УДК 519.872

АНАЛИЗ СИСТЕМЫ ОБСЛУЖИВАНИЯ С РАЗЛИЧНЫМИ УРОВНЯМИ ПРОСТРАНСТВЕННЫХ И ВРЕМЕННЫХ ПРИОРИТЕТОВ

Пономаренко Л.А., Меликов А.З., Нагиев Ф.Н.

Введение

Приоритеты, определяющие процедуры принятия в буфер (очередь) разнотипных заявок, зачастую называются пространственными приоритетами (*Space Priorities, SP*), а приоритеты, задающие правила выбора типа заявки из очереди (буфера), получили название временных (*Time Priorities, TP*). В классических схемах приоритетного обслуживания в системах с очередями, как правило, предполагается, что заявки определенного типа обладают (по сравнению с заявками другого типа) одновременно высокими (или низкими) приоритетами обоих видов.

Вместе с тем, интерес представляют модели обслуживания, в которых заявки одного типа являются чувствительными к возможным потерям из-за переполненности буфера и одновременно не предъявляют жестких требований к потерям, в то время как заявки другого типа, наоборот, не предъявляют жестких требований к потерям, а являются очень чувствительными к возможным задержкам в очереди. Иными словами, целесообразно исследовать модели, в которых заявки одного типа имеют высокие *SP* и низкие *TP*, а заявки другого типа, наоборот, имеют низкие *SP* и высокие *TP*. Модели такого типа назовем моделями обслуживания с различными уровнями пространственных и временных приоритетов (можно также использовать термин модели обслуживания с множественными приоритетами).

Подобные модели достаточно точно описывают работу узлов сетей коммутации пакетов, в которых пакеты трафика реального времени (например, пакеты речевой информации) имеют высокие *TP* и низкие *SP* по сравнению с пакетами трафика нереального времени (например, пакетами данных), имеющими более высокие *SP* и низкие *TP* по сравнению с пакетами реального времени.

Несмотря на то, что указанные модели представляют большой теоретический и практический интерес, они в доступной литературе не достаточно исследованы. Лишь в последние годы в работах [1-3] исследованы такие модели. При этом в работах [1, 2] используется метод имитационного моделирования, а в работе [3] предлагается аналитический метод для нахождения характеристик системы (вероятностей потерь разнотипных заявок и средних времен их ожидания в очереди). Отметим, что в работе [3] для ожидания заявок с различными требованиями *TP* используются различные буфера и внутри каждого буфера *SP* реализуются на основе стратегии вытеснения.

В настоящей работе также исследуется модель с различными уровнями *TP* и *SP*. Однако, данная модель отличается от [3] двумя моментами. Во-первых, здесь рассматривается модель с общим буфером для ожидания разнотипных заявок. Во-вторых, для нахождения характеристик модели здесь предлагается численный метод, основанный на принципах теории фазового укрупнения [4]. Использование данного подхода позволяет разработать простые вычислительные процедуры для нахождения искомых характеристик. Ранее данный подход был успешно применен для модели систем со специализированными каналами обслуживания и при наличии лишь пространственных приоритетов различного вида [5-7].

1. Описание модели и расчет ее характеристик

Буферное пространство размера B и канал передачи используются совместно заявками двух типов, при этом заявки первого типа представляют трафик нереального времени, заявки второго типа – трафик реального времени. Используется обычное

 ISBN 7-776-8361-7

предположение о пуассоновском характере входящих трафиков, т.е. предполагается, что процесс поступления заявок i-го типа подчиняется закону Пуассона с параметром λ_i, =1,2.

Из-за сложности рассматриваемой модели и с целью получения достаточно обозримых результатов здесь предполагается, что для обоих трафиков время обслуживания распределено показательно со средним μ^{-1}. Считается, что заявка любого типа освобождает свое место в буфере в момент выбора ее для обслуживания.

Множественные приоритеты определяются следующим образом. Поскольку трафик первого типа является более чувствительным к возможным потерям пакетов из-за переполненности общего буфера, чем трафик второго типа, то они (т.е. заявки первого типа) имеют высокие пространственные приоритеты. Эти приоритеты осуществляются с помощью стратегии доступа с вытеснением, т.е. поступившая заявка первого типа теряется лишь тогда, когда буфер полностью заполнен и число текущих заявок данного типа в буфере не меньше, чем заданное число c, $1 \le c \le B$. Иными словами, поступившая заявка первого типа вытесняет из полностью заполненного буфера заявки второго типа, если текущее число заявок первого типа меньше, чем c; в противном случае, т.е. когда в полностью заполненном буфере число заявок первого типа не меньше чем c, поступившая заявка первого типа теряется. Заявки второго типа теряются тогда, когда буфер полностью заполнен. Параметр c назовем пороговым параметром для заявок первого типа.

Замечание 1. При $c = B$ рассматриваемые *SP* приводят к схеме приоритетизации, исследованной в [3], т.е. введение порогового значения c придает рассматриваемым здесь *SP* элемент адаптивности.

Поскольку трафик второго типа является более чувствительным к возможным задержкам в очереди, чем трафик первого типа, то заявки второго типа имеют высокие относительные временные приоритеты, т.е. заявки первого типа принимаются для обслуживания лишь тогда, когда в момент освобождения канала в буфере отсутствуют заявки второго типа.

Функционирование буфера описывается двумерной цепью Маркова (а более точно – двумерным процессом рождения и гибели) с состояниями типа $\boldsymbol{n} = (n_1, n_2)$, где n_i указывает число заявок i-го типа в очереди, $i = 1,2$. Фазовое пространство состояний (ФПС) буфера задается так:

$$E := \left\{\boldsymbol{n} : n_1 = \overline{0, B}, n_1 + n_2 \le B\right\}. \tag{1}$$

Согласно определению введенных множественных приоритетов интенсивности переходов между состояниями $\boldsymbol{n}, \boldsymbol{n}' \in E$ определяются следующим образом:

$$q(\boldsymbol{n}, \boldsymbol{n}') = \begin{cases} \lambda_i, \textit{если } n_1 + n_2 < B, \boldsymbol{n} = \boldsymbol{n}' + \boldsymbol{e_i}, \\ \lambda_1, \textit{если } n_1 + n_2 = B, n_1 < c, n_2 > 0, \boldsymbol{n} = \boldsymbol{n}' + \boldsymbol{e_1} - \boldsymbol{e_2}, \\ \mu, \textit{если } n_2 = 0, \boldsymbol{n}' = \boldsymbol{n} - \boldsymbol{e_1} \textit{ или } n_2 > 0, \boldsymbol{n}' = \boldsymbol{n} - \boldsymbol{e_2}, \\ 0, \textit{в остальных случаях}, \end{cases} \tag{2}$$

где $\boldsymbol{e_1} = (1,0)$, $\boldsymbol{e_2} = (0,1)$.

Граф модели показан на рис. 1.

Стационарную вероятность состояния $\boldsymbol{n} \in E$ обозначим через $p(\boldsymbol{n})$. Тогда, согласно известной теореме PASTA [8] стационарные вероятности потери заявок i-го типа (*Call Loss Probability, CLP*$_i(B,c)$), i=1,2 определяются так:

$$CLP_i(B,c) = \sum_{\boldsymbol{n} \in E_i^i} p(\boldsymbol{n}) \tag{3}$$

Здесь E_i – множество блокирующих состояний из ФПС (1) для трафика i-го типа, $i = 1,2$. При этом под множеством блокирующих состояний для трафика i-го типа понимается множество тех состояний из ФПС E, в которых поступление заявки данного типа означает его потерю безотносительно того, поступает ли в этот момент заявка данного типа или нет.

