ИССЛЕДОВАНИЕ ПАРАМЕТРОВ ДИНАМИЧЕСКИХ ПРИОРИТЕТОВ В ВЫСОКОСКОРОСТНЫХ СЕТЯХ С ДВУМЯ ТИПАМИ ЗАЯВОК

Агаси Меликов[1], Балами Исмайлов[2]

[1]Институт кибернетики НАН Азербайджана
AZ 1141, Баку, ул.Ф. Агаева 9
[2]Сумгаитский государственный университет
43-й квартал, AZ-5008 Сумгаит, Азербайджан, Email: balemi@rambler.ru

Аннотация

Разработаны вычислительные процедуры для расчета оптимальных значений параметров динамических приоритетов в высокоскоростных сетях с двумя классами заявок, где функция приоритетности имеет мультипликативный вид. Критерием оптимальности является суммарная длина очереди разнотипных заявок, при этом имеются ограничения на времени ожидания заявок каждого типа. Задача сформулируется в виде некоторой задачи дробно-линейного программирования и решены с применением метода искусственного базиса. Проведены объемные вычислительные эксперименты и получены численные результаты.

Введение

В коммутаторах высокоскоростных мультимедийных сетей с общим буферным пространством (Shared Buffering) в целях удовлетворения заданных уровней качества обслуживания (Quality of Service) используются различные приоритеты. При этом приоритеты, определяющие процедуры принятия в буфер разнотипных заявок, называются пространственными приоритетами (Space Priorities), а приоритеты, задающие правила выбора типа заявки из буфера, получили название временных приоритетов (Time priorities)[1].

В классических схемах приоритетного обслуживания, как правило, предполагается, что заявки определенного типа обладают одновременно высокими (по сравнению с заявками другого типа) приоритетами обоих видов. Вместе с тем, в доступной литературе известны (хотя в очень незначительном количестве) работы, в которых изучаются схемы с различными уровнями пространственных и временных приоритетов (см. например [2-4]).

Пространственные приоритеты позволяют, главным образом, управлять интенсивностями (вероятностями) потери разнотипных заявок, в то время как временные приоритеты существенным образом влияют на времени их задержки (ожидания) в буфере. В литературе предложены различные схемы определения пространственных приоритетов с целью нахождения оптимальных (в известном смысле) размеров буферной памяти в узлах различных сетей обслуживания (см. например, [5-7])

Следует отметить, что исходя из реальных условий работы высокоскоростных мультимедийных сетей обслуживания целесообразно в качестве временных приоритетов использовать динамические приоритеты, изменяющиеся с течением времени. В системах с динамическими приоритетами решение об определении типа заявки, выбираемой на обслуживание, зависит от значения некоторой функции $J_i(t)$, определяющий мгновенный приоритетный индекс заявки $i-$го типа в момент t, $i=\overline{1,n}$, где $n-$общее количество типов заявки. Эту функцию аналитически можно определить различными способами. Так, функция мультипликативного характера $J_i(t)=b_i\tau_{q_i}(t)$ определена в работе [8], где b_i-некоторый коэффициент, определяющий скорость изменения приоритетности заявки $i-$го типа при ее ожидании в очереди; $\tau_{q_i}(t)-$ случайное время ожидания заявки $i-$ го типа от момента поступления до текущего момента t.

Исследование динамических приоритетов, зависящих от времени, особенно актуально при организации обслуживания заявок с ограниченным временем ожидания [5], а также в системах с конечным временем старения заявок [7].

Одно из обстоятельств, сдерживающей широкое применение динамических приоритетов является отсутствие методики определения коэффициентов $b_i, i=\overline{1,n}$ в указанном выше формуле функции приоритетности $J_i(t)$. Вместе с тем, именно за счет надлежащего выбора этих коэффициентов b_i можно повысить эффективность функционирования системы относительно выбранного критерия качества.

В настоящей работе предложен метод решения задачи выбора оптимальных значений указанных коэффициентов, минимизирующих суммарное значение длины очереди в системах с двумя типами заявок.

Описание модели с динамическими приоритетами

Рассматривается узел распределенной компьютерной сети, в которой для обслуживания поступают пуассоновские потоки заявок, при этом интенсивность $i-$го потока равна $\lambda_i, i=\overline{1,n}$. Предполагается, что время обслуживания заявки $i-$го типа имеет экспоненциальное распределение со средним значением μ_i^{-1}, $i=\overline{1,n}$.

