

УДК 621.391

ДВУХУРОВНЕВЫЙ МЕТОД ИЕРАРХИЧЕСКО-КООРДИНАЦИОННОГО ОБСЛУЖИВАНИЯ ОЧЕРЕДЕЙ НА УЗЛАХ ТЕЛЕКОММУНИКАЦИОННОЙ СЕТИ

М.В. Семеняка^а^а Харьковский национальный университет радиозлектроники, Харьков, Украина, maxisemen@gmail.com

Аннотация. Предложен метод иерархическо-координационного обслуживания очередей. В рамках предложенного метода задача обслуживания очередей была сведена к решению оптимизационной задачи, представленной двухуровневой иерархической структурой. На первом уровне решалась задача распределения потоков между очередями, на втором уровне решения, полученные на первом уровне, координировались с целью предотвращения вероятной перегрузки канала связи. При управлении многоуровневой структурой в основу был положен принцип целевой координации, который определял порядок учета ограничений на взаимодействие очередей и распределения расчетных задач между уровнями иерархии. Координация решений осуществлялась с использованием метода множителей Лагранжа. Методами аналитического моделирования проведено исследование сходимости предложенного метода.

Ключевые слова: обслуживание очередей, метод целевой координации, качество обслуживания, иерархическая структура.

TWO-LEVEL HIERARCHICAL COORDINATION QUEUING METHOD FOR TELECOMMUNICATION NETWORK NODES

M.V. Semenyaka^а^а Kharkiv National University of Radioelectronics, Kharkiv, Ukraine, maxisemen@gmail.com

Abstract. The paper presents hierarchical coordination queuing method. Within the proposed method a queuing problem has been reduced to optimization problem solving that was presented as two-level hierarchical structure. The required distribution of flows and bandwidth allocation was calculated at the first level independently for each macro-queue; at the second level solutions obtained on lower level for each queue were coordinated in order to prevent probable network link overload. The method of goal coordination has been determined for multilevel structure managing, which makes it possible to define the order for consideration of queue cooperation restrictions and calculation tasks distribution between levels of hierarchy. Decisions coordination was performed by the method of Lagrange multipliers. The study of method convergence has been carried out by analytical modeling.

Keywords: queue management, goal coordination method, service quality, hierarchical structure.

Введение

Важной задачей развития современных телекоммуникационных сетей является повышение качества обслуживания (Quality of Service, QoS) возрастающего количества запросов пользователей. Среди множества средств повышения качества обслуживания значимое место занимают методы управления очередями, позволяющие эффективно улучшить такие показатели, как средняя задержка пакетов, джиттер, количество отброшенных пакетов без существенных затрат на модернизацию существующей инфраструктуры сети [1]. В свою очередь, особенности и эффективность технологических решений в области обслуживания очередей зависят от тех математических моделей и методов, которые в них заложены.

Применяемые сегодня механизмы обслуживания очередей (FIFO Queueing (First In, First Out), приоритетные очереди (Priority Queueing, PQ), заказные очереди (Custom Queueing, CQ), взвешенное справедливое обслуживание (Weighted Fair Queueing, WFQ), взвешенные справедливые очереди, основанные на классах (Class-based Weighted Fair Queueing, CBWFQ), очереди с малой задержкой (Low Latency Queueing, LLQ)) требуют вмешательства администратора при настройке, не способны к адаптации и тем самым не способны обеспечить возрастающие потребности в обеспечении качества обслуживания нового информационного общества [2]. В связи с этим актуальным видится исследование новых подходов обслуживания очередей на узлах как проводной, так и беспроводной телекоммуникационной сети [3–6].

Современная телекоммуникационная сеть имеет многоуровневую структуру, происходит переход к распределенным, облачным технологиям, которые требуют соблюдения принципов согласования, координации элементов [7, 8]. В области обслуживания очередей это распределенные алгоритмы обслуживания DWFQ (Distributed Weighted Fair Queueing), DCBWFQ (Distributed Class-based Weighted Fair Queueing, DLLQ (Distributed Low Latency Queueing)) и др. Математическим аппаратом, способным решить поставленную задачу согласования и координации распределенной сети для повышения качества обслуживания потоков пользователей, является теория многоуровневых иерархических систем [9, 10].

Иерархические структуры (очереди) подходят для масштабирования сети в случае увеличения количества обслуживаемых потоков. Этот математический аппарат позволяет более эффективно использовать многоядерные (многопроцессорные) системы, когда общий ресурс распределяется параллельно и каждый процессор занимается обслуживанием своей группы очередей.

