УДК 681.3

В.А. Артеменко, Н.В. Богданова

ФОРМИРОВАНИЕ ДИНАМИЧЕСКИХ ПРИОРИТЕТОВ В УЗЛАХ ТЕЛЕКОММУНИКАЦИОННОЙ СЕТИ

Аннотация: Предложенный способ сокращения времени передачи данных за сет использования гибкой динамической системы приоритетов, который позволяет повысить качество обслуживания в телекоммуникационных сетях.

Ключевые слова: телекоммуникационная сеть, динамические системы, приоритеты, качество обслуживания.

Введение

В современных телекоммуникационных сетях (ТС) качество обслуживания (QoS) обеспечивается за счет статического резервирования ресурсов [1]. В связи с тем, что в ТС изменение нагрузки осуществляется динамически, то статические методы оказываются неэффективными, т.к. требуют большой объем резервирования сетевых ресурсов.

Наиболее общими характеристиками сетевого трафика являются “взрывообразность”, терпимость к задержкам, время ответа, емкость сети и пропускная способность.

Эти характеристики, с учетом маршрутизации, приоритетов, соединений и т.д. определяют характер работы приложений в сети.

По требованиям к задержке передачи можно условно разделить трафик на три категории: трафик реального времени, трафик транзакций и трафик даннях [2].

В пределах каждой категории трафик может быть распределен по приоритетам. Высокоприоритетный трафик имеет преимущества при обработке из-за его важности для данного предприятия. Примером приоритетного трафика может быть транзакция с заказом. Введение приоритетов неизбежно при недостаточности ресурсов сети. Приоритеты могут использоваться для выделения групп, прикладных программ и отдельных пользователей в группах.

Метод приоритетных очередей наиболее часто используется для предоставления временных гарантий чувствительным к задержкам приложениям. Данный метод может применяться для передачи аудио- и видеоинформации, когда не требуется высокое качество. Для доставки аудио- и видеоинформации с высоким качеством необходимо гарантировать низкую задержку и небольшой эффект дрожания. Этого трудно добиться в сетях без значительных накладных расходов при резервировании буферного пространства маршрутизаторов и без реализации сложных алгоритмов обработки очередей.

В динамически реконфигурируемых ТС традиционные методы обеспечения заданных параметров QoS, основанные на резервировании ресурсов не эффективны, так как ресурсы, которые были зарезервированы

для виртуального соединения, в результате его реконфигурации могут оказаться недоступными. Гарантирование же постоянного уровня QoS в реконфигурируемой среде за счет максимально возможного резервирования ресурсов для каждого виртуального соединения существенно снижает эффективность использования компьютерной сети. Поэтому в ряде работ [3, 4, 5] по обеспечению заданного уровня QoS рассматривается частичное или динамическое резервирование ресурсов. В настоящее время наиболее исследован частный случай [6] задачи маршрутизации с учетом QoS, при котором структура сети остается постоянной, а перемещаются только абонентские системы. В этом случае предположение о предсказуемости перемещения абонентских систем позволяет повысить эффективность решения задачи маршрутизации с учетом QoS [7].

Постановка задачи

Эффективным способом обеспечения требуемых параметров трафика в сети является организация динамической системы приоритетов в зависимости от ее нагрузки. В данной статье предложен способ повышения эффективности функционирования коммутационных узлов в ТС. За счет организации динамической системы приоритетов обработки входного трафика.

Решение

В большинстве случаев для обслуживания поступающих пакетов, например, в стеке протоколов TCP/IP, используется механизм без приоритетов. Для такого случая время ожидания обслуживания для некоторого пакета определяется следующим образом:

$$T_{ож} = T_0 + T_n,$$

где T_0 – время окончания обслуживания текущего пакета,
T_n – время обслуживания пакетов, находящихся в очереди на момент поступления данного.

Для дальнейших преобразований необходимо перейти к математическим ожиданиям указанных величин. Второе слагаемое представляется в виде:

$$E(T_n) = E(m)/\mu,$$

где $E(m)$ – математическое ожидание числа пакетов в очереди,
μ – интенсивность обслуживания.

Математическое ожидание числа пакетов можно выразить через среднее время ожидания пакета:

$$E(m) = \lambda E(T_{ож}),$$

где λ – интенсивность входного потока.

Выполнив подстановку, получаем следующее выражение:

 ISSN 1560-8956

$$E(T_{ож}) = \frac{E(T_0)}{1-\rho},$$

где ρ – так называемая – трафик-интенсивность.

Динамические приоритеты изменяются с течением времени. Для определения приоритета заявки обычно задается функция вида:

$$q_k(t) = T_c + C_k,$$

где $q_k(t)$ – приоритет заявки класса k в момент времени t,
T_n – момент поступления заявки,
C_k – коэффициент стоимости заявок k-го класса.

