УДК 004.7.052

Меликов А.З., д.т.н.,
Джафар-заде Т.И.

АНАЛИЗ МОДЕЛИ ДВУХПОТОКОВОЙ СИСТЕМЫ ОБСЛУЖИВАНИЯС ОБЩЕЙ ОЧЕРЕДЬЮ И СКАЧКООБРАЗНЫМИ ПРИОРИТЕТАМИ

Национальная Академия Авиации
Баку, Азербайджан

Рассмотрена модель системы обслуживания с двумя типами вызовов и общей очередью при наличии скачкообразных приоритетов. Переход вызовов низкого приоритета в очередь вызовов высокого приоритета зависит от состояния очереди. Разработан метод расчета показателей качества обслуживания в данной системе.

Введение

Классические приоритеты делятся на следующие классы: статические, динамические по времени и динамические по состояниям (ситуационные). Статическими (относительными) называются приоритеты, которые устанавливаются до начало работы системы и не изменяются за все время ее работы. Здесь каждый трафик имеет свой приоритет и в момент освобождения канала выбирается вызов из головы той очереди, которая имеет наивысший приоритет среди всех непустых очередей. Эти приоритеты в англоязычной литературе называются *HOL*-приоритеты (*Head-Of-Line*). При использовании динамических (по времени) приоритетов, приоритет вызова каждого типа изменяется в зависимости от времени его пребывания в очереди. При этом могут быть использованы различные функции, которые определяют закон изменения этих приоритетов. Подобные приоритеты впервые были введены в [1]. Ситуационные приоритеты предполагают, что приоритет вызова каждого типа определяется в зависимости от состояния очереди, при этом состояния очереди обычно задается с помощью вектора, компоненты которых указывают число разнотипных вызовов в очереди. Эти приоритеты впервые были введены в [2]. Теория и области приложения ситуационных приоритетов далее были развиты в [3].

В последние годы появились работы [4-9], в которых изучаются новый тип *HOL*-приоритетов. Первая работа в этом направлении является [4]. Авторы указанной работы дали этим приоритетам название *Head-Of-Line with Priority Jumps* (*HOL-PJ*). При использовании этих приоритетов вызовы низкого приоритета могут переходить в очередь вызовов высокого приоритета, при этом условия перехода могут быть разными. Так, например, в работе [4] переход из одной очереди в другой определяется исходя из времени ожидания в очереди. В работах [5-9] предложены другие виды *HOL-PJ* для системы обслуживания с дискретным временем (т.е. системы, в которой время разделены на слоты) и с двумя типами вызовов – вызовы высокого приоритета (*H*-вызовы) и вызовы низкого приоритета (*L*-вызовы). В работе [5] предложена схема *HOL-MBP* (*Head-Of-Line Merge-By-Probability*), согласно которой в конце каждого временного слота все *L*-вызовы переходят в конец очереди *H*-вызовов с вероятностью β, $0<\beta<1$. Модификация схемы *HOL-MBP* была изучена в работе [6]. Она получила название HOL-JOS (*Head-Of-Line Jump-Or-Serve*) и в отличие от предыдущей схемы здесь только один *L*-вызов из головы очереди переходит в *H*-очередь. В схемах *HOL-JIA* (*Head-Of-Line Jump-If-Arrival*) [7, 8] в отличие от схемы *HOL-JOS*, возможный переход *L*-вызова в *H*-очередь зависит не только от содержания *H*-очереди в начале слота, но этот переход зависит также от числа поступлений *L*-вызовов в период данного слота.

В [4-9] авторы разработали формулы для производящих функций длины очереди вызовов обоих типов и времени ожидания в очереди *H*-вызовов, а также их моменты. Кроме того им удалось найти среднее время ожидания в очереди *L*-вызовов.