Таким образом, актуальной является задача исследования новых методов обслуживания очередей, представленных иерархическими структурами. Предложенный в настоящей работе подход также исследован в области управления трафиком для решения задач маршрутизации, представлен математическими

моделями дифференциальных и алгебраических уравнений [11, 12]. В работе предложена потоковая модель иерархических очередей в декомпозиционной форме в виде алгебраических уравнений.

Двухуровневый метод иерархическо-координационного обслуживания очередей на узлах телекоммуникационной сети

Базовыми принципами иерархической системы являются принципы согласованности и координации решений, принимаемых на всех уровнях управления. В рамках задачи обслуживания очередей эти принципы заключаются в том, что все сетевые элементы управления буферным ресурсом (как по горизонтали, так и по вертикали) должны быть согласованы между собой по всем показателям с целью достижения заданных значений средней задержки, джиттера, количества отброшенных пакетов. Например, при проектировании системы обслуживания очередей должна быть предусмотрена координация элементов по расходу ресурсов пропускной способности (ПС) исходящего канала связи, загруженности сетевого интерфейса, по целям функционирования, программному и алгоритмическому обеспечению и т.д.

В сетевом узле в ходе решения задачи обслуживания очередей необходимо определить распределение множества потоков между очередями, а также порядок обслуживания потоков пакетов заданной интенсивности [13]. Для повышения масштабируемости задачи обслуживания очередей определим двухуровневую функциональную иерархию расчетов:

- на нижнем уровне искомое распределение потоков подлежит расчету независимо для каждой макроочереди для потоков, которые поступают от сетей доступа или других сетевых узлов;
- на верхнем уровне полученные на нижнем уровне решения задач распределения потоков и выделения ПС для каждой очереди координируются с целью предотвращения вероятной перегрузки канала связи ввиду децентрализации принимаемых решений на каждой конкретной макроочереди.

Общая схема предлагаемого метода представлена на рис. 1.

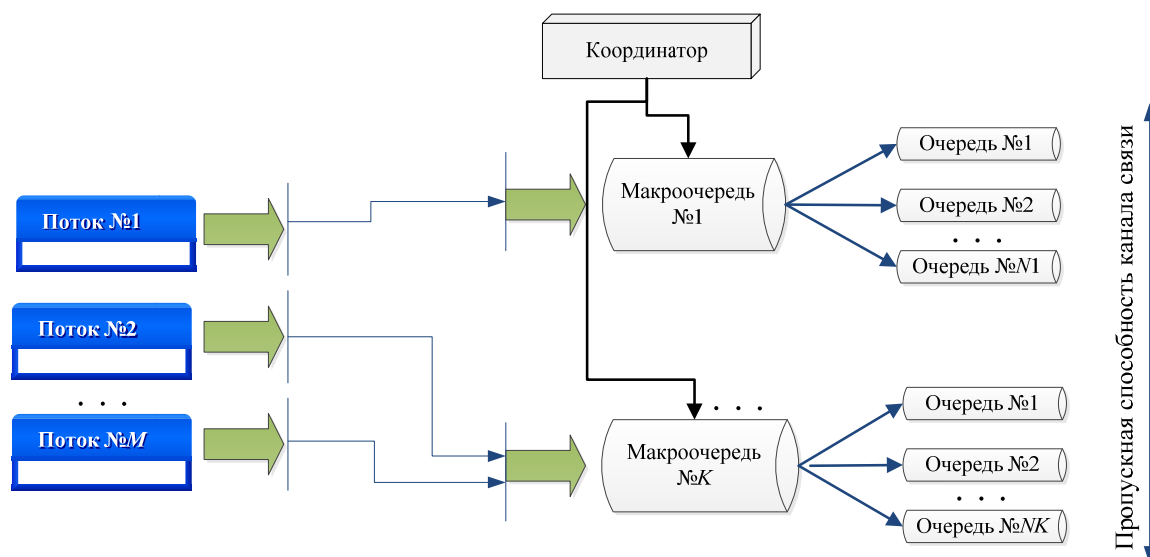


Рис. 1. Общая схема двухуровневого метода иерархическо-координационного обслуживания очередей

На вход сетевого узла поступают M потоков с различными параметрами интенсивности, приоритета, длины и распределения времени поступления пакетов. В зависимости от настроек списков доступа на сетевом узле потоки распределяются между K макроочередями. Каждая макроочередь обслуживается отдельным процессором или ядром сетевого узла. В рамках макроочереди параллельно рассчитывается распределение потоков между очередями и выделение для каждой очереди ПС. Координатор следит за тем, чтобы сумма ПС, выделенных каждой макроочереди, не превышала ПС канала связи.

