УДК 519.872

А.З. Меликов, Л.А. Пономаренко, С.А. Багирова

АНАЛИЗ СИСТЕМ ОБСЛУЖИВАНИЯ-ЗАПАСАНИЯ С НЕТЕРПЕЛИВЫМИ РАСХОДУЮЩИМИ ЗАЯВКАМИ

Введение

Системы обеспечения материальными ресурсами одновременно обладают свойствами систем массового обслуживания (СМО) и систем управления запасами (СУЗ), так как в них канал обслуживания подключен к складу системы и уровень ресурсов на складе уменьшается лишь в моменты завершения обслуживания расходующих заявок (*p*-заявок). Поэтому при исследовании таких систем необходимо учитывать эти факторы. Вместе с тем анализ доступной литературы показал, что зачастую в известных моделях систем обеспечения материальными ресурсами они полностью не учитываются (см., например, [1–3] и списки литературы к ним). Как правило, в классических моделях СМО предполагается, что процесс обслуживания заявок не приводит к уменьшению ресурсов системы; иными словами, обычно предполагается, что СМО имеют неограниченные ресурсы, а возможные потери заявок связаны, в основном, с ограничением на число каналов или количество мест для ожидания в очереди, а также с нетерпеливостью заявок в очереди. С другой стороны, в классических моделях СУЗ обычно не учитываются возможности потери и/или образования очереди *p*-заявок даже при наличии необходимого количества запасов системы, т.е. в них предполагается, что система имеет неограниченное число каналов для отпуска требуемых ресурсов. Кроме того, в моделях СУЗ, как правило, предполагают, что уровень ресурсов системы уменьшается в моменты поступления *p*-заявок, а не в моменты завершения их обслуживания.

Следовательно, при построении единой модели системы обеспечения материальными ресурсами наряду с числом *p*-заявок в системе необходимо учитывать еще и уровень ее запасов (ресурсов). Такие системы в англоязычной литературе называются системами обслуживания-запасания (СОЗ, queueing-inventory systems) [4, 5]. Сразу же отметим, что подобные модели ранее были изучены в работах [6–9], где они назывались системами обслуживания со встречными потоками [6, 9], а также системами транспортно-складского типа [7, 8].

Отметим, что модели СОЗ, в которых учитываются указанные выше факторы, недостаточно исследованы. Исходя из этого, здесь рассматривается модель СОЗ при достаточно общих предположениях относительно политики пополнения запасов и поведения *p*-заявок.

Краткий обзор известных результатов

Исходя из хронологической последовательности, в первую очередь следует отметить работу [6]. В ней изучена модель СОЗ, в которой с постоянной скоростью (непрерывно) происходит убывание запасов и по пуассоновскому закону их пополнение порциями, имеющими известную функцию распределения (ф.р.). В указанной работе получены формулы для нахождения моментов достижения запасами нулевого или некоторого положительного уровня, времени пребывания системы на нулевом уровне запасов и т.д.

© А.З. МЕЛИКОВ, Л.А. ПОНОМАРЕНКО, С.А. БАГИРОВА, 2016

Отметим, что в данной работе рассматриваются модели СОЗ, в которых убывание запасов происходит не непрерывно, а лишь после завершения обслуживания очередной p-заявки. Подобные модели изучены в работах [7–9]. Так, в работе [7] изучена марковская модель СОЗ с ограниченной очередью терпеливых p-заявок, в которой при достижении уровнем запасов некоторой критической величины s, $0 \le s \le S-1$, где S — максимальный объем склада системы, система делает заказ на вышестоящий склад на поставку ресурсов (здесь и далее модель СОЗ будем называть марковской, если поток p-заявок пуассоновский, а случайные времена их обслуживания и выполнения заказов имеют экспоненциальные ф.р.). Там предполагалось, что заказ выполняется с некоторой случайной задержкой с известным (постоянным) средним. При этом с вероятностью $\alpha_s(i)$ поступают ресурсы объема i, $0 \le i \le S-s$, и допускается, что вероятности $\alpha_s(i)$ — управляемые параметры. Решена задача нахождения оптимальных значений этих вероятностей, при которых минимизируются суммарные затраты, связанные с доставкой и хранением ресурсов, ожиданием p-заявок в очереди и неудовлетворенным спросом p-заявок. Для этой цели использованы методы теории марковских процессов принятия решений.

Аналогичная модель изучена в [8], однако, в отличие от [7], в ней предполагалось, что вероятности $\alpha_s(i)$ постоянные, т.е. не являются управляемыми. Для описания работы системы построена соответствующая двумерная цепь Маркова (ЦМ) и найдено совместное распределение уровня ресурсов системы и числа p-заявок в ней. Здесь решена задача нахождения оптимального значения критического уровня ресурсов (s), при котором минимизируются суммарные штрафы из-за потери p-заявок, их ожидания в очереди, а также из-за хранения определенного уровня запасов в единицу времени.

В работе [9] изучена марковская модель СОЗ, в которой принимается, что p-заявки не являются идентичными, а имеют случайные размеры (здесь и в дальнейшем под размером p-заявки понимается объем требуемых ею ресурсов). В работе найдена оптимальная политика пополнения запасов, которая учитывает текущую ситуацию в системе, при этом ситуация определяется уровнем запасов на складе и количеством разнотипных p-заявок в системе.

Отметим, что в работах [7–9] предполагалось, что во время выгрузки ресурсов их отпуск по p-заявкам прекращается (в исследуемых в данной работе моделях это ограничение снимается).

В [10] изучена марковская модель одноканальной СОЗ с повторными и идентичными по размеру p-заявками и (s, S)-политикой (схемой) пополнения запасов. Последнее означает, что когда уровень ресурсов на складе системы становится меньшим или равным некоторому критическому уровню s, отправляется заказ на вышестоящий склад на поставку ресурсов объема $S-s$. Предполагается, что если в момент поступления некоторой p-заявки в системе отсутствуют необходимые ресурсы, то она не теряется, а поступает в орбит бесконечного размера; из орбита p-заявки повторяют попытку начать обслуживание после некоторого случайного времени, которое также имеет экспоненциальную ф.р. В указанной публикации в качестве математической модели исследуемой системы используется двумерная ЦМ и найдено условие эргодичности системы, а также найдены явные формулы для вычисления средних значений характеристик системы: среднего периода занятости системы, среднего времени ожидания в орбите, среднего уровня запасов системы и т.д. Кроме того, здесь также решена задача нахождения оптимальных значений