Отметим, что основной целью введения

скачкообразных приоритетов является разрешения проблемы старения L-вызовов в системах с HOL-приоритетами. Эта проблема особенно актуально в системах, где нагрузка H-вызовов намного превышают нагрузки L-вызовов. При этом очевидно, что введения скачкообразных приоритетов позволяет разрешить указанную проблему за счет увеличения времени ожидания в очереди H-вызовов. Следовательно, при введении скачкообразных приоритетов необходимо учитывать допустимые границы увеличения времени ожидания в очереди H-вызовов.

Следует отметить, что указанные выше работы [4-9] посвящены исследованию моделей систем с бесконечными очередями, которые не могут быть приняты в качестве адекватных моделей реальных систем телекоммуникации, так как реальные системы, как правило, имеют ограниченные буферные накопители для временного хранения разнотипных вызовов (пакетов). Иными словами, для широкого внедрения приоритетов типа *HOL-PJ* потребуется определить их эффективность в реальных системах.

В настоящей работе вводится новый класс скачкообразных приоритетов в системе обслуживания с непрерывным временем и общей ограниченной очередью для разнотипных вызовов. Введенные приоритеты имеют рандомизированный характер, т.е. они позволяют осуществить переход из L-очереди в H-очередь лишь в моменты поступления L-вызовов, при этом вероятность такого перехода зависит от числа разнотипных вызовов в очередях. Введение ограничений на размер общего буфера для ожидания разнотипных вызовов приводит к необходимости определения нового показателя качества обслуживания (*Quality of Service*, *QoS*) – вероятности потери пакетов (*Cell Loss Probability*, *CLP*). Другим отличающимся моментом этой работы от работ [4-9] состоит в том, что для анализа системы используется подход, основанный на теории фазового укрупнения состояний двумерных цепей Маркова [10]. С использованием этого подхода разработаны простые вычислительные процедуры для нахождения всех показателей *QoS* изучаемой системы.

Определение скачкообразных приоритетов

Подробное описание исследуемой системы обслуживания состоит в следующем. На вход одноканальной системы поступают два пуассоновские потоки разнотипных вызовов, при этом интенсивность i-го потока равна λ_i, i=1,2. Первый поток представляет собой поток вызовов реального времени (H-вызовы), в то время как второй поток является потоком вызовов нереального времени (L-вызовы). Время занятия канала является случайной величиной, подчиненной показательному закону распределения с параметром μ для вызовов обоих типов.

Для ожидания в очереди разнотипных вызовов имеется общий буфер с максимальным размером R, $0<R<\infty$, при этом предполагается, что в буфере производится виртуальное разделение очереди разнотипных вызовов. Ограниченность буфера означает, что если в момент поступления вызова любого типа общий буфер полностью заполнен, то он теряется независимо от числа разнотипных вызовов в буфере. Вызовы реального времени имеют высокие относительные приоритеты перед вызовами нереального времени. Это означает, что при освобождении канала на обслуживание из очереди всегда выбирается вызов первого типа независимо от числа вызовов второго типа в очереди, а также времени их ожидания в очереди. Внутри каждого потока используется дисциплина "первый пришел – первым обслужился" (*First Come First Serviced*).

Скачкообразные приоритеты вводятся с целью увеличения шансов L-вызовов быть обслуженными, при этом, очевидно, что ожидается несущественное ухудшение показатели качества обслуживания H-вызовов. Основными вопросами при введении скачкообразных приоритетов являются определение а) момента перехода от L-очереди к H-очереди и б) числа L-вызовов, переходящих к H-очереди. Здесь скачкообразные приоритеты определяются следующим образом. Прежде всего, отметим, что H-вызовы всегда принимаются с вероятностью 1, если в момент их поступления имеется хотя бы одно свободное место в H-буфере; в противном случае они теряются с вероятностью 1. Если в момент поступления L-вызова числа вызовов данного типа в буфере равно k, и при этом имеется свободное место в буфере, то с вероятностью $\alpha(k)$ один L-вызов мгновенно переходит в H-очередь (для определенности изложения предположим, что в H-вызовом становится L-вызов, стоящий в голове очереди L-вызовов); с допо-

лнительной вероятностью $1-\alpha(k)$ поступивший L-вызов присоединяется к очереди, если там имеется свободное место, и никаких переходов не осуществляются. В случае успешного "прыжка" L-вызов становится H-вызовом, и в дальнейшем обслуживается как H-вызов согласно HOL-приоритеты. Если в момент поступления L-вызова не имеется свободное место в общей очереди, то с вероятностью 1 он теряется.

