МОДЕЛИ ОБСЛУЖИВАНИЯ ОЧЕРЕДЕЙ НА УЗЛАХ СЕТИ ТЕЛЕКОММУНИКАЦИИ

А.А. Кодиров

Ташкентский университет информационных технологий имени Мухаммада ал-Хорезми
e-mail: azamattuit2013@gmail.com

KEYWORDS

ABSTRACT

Рассмотрены существующие статистические алгоритмы и потоковые модели обслуживания очередей на узлах сети телекоммуникации, приведены основные требования к методам обслуживания очередей, предложена модель динамического гарантированного обслуживания очередей, проведен сравнительный анализ вероятностно-временных характеристик предложенной и известной (приоритетной) модели обслуживания очередей.

Введение

Базовыми принципами иерархической системы являются принципы соглосованности и координации, принимаемых на всех уровнях управления. В узлах сети телекоммуникации эти принципы заключаются, что все сетевые элементы управления буферным и канальным ресурсом должны быть согласованными между собой с целью достижения заданных показателей качество обслуживания (QoS). Характерной чертой существующих решений по управления буферным (очередями) и канальным (пропускной способностью канала) ресурсом является статистическая стратегия их распределения. Порядок распределения буферного пространства и канальной емкости по-прежнему преимущественно устанавливается административно, т.е. вне реального времени, хотя динамика изменения состояния отдельного сетевого узла выше чем динамика загруженности сети телекоммуникации. В связи с этим актуальной представляется задача о придании динамического характера решениям по управления канальным и буферным ресурсов в зависимости от характеристик поступающего на узел трафика, величины доступного ресурса и требуемых показателей качества обслуживания

Анализ существующих алгоритмов обслуживания очередей

Алгоритмы обслуживания очередей должны отвечать следующим основным требованиям:

- поддержка функций дифференцированного обслуживания;
- обеспечение гарантий по качеству обслуживания пакетов различных потоков (классов);
- обеспечение справедливого обслуживания пакетов различных приоритетов, недопущения перегрузки каждой отдельной очереди и сетевого узла в целом;
- согласованное функционирование с другими механизмами управления ресурсами сети.

Однако не все из перечисленных требований в должной мере учтены в существующих алгоритмах обслуживания очередей, которые можно разделить на несколько классов [1-3]:

- без приоритетного обслуживания (FIFO – First In – First Out);
- приоритетного обслуживания (PQ – Priority Queuing, CQ – Custom Queuing);
- взвешенного обслуживания (FQ- Fair Queuing, WFQ – Weighted Fair Queuing);
- гибридные (LLQ – Low Latency Queuing, CBWFQ – Class Based Weighted Fair Queuing);

В алгоритме FIFO все поступившие пакеты ставятся в одну очередь и обслуживаются в порядке поступления. Поэтому алгоритм FIFO не обеспечивает дифференцированного качества обслуживания трафика.

Алгоритм приоритетного обслуживания трафика предусматривает разделение всего сетевого трафика на небольшое количество классов с назначением каждому классу некоторого числового признака — приоритета. Классификация пакетов может производиться разными способами. Например, в пакете IP для этой цели предусмотрено трехразрядное подполе IP Precedence в поле Type Of Service (TOS). В узле сети имеется несколько очередей в соответствии с количеством классов. В начале обслуживается очередь с пакетами наибольшего приоритета. После обслуживания всех пакетов данной очереди, обслуживание получает очередь, приоритет у которой ниже и т.д. Из принципа работы приоритетного обслуживания очевидно, что постоянное наличие высокоприоритетного трафика в очереди приведет к существенным задержкам обслуживания и потерям низкоприоритетных трафиков. С целью исключения данного недостатка приоритетного обслуживания разработаны алгоритмы взвешенного обслуживания.

В алгоритмах взвешенного обслуживания очередям выделяется сетевой (канальный) ресурс, пропорциональный назначенным им весам. При этом все очереди гарантированно обслуживаются. Например в OpenFlow – коммутаторе по умолчанию имеются восемь очередей на физический порт с минимально гарантированными долями полосы пропускания канала.

Выделение постоянного канального ресурса при отсутствии пакетов в очереди приводит к неэффективному использованию пропускной способности канала.