Базовая модель обслуживания очередей на узлах телекоммуникационной сети

Предположим, что для каждой r -й ($r = \overline{1, K}$) макроочереди в качестве искомого выступают булевые переменные x_{ij}^r , которые характеризуют возможность обслуживания i -го ($i = \overline{1, M}$) потока j -й ($j = \overline{1, N_r}$) очередью r -й макроочереди, а также переменные b_j^r , под которыми понимается ПС, выделенная для j -й очереди r -й макроочереди.

Согласно физическому содержанию, искомая переменная x_{ij}^r может принимать только два значения – или 0, или 1, так как поток одного класса может быть обслужен только в рамках одной очереди:

$$x_{ij}^r \in \{0, 1\}. \quad (1)$$

С целью обеспечения управляемости предложенного метода необходимо обеспечить выполнение условия

$$\sum_{r=1}^K \sum_{j=1}^{N_r} x_{ij}^r = 1 \quad (i = \overline{1, M}). \quad (2)$$

Условие (2) гарантирует обслуживание потоков пользователей исходящими очередями и должно выполняться для каждого i -го потока, поступающего в произвольную макроочередь.

Основной причиной потери пакетов в телекоммуникационной сети является переполнение буфера сетевых устройств, вызванное недостаточной ПС канала связи для обслуживания поступивших потоков. Отсутствие потерь на сетевом узле должно способствовать отсутствию потерь в сети в целом. Интенсивность потока, направленного на обслуживание в очередь, не должна превышать ПС, выделенной для этой очереди. С целью предотвращения возможной перегрузки очередей в ходе расчета распределения потоков важно выполнить условие

$$\sum_{i=1}^M a_i x_{ij}^r < b_j^r, \quad (3)$$

где a_i – интенсивность потока i -го класса, поступающего на обслуживание сетевого узла.

Для решения задачи распределения трафика между очередями к использованию предлагается целевая функция минимизации средней длины очереди:

$$F = \sum_{r=1}^K \sum_{j=1}^{N_r} f(p_j^r, d_j^r) \bar{n}_j^r, \quad (4)$$

где $f(p_j^r, d_j^r)$ – некоторая функция от характеристик пакетов j -й очереди, например их IP-приоритета p_j^r или длины d_j^r ; \bar{n}_j^r – средняя длина j -й очереди r -й макроочереди. Стоит отметить, что поток с более высоким приоритетом обслуживается лучше, чем поток с низким приоритетом. В результате значение функции $f(p_j^r, d_j^r)$ должно быть тем больше, чем выше приоритет и чувствительность к задержке данного потока. В связи с этим в (4) в качестве функции характеристик потока можно использовать следующее выражение:

$$f(p_j^r, d_j^r) = \frac{p_j^r}{d_j^r}. \quad (5)$$

В зависимости от параметров входящего потока \bar{n}_j^r может быть рассчитана с использованием следующих выражений:

$M/M/1$:

$$\bar{n}_j^r = \frac{(\rho_j^r)^2}{1 - \rho_j^r}, \quad (6)$$

$M/D/1$:

$$\bar{n}_j^r = \frac{(\rho_j^r)^2}{2(1 - \rho_j^r)}, \quad (7)$$

$M/M/1/N$:

$$\bar{n}_j^r = \frac{\rho_j^r}{1 - (\rho_j^r)^{N+1}} \cdot \frac{1 - (N+1)(\rho_j^r)^N + N(\rho_j^r)^{N+1}}{1 - \rho_j^r} - \rho_j^r, \quad (8)$$

$fBM/M/1/N$:

$$\bar{n}_j^r = \frac{\frac{\rho_j^r}{\beta} f(H)}{1 - \left[\frac{\rho_j^r}{\beta} f(H) \right]^{N+2}} \cdot \frac{\left\{ 1 - (N+1) \left[\frac{\rho_j^r}{\beta} f(H) \right]^N \right\} + N \left[\frac{\rho_j^r}{\beta} f(H) \right]^{N+1}}{1 - \frac{\rho_j^r}{\beta} f(H)} - \frac{\rho_j^r}{\beta} f(H), \quad (9)$$