$\boldsymbol{\lambda_2}$

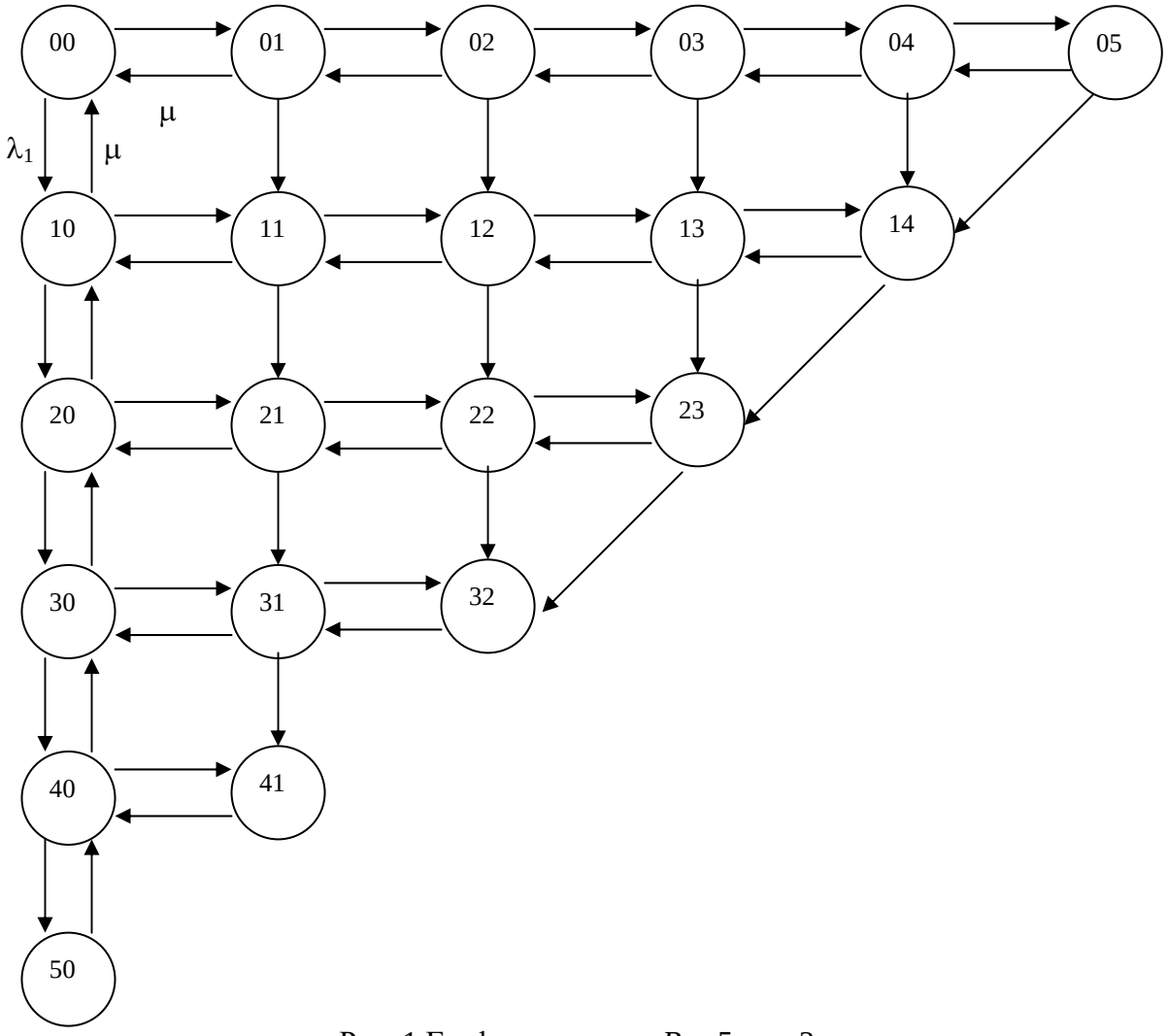

Рис. 1 Граф модели при $B = 5$, $c = 2$.

Формально указанные выше множества определяются так:

$E_1 = \{\boldsymbol{n} \in E_2 : n_1 \geq c\}$;

$E_2 = \{\boldsymbol{n} \in E: n_1+n_2=B\}$ – множество диагональных состояний.

Отсюда с учетом (3) получаем, что

$CLP_1(B, c) < CLP_2(B, c)\ \forall c \in [1, B]$.

Среднее время задержки в очереди заявок i-го типа (*Call Transfer Delay*, $CTD_i(B,c)$) определяется с помощью формулы Литтла для систем обслуживания с конечной очередью:

$$CTD_i(B, c) := Q_i(B, c)/\lambda_i(1\text{-}CLP_i(B, c)),\ \ i = 1,2 \qquad (4)$$

Здесь $Q_i(B,c)$ обозначает среднее число заявок i-го типа в очереди, $i = 1,2$. Эти величины определяются так:

$$Q_i(B,c) := \sum_{k=1}^{B} k \sum_{\boldsymbol{n} \in E_k} p(\boldsymbol{n}), \qquad (5)$$

где $E_k^{\,i} = \{\boldsymbol{n} \in E : n_i = k\}$, $i = 1,2$.

Следовательно, искомые характеристики модели (3)-(5) определяются через стационарное распределение $p(\boldsymbol{n})$, $\boldsymbol{n} \in E$ исходной цепи Маркова. С использованием известного критерия Колмогорова для двумерных цепей Маркова [9] можно показать, что для стационарного распределения данной модели не существует мультипликативное решение. Этот факт существенно осложняет проблемы расчета характеристик модели (3)-(5), особенно при больших размерах буфера. Для преодоления этих трудностей предлагается использовать приближенный метод расчета указанных характеристик.

 ISBN 7-776-8361-7

Поскольку предложенный подход достаточно подробно описан в литературе (см., например, [5-7] и библиографию при них), то здесь приводится лишь его краткое изложение применительно к данной модели.

Отметим, что предложенный подход имеет высокую точность и является особенно полезным для расчета стационарного распределения двумерных процессов размножения и гибели, которые являются непрерывными хотя бы по одной компоненте. Напомним, что двумерный процесс размножения и гибели является непрерывным, скажем по первой компоненте, если существует положительные двусторонние вероятности переходов между любыми двумя возможными состояниями типа (i, j) и $(i+1, j)$.

Из соотношений (2) легко видно, что изучаемый двумерный процесс рождения и гибели с ФПС (1) является непрерывным лишь по второй компоненте (см. также рис. 1).

Теперь перейдем к краткому изложению предложенного подхода расчета стационарного распределения изучаемой двумерной цепи Маркова.

Рассматривается следующее разбиение ФПС E:

$$E = \bigcup_{k=0}^{B} E_k,\ E_k \bigcap E_{k'} = \varnothing,\ k \neq k', \tag{6}$$

где $E_k := E_k^{\,1}$ (см.(5)).