В гибридном алгоритме обслуживания LLQ для трафика, чувствительного к задержкам, выделяется одна очередь, для обслуживания которой резервируется определенная пропускная способность канала. Остальные очереди обслуживаются в соответствии с алгоритмом взвешенного обслуживания. Поэтому гибридные алгоритмы также имеют недостаток алгоритмов взвешенного обслуживания.

Во всех вышерассмотренных алгоритмах решения задачи распределения пропускной способности между очередями имеет статистический характер, т.е. решается администратором путем настройки. Например, в алгоритме *CQ* распределение пропускной способности канала определяется значением задаваемого вручную счетчиком байт (*byte count*) для каждой очереди. В алгоритмах *CBWFQ* и *LLQ* порядок распределения пропускной способности задается также вручную командами *bandwidth* и *priority*.

Недостатки алгоритмов статистического обслуживания очередей определили актуальность разработки потоковых моделей (flow-based model) обслуживания очередей, в рамках которых учитывается интенсивность трафика наряду с другими с другими важными параметрами.

Анализ существующих потоковых моделей обслуживания очередей

В потоковых моделях рассматривается следующая задача распределения трафика [4-6]. Имеются M классов трафиков, различаемых типом приоритета и N обслуживаемых очередей. Интенсивность трафика i – го класса равна d_i $(i=\overline{1,M})$. Часть пропускной способности исходящего канала, закрепленная за j -й очередью равна c_j $(j=\overline{1,N})$. Необходимо выполнить следующие условия:

$$\sum_{j=1}^{N} c_j \le c, \qquad (1)$$

$$\sum_{i=1}^{M} d_i \le \sum_{j=1}^{N} c_j . \qquad (2)$$

Динамический характер распределения пакетов трафика в очереди узла сети осуществляется за счет введения переменной x_{ij}, под которой подразумевается часть i - го трафика, который будет направлен для

обслуживания в j – ю очередь. Для предотвращения перегрузки очередей вводятся условия:

$$\sum_{i=1}^{M}\sum_{j=1}^{N} x_{ij} < c_j, \qquad (3)$$

$$\sum_{i=1}^{M} d_i x_{ij} < c_j. \qquad (4)$$

Так как поток одного класса может быть обслужен только в рамках одной очереди, то искомая переменная x_{ij} принимает только два значения: 0 или 1, $x_{ij} = \{0,1\}$.

Задача определения значений переменных x_{ij} формулируется в виде оптимизационной задачи с различными целевыми функциями.

В [4] минимизируется следующая целевая функция:

$$F(x) = \sum_{i=1}^{M}\sum_{j=1}^{N} f_{ij} x_{ij}, \qquad (5)$$

где f_{ij} – характеризует относительную стоимость использования пакетами i -го трафика ресурсов j – й очереди.

В [5] в качестве искомой переменной выбирается вектор:

$$\vec{x} = \left[\frac{x_{ij}}{c_j}\right], \ (i = \overline{1,M}; j = \overline{1,N}). \qquad (6)$$

Минимизируется следующая целевая функция:

$$F(x) = \vec{s}^{\,t}\vec{x}, \ (7)$$

где координаты вектора весовых коэффициентов:

$$\vec{s} = \left[\frac{s_{ij}}{s_j}\right], \qquad (8)$$

характеризуют условную стоимость (s_{ij}) использования пакетами i – го трафика ресурсов j – й очереди, а также стоимость (s_j) выделения j – й очереди того или иного объема пропускной способности исходящего канала передачи данных.

В [6] с целью сбалансированной загрузки буферных ресурсов дополнительно вводится следующее условие:

$$f(p_j)Q_j \le \alpha, \ (j = \overline{1,N}), \quad (9)$$

где α – верхний динамически управляемый предел загруженности очередей узла сети.

Решается задача минимизации следующей целевой функции:

$$\min \alpha, \qquad (10)$$

где управляемыми переменными являются x_{ij}, c_j и α.

В [7] используется целевая функция минимизации суммы средних длин очередей:

$$F = \sum_{j=1}^{N} f(p_j)Q_j, \qquad (11)$$

где $f(p_j)$ – некоторая функция от характеристик пакетов j – ой очереди с приоритетом p_j.

Значение функции $f(p_j)$ должно быть тем больше, чем выше приоритет. При этом поток с более высоким приоритетом обслуживается лучше, чем поток с низким приоритетом.

В [8] условия (3) и (4) дополняются условиями предотвращения перегрузки очередей по их длине:

$$Q_j \le Q_j^{\max}, \qquad (12)$$

где $Q_j^{\max}$ – максимальная емкость очереди.