где $\rho_j^r = \frac{\sum_{i=1}^M a_i x_{ij}^r}{b_{ij}^r}$ – коэффициент использования пропускной способности, выделенной для j -й очереди;

β – вероятность отсутствия повторной передачи пакетов; H – коэффициент Херста; $f(H)$ – функция, которая учитывает влияние самоподобия поступающего в узел потока. Из (9) видно, что для определения средней длины очереди необходимо определиться с функцией $f(H)$. Например, если использовать линейный закон и значение данной функции для $H = 0,5$, равное 1, то можно представить $f(H)$ в виде $f(H) = 2H$ [14].

Использование целевой функции (4) позволяет наиболее полно сохранить физический смысл исходных переменных, адаптировав метод под характеристики различных передаваемых потоков. В зависимости от модели входящего потока и модели обслуживания решаемую задачу можно аппроксимировать одной из систем массового обслуживания, воспользовавшись для вычисления средней длины очереди известными выражениями (6)–(9). При этом учитывается IP-приоритет пакетов и их длина (5).

Переход от централизованного решения к декомпозиционной форме

Для перехода к более гибкой, масштабируемой иерархической структуре обслуживания очередей описанную выше централизованную постановку задачи распределения потоков между очередями и выделения ПС необходимо представить в декомпозиционной форме по количеству макроочереди с возможностью координации результатов решения на каждой макроочереди.

Стоит учесть, что распределение потоков для каждой макроочереди рассчитывается только на основе имеющейся у нее информации о поступивших в нее пакетах и ПС канала связи, без данных о результатах распределения потоков и выделения ПС на других макроочередях. В связи с этим в дополнение к условию предотвращения перегрузки очереди (3) целесообразно определить условие на непереполнение канала связи:

$$\sum_{i=1}^M \sum_{j=1}^{N_r} a_i x_{ij}^r < b - \sum_{\substack{s=1 \\ s \neq r}}^K \sum_{i=1}^M \sum_{j=1}^{N_s} a_i x_{ij}^s \quad (r = \overline{1, K}). \quad (10)$$

В левой части неравенства (10) приведена суммарная интенсивность потоков пакетов, поступивших на обслуживание в r -ю макроочередь. В правой части неравенства (10), кроме ПС канала связи b , присутствует суммарная интенсивность потоков, поступивших на обслуживание в другие (отличные от r -й) макроочереди. Смысл выражения (10) заключается в том, что поток, поступивший на обслуживание в r -ю макроочередь, не должен по своей интенсивности превышать доступную ПС канала связи, которая осталась после обслуживания потоков, обслуженных в других макроочередях. Выполнение условия (10) будет положено в основу координирующего уровня предложенной модели. В дальнейшем условие (10) удобно будет использовать в векторно-матричной форме:

$$\mathbf{B}_r \mathbf{x}_r < b - \sum_{\substack{s=1 \\ s \neq r}}^K \mathbf{B}_s \mathbf{x}_s \quad (r = \overline{1, K}), \quad (11)$$

где $\mathbf{B}_r, \mathbf{B}_s$ ($r, s \in K; r \neq s$) – согласующие матрицы, координаты которых выбираются в соответствии с выражением (10) и интенсивностями потоков, поступающих в сетевой узел.

Решение оптимизационной задачи, связанной с минимизацией функции (4) при наличии ограничений (1)–(3), (11), является основой предлагаемого метода иерархическо-координационного обслуживания очередей. Тогда, используя теорему о строгой дуальности задач на условный экстремум и переходя к задаче на безусловный экстремум, имеем

$$\min_{x, b} F = \max_{\mu} \Phi,$$

определив в соответствии с формулой (8) двойственную функцию $\Phi(\mu)$ следующим образом:

$$\Phi(\mu) = \{ \min_{x, b} L(x, b, \mu) \},$$

где

$$L = \sum_{r=1}^K \sum_{j=1}^{N_r} f(p_j^r, d_j^r) \bar{n}_j^r + \sum_{r=1}^K \mu_r^t (\mathbf{B}_r \mathbf{x}_r + \sum_{\substack{s=1 \\ s \neq r}}^K \mathbf{B}_s \mathbf{x}_s - b). \quad (12)$$

Применение иерархическо-координационного подхода к решению поставленной задачи предполагает обоснованный выбор координирующих векторов (вектора). При решении оптимизационной задачи по управлению многоуровневой структурой в основу иерархическо-координационных методов могут

быть положены принципы координации, такие, как принцип целевой координации, принцип оценки взаимодействий и принцип прогнозирования взаимодействия [9, 15].