Как видно из описания скачкообразных приоритетов, они включаются в моменты поступления L-вызовов, зависят лишь от числа таких вызовов в буфере и только один L-вызов может осуществить переходит в H-очередь. Такая схема определения приоритетов объясняется по следующим причинам. Во-первых, поскольку эти приоритеты вводятся с целью увеличения шансов L-вызовов обслужиться в приемлемые сроки (т.е. не дать им возможность стареть в очереди), то естественно полагать, что они должны включаться в моменты их поступления. Во-вторых, разрешение только одному L-вызову перейти в H-очередь объясняется тем, что при таком предположении промежуточные математические выкладки окажутся достаточно простыми, и потому удается получить аналитически трактуемые результаты. В-третьих, как будет видно из дальнейшего изложения, разработанный подход позволяет при определении вводимых приоритетов учитывать и число H-вызовов в очереди.

Отметим некоторые важные (частные) схемы введенных выше скачкообразных приоритетов.

1. *Равномерная схема.* Так назовем схему, в которой вероятности $\alpha(k)$ являются постоянными и не зависят от числа L-вызовов в буфере, т.е. $\alpha(k)=\alpha$ для любого $k=0,1,\ldots, R-1$. В случае $\alpha=0$ получаются классические HOL-приоритеты.

2. *Пороговая схема.* В данной схеме вводятся пороговые параметры L_i, $i=1,\ldots,r$, и вероятности $\alpha(k)$ определяются так:

$$\alpha(k)=\begin{cases}\alpha_i, & \text{если } L_{i-1}\le k<L_i, i=1,2,..,r-1,\\ \alpha_r, & \text{если } L_{r-1}\le k\le L_r.\end{cases}$$

Здесь полагается, что $L_0:=0$, $L_r:=R$. При этом вероятности α_i, $i=1,\ldots,r$ могут быть определены различными способами, т.е. они могут быть возрастающими (относительно i), убывающими, случайными и т.д.

Методы расчета

Рассмотрим задачу нахождения показателей QoS этой модели. Основными показателями QoS являются стационарная вероятность блокировки вызовов i-го типа (CLP_i), среднее число вызовов каждого типа в буферах (Q_i), а также среднее времени их ожидания в буфере (CTD_i), $i=1,2$.

Поскольку время обслуживания разнотипных вызовов имеют одинаковые средние значения, то состояние буферов в произвольный момент времени может быть описано с помощью двумерного вектора $\boldsymbol{n}=(n_1,n_2)$, где n_i означает число i-вызовов в буфере, $i=1,2$. Иными словами, функционирование данной системы описывается двумерной цепью Маркова со следующим фазовым пространством состояний (ФПС):

$$S:=\{\boldsymbol{n}: n_i=0,1,\ldots,R,\ i=1,2;\ n_1+n_2\le R\}. \qquad (1)$$

Переходы между состояниями системы происходят лишь в моменты поступления вызовов и ухода их из системы после завершения обслуживания. С учетом выше изложенного заключаем, что неотрицательные элементы Q-матрицы данной многомерной цепи определяются из следующих соотношений:

$$q(\boldsymbol{n},\tilde{\boldsymbol{n}})=\begin{cases}\lambda_1+\lambda_2\alpha(n_2), & \text{если } \tilde{\boldsymbol{n}}=\boldsymbol{n}+\boldsymbol{e}_1,\\ \lambda_2, & \text{если } \tilde{\boldsymbol{n}}=\boldsymbol{n}+\boldsymbol{e}_2,\\ \mu_i & \text{если } n_1>0, \tilde{\boldsymbol{n}}=\boldsymbol{n}-\boldsymbol{e}_1 \text{ или } n_1=0, \tilde{\boldsymbol{n}}=\boldsymbol{n}-\boldsymbol{e}_2,\\ 0 & \text{в остальных случаях,}\end{cases} \qquad (2)$$

где $\boldsymbol{e}_1=(1,0)$, $\boldsymbol{e}_2=(0,1)$.