Решается оптимизационная задача, связанная с минимизацией целевой функции вида:

$$F(x)=\sum_{i=1}^{M}\sum_{j=1}^{N} f_{ij}x_{ij}+\alpha. \qquad (13)$$

Целевая функция (13) включает себя функции (5) и (10).

Оптимизационные задачи (5) и (7) относятся к задачам линейного программирования, а задачи (10), (11) и (13) – нелинейного программирования. В задачах (5) и (7) результаты оптимизации существенно зависят от метрик f_{ij} и $\vec{s}^{t}$. Чем меньше метрика f_{ij} и $\vec{s}^{t}$, тем больше данная очередь загружаться. Минимизация целевых функций (10), (11) и (13) обеспечивает сбалансированную загрузку очередей узла сети [7,8].

Общими недостатками рассмотренных потоковых методов обслуживания очередей являются:

- объем буферных ресурсов задается и не распределяется между очередями;

- на практике каждому трафику выделяется своя очередь и нет необходимости распределения трафика между очередями.

Предлагаемая модель динамического приоритетного обслуживания очередей

Для устранения вышеуказанных недостатков предлагается следующий метод. Каждому трафику i – го приоритета выделяется j – я очередь, при этом $i=\overline{1,M}$, $j=\overline{1,N}$. Так как $M=N$, далее будем использовать N. Общий буферный ресурс с объемом L распределяется между очередями в зависимости от интенсивностей трафиков. Определяются суммарная интенсивность трафиков

$$D=\sum_{i=1}^{N} d_i \qquad (14)$$

и доля i – го трафика в общем потоке

$$\gamma_i=\frac{d_i}{D},\ i=\overline{1,N}. \qquad (15)$$

Объем буфера, выделяемого для i -й очереди, прямо пропорционален доле i – го трафика

$$l_i=\gamma_i L,\ i=\overline{1,N}. \qquad (16)$$

Управляемой переменной служит x_i, которая характеризует долю пропускной способности канала выделяемого для обслуживания i – й очереди. При этом необходимо выполнить условия предотвращения перегрузки очередей:

$$d_i=x_i c,\ i=\overline{1,N}, \qquad (17)$$

$$\sum_{i=1}^{N} x_i=1,\ 0\le x_i\le 1. \qquad (18)$$

Вводится коэффициент важности трафика i – приоритета - υ_i $(\upsilon_i\ge 0)$.

Минимизируется сумма длин всех очередей с учетом коэффициентов важности трафика:

$$\min\sum_{i=1}^{N}\upsilon_i Q_i. \qquad (19)$$

Если каждая очередь представляет собой экспоненциальную систему массового обслуживания с ограниченной очередью, то вероятность потери пакетов из-за переполнения буфера определяется выражением:

$$p_i=\frac{1-\rho_i}{1-\rho_i^{l_i+2}}\rho_i^{l_i+1}, \qquad (20)$$

где ρ_i – загрузка i – й очереди $(i=\overline{1,N})$:

$$\rho_i=\frac{d_i}{x_i c} \qquad (21)$$

Средняя длина очереди определяется по формуле [6]:

$$Q_i = \frac{\rho_i^2}{(1-\rho_i)^2} p_{0i} \left\{1 - \rho_i^{l_i} \left[l_i (1-\rho_i) + 1\right]\right\}, \qquad (22)$$

где p_{0i} – вероятность отсутствия пакетов в i – й очереди:

$$p_{0i} = \frac{1-\rho_i}{1-\rho_i^{l_i+2}}. \qquad (23)$$

Среднее время ожидания (W_i) и задержка (T_i) пакетов:

$$W_i = \frac{Q_i}{d_i}, i = \overline{1,N}. \qquad (24)$$

$$T_i = W_i + 1/x_i c, \; i = \overline{1,N}. \qquad (25)$$

На рисунке 1 приведен график зависимости средней задержки пакетов от интенсивности трафика каждого приоритета при коэффициентах важности трафика $\upsilon = [18,14,11,8,5,4,2,1]$, $N = 8$, и $L = 80$.