В сети необслуженные заявки образуют, неограниченную очередь в буферной памяти. Приоритеты обслуживания заявок изменяются в зависимости от длительности их ожидания в очереди, т.е. выбор типа заявки для обслуживания осуществляется с учетом текущего значения функции приоритетности $J_i(t)=b_i\tau_{q_i}(t)$, зависящей для каждой заявки от времени ее ожидания в очереди. При этом обслуживания заявок осуществляется в порядке относительного приоритета, т.е. в момент освобождения канала на обслуживание поступает заявка, обладающая в данный момент времени максимальным значением функции приоритетности J_i (t) из числа заявок, находящихся в буферной памяти.

Здесь исследуется два случая - относительные приоритеты при линейно возрастающей и линейно убывающей функции приоритетности заявок в распределенной сети.

Случай 1. Динамические приоритеты при линейно возрастающей функции приоритетности заявок.

Такие приоритеты особенно актуальны в системах с ограниченным временем ожидания заявок в очереди.

В этом случае в выражении функции приоритетности $J_i(t)$ коэффициенты b_i, определяющие скоростей изменения приоритетности разнотипных заявок при их пребывания в буферной памяти сети, являются положительными величинами, $b_i>0, i=\overline{1,n}$. При использовании приоритетов данного типа, если в некоторый момент T_i поступила заявка i - го типа, а в момент T_k заявка k- го типа, $T_k>T_i$, и при этом выполняется неравенство $b_k>b_i$, то в соответствии с рассматриваемой дисциплиной динамического назначения приоритетов до момента $T_{ik}=(b_kT_k-b_iT_i)/(b_k-b_i)$ более высоким приоритетом обладает заявка i - го типа, а при t > T_{i_k} заявка k -го типа, несмотря на то, что она находится в буферной памяти сети меньшее время, чем заявка i -го типа. Это означает, что если компьютер сети освободится от обслуживания некоторой заявки до момента T_{ik}, то на обслуживание будет выбрана заявка i -го типа; в противном случае, на обслуживание первой поступит заявка k -го типа. Таким образом, при данной дисциплине обслуживания даже заявки с малым коэффициентом b_i при достаточно больших значениях времени ожидания в очереди могут получить преимущество на обслуживание перед всеми другими типами заявок.

В стационарном режиме (т.е. при $\rho=\lambda/\mu<1$) средняя время ожидания в очереди заявок i - го типа определяется так [8]:

$$\tau_{q_i}=\frac{\tau_{q_0}/(1-\rho)-\sum_{j=i+1}^{n}\rho_j\tau_{q_j}\left(1-b_j/b_i\right)}{1-\sum_{j=1}^{n}\rho_j\left(1-b_i/b_j\right)}. \qquad (1)$$

где $\tau_{q_0}=\frac{1}{2}\sum_{i=1}^{n}\lambda_iT_i^2\left(1+\upsilon_i^2\right)$ - средняя время ожидания завершения начатого обслуживания одной заявки, ϑ_i-коэффициент вариации времени обслуживания заявок $i-$ го приоритета, $\rho_i=\lambda_i/\mu_i$ - загрузка сети обслуживанием заявок $i-$го приоритета, $\rho=\lambda/\mu$ - общая загрузка сети всеми потоками заявок.

Из формулы (1) видно, что зависимость τ_{q_i} от b_i определяется только отношениями этих параметров, а не их собственными значениями.

Случай 2. Динамические приоритеты при линейно убывающей функции приоритетности заявок.

В этом случае функция приоритетности заявок линейно убывает при их ожидании в очереди, т.е. коэффициенты, определяющие скоростей убывания приоритетности заявок, являются отрицательными величинами, $b_i < 0, i = \overline{1,n}$.

Отметим, что такая ситуация характерна, в частности, для сетей с быстро стареющей информации, ценность которой для функционирования сети снижается пропорционально времени задержки перед обслуживанием.

В этом случае также выбор заявки из числа ожидающих в очереди осуществляется, таким образом, чтобы на обслуживание поступила заявка, для которой в момент освобождения компьютера функция приоритетности имеет максимальное значение.

В этом случае время ожидания в очереди заявок $i-$го типа определяются так [8]:

$$\tau_{q_i} = \frac{\tau_{q_0} + \sum_{j=1}^{i-1} \rho_i \tau_{q_j} \left(1 - \frac{b_j}{b_i}\right)}{1 - \sum_{j=1}^{i} \rho_j - \sum_{j=i+1}^{n} \rho_j \frac{b_i}{b_j}}. \quad (2)$$

Так же как и в предыдущем случае, значение τ_{q_i} зависят лишь от отношения параметров b_i.

Оптимальное назначение динамических приоритетов при $n = 2$.