Далее классы состояний E_k объединяются в отдельные укрупненные состояния $<k>$ и вводится функция укрупнения на исходном ФПС E:

$$U(\boldsymbol{n}) =< k >, \boldsymbol{n} \in E_k, k = \overline{0, B} \tag{7}$$

Функция укрупнения (7) определяет укрупненную цепь Маркова с ФПС

$E := \{< k >: k = \overline{0, B}\}$.

Стационарное распределение исходной модели определяется так:

$$p(n_1, n_2) \approx \rho_{n_1}(n_2)\pi(< n_1 >) \tag{8}$$

где $\rho_{n_1}(n_2), (n_1, n_2) \in E_{n_1}$ и $\pi(< n_1 >), < n_1 > \in E$ являются стационарными распределениями внутри класса E_{n_1} и укрупненной модели, соответственно.

Стационарное распределение внутри класса E_i с учетом (2) определяется как распределение одномерного процесса размножения и гибели:

$$\rho_i(k) = \upsilon_2^k(1-\nu_2)/(1-\nu_2^{B+1-i}),\ i = \overline{0, B},\ k = \overline{0, B-i}, \tag{9}$$

где $\nu_2 := \lambda_2/\mu$ (для краткости изложения здесь приводятся формулы лишь для случая $\nu_2 \neq 1$).

Далее на основе (2) и (9) определяется стационарное распределение укрупненной модели:

$$\pi(k) = \begin{cases} \nu_1^k \left(\prod_{i=1}^{k} \rho_i(0) \right)^{-1} \pi < 0 >,\ \textit{если}\ k = \overline{1, c}, \\ \nu_1^k \left(\prod_{i=1}^{k} \rho_i(0) \right)^{-1} \left(\prod_{j=B-k+1}^{B-c} (1 - L(\nu_2, j)) \right) \pi < 0 >,\ \textit{если}\ k = \overline{c+1, B}, \end{cases} \tag{10}$$

где

$$\pi < 0 >= \left(\sum_{k=0}^{c} \nu_1^k \left(\prod_{i=1}^{k} \rho_i(0) \right)^{-1} + \sum_{k=c+1}^{B} \nu_1^k \left(\prod_{i=1}^{k} \rho_i(0) \right)^{-1} \prod_{j=B-k+1}^{B-c} (1 - L(\nu_2, j)) \right)^{-1}, \tag{11}$$

$$\nu_1 = \lambda_1/\mu,\ L(\nu, k) = \nu^k(1-\nu)/(1-\nu^{k+1}).$$

Тогда, после определенных математических преобразований получим следующие формулы для вычисления искомых характеристик модели:

$$CLP_1(B,c) = \sum_{k=0}^{B-c} L(\nu_2, k)\pi(< B-k >) \tag{12}$$

$$CLP_2(B,c) = \sum_{k=0}^{B} L(\nu_2, k)\pi(< B-k >) \tag{13}$$

$$Q_1(B,c) = \sum_{k=1}^{B} k\pi(< k >) \tag{14}$$

$$Q_2(B,c) = \sum_{k=1}^{B} k \sum_{i=0}^{B-k} \pi(< i >)\rho_i(k) \tag{15}$$

Далее, с помощью (12)-(15) вычисляются характеристики (4).

Из формул (9)-(15) видно, что расчет характеристики данной системы обслуживания со сложной схемой приоритетизации сводится к простым вычислительным процедурам. Особенность этих процедур состоит в том, что они подразумевают использование табулированных величин типа $L(\nu, k)$ (см. также формулы 9).

3. Численные эксперименты по расчету модели

Некоторые результаты численных экспериментов, проведенных с помощью разработанных выше расчетных формул, показаны на рис. 2-13. Во всех экспериментах для простоты принято, что $\mu := 1$.

Целью их выполнения является изучение поведения характеристик модели в зависимости от изменения ее параметров (нагрузочных и структурных), а также от порогового параметра для заявок первого типа. Для сравнительного анализа в графиках также показано поведение соответствующих характеристик бесприоритетной модели. Эти характеристики обозначены аналогично приоритетной модели, но только без индексов.

Зависимости характеристик модели от порогового параметра c показаны на рис. 2-4. Как и следовало ожидать, с ростом параметра c шансы для принятия в буфер заявок первого типа растут, и, таким образом, вероятность их потери уменьшается, а одновременно с этим увеличивается вероятность потери пакетов второго типа. При этом скорости их изменения существенным образом зависят от значений параметров модели ν_1, ν_2 и B (см. рис. 2).

Число заявок первого типа (Q_1) с увеличением параметра c также растет вследствие того, что при этом увеличиваются их шансы попасть в буфер, и одновременно в результате вытеснения из буфера уменьшается число заявок второго типа (Q_2) в буфере (см. рис. 3). Несколько неожиданным оказалось поведение функции CTD_1 от параметра c (рис. 4), т.е. с ростом данного параметра CTD_1 уменьшается. Это объясняется тем, что при указанных исходных данных с ростом параметра c скорость уменьшения функции CLP_1 во много раз превосходит скорость увеличения функции Q_1 (см. формулы (4)), а CTD_2 при тех же данных является почти постоянной.

Зависимости характеристик модели от размера буфера (B) показаны на рис. 5-7. Характер этих зависимостей во всех диапазонах изменения исходных данных моделей полностью совпадают с теоретическими ожиданиями, т.е. функции CLP_1 и CLP_2 являются убывающими, а функции Q_k и CTD_k, k = 1,2, наоборот, возрастают относительно аргумента B.

Интерес представляет также изучение зависимости поведения характеристик модели от нагрузки трафиков нереального (λ_1) и реального (λ_2) времени. Соответствующие трафики показаны на рис. 8-9 и 10-12. Как и следовало ожидать, с ростом нагрузки трафика любого типа все характеристики модели имеют тенденции к возрастанию, при этом скорости их изменения отличаются друг от друга, и существенным образом зависят от исходных данных модели.