При любых положительных значениях параметров входящих трафиков все состояния являются сообщающимися, и, следовательно, система является эргодической. Стационарную вероятность состояния $\boldsymbol{n}\in S$ обозначим через $p(\boldsymbol{n})$. Стандартным путем нахождения

стационарных вероятностей состояний является составление и решение соответствующей системы уравнений равновесия (СУР). Она составляется с учетом соотношений (2), при этом к ней добавляется нормирующее условие:

$$\sum_{n\in S} p(n)=1. \qquad (3)$$

После нахождения вероятностей состояний системы можно определить ее показатели *QoS*. Так, вероятности потери разнотипных вызовов равны друг другу и определяются так:

$$CLP_1 = CLP_2 = \sum_{k=0}^{R} p(R-k,k). \qquad (4)$$

Для нахождения среднего числа разнотипных пакетов в очереди (Q_k, k=1,2) используется стандартный способ определения среднего значения дискретной случайной величины:

$$Q_k = \sum_{i=1}^{R} i\,\xi_k(i), \qquad (5)$$

где $\xi_k(i) = \sum_{n\in S} p(n)\delta(n_k,i), k = 1,2,$ являются маргинальными распределениями исходной модели.

После нахождения показателей *QoS* (3)-(5) с помощью модифицированной формулы Литтла определяются среднее время задержки передачи разнотипных вызовов:

$$CTD_k = \frac{Q_k}{\lambda_k(1-CLP_k)}, k = 1,2. \qquad (6)$$

Указанный выше точный метод нахождения показателей *QoS* данной модели, основанный на решения СУР для вероятностей состояний, может быть использован лишь при небольших размерностях ФПС данной модели и он сталкивается известными вычислительными трудностями при больших размерностях ФПС (1). Поэтому возникает необходимость использования приближенных методов.

Теперь перейдем к описанию приближенного метода к решению этой проблемы. Предложенный метод имеет высокую точность для моделей с высокой нагрузкой *H*-вызовов, т.е. ниже принимается следующее допущение: $\nu_1 >> \nu_2$ (иными словами, $\lambda_1 >> \lambda_2 >> \mu$). Отметим, что это допущение не является экстраординарным, так как именно в системах с высокими нагрузками *H*-вызовов имеет смысл введения скачкообразных приоритетов для *L*-вызовов.

Рассматривается следующее расщепление ФПС (1) модели:

$$S = \cup_{i=0}^{R} S_i, S_i \cap S_j = \emptyset, i \neq j, \qquad (7)$$

где $S_i = \{n \in S: n_2 = i\}, i = 0,1,2,\dots,R.$

Заметим, что принятое выше допущение относительно соотношения нагрузок разнотипных вызовов обеспечивает выполнение условие корректного применения алгоритмов фазового укрупнения двумерных цепей Маркова [10].

Классы микросостояний S_i объединяются в отдельные укрупненные состояний $<i>$, и вводится функция укрупнения на исходном пространстве состояний S:

$$U(\boldsymbol{n}) = < i >, \text{если } \boldsymbol{n} \in S_i. \qquad (8)$$

Функция укрупнения (8) определяет укрупненную модель с пространством состояний $\Omega = \{< i >: i = 0,1,\dots,R\}$. Стационарную вероятность состояния (k,i) в расщепленной модели с пространством состояний S_i обозначим $\rho_i(k)$, $i=0,1,\dots R$, $k=0,1,\dots R-i$. Каждая расщепленная модель с ФПС S_i является одномерный процесс размножения и гибели. Следовательно, для нахождения стационарных вероятностей состояний внутри расщепленных моделей с ФПС S_i могут быть использованы формулы расчета стационарных вероятностей состояний СМО с зависящей от состояния интенсивностью входящего трафика $M(\lambda_1+\lambda_2\alpha(i))/M(\mu)/1/R-i$. Для нахождения искомых параметров могут быть использованы следующие формулы:

