k = \lambda_k u_k$ — среднее число запросов в системе, f_k — коэффициент, зависящий от закона распределения числа запросов в системе и допустимой вероятности потерь запросов из-за ограниченной емкости памяти $(k=\overline{1,H})$. В частности, в случае геометрического закона и допустимой вероятности 10^{-3} коэффициент $f_k = 10$ для систем, загрузка которых 40 % и более.

Таким образом, стоимость системы составит:

$$S = \gamma V^{\chi} + s_0 \sum_{k=1}^{H} d_k f_k \lambda_k u_k \,. \tag{2}$$

Задача проектирования систем с приоритетами формулируется следующим образом: найти ДО и определить производительность системы, которые обеспечивают выполнение ограничений (1) при минимальной стоимости системы (2).

Проектирование систем с приоритетами реализуется в три этапа:

1) определение нижней границы производительности системы, начиная с которой можно искать ДО, обеспечивающую выполнение ограничений (1);

2) синтез ДО, обеспечивающей выполнение заданных ограничений при наименьшей производительности системы;

3) определение оптимальной производительности системы, обеспечивающей минимальную стоимость системы при выбранной ДО.

Нижняя граница производительности системы V_0 соответствует значению, начиная с которого может решаться задача выбора ДО. Другими словами, если $V < V_0$, то не может быть найдена ДО, обеспечивающая требуемое качество функционирования системы.

При отсутствии ограничений на время пребывания запросов в системе нижняя граница производительности V_0' определяется из условия отсутствия перегрузки: $R < 1$, где

$R = \sum_{i=1}^{H} \rho_i$ — суммарная загрузка системы, $\rho_i = \frac{\lambda_i \theta_i}{V}$ — загрузка, создаваемая i-запросами, откуда $V > \sum_{i=1}^{H} \lambda_i \theta_i$, $V_0' = \sum_{i=1}^{H} \lambda_i \theta_i$.

Для систем с ограничениями (1) под нижней границей производительности V_0'' понимается значение, начиная с которого может существовать ДО, обеспечивающая выполнение заданных ограничений. Для определения V_0'' воспользуемся законом сохранения времени пребывания [2], который запишем в следующем виде:

$$\sum_{i=1}^{H} \rho_i u_i = \frac{R}{2(1-R)} \sum_{i=1}^{H} \lambda_i b_i^2 (1 + \nu_i^2) + \sum_{i=1}^{H} \rho_i b_i .$$

Заменив в этом выражении среднее значение времени пребывания u_i на ограничение u_i^* и учитывая неравенство (1), получим:

$$\sum_{i=1}^{H} \rho_i u_i^* \geq \frac{R}{2(1-R)} \sum_{i=1}^{H} \lambda_i b_i^2 (1 + \nu_i^2) + \sum_{i=1}^{H} \rho_i b_i . \qquad (3)$$

Выражение (3) представляет собой необходимое условие существования ДО, обеспечивающей выполнение ограничений (1).

Заменив в (3) все величины, зависящие от производительности V, на соответствующие выражения: $b_i = \frac{\theta_i}{V}$; $\rho_i = \frac{\lambda_i \theta_i}{V}$; $R = \frac{1}{V} \sum_{i=1}^{H} \lambda_i \theta_i$ и решив квадратное неравенство относительно V, получим:

$$V > \frac{1}{2}\left(\sum_{i=1}^{H} \lambda_i \theta_i + \sum_{i=1}^{H} \lambda_i \theta_i^2 \Big/ \sum_{i=1}^{H} \lambda_i \theta_i u_i^* \right) + \left[\frac{1}{4}\left(\sum_{i=1}^{H} \lambda_i \theta_i + \sum_{i=1}^{H} \lambda_i \theta_i^2 \Big/ \sum_{i=1}^{H} \lambda_i \theta_i u_i^* \right)^2 - \right.$$

$$\left. - \sum_{i=1}^{H} \lambda_i \theta_i \left(2 \sum_{i=1}^{H} \lambda_i \theta_i^2 - \sum_{i=1}^{H} \lambda_i \theta_i^2 (1 + \nu_i^2) \right) \Big/ 2 \sum_{i=1}^{H} \lambda_i \theta_i u_i^* \right]^{1/2} . \qquad (4)$$