 ISBN 7-776-8361-7

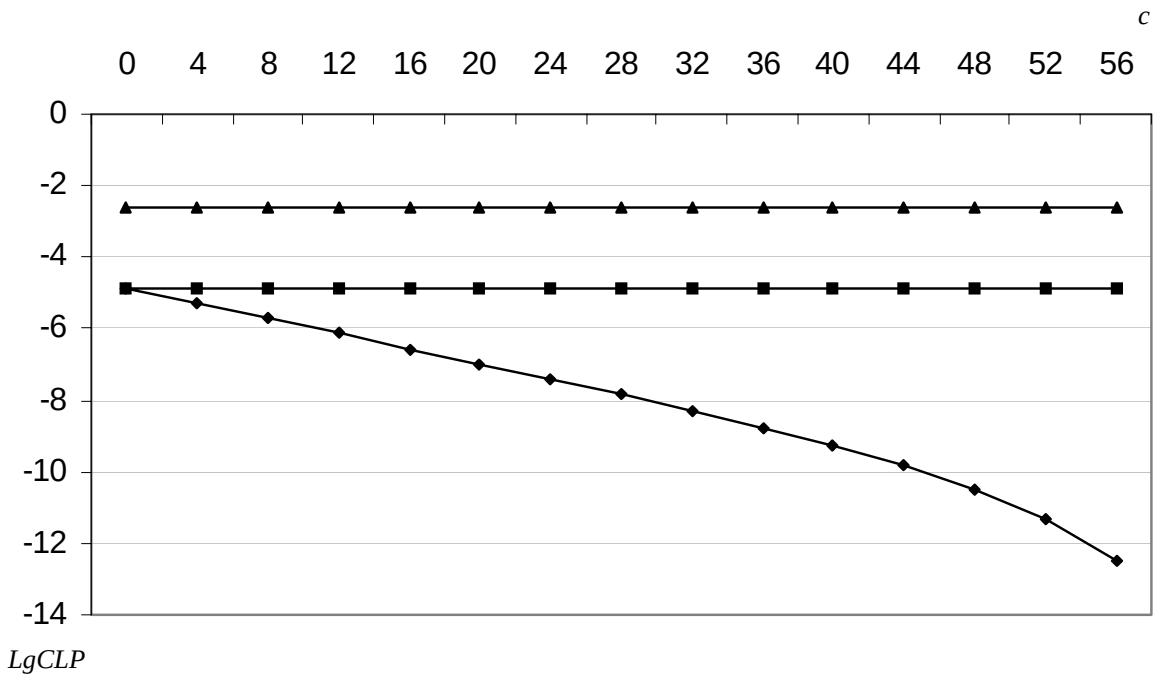

Рис. 2 Зависимость вероятностей потерь разнотипных заявок от параметра *c* при B=60; λ_1=0,1; λ_2=0,85;
▲ –*CLP;* ■ – CLP_2; ♦ – CLP_1

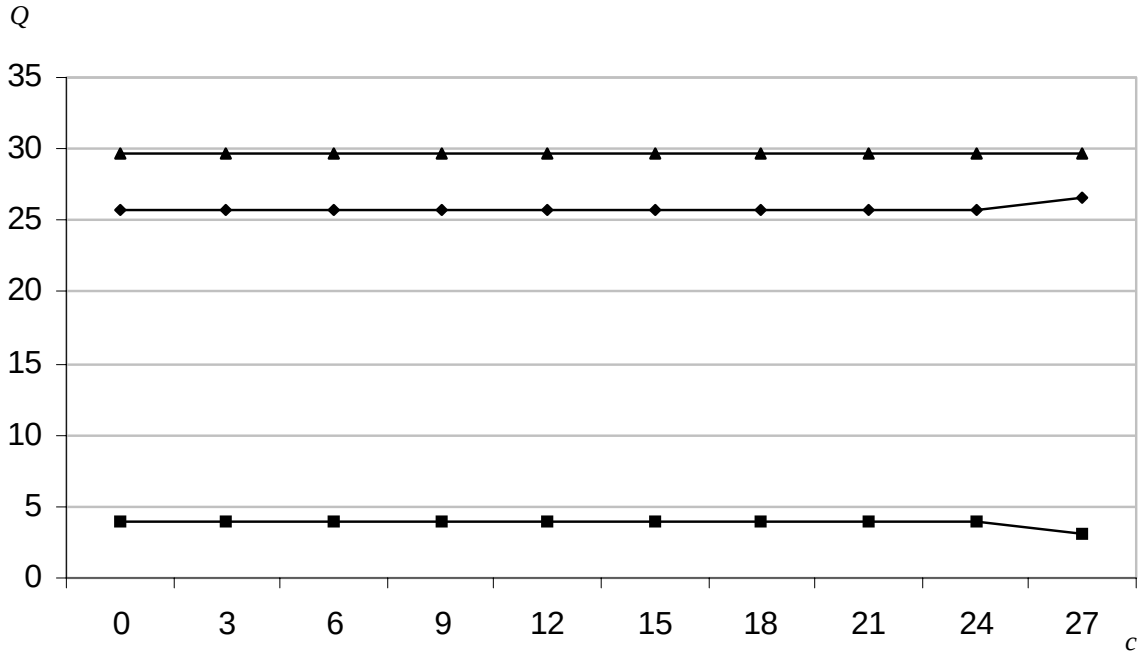

Рис. 3 Зависимость средних длин очередей заявок от параметра c при B=30; λ1=0,01; λ 2=5;
▲ – Q; ■ – Q2; ♦ – Q1

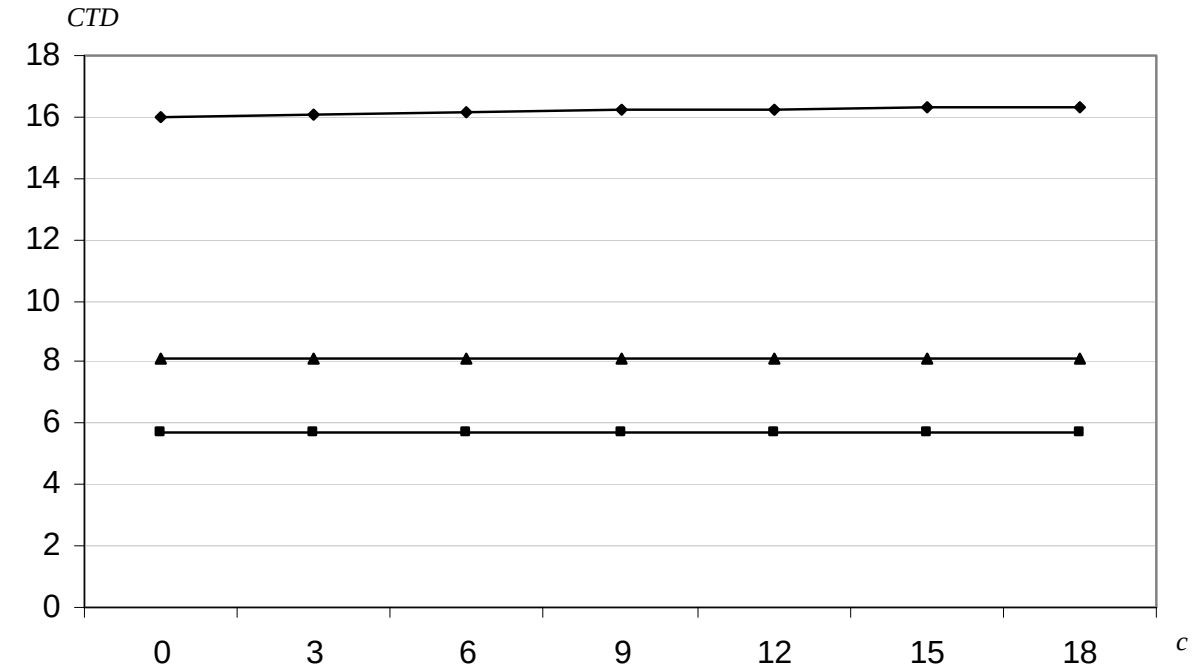

Рис. 4 Зависимость среднего времени ожидания разнотипных заявок от параметра с при B=20; λ1=0,1; λ 2=0,85;
▲ – CTD; ■ CTD 2; ♦ – CTD 1