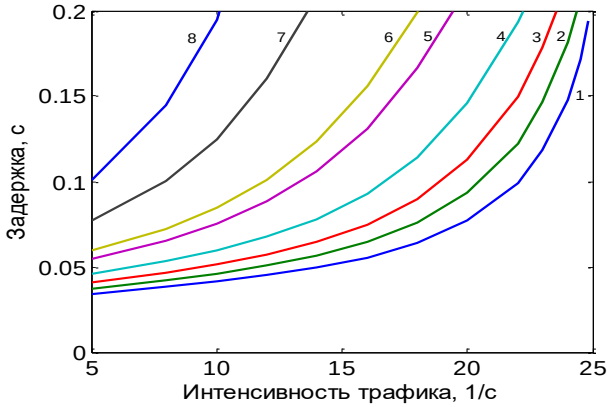

Рис.1. Зависимость задержки пакетов от интенсивности трафика каждого приоритета

Наименьшую задержку имеют пакеты трафика высокого приоритета. С уменьшением приоритета увеличивается средняя задержка пакета.

На рисунке 2 приведен график зависимости вероятности потери пакетов от интенсивности трафика каждого приоритета. Из рисунка 2 следует, трафик с более высоким приоритетом по вероятности потерь также имеет преимущество по сравнению с низкими приоритетами.

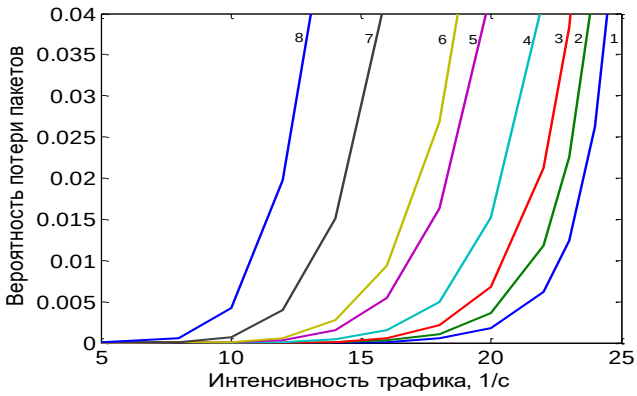

Рис. 2. График зависимости вероятности потери пакетов от интенсивности поступления трафика каждого приоритета

Как указано выше, алгоритмы обслуживания очередей должны обеспечить гарантированное обслуживание очередей с низкими приоритетами. Далее с целью проверки выполнения данного требования проводится сравнительный анализ предложенного алгоритма обслуживания с алгоритмом относительного приоритетного обслуживания.

Для определения характеристик обслуживания трафика с относительным приоритетом будем использовать формулы, приведенные в работе [9].

Вероятность потери пакетов i – го приоритета:

$$p_i = P(N) \frac{1-\rho_i}{1-\rho_i^{L+1}} \rho_i^L, \qquad (26)$$

где

$$\rho_i = \sum_{j=1}^{i} d_i / \mu_i, \; \mu_i = c_i, \; i = \overline{1,N}, \qquad (27)$$

$$P(N) = \frac{\rho_N - \rho_N^{L+2}}{1-\rho_N^{L+2}}. \qquad (28)$$

Среднее время ожидания обслуживания пакетов i – го приоритета:

$$W_i = \begin{cases} W_i^*, i = 1; \\ \frac{\wedge_i}{d_i}(W_i^* - \frac{\wedge_{i-1}}{\wedge_i} W_{i-1}^*), i = \overline{2, N}, \end{cases} \quad (29)$$

где:

$$\wedge_i = \sum_{j=1}^{i} d_j, \quad (30)$$

$$W_i^* = P(N)\frac{1-(L+1)\rho_i^L + L\rho_i^{L+1}}{\mu_i(1-\rho_i)(1-\rho_i^{L+1})}. \quad (31)$$

Среднее время задержки пакетов i – го приоритета:

$$T_i = W_i + 1/\mu_i. \quad (32)$$

Для сравнительного анализа предложенного метода обслуживания и приоритетного обслуживания на рисунке 3 приведен график зависимости среднего времени задержки пакетов от интенсивности трафика первого приоритета при заданных интенсивностях других видов трафика d_i $(d_i = 5, i = \overline{2, N})$.