В рамках потоковой модели наиболее обоснованным является принцип целевой координации, который определяет порядок учета ограничений на взаимодействие макроочереди и распределения расчетных задач между уровнями иерархии. При этом в ходе декомпозиции решения задачи (12) векторы искомых переменных \mathbf{x}_{ij}^r и \mathbf{b}_j^r будут рассчитываться непосредственно на процессорах каждой макроочереди, а множители Лагранжа μ_r будут рассчитываться на верхнем уровне управления для предотвращения перегрузки каналов связи.

Таким образом, для решения сформулированной оптимизационной задачи используем принцип целевой координации [9, 15], в рамках которого лагранжиан (12) представим в следующем виде:

$$L = \sum_{r=1}^K \sum_{j=1}^{N_r} f(p_j^r, d_j^r) \bar{n}_j^r + \sum_{r=1}^K \mu_r^t \mathbf{B}_r \mathbf{x}_r + \sum_{r=1}^K \mu_r^t \sum_{\substack{s=1 \\ s \neq r}}^K \mathbf{B}_s \mathbf{x}_s - \sum_{r=1}^K \mu_r^t b. \quad (13)$$

Предположив, что величины μ_r являются фиксированными, можно предпоследнее слагаемое в выражении (8) привести к виду [15]

$$\sum_{r=1}^K \mu_r^t \sum_{\substack{s=1 \\ s \neq r}}^K \mathbf{B}_s \mathbf{x}_s = \sum_{r=1}^K \sum_{\substack{s=1 \\ s \neq r}}^K \mu_r^t \mathbf{B}_s \mathbf{x}_s, \quad (14)$$

и с учетом (14) выражение (8) можно записать следующим образом:

$$L = \sum_{r=1}^K L_r,$$

где

$$L_r = \sum_{j=1}^{N_r} f(p_j^r, d_j^r) \bar{n}_j^r + \mu \mathbf{B}_r \mathbf{x}_r - \mu^t b. \quad (15)$$

Таким образом, функция (8) приобретает сепарабельную форму, а общая проблема распределения нагрузки между очередями и выделения ПС оказалась декомпозированной на ряд подзадач (по числу макроочереди), где каждая задача распределения на r -й макроочереди, состоящая в расчете векторов \mathbf{x}_{ij}^r и \mathbf{b}_j^r , свелась к оптимизации лагранжиана L_r (15). При этом лагранжиан (15) является функцией переменных только одного индекса, определяющего номер макроочереди.

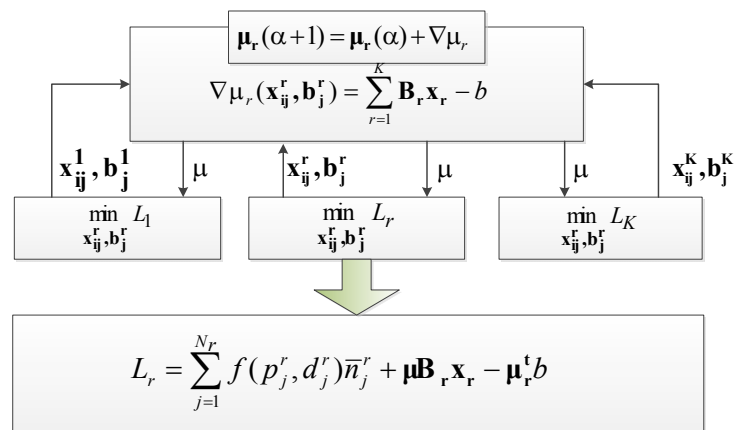


Рис. 2. Схема двухуровневого обслуживания очередей

Решение задачи минимизации выражения (15) определяет нижний уровень расчетов, результатом которого является формирование распределения трафика и выделения ПС каждой конкретной макроочереди согласно содержанию искомых векторов \mathbf{x}_{ij}^r и \mathbf{b}_j^r . На верхнем уровне, основной задачей которого является координация решений, полученных на нижнем уровне с целью недопущения перегрузки канала связи (11), осуществляется модификация вектора множителей Лагранжа в ходе выполнения следующей градиентной процедуры:

$$\mu_r(\alpha + 1) = \mu_r(\alpha) + \nabla \mu_r, \quad (16)$$

где $\nabla \mu_r$ – градиент функции, который рассчитывается исходя из получаемых на нижнем уровне результатов решения задачи распределения потоков и выделения ПС каждой конкретной r -й макроочереди ($r \in K$), т.е.