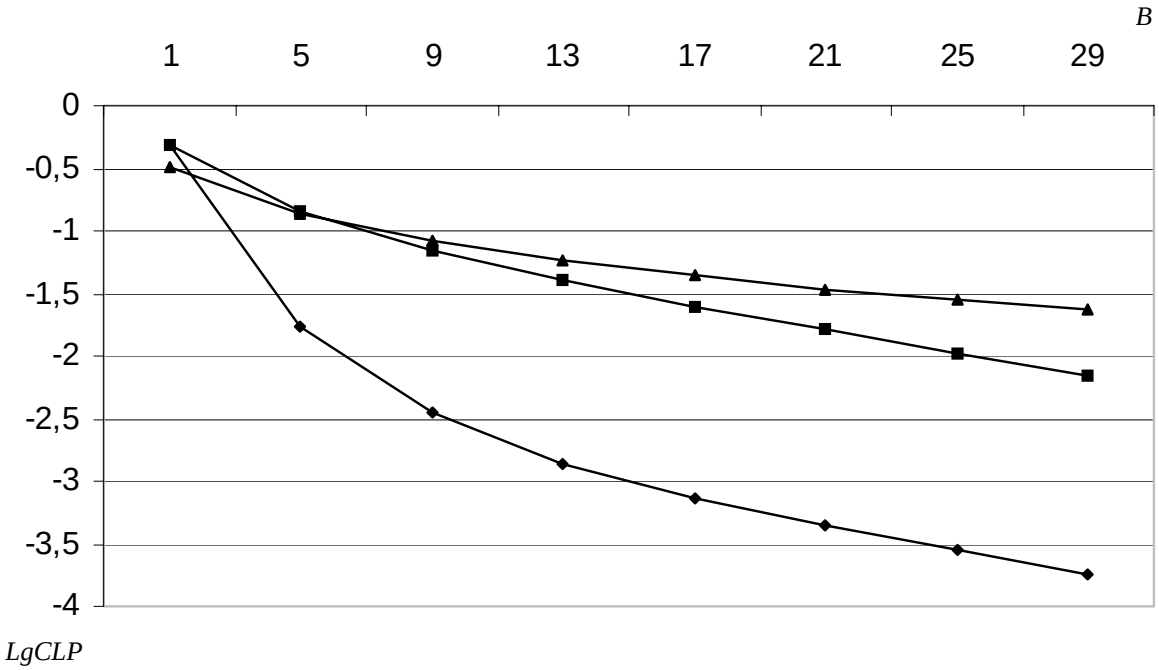

Рис. 5 Зависимость вероятностей потерь разнотипных заявок от параметра B при c=[B/2]; λ1=0,08; λ 2=0,9;
▲ –CLP; ■ – CLP 2; ♦ – CLP 1

ISBN 7-776-8361-7

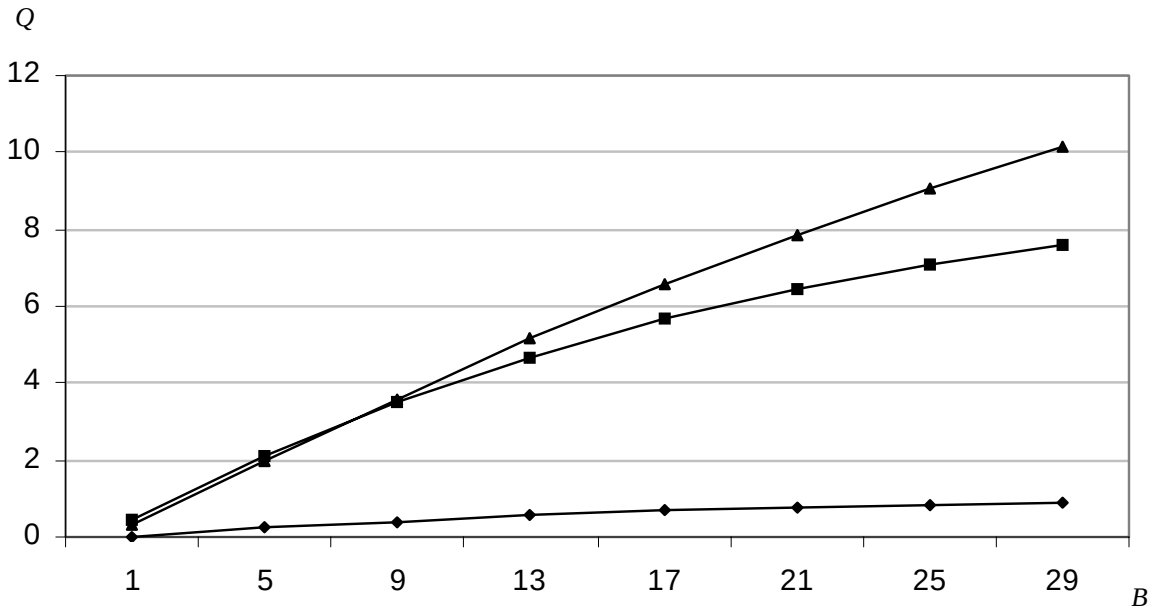

Рис. 6 Зависимость средних длин очередей заявок от параметра B
при c=[B/2]; λ1=0,05; λ 2=0,9;
▲ – Q; ■ – Q2; ♦ – Q1

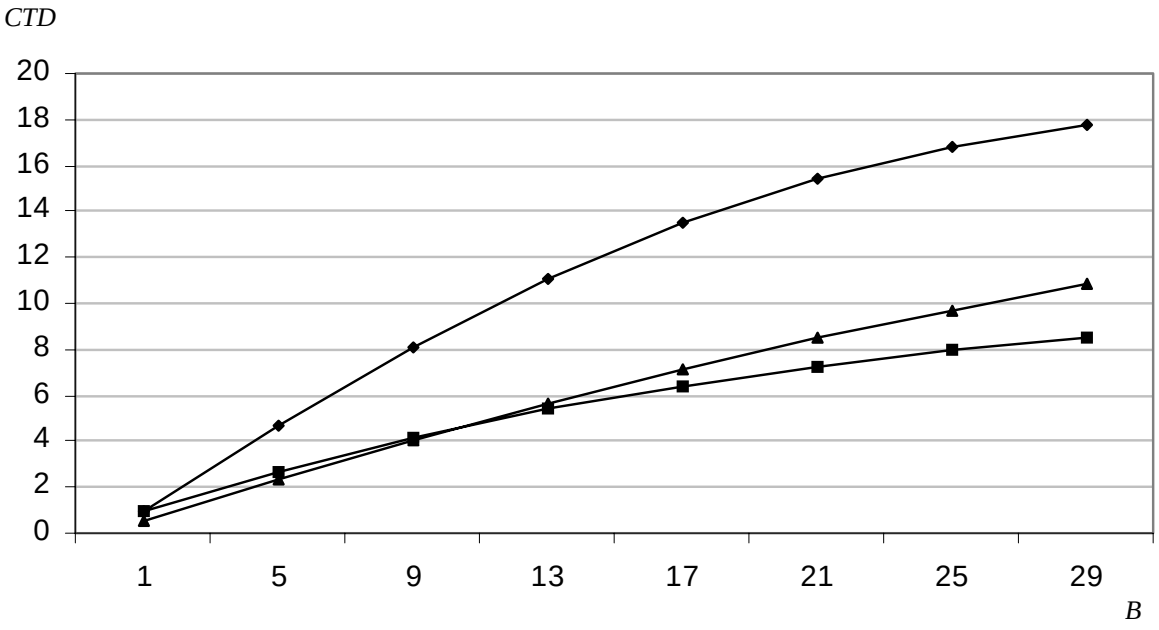

Рис. 7 Зависимость среднего времени ожидания разнотипных заявок
от параметра B при c=[B/2]; λ1=0,05; λ 2=0,9;
▲ – CTD; ■ CTD 2; ♦ – CTD 1