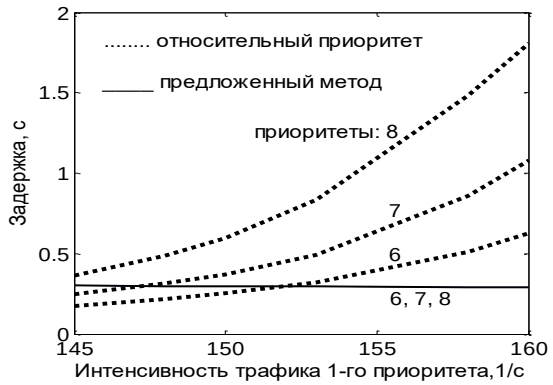

Рис.3. График зависимости средней задержки пакетов от интенсивности трафика первого приоритета при заданных интенсивностях других видов трафика

Из рисунка 3 следует, что при алгоритме относительного приоритетного обслуживания при высоких интенсивностях трафика высокого приоритета, средняя задержка пакетов низких приоритетов резко возрастает, а в предложенном алгоритме обслуживания средняя задержка пакетов низкого приоритета почти не изменяется. При этом, предложенный метод обслуживания уменьшает среднюю задержку низкоприоритетных пакетов до 3-х раз по сравнению с алгоритмом относительного приоритетного обслуживания.

Выводы

Проведенный анализ существующих алгоритмов управления очередями, позволил констатировать следующие недостатки: применение преимущественно неадаптивных схем управления, когда за каждой очередью закрепляется строго определенная канальная емкость и статистическое распределение буферной емкости. Повышение показателей качества обслуживания неразрывно связано с совершенствованием механизмов управления очередями на основе минимизации свойственных им выше перечисленных недостатков. Одним из основных направлением обеспечения качества обслуживания является разработка потоковых моделей обслуживания очередей, которые позволят в полной мере реализовать преимущества динамического управления трафиком на основе текущей загруженности канальных и буферных ресурсов. Предложена модель динамического приоритетного обслуживания очередей, отличающейся от известных тем, что реализуется механизмы динамического распределения канальных и буферных ресурсов, и позволяющая дифференцированное гарантированной обслуживание потоков различного класса.

Развитие предложенного подхода видится в повышение согласованности в решении задач управления очередями другими задачами управления трафиком, например, управления доступом, маршрутизация потоков и резервирования ресурсов.

Литература

1. 1.Tsai T.-Y., Chung Y.-L., Tsai Z. Introductiom to Packet Scheduling Allgorithms for Communication Networks // Communications and Networking, Published 2010, Computer Sciece. - pp. 264-288. https://pdfs.semanticscholar.org/9358.

2. 2.Vegesna S. IP Quality of Service // Cisco press, 2001.-368 p.
3. Яновский, Г.Г. Качество обслуживания в сетях IP. / Г.Г. Яновский. // Вестник связи. №1, 2008. – С. 65-74.
4. Лемешко А.В., Ватти М., Симоненко А.В. Управление очередями на узлах активной сети. Радиотехника: Всеукр. межвед. науч.-техн. сб. 2007. Вып. 151. С. 92-97. http://openarchive.nure.ua/handle/document/2847
5. Лемешко А.В., Симоненко А.В. Математическая модель динамического управления канальными и буферными ресурсами на узлах телекоммуникационной сети. Радиотехника: Всеукр. межвед. науч.-техн. сб. 2009. Вып. 156. С. 36-41. http://openarchive.nure.ua/handle/document/2608
6. Али С. Али, Симоненко А.В. Потоковая модель динамической балансировки очередей в MPLS- сети с поддержкой traffic engineering queues. Электронное научно специализированное издание – журнал «Проблемы телекоммуникаций», № 1(1), 2010. С.60-67. *(http://pt.jornal.kh.ua).*
7. Семеняка М.В. Двухуровневый метод иерархически-координационного обслуживания очередей на узлах телекоммуникационной сети. Научно-технический вестник информационных технологий, механики и оптики. 2014, №4 (92). С.98-105.
8. Симоненко А.В., Андрушко Д.В. Математическая модель управления очередями на маршрутизаторах телекоммуникационной сети на основе оптимального агрегирования потоков и распределения пакетов по очередям. Электронное научно специализированное издание – журнал «Проблемы телекоммуникаций», № 1(16), 2015. С.94-102. *(http://pt.jornal.kh.ua.*
9. Назаров А.Н., Сычев К.И. Модели и методы исследования процессов функционирования узлов коммутации сетей связи следующего поколения при произвольных распределениях поступления и обслуживания заявок различных классов качества. T-Comm, №7, 2012. С.135-140.