$$\nabla_{\mu_r}(\mathbf{x}_{ij}^r, \mathbf{b}_j^r) = \sum_{r=1}^K \mathbf{B}_r \mathbf{x}_r - b. \quad (17)$$

Результаты решения задач на каждой отдельной макроочереди собираются координатором (верхний уровень иерархии), осуществляется их анализ и координация путем расчета (корректировки) векторов множителей Лагранжа μ_r ($r \in K$) (16)–(17) в ходе оптимизации лагранжиана (13). Полученные новые решения снова «спускаются» на первый уровень управления для итерационной оптимизации процесса распределения трафика и выделения пропускной способности (рис. 2).

Исследование и оценка эффективности предложенного двухуровневого метода иерархическо-координационного обслуживания очередей на узлах телекоммуникационной сети

Исследование предложенного метода основывалось на анализе сходимости результатов аналитического моделирования. Для лабораторного исследования метода можно использовать подход, описанный в [16]. Оптимизационная задача решалась с использованием пакета Optimization Toolbox среды MATLAB 7, представленного программой fmincon, в рамках которой поддерживается решение задач нелинейного программирования с произвольно заданной целевой функцией. Проводился анализ для различного количества потоков на входе сетевого узла, числа макроочереди, ПС канала. Для примера решения задачи обслуживания очередей с использованием предложенного метода взяты следующие исходные данные. На вход сетевого узла поступал трафик, состоящий из девяти потоков интенсивностью $a_1 = 8 \text{ с}^{-1}$, $a_2 = 12 \text{ с}^{-1}$, $a_3 = 10 \text{ с}^{-1}$, $a_4 = 24 \text{ с}^{-1}$, $a_5 = 22 \text{ с}^{-1}$, $a_6 = 18 \text{ с}^{-1}$, $a_7 = 38 \text{ с}^{-1}$, $a_8 = 22 \text{ с}^{-1}$, $a_9 = 26 \text{ с}^{-1}$. Обслуживание трафика производилось с помощью модели $M/M/1$. Для обслуживания очередей предложено использование трех макроочереди с двумя очередями в каждой макроочереди. ПС канала связи составляла 250 с^{-1} . Результаты исследования сходимости метода представлены на рис. 3.