В работе [8] предлагается вычислять приоритет заявок мультипликативно, связывая приоритет заявки каждого класса с некоторым параметром (в наших обозначениях это стоимость заявки), причем значение данного параметра увеличивается с увеличением приоритета данного класса. Таким образом, приоритет заявки в момент времени t определяется следующим образом:

$$q_k(t) = q_k(T) + C_k(t-T) = (t-T)\cdot C_k,$$

причем $C_1 \geqslant C_2 \geqslant \ldots \geqslant C_n \geqslant 0$.

Изменение приоритета для нескольких заявок изображено на рисунке 1.

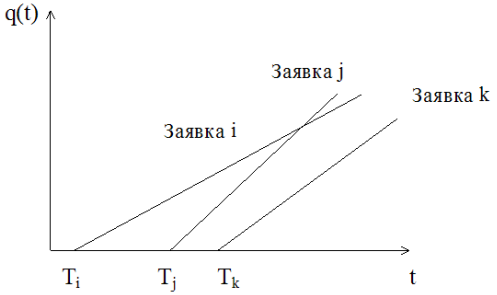

Рис. 1 – Изменение приоритета для нескольких заявок

Предполагаем, что в системе имеется n классов приоритетов. Следовательно, в очереди находится некоторое количество заявок, которые можно разделить на n классов. Интенсивности создаваемых ими потоков равны соответственно $\lambda_1,\ \lambda_2, \ldots, \lambda_n$, для каждого из классов. Считаем, что входной поток каждого из классов имеет пуассоновское распределение. Среднее время обслуживания для некоторого i-го класса равно $1/\mu_i,\ i = 1..n$. Будем также считать, что наивысшим приоритетом обладают заявки класса 1, а наиболее низким – заявки класса n, т.е. с повышением номера класса заявки ее приоритет понижается. Рассчитаем среднее время ожидания для любого класса в предположении, что приоритеты являются относительными. Рассмотрим, в частности некоторый

класс k $(1 \leqslant k \leqslant n)$. Пусть типичная заявка данного класса поступает в некоторый произвольный момент времени t_0. Ее случайное время ожидания $T_{ожk}$, измеряемое от момента поступления до начала обслуживания, зависит от трех параметров. Во-первых, поступившая заявка должна в течение некоторого случайного промежутка времени T_0 ждать завершения обслуживания текущей заявки. Во-вторых, она должна ждать некоторое случайное число T_i единиц времени, пока закончится обслуживание всех клиентов класса i, высшего или равного классу k, которые уже находились в очереди в момент поступления данной заявки. И, наконец, в-третьих, она должна ждать случайное время T_i' обслуживания заявок каждого из классов, которые выше класса k, поступивших в течение времени ожидания $T_{ожk}$. Все вышесказанное можно записать как:

$$T_{ожk} = T_0 + \sum_{i=1}^{n} T_i + \sum_{i=1}^{n} T_i',$$

где T_i – время обслуживания заявок с более высоким или равным приоритетом, уже находившихся в очереди,
T_i' – время обслуживания заявок с более высоким приоритетом, поступающих за время $T_{ож}$.

Здесь также необходимо выполнить переход к математическим ожиданиям указанных величин. Тогда среднее время ожидания $E(T_{ожk})$ для класса k представляется в виде:

$$E(T_{ожk}) = E(T_0) + \sum_{i=1}^{n} E(T_i) + \sum_{i=1}^{n} E(T_i\prime), \qquad (1)$$

где $E(T_0)$ – среднее значение времени обслуживания текущей заявки,
$E(T_i)$ – среднее значение времени обслуживания заявок с более высоким или равным приоритетом, уже находившихся в очереди,
$E(T_i')$ – среднее значение времени обслуживания заявок с более высоким приоритетом, поступающих за время $T_{ожk}$.

Слагаемое $E(T_0)$ – остаточное время обслуживания заявки, находящейся на обслуживании в момент поступления рассматриваемой. Для предлагаемой системы обслуживания с относительными приоритетами это время не зависит от дисциплины обслуживания. Оно должно быть одинаковым для всех классов, если учесть, что заявки всех классов обслуживаются с одинаковым приоритетом в порядке поступления. Слагаемое $E(T_0)$ определяется на основании среднего времени ожидания для системы массового обслуживания типа M/G/1:

$$E(T_0) = \lambda E(\tau^2)/2,$$

где $E(\tau^2)$ – второй момент распределения времени обслуживания, определяемый на основании формулы Поллачека - Хинчина.