$$\rho_i(k) = \theta_i^k \cdot \frac{1-\theta_i}{1-\theta_i^{R+1-i}}, \; i = 0,1,\dots R, k = 0,1,\dots R-i, \qquad (9)$$

где $\theta_i := \nu_1+\nu_2\alpha(i)$ (для краткости здесь приводятся формулы лишь для случая $\theta_i \neq 1$).

Согласно алгоритмы фазового укрупнения двумерных цепей Маркова [10] элементы производящей матрицы укрупненной модели $q(< i >,< j >), < i >,< j > \in \Omega$

определяются так:

$$q(<i>,<j>)=\begin{cases}\lambda_2\big(1-\alpha(i)\big)\big(1-\rho_i(R-i)\big), & j=i+1,\\ \mu\rho_i(0), & j=i-1,\\ 0 & \text{в остальных случаях.}\end{cases} \quad (10)$$

Следовательно, стационарные вероятности состояний укрупненных состояний определяются так:

$$\pi(<i>)=\prod_{j=1}^{i}A_j\pi(0),\ i=1,2,\dots,R, \quad (11)$$

где:

$$A_j=\nu_2\cdot\frac{\big(1-\alpha(j-1)\big)\big(1-\rho_{j-1}(R-j+1)\big)}{\rho_j(0)},\ \pi(0)=1/\Big(1+\sum_{k=1}^{R}\prod_{i=1}^{k}A_i\Big).$$

С учетом формул (9)-(11) после определенных преобразований получим следующие формулы для приближенного вычисления показателей *QoS* исследуемой модели:

$$CLP\approx\sum_{k=0}^{R}\rho_k(R-k)\pi(<k>); \quad (12)$$

$$Q_1\approx\sum_{k=1}^{R}k\sum_{i=0}^{R-k}\rho_i(k)\pi(<i>); \quad (13)$$

$$Q_2\approx\sum_{k=1}^{R}k\pi(<k>). \quad (14)$$

После нахождения параметров CLP_k и Q_k из формулы (12)-(14) определяются и параметры CTD_k, k=1,2 для данной модели.

В частном случае, когда $\alpha(k)$=0 для любого k (классические *HOL*-приоритеты), разработанные формулы (12)-(14) существенным образом упрощаются. Так, в этом случае стационарные вероятности состояний внутри классов S_i определяются так:

$$\rho_i(k)=\nu_1^k\cdot\frac{1-\nu_1}{1-\nu_1^{R+1-i}},\ i=0,1,\dots,R,\ k=0,1,\dots,R-i, \quad (15)$$

Вероятности укрупненных состояний определяются из следующих простых формул:

$$\pi(<i>)=\pi(0)\,\nu_2^i\prod_{k=0}^{i-1}G(k),\ i=1,2,\dots,R, \quad (16)$$

где:

$$G(k)=\big(1-\nu_1^{R-k}\big)/(1-\nu_1),\ \pi(0)=1/\Big(1+\sum_{i=1}^{R}\nu_2^i\prod_{k=0}^{i-1}G(k)\Big).$$

Далее с помощью формул (12)-(14) вычисляются показатели *QoS* данной модели при использовании классических *HOL*-приоритетов. Формулы (12)-(16) полностью совпадают с результатами, полученные ранее в работе [11]. Там же с помощью объемных вычислительных экспериментов показаны высокая точность этих формул [11]. Иными словами, эти формулы являются косвенными подтверждениями факта о высокой точности разработанного приближенного метода. Аналогичные формулы могут быть записаны и для пороговой схемы определения скачкообразных приоритетов.