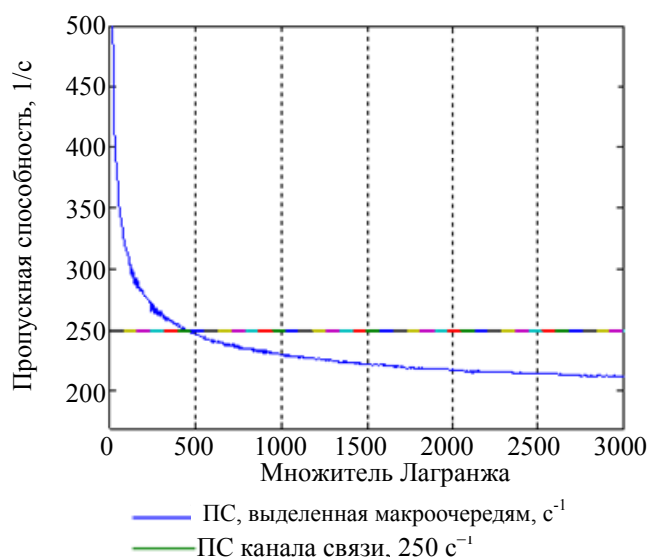


Рис. 3. Результаты исследования сходимости предложенного метода

На рис. 3 показан координирующий эффект введения множителей Лагранжа при переходе к иерархической постановке обслуживания очередей. Изначально, при малых значениях множителя Лагранжа, каждая из макроочереди для обслуживания поступавшего в нее трафика выделяла максимально возможную ПС, что в сумме в разы превышало ПС канала связи. С увеличением множителя Лагранжа, который является штрафующим (координирующим) элементом метода, суммарная ПС, выделяемая макроочередям, уменьшалась. Для исходных данных, представленных выше, значение множителя Лагранжа, обеспечивающее удовлетворение условия непереполнения канала связи, (11) лежит в пределах от 450 и выше. Достаточно большое значение множителя Лагранжа позволит сразу выполнить ограничение непереполнения канала связи независимо от исходных данных, но это приведет к увеличению длины очереди, ведь чем меньше ПС выделено для потока, тем выше будет значение средней длины очереди, что отрицательно скажется на качестве обслуживания потоков пользователей. Наша задача заключается в поиске такого значения координирующих элементов (множителей Лагранжа), которое позволит максимально использовать доступную ПС канала связи, не переполнив его, т.е. на пересечении графиков на рис. 3.

Порядок решения сформулированной задачи с исходными данными, представленными выше, показан на рис. 4.

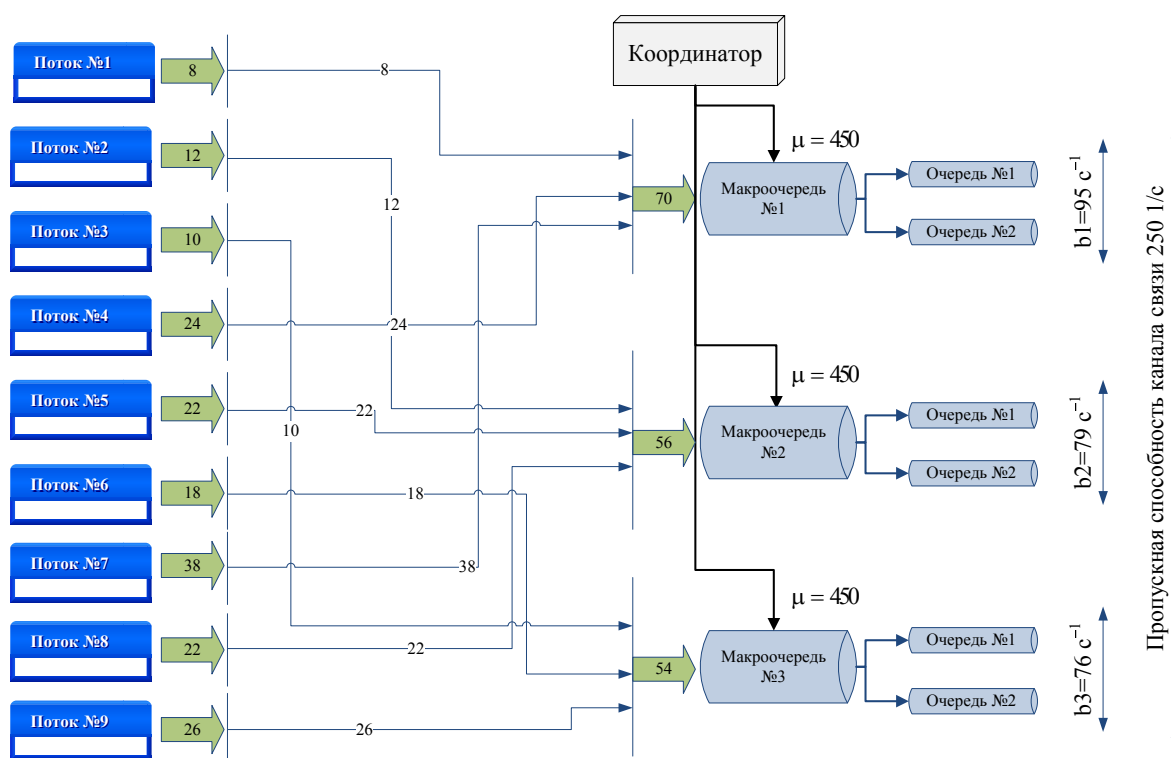


Рис. 4. Пример результатов расчета предложенного метода

В результате решения оптимизационной задачи получено распределение потоков между очередями, выделенная пропускная способность \mathbf{b}^r для каждой макроочереди составила 95 c^{-1} , 79 c^{-1} , 76 c^{-1} соответственно, множитель Лагранжа равнялся 450.