Учитывая, что входной поток рассматриваемой системы представляет собой сумму нескольких входных потоков с одинаковым распределени-

ем, среднее значение времени обслуживания текущей заявки определяется как:

$$E(T_0) = \sum_{i=1}^{n} \lambda_i E(\tau_i^2)/2 = \sum_{i=1}^{n} \lambda_i \cdot \left(\sigma_i^2 + 1/\mu_i^2\right)/2,$$

где λ_i – интенсивность i-го входного потока,
σ_i^2 – дисперсия распределения времени обслуживания заявок i-го класса,
μ_i – интенсивность потока обслуживания заявок i-го класса (выходного потока).

Значение $E(T_i)$ возникает за счет среднего числа $E(m_i)$ заявок класса i, которые ожидают в системе и получат обслуживание раньше данной. Каждая из них требует в среднем $1/\mu_i$ единиц времени для обслуживания, поэтому получаем:

$$E(T_i) = f_{ik} E(m_i)/\mu_i, \tag{2}$$

где f_{ik} – ожидаемая часть заявок i-го класса, которые получат обслуживание раньше рассматриваемой,
$E(m_i)$ – среднее количество заявок i-го класса.

На основании формулы Литтла :

$$E(T_i) = f_{ik}\lambda_i E(T_{ожi})/\mu_i = \rho_i f_{ik} E(T_{ожi}), \tag{3}$$

где ρ_i – трафик-интенсивность i-го потока.

Последнее слагаемое в (1) возникает за счет поступления в среднем $E(m_i')$ заявок класса i в течение промежутка времени $E(T_{ожk})$. Поскольку, как оговаривалось выше, интенсивность поступлений равна λ_i и каждая заявка требует в среднем $1/\mu_i$ единиц времени на обслуживание, получаем:

$$E(T_i') = \lambda_i g_{ik} E(T_{ожk})/\mu_i = \rho_i g_{ik} E(T_{ожk}),$$

где g_{ik} – ожидаемая часть из поступивших за время $T_{ожk}$ заявок i-го класса, которые получат обслуживание раньше рассматриваемой.

Из определений f_{ik} и g_{ik} очевидно, что:
$f_{ik} = 1$ при $i \leqslant k$,
g_{ik} при $i \leqslant k$.

Подставив в (1) результаты вычисления каждого слагаемого, получаем рекурсивное уравнение:

$$E(T_{ожk}) = \frac{1}{2}\sum_{i=1}^{n} \lambda_i E\left(\tau_i^2\right) + \sum_{i=1}^{n} \rho_i f_{ik} E(T_{ожi}) + \sum_{i=1}^{n} \rho_i g_{ik} E(T_{ожk}).$$

Решая это уравнение, относительно времени ожидания k-го класса заявок, получаем следующий результат:

$$E(T_{ожk}) = \frac{\frac{1}{2}\sum\limits_{i=1}^{n}\lambda_i E\left(\tau_i^2\right) + \sum\limits_{i=1}^{k}\rho_i E(T_{ожi}) + \sum\limits_{i=k+1}^{n}\rho_i f_{ik} E(T_{ожi})}{1 - \sum\limits_{i=1}^{k-1}\rho_i g_{ik}}. \qquad (4)$$

Для решения данного уравнения необходимо определить коэффициенты f_{ik} и g_{ik}. Учитывая, что f_{ik} определяет ожидаемую часть заявок, которые находятся в очереди на момент поступления рассматриваемой заявки и получают обслуживание раньше нее, удобно воспользоваться рисунок 2.

Считаем, что некоторая заявка i-го класса поступает в произвольный момент времени и находится в очереди $w(T_1)$ единиц времени до поступления рассматриваемой заявки в момент T_1. Нетрудно заметить, что, если $w(T_1) > T_1 + T_2$, то заявка i-го класса будет иметь более низкий приоритет, чем рассматриваемая. В момент T_2 приоритеты обеих заявок становятся равными. Очевидно, что коэффициенты стоимости рассматриваемых классов заявок определяют тангенс угла наклона функции приоритета.

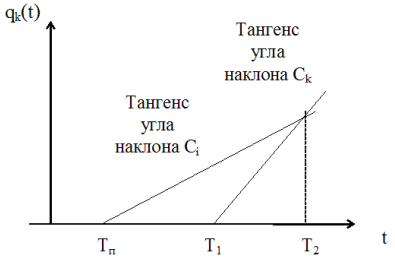

Рис. 2 – Определение коэффициентов стоимости для различных классов заявок

Воспользовавшись довольно простыми геометрическими соотношениями, можно получить следующее уравнение:

$$T_1 + T_2 = \frac{C_k}{C_k - C_i} T_1.$$

Математическое ожидание числа заявок i-го класса, получающих обслуживание раньше рассматриваемой, определяется как

$$E(n_i) f_{ik} = \int_0^{\infty} \lambda_i P\left[t \leqslant \omega_i(t) \leqslant \frac{C_k}{C_k - C_i} t\right] dt,$$

где $\lambda_i dt$ – математическое ожидание числа заявок i-го класса, поступающих в интервале времени dt;

$P\left[t \leqslant \omega_i(t) \leqslant \frac{C_k}{C_k - C_i} t\right]$ – вероятность того, что заявка, прибывшая за этот интервал, проведет в очереди по крайней мере t, но не более $\left[\frac{C_k}{C_k - C_i} t\right]$ единиц времени.