Обозначим через V_0'' выражение в правой части неравенства (4). Значение V_0'' позволяет учитывать ограничения u_i^* $(i = \overline{1, H})$ на время пребывания запросов по всем классам. С уменьшением $u_1^*, \ldots, u_H^*$ требуемое значение V_0'' растет, а при больших значениях $u_1^*, \ldots, u_H^*$ — стремится к V_0'.

Значение V_0'' было получено из необходимого условия (3) существования ДО, обеспечивающей выполнение ограничений (1), и может рассматриваться как необходимое, но не достаточное для синтеза ДО. Кроме того, в (4) ограничения (1) учитываются по всем классам в совокупности, при этом не учитываются особенности каждого класса.

Значения производительности $V_1''', \ldots, V_H'''$, учитывающие ограничения на времена пребывания запросов по каждому классу, могут быть получены на основе следующих рассуждений.

Минимальное время пребывания в системе для k-запросов может быть обеспечено за счет присвоения этому классу самого высокого абсолютного приоритета [2]. Это время, согласно (1), должно быть меньше u_k^*:

$$\frac{\lambda_k b_k^2(1+\nu_k^2)}{2(1-\rho_k)} + b_k < u_k^* \qquad (k=\overline{1,H}). \tag{5}$$

Заменив в (5) параметры, зависящие от производительности, после некоторых преобразований получим:

$$V > \frac{1}{2}\left(\lambda_k\theta_k + \frac{\theta_k}{u_k^*}\right) + \left[\frac{1}{4}\left(\lambda_k\theta_k + \frac{\theta_k}{u_k^*}\right)^2 - \frac{\lambda_k\theta_k^2(1-\nu_k^2)}{2u_k^*}\right]^{1/2} \qquad (k=\overline{1,H}). \tag{6}$$

Обозначим через V_k''' выражение, стоящее в правой части неравенства (6). Значение V_k''' представляет собой нижнюю границу производительности для *k*-запросов, учитывающую ограничение $u_k \le u_k^*$ $(k=\overline{1,H})$.

Окончательно нижняя граница производительности системы V_0 при ограничениях (1) определяется как $V_0 = \max(V_0'', V_1''', \ldots, V_H''')$.

Синтез дисциплины обслуживания. Очевидно, что требуемое качество функционирования системы, заданное в виде ограничений (1), может быть достигнуто при любой ДО только за счет производительности системы, при этом наилучшим решением следует считать ДО, обеспечивающую эти ограничения при производительности, близкой к нижней границе. Разработка высокоэффективных и хорошо формализованных алгоритмов синтеза приоритетных ДО, позволяющих получить однозначное оптимальное решение, представляет собой сложную задачу. В этом случае целесообразно использовать эвристические алгоритмы, предполагающие целенаправленный перебор множества ДО в классе ДО со смешанными приоритетами (СП) [3]. Дисциплина обслуживания СП задается матрицей приоритетов (МП) $Q = [q_{ij}\ (i,j=1,\ldots,H)]$, где q_{ij} описывает приоритет *i*-запросов по отношению к *j*-запросам: 0 — нет приоритета, 1 — приоритет относительный (ОП) и 2 — приоритет абсолютный (АП). Тогда среднее время пребывания *k*-запросов [4]:

$$u_k = \frac{\sum_{i=1}^{H}(2-q_{ki})(1+q_{ki})\lambda_i b_i^2(1+\nu_i^2)}{[2-\sum_{i=1}^{H}q_{ik}(3-q_{ik})\rho_i][2-\sum_{i=1}^{H}(1-q_{ki})(2-q_{ki})\rho_i]} + \frac{2b_k}{2-\sum_{i=1}^{H}q_{ik}(q_{ik}-1)\rho_i}.$$