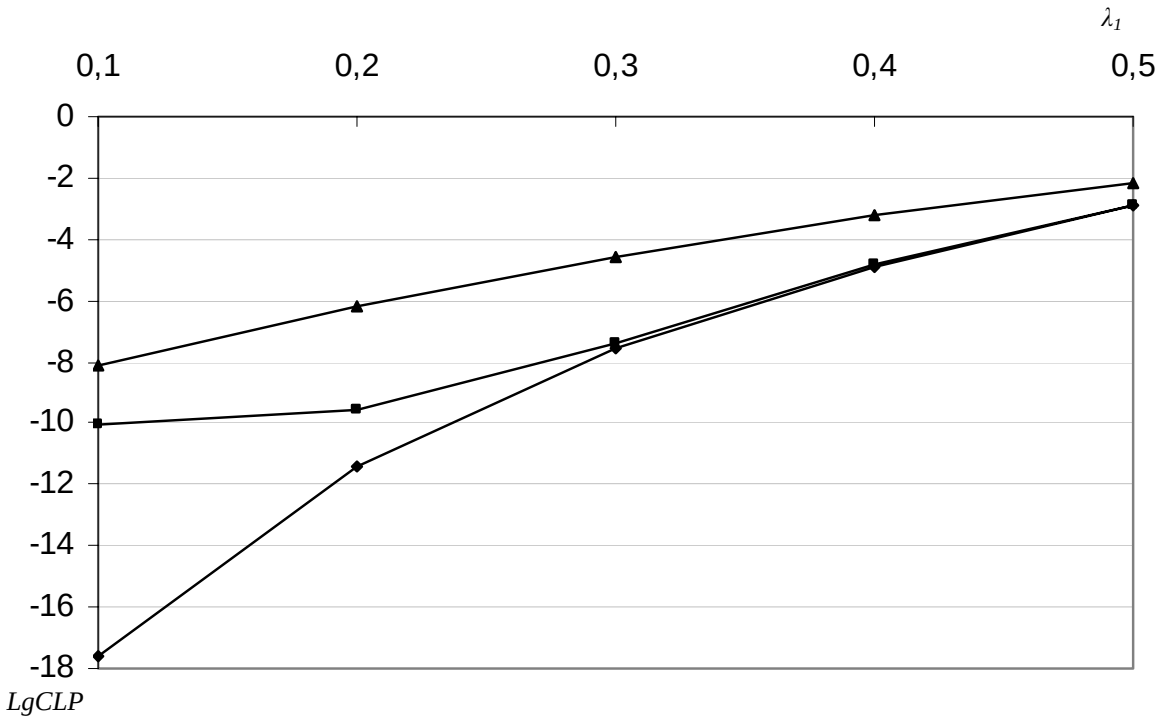

Рис. 8 Зависимость вероятностей потерь разнотипных заявок от параметра λ1 при B=25; c=20; λ 2=0,4;
▲ – CLP; ■ – CLP 2; ♦ – CLP 1

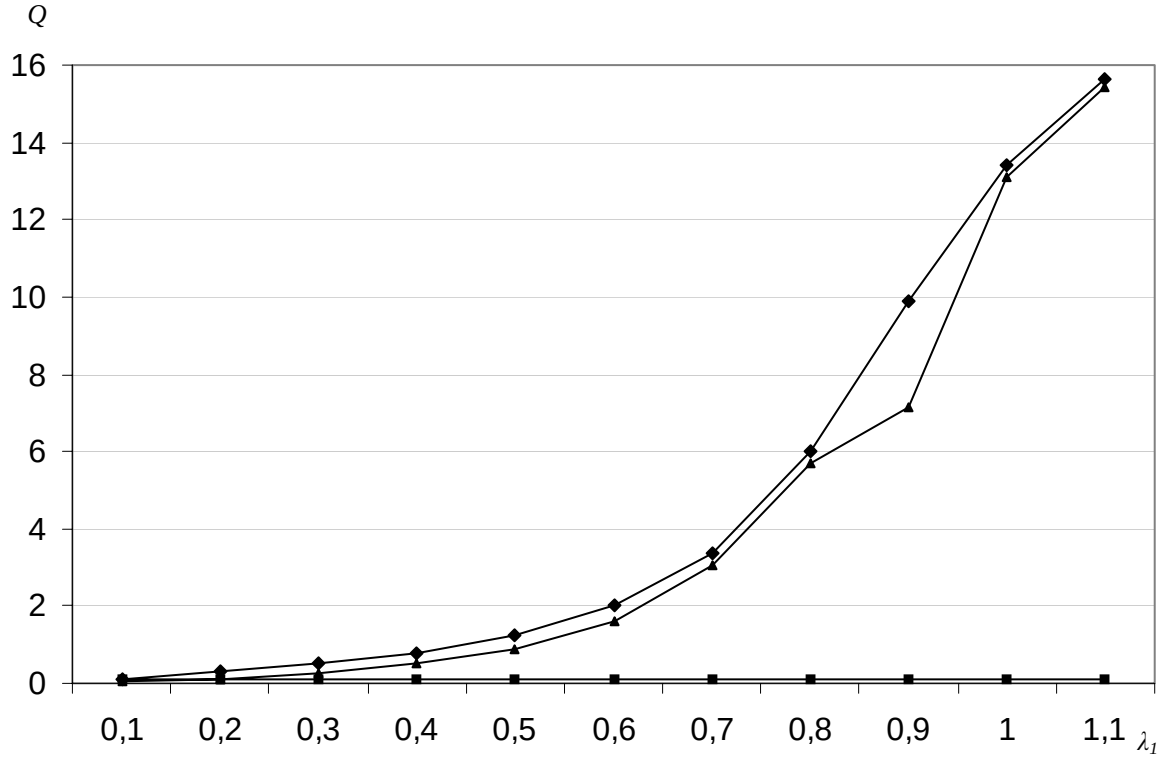

Рис. 9 Зависимость средних длин очередей заявок от параметра λ1 при B=20; c=10; λ 2=0,1;
▲ – Q; ■ – Q2; ♦ – Q1