Здесь рассматривается задача выбора оптимальных значений коэффициентов b_i для случая двух типов заявок. При этом в качестве критерия оптимальности выбирается суммарная длина очереди разнотипных заявок, иными словами требуется найти такие значение, b_i^*, $i = \overline{1,2}$ чтобы суммарная длина очереди была минимальной при заданных ограничениях на времени ожидания заявок каждого типа. Математически эта задача записывается так:

$$L_q = \lambda_1 \tau_{q_1} + \lambda_2 \tau_{q_2} \to \min. \quad (3)$$

при ограничениях $\tau_{q_1} \le \bar{\tau}_{q_1}$. (4)

$$\tau_{q_2} \le \bar{\tau}_{q_2}, \quad (5)$$

где $\bar{\tau}_{q_1}$ и $\bar{\tau}_{q_2}$ являются известные величины.

Для первого случая задача (3)-(5) в явном виде записывается так:

$$L_q = \frac{\tau_{q_0}}{1-\rho} \cdot \frac{(\lambda_1 + \lambda_2) b_1 + (\lambda_1 - \rho\lambda_1) b_2}{(1-\rho_1) b_1 + \rho_1 b_2} \longrightarrow \min, \quad (6)$$

при ограничениях $\frac{\tau_{q_0}}{1-\rho} \cdot \frac{(\lambda_1 - \rho\lambda_1) b_1 + \lambda_1 b_2}{(1-\rho_1) b_1 + \rho b_2} \le \bar{\tau}_{q_1}$. (7)

$$\frac{\tau_{q_0}}{1-\rho} \cdot \frac{\lambda_2 b_1}{(1-\rho_1) b_1 + \rho_1 b_2} \le \bar{\tau}_{q_2}. \quad (8)$$

Для второго случая задача (3)-(5) имеет следующий вид:

$$L_q = \frac{\lambda_1 \tau_{q_0} b_2 (1-\rho) + \lambda_2 \tau_{q_0} (b_2 - (\rho_2 + \rho_1) b_1)}{(b_2 - b_2\rho_1 - \rho_2 b_1)(1-\rho)} \longrightarrow \min, \quad (9)$$

при ограничениях $\frac{\tau_{q_0} b_2}{b_2(1-\rho_1) - \rho_2 b_1} \le \bar{\tau}_{q_1}$. (10)

$$\frac{\tau_{q_0}(b_2 - \rho_2 b_1 - \rho_1 b_1)}{(1-\rho)(b_2 - b_2\rho_1 - \rho_2 b_1} \le \bar{\tau}_{q_2} . \qquad (11)$$

Задачи (6)-(8) и (9)-(11) являются задачами дробно–линейного программирования (ДЛП). С использованием методов ДЛП эти задачи могут сведены к основной задаче линейного программирования (ОЗЛП).Эти задачи решены с применением метода искусственного базиса и получены численные результаты.

Выводы

В работе предложены вычислительные процедуры для нахождения оптимальных значений параметров динамических приоритетов, зависящих от времени ожидания заявок в очереди. Решение указанных задач в случае двух типов заявок сводится к решению определенных задач дробно-линейного программирования. В настоящее время проводятся исследования по обобщению разработанных процедур для решения рассматриваемых проблем для случае произвольного числа типов заявок.

Литература:

[1] Pitts J.M., Schormans J.A. Introduction to ATM design and performance.N.Y.: John Wiley & Sons., 1997.
[2] Chao H.J., Peckan I.H. Queue management with multiple delay and loss priorities for ATM switches // Proc. ICC'94,1994.-pp.1184-1189.
[3] Lee Y., Choi B. D.Queueing system with multiple delay and loss priorities for ATM networks // Information systems .-2001.- vol. 138. – pp.7-29.
[4] Kim C.S., Melikov A.Z., Ponomarenko L.A. Approximation method for performance analysis of queuing system with multimedia traffics, // Applied and Comput. Math.-2007.-vol.6, no.2.- pp.218-226.
[5] .Пономаренко, Л.А., Меликов А.З., Исмайлов Б.Г. Оптимизация структуры сетей передачи данных, //Проблемы информатизации и управления. (Киев),1997г., Выпуск1, с.155-160
[6] .Меликов А.З. ,Исмайлов Б.Г. Оптимизация иерархических структур сетей передачи данных, // Изв. НАНА, Серия физико-технических и математических наук,№6,1997г. с.69-73.
[7] . Исмайлов Б.Г Проектирование распределенной сети обслуживания объектов добычи и подготовки газа.// М.: Научно-технический сборник. 2001,№1, с.29-35.
[8] Клейнрок Л. Коммуникационные сети.- М.: Наука,1970, с.256.