Заключение

В работе предложен двухуровневый метод иерархическо-координационного обслуживания очередей на узлах телекоммуникационной сети. В основу метода положен принцип целевой координации, а задача обслуживания очередей сведена к решению оптимизационной задачи, представленной в декомпозиционной форме. При решении задачи обслуживания очередей это позволило обеспечить более высокий уровень адекватности (системности) процесса управления и получить более высокое качество решения задач управления перегрузками на узлах сети, адаптировав их под современную многопроцессорную архитектуру сетевых узлов. Адекватность полученных результатов обеспечивалась использованием положений известного и хорошо апробированного математического аппарата – теории управления многоуровневыми системами, теории массового обслуживания, а также четкой физической трактовкой искомых переменных и полученных результатов исследования. Предложенный математический аппарат подходит для решения задач распределения ресурсов в распределенных, многоядерных, многопроцессорных, многомодульных системах. Использование параллельно работающих процессоров для вычислений на первом уровне иерархии позволит увеличить скорость нахождения оптимального решения. Исследование сходимости предложенного метода показало характер влияния координирующего вектора на результаты расчета, а также на способ нахождения таких координирующих значений, которые позволят максимально использовать доступный ресурс пропускной способности, тем самым повысив качество обслуживания потоков пользователей.

Литература

1. Вегенша Ш. Качество обслуживания в сетях IP: Пер. с англ. М.: Вильямс, 2003. 386 с.
2. Valencia C. Scheduling alternatives for mobile WiMAX end-to-end simulations and analysis. Master Thesis. Ottawa, Canada, Carleton University, 2009. 116 p.
3. Victoria R.S., Kumar S.S. Efficient bandwidth allocation for packet scheduling // International Journal of Future Computer and Communication. 2012. V. 1. N 4. P. 381–384.
4. Luo Y., Yin S., Wang T., Suemura Y., Nakamura S., Ansari N., Cvijetic M. QoS-aware scheduling over hybrid optical wireless networks // Proc. of OFC/NFOEC 2007. Anaheim, USA, 2007. Art. N 4348497.
5. Down D., Lewis M. Dynamic load balancing in parallel queueing systems: stability and optimal control // European Journal of Operational Research. 2006. V. 168. N 2 spec. P. 509–519.

6. Yang L., Kwok Y.-K., Wang J. An adaptive packet scheduling algorithm for efficient downlink bandwidth allocation in UWB based wireless infrastructure networks // *Computer Communications*. 2007. V. 30. N 9. P. 2087–2095.
7. Badger M., Grance T., Patt-Corner R., Voas J. Cloud computing synopsis and recommendations (Special Publication 800-146). NIST, 2012. 82 p.
8. Cisco cloud computing – data center strategy, architecture, and solutions, 2009 [Электронный ресурс]. Режим доступа: http://www.cisco.com/web/strategy/docs/gov/CiscoCloudComputing_WP.pdf свободный. Яз. англ. (дата обращения 12.12.2013).
9. Месарович М., Мако Д., Такахара И. Теория иерархических многоуровневых систем: Пер. с англ. М.: Мир, 1973. 344 с.
10. Dandamudi S.P., Ayachi S. Performance of hierarchical processor scheduling in shared-memory multiprocessor systems // *IEEE Transactions on Computers*. 1999. V. 48. N 11. P. 1202–1213.
11. Лемешко А.В. Хайлан А.М. Многоуровневое управление трафиком в сети MPLS-TE DiffServ на основе координационного принципа прогнозирования взаимодействий // *Проблемы телекоммуникаций*. 2010. № 1 (1). С. 35–44.
12. Беленков А.Г., Евсеева О.Ю., Лемешко А.В. Метод распределения нагрузки в иерархических телекоммуникационных сетях на базе декомпозиционных принципов предсказания взаимодействий и целевой координации // *Праці УНДІРТ*. 2005. № 2 (42). С. 11–16.
13. Tsai T.-Y., Chung Y.-L., Tsai Z. Introduction to packet scheduling algorithms for communication networks. In: Jun Peng (ed.) *Communications and Networking*. Sciyo, Croatia, 2010. 434 p.
14. Петров М.Н., Пономарев Д.Ю. Самоподобие в системах массового обслуживания с ограниченным буфером // *Электросвязь*. 2002. № 2. С. 35–39.
15. Сингх М., Титли А. Системы: декомпозиция, оптимизация и управление: Пер. с англ. М.: Машиностроение, 1986. 496 с.
16. Лемешко А.В., Семеняка М.В., Али С. Особенности проведения лабораторного эксперимента по исследованию процессов управления очередями в телекоммуникационных сетях // *Цифрові технології*. 2011. № 10. С. 5–17.