Разработанные формулы позволяют легко осуществить изучения поведения показателей *QoS* исследуемой системы относительно изменения ее структурных и нагрузочных параметров. Однако из-за ограниченности объема работы эти результаты здесь не приводятся. По этой же причине здесь не приводятся результаты о точности разработанных приближенных формул. При этом точные значения искомых показателей *QoS* для моделей умеренной размерности находились из соответствующих СУР (некоторые соображения о высокой точности разработанных формул приведены в конце предыдущего раздела). Здесь лишь отметим, что точные и приближенные значения искомых показателей *QoS* в худшем случае отличаются друг от друга в третьем знаке после десятичной точки.

Выводы

Предложен новый подход для вычисления показателей качества обслуживания разнотипных вызовов в системах обслуживания с двумя типами вызовов и общими очередями

при наличии скачкообразных приоритетов. Важным достоинством предложенного подхода состоит в том, что он может быть использован для моделей с любой размерностью общего буфера, так как искомые показатели вычисляются с помощью явных формул. Предложенный подход может быть использован для исследования моделей, в которых вероятности $\alpha(i)$ зависят еще и от числа вызовов первого типа. Кроме того, он может быть использован для изучения моделей с раздельными очередями, а также для моделей, в которых возможны перехода случайного числа L-вызовов в H-очередь.

Предложенный подход позволяет найти надлежащую схему в системах со скачкообразными приоритетами при заданных ограничениях на показатели *QoS* исследуемых моделей. Однако решения таких задач в строгой математической формулировке представляют собой достаточно сложной проблемы из-за многочисленности изучаемых показателей качества обслуживания. Эти проблемы представляют собой предмет отдельных исследований.

Список літератури

1. Kleinrock L. A delay dependent queue discipline // Naval Res. Logist. Quart. – 1964. – Vol. 11. – P. 329-341.

2. Мова В.В., Пономаренко Л.А., Калиновский А.М. Организация приоритетного обслуживания в АСУ. – К.: Техника, 1977. – 648 с.

3. Меликов А.З., Пономаренко Л.А., Рюмшин Н.А. Математические модели многопотоковых систем обслуживания. – К.: Техника, 1991. – 580 с.

4. Lim Y., Kobza J.E. Analysis of delay dependent priority discipline in an integrated multiclass traffic fast packet switch // IEEE Transactions on Communications. – 1990. – Vol. 38, – No. 5. – P. 659-665.

5. Maertens T., Walraevens J., Bruneel H. On priority queues with priority jumps // Performance Evaluation. – 2006. – Vol. 63, – No. 12. – P. 1235-1252.

6. Maertens T., Walraevens J., Bruneel H. A modified HOL priority scheduling discipline: performance analysis // European Journal of Operational Research. – 2007. – Vol. 180, – No. 3. – P. 1168-1185.

7. Maertens T., Walraevens J., Moeneclaey M., Bruneel H. A new dynamic priority scheme: performance analysis // In Proceedings of the 13th International Conference on Analytical and Stochastic Modeling Techniques and Applications (ASMTA). – 2006. – P. 74-84.

8. Maertens T., Walraevens J., Bruneel H. Performance comparison of several priority schemes with priority jumps // Annals of Operations Research. – 2008. – Vol. 162. – P. 109-125.

9. Walraevens J., Steyaert B., Bruneel H. Performance analysis of single-server ATM queue with priority scheduling // Computers and Operations Research. – 2003. – Vol. 30, – No. 12. – P. 1807-1829.

10. Ponomarenko L., Kim C.S., Melikov A. Performance analysis and optimization of multi-traffic on communication networks. – Heidelberg, Dordtrecht, London, New York: Springer, – 2010. – 208 p.

11. Меликов А.З., Пономаренко Л.А., Фаттахова М.И. Управление мультисервисными сетями связи с буферными накопителями. – К.: НАУ-друк, – 2008. – 156 с.