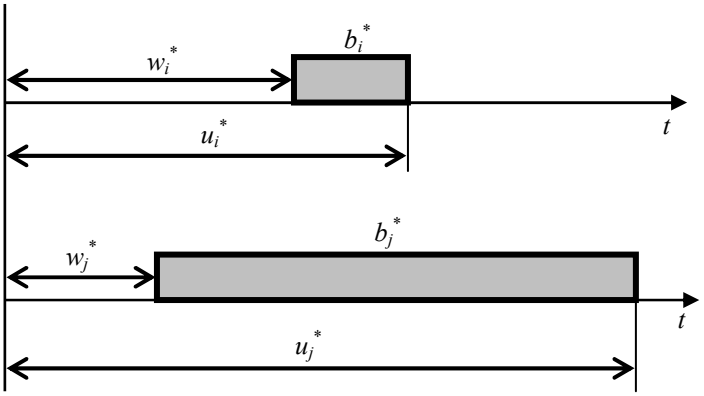
Задача синтеза ДО сводится к определению значений q_{ij}, при которых выполняются ограничения (1).

Аналитическое решение системы неравенств (1) не представляется возможным, поскольку число различных ДО СП даже при небольшом количестве классов запросов значительно. Так, в случае пяти классов $(H=5)$ число корректных ДО более 4,5 тыс, а при $H=10$ — более 100 млн. Последовательный перебор всех возможных МП приводит к большим затратам времени, снизить которые можно, используя эвристические алгоритмы.

Алгоритм распределения приоритетов, основанный на целенаправленном переборе различных ДО, предполагает задание начального варианта назначения приоритетов. Классы запросов должны быть расположены в порядке убывания отношения θ_k / u_k^*, т.е. по правилу: $\theta_\alpha / u_\alpha^* > \theta_\beta / u_\beta^* > \ldots > \theta_\omega / u_\omega^*$. Для последовательности номеров $\alpha,\beta,\ldots,\omega$ формируется начальный вариант распределения приоритетов по правилу: классам запросов, расположенным правее в указанной последовательности, назначается приоритет не выше, чем классам, расположенным левее. В нашем случае наиболее высокий приоритет назначается запросам класса

α и самый низкий — запросам класса ω. В качестве начального варианта может быть выбрана ДО ОП или ДО АП.

Необходимость упорядочения классов запросов по указанному правилу иллюстрирует рисунок, из которого видно, что из-за большой ресурсоемкости обслуживания θ_j ($b_j = \theta_j / V$) j-запросов, по сравнению с i-запросами, целесообразно более высокий приоритет назначить j-запросам, несмотря на то что $u_j^* > u_i^*$. Это необходимо для обеспечения меньшего времени ожидания j-запросов: $w_j^* < w_i^*$.



Каждый последующий вариант назначения приоритетов формируется на основании показателя, определяющего необходимость изменения приоритета k-запросов. Таким показателем может служить относительное отклонение ζ_k времени пребывания u_k, полученного для рассматриваемого варианта назначения приоритетов, от заданного ограничения: $\zeta_k = (u_k^* - u_k)/u_k^*$ $(k = \overline{1,H})$. При этом в первую очередь необходимо повышать приоритет класса с минимальным значением ζ_k. Приоритет может изменяться путем изменения приоритета данного класса относительно других классов или других классов относительно данного. При этом необходимо стремиться к тому, чтобы значения ζ_k для всех классов были одинаковы.

Определение оптимальной производительности. На последнем этапе определяется оптимальное значение производительности, которое, обеспечивая при выбранной ДО выполнение заданных ограничений (1), позволяет минимизировать стоимость S системы (см. (2)).