ISBN 7-776-8361-7

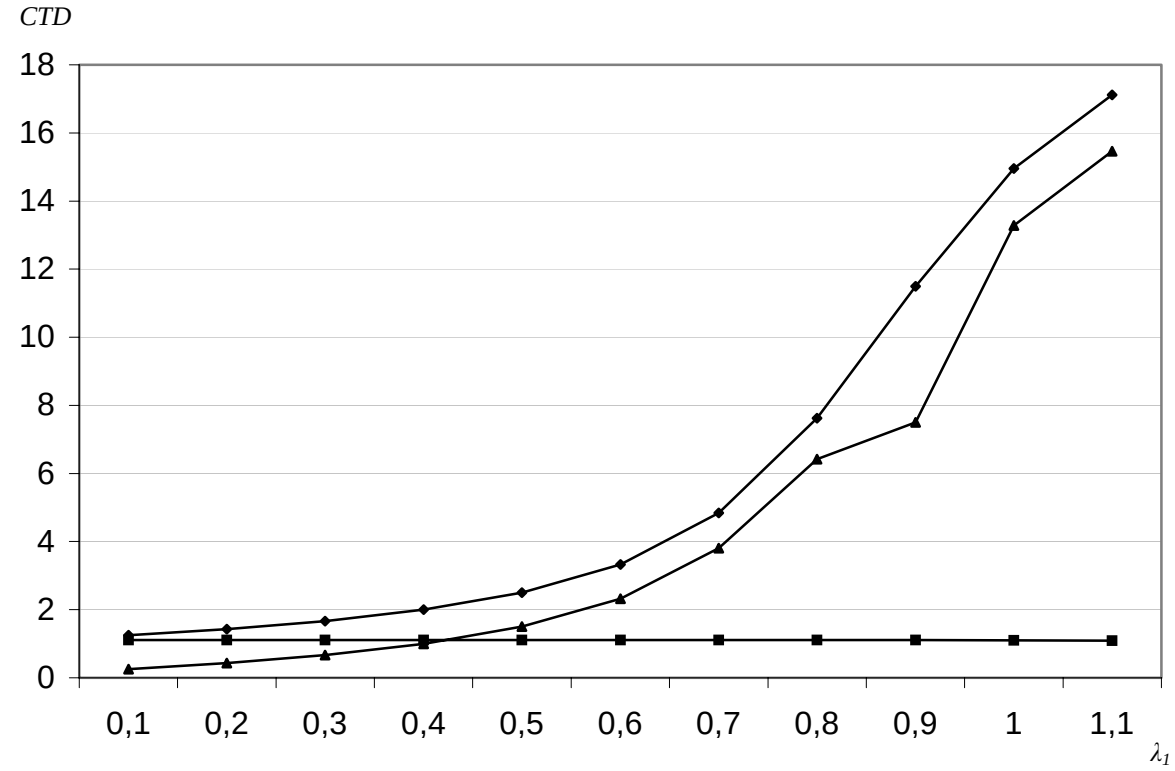

Рис. 10 Зависимость среднего времени ожидания разнотипных заявок от параметра λ1 при B=20; c=10; λ 2=0,1;
▲ – CTD; ■ CTD 2; ♦ – CTD 1

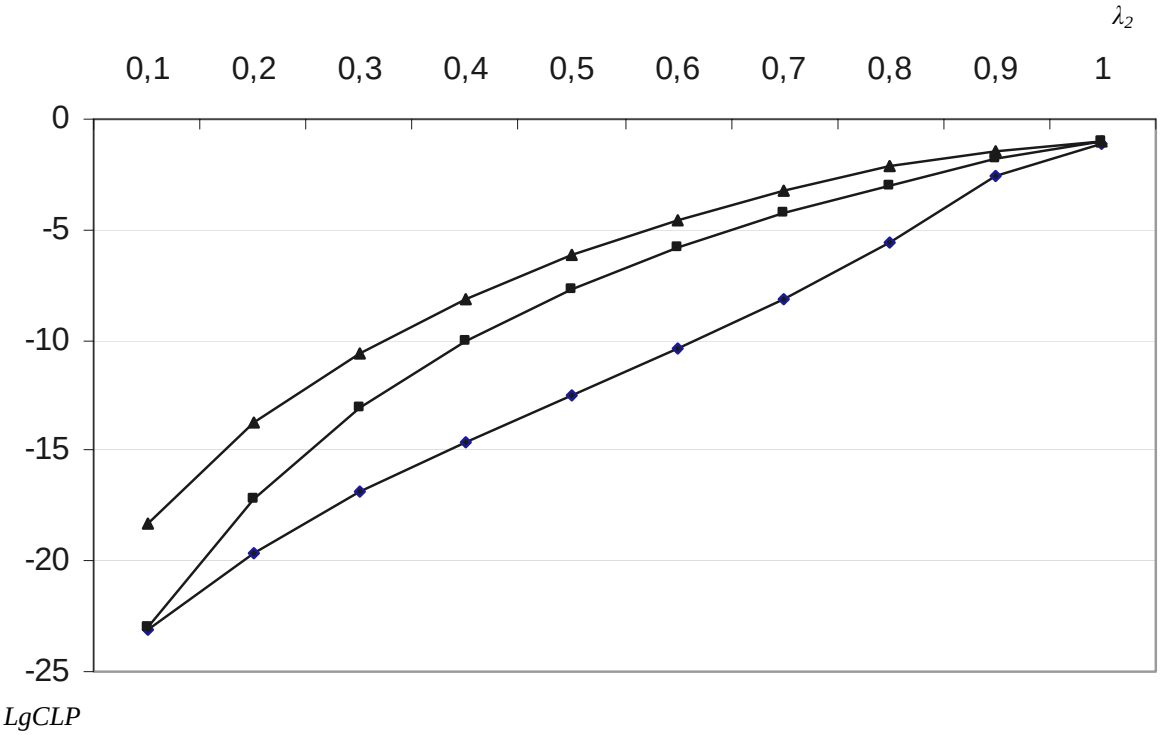

Рис. 11 Зависимость вероятностей потерь разнотипных заявок от параметра λ2 при B=15; c=7; λ1=0,2;
▲ – CLP; ■ – CLP 2; ♦ – CLP 1