Используя (2) и несколько преобразовав (3), приходим к следующему результату для математического ожидания:

$$E(n_i) f_{ip} = \lambda_i E(T''_{ci}) - \lambda_i \left(1 - \frac{C_i}{C_k}\right) E(T''_{ci}).$$

Учитывая, что $E(n_i) = \lambda_i E(T_{ожi})$, получаем искомое значение: $f_{ip} = \frac{C_i}{C_k}$.

Теперь определим g_{ik}. Для этого воспользуемся рисунок 3.

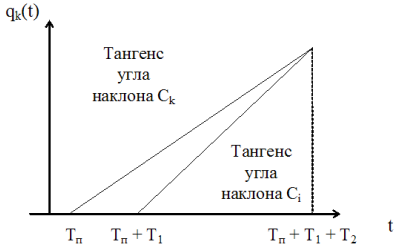

Рис. 3 – Определение коэффициентов стоимости для различных классов заявок

Можно заметить, что для i-го класса заявок справедливо следующее соотношение:

$$E(m_i) g_{ik} = \lambda_i T_1,$$

где $E(m_i)$ – математическое ожидание числа заявок i-го класса в очереди.

Значение T_1 можно определить из уравнения:

$$C_k E(T_{ожk}) = C_i \left(E(T_{ожk}) - T_1\right).$$

Тогда искомый коэффициент можно определить в виде:

$$g_{ik} = 1 - \frac{C_k}{C_i}.$$

Подставив найденные значения в (4), получаем:

$$E(T_{ожk}) = \frac{E(T_0)/(1-\rho) - \sum\limits_{i=k+1}^{n} \rho_i E(T_{ожi}) \left(1 - \frac{C_i}{C_k}\right)}{1 - \sum\limits_{i=1}^{k-1} \rho_i \left(1 - \frac{C_k}{C_i}\right)},$$

где ρ определяется в виде $\rho = \sum\limits_{i=1}^{n} \rho_i$.

Характеристика длины очереди в системе обслуживания определяет размер буфера для поступающих пакетов в узле. Средняя длина $\bar{L}$ очереди может быть вычислена как: $\bar{L} = \sum\limits_{k=1}^{n} \lambda_k E(T_{ожk})$.

Вывод

В данной статье предложены математические модели формирования динамических приоритетов в узлах ТС, с целью определения основных факторов и степень их влияния на эффективность функционирования коммутационных узлов сети. Проанализирована зависимость среднего времени ожидания для пакетов с разными приоритетами. Определена степень зависимости среднего времени ожидания от состояния для различного значения параметра интенсивности.

Таким образом, предложенный способ позволяет сократить время передачи данных за сет использования гибкой динамической системы приоритетов и таким образом позволяет повысить качество обслуживания в ТС.

Литература

1. Liu J. FRR for Latency Reduction and QoS Provisioning in OBS Networks//IEEE Journal on Selected Areas In Communications. – September 2003.-Vol. 21. – No.7.
2. Кулигин М. Технологии корпоративных сетей: [энциклопедия] /Питер: 2000.-704 с.
3. Щербо В.К. Стандарты вычислительных сетей. Взаимосвязи сетей: справочник./М.:КУДИЦ-ОБРАЗ 2000.-272 с.
4. Chen S, Nahrstedt K. Distributed Quality-of-Service Routing in Ad Hoc Networks//IEEE Journal on Selected Areas in Communications. August 1999.-Vol.17.-No.8
5. Lin C. QoS Routing in Ad Hoc Wireless Networks//IEEE Journal on Selected Areas in Communications. August 1999.-Vol.17.-No.8
6. Sivakumar R. CEDAR//IEEE Journal on Selected Areas in Communications. August 1999.-Vol.17.-No.8
7. Talukdar A. C&C Research Labs, On Accommodating Mobile Hosts in an Integrated Services Packet Network, NEC USA, Princeton, NJ Proceedings of INFOCOM’97//Kobe, Japan.-April 1997.
8. ATM Forum Technical Committee Traffic Management Specification.-V. 4.1 af-tm-0121.000.

Отримано 10.11.2011 р.

 ISSN 1560-8956