Задача отыскания минимума функции S сводится к решению уравнения, полученного путем приравнивания нулю производной по V от функции S:

$$\gamma \chi V^{\chi-1} + s_0 \sum_{k=1}^{H} d_k f_k \lambda_k \frac{\mathrm{d}u_k}{\mathrm{d}V} = 0,$$

где u_k определяется для выбранной на предыдущем этапе ДО СП. Это уравнение решается с учетом ограничений (1), к которым на данном этапе могут быть добавлены требования по производительности и надежности системы [4, 5]. Ограничения определяют область допустимых значений V, в которой ищется оптимальное значение производительности. В частности, решив систему неравенств $u_k(V) \le u_k^*$, находим, что V должно определяться из условия: $V > \max\{V_1, \ldots, V_H\}$, где $u_k(V_k) = u_k^*$ $(k = \overline{1,H})$.

Заключение. Предлагаемый подход к проектированию систем с приоритетами при наличии ограничений на среднее время пребывания в системе запросов разных классов позволяет решить задачу выбора наилучшей дисциплины обслуживания в классе дисциплин со смешанными приоритетами и обеспечить минимальную стоимость системы.

СПИСОК ЛИТЕРАТУРЫ

1. Рекомендация МСЭ-Т Y.1541 (02/2006 г.). Требования к сетевым показателям качества для служб, основанных на протоколе IP. 2006.
2. *Алиев Т. И.* Основы моделирования дискретных систем. СПб: СПбГУ ИТМО, 2009. 363 с.
3. *Алиев Т. И.* Характеристики дисциплин обслуживания заявок с несколькими классами приоритетов // Изв. АН СССР. Техническая кибернетика. 1987. № 6. С. 188—191.
4. *Алиев Т. И.* Задачи синтеза систем с потерями // Изв. вузов. Приборостроение. 2012. Т. 55, № 10. С. 57—63.
5. *Богатырев В. А., Богатырев С. В., Богатырев А. В.* Функциональная надежность вычислительных систем с перераспределением запросов // Изв. вузов. Приборостроение. 2012. Т. 55, № 10. С. 53—56.

Сведения об авторе

Тауфик Измайлович Алиев — д-р техн. наук, профессор; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра вычислительной техники; заведующий кафедрой E-mail: aliev@d1.ifmo.ru

Рекомендована кафедрой вычислительной техники

Поступила в редакцию 23.12.13 г.

УДК 004.89: 002.53

Л. А. Муравьева-Витковская

МЕТОД РАСЧЕТА ХАРАКТЕРИСТИК ЗАМКНУТЫХ ДЕТЕРМИНИРОВАННЫХ МОДЕЛЕЙ МУЛЬТИСЕРВИСНЫХ КОМПЬЮТЕРНЫХ СЕТЕЙ

Рассматривается метод расчета характеристик мультисервисных компьютерных сетей, моделями которых являются замкнутые сети с детерминированным временем обслуживания заявок в узлах. Предположение о детерминированном обслуживании позволяет учитывать реальную статистику распределений фиксированных длин пакетов в мультисервисных компьютерных сетях.

Ключевые слова: *мультисервисные компьютерные сети, детерминированное обслуживание, замкнутые сети массового обслуживания.*

Введение. Для анализа процесса функционирования мультисервисных компьютерных сетей (КС) широко используются сетевые модели массового обслуживания, учитывающие наличие множества ресурсов. Замкнутые сети массового обслуживания (МО), содержащие постоянное число заявок, успешно применяются для моделирования работы мультисервисных КС [1].

Хорошо известны методы расчета характеристик замкнутых сетей МО при распределении по экспоненциальному закону временных интервалов обслуживания заявок в узлах сети. В настоящей статье предлагается метод расчета характеристик мультисервисных КС, моделями которых являются замкнутые сети с детерминированным временем обслуживания заявок в узлах. Предположение о детерминированном обслуживании позволяет учитывать реальную статистику распределений фиксированных длин пакетов [2].

Постановка задачи. В качестве моделей функционирования мультисервисных КС рассматриваются замкнутые сети МО, содержащие n узлов с произвольным количеством обслуживающих приборов, время обслуживания в которых детерминировано. Пакетам в моделях мультисервисных КС будут соответствовать заявки.