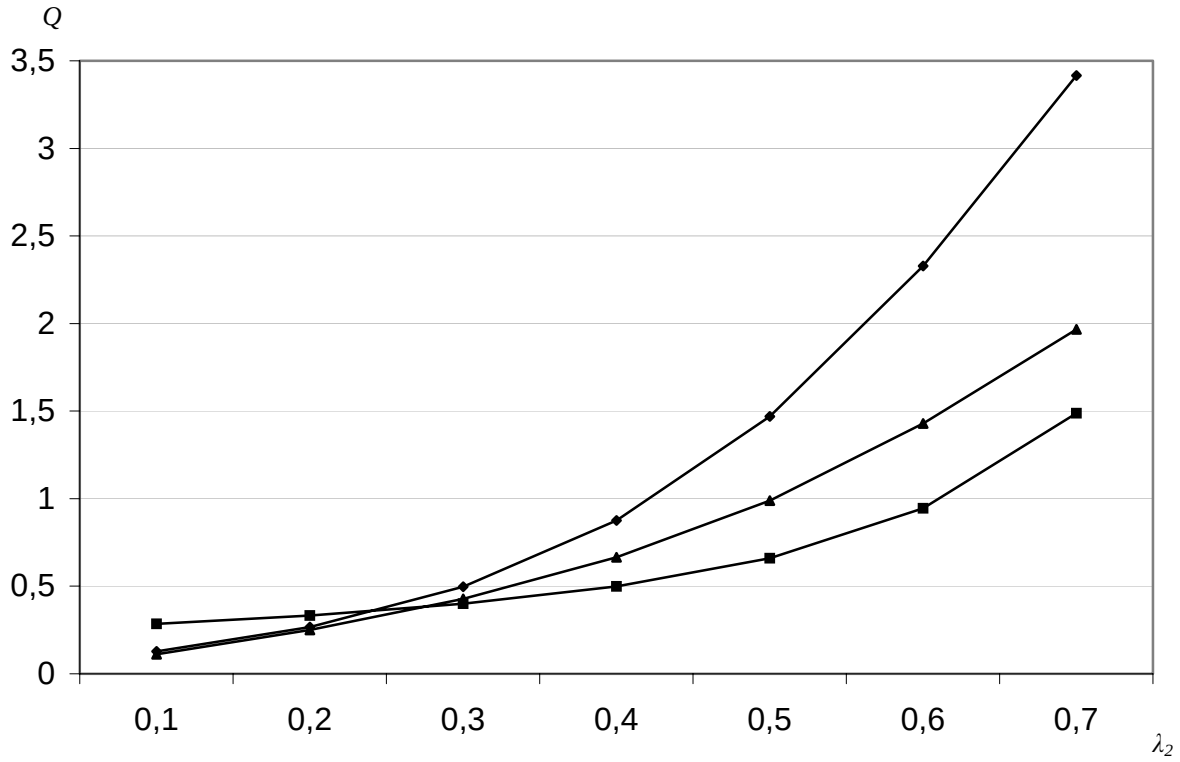

Рис. 12 Зависимость средних длин очередей заявок от параметра λ2
при B=15; c=7; λ1=0,2;
▲ – Q; ■ – Q2; ♦ – Q1

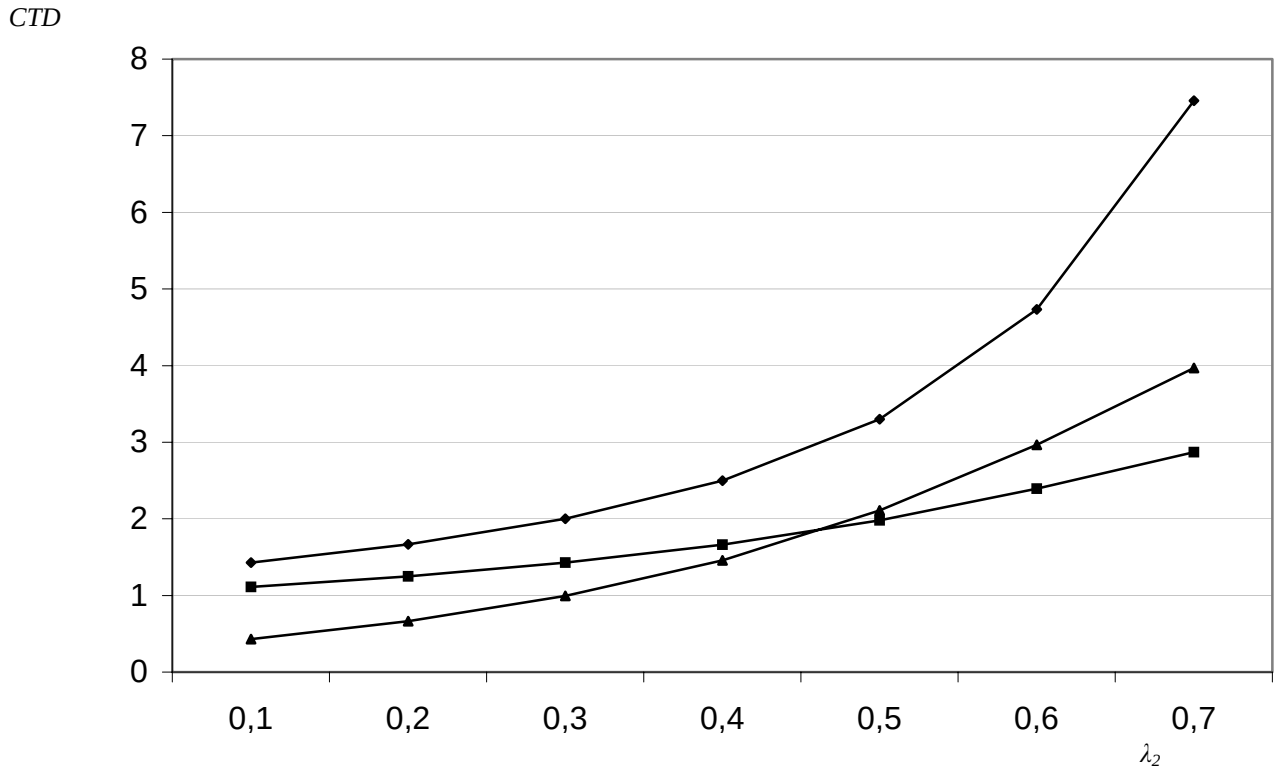

Рис. 13 Зависимость среднего времени ожидания разнотипных заявок
от параметра λ2 при B=15; c=7; λ1=0,2;
▲ – CTD; ■ CTD 2; ♦ – CTD 1

ISBN 7-776-8361-7

Отметим, что представление в графиках на рис. 2-13 также поведения характеристик бесприоритетной модели позволяют производить сравнительный анализ изучаемых характеристик, и, таким образом, принимать решения относительно целесообразности введения множественных приоритетов. Из представленных графиков видно, что последняя проблема (т.е. принятие решения относительно выбора схемы приоритетизации) не является тривиальной и представляет собой многокритериальную проблему (здесь эти проблемы не рассматриваются).

Заключение

Предложенный в данной работе подход предлагает достаточно простые алгоритмы расчета характеристик модели обслуживания со сложной схемой приоритетизации разнотипных трафиков, представляющих различные требования к возможным потерям и задержкам в очереди. Разработанные алгоритмы могут быть использованы и для выбора оптимальных (в заданном смысле) параметров модели с целью достижения искомых значений исследуемых характеристик. Эти задачи представляют собой предмет специальных исследований.

Numerical method to calculate the characteristics of single channel queuing system with priorities of different level is proposed. Here real time calls have low space priorities and high time priorities while non-real time calls have high space priorities and low time priorities. Results of computational experiments are given.

1. Chao H.J., Peckan I.H. Queue management with multiple delay and loss priorities for ATM switches // Proc. ICC'94. – 1994. – P.1184 – 1189.

2. Shan Zhi C., Liemin Y. A new priority control of ATM output buffer // Telecommunication Systems. – 1995. – № 4. – P. 61 – 69.

3. Lee Y., Choi B.D. Queueing system with multiple delay and loss priorities for ATM networks // Information Systems. – 2001. – № 138. – P. 7 – 29.

4. Королюк В.С. Стохастические модели систем. – К.: Наук. думка, 1989. – 208 с.

5. Меликов А.З., Фаттахова М.И., Нагиев Ф.Н. Подход фазового укрупнения для оптимизации стратегий доступа с вытеснением в сетях коммутации пакетов // Кибернетика и системный анализ. – 2004. – № 2. – С.107 – 115.

6. Пономаренко Л.А., Меликов А.З., Фаттахова М.И. Стратегия вытеснения с виртуальным порогом для доступа в буфер узла интегральной сети // Проблемы управления и информатики. – 2004. – № 4. – С. 106 – 115.

7. Пономаренко Л.А., Меликов А.З., Фаттахова М.И. Численные методы исследования многопотоковых систем обслуживания с виртуальным разделением буфера // Кибернетика и системный анализ. – 2004. – № 6. – С. 168 – 172.

8. Wolff R.W. Poisson arrivals see time averages // Oper. Res. – 1992. – **30**, № 2. – P. 223 – 231.

9. Kelly F.P. Reversibility and stochastic networks. – London.: John Wiley & Sons, 1979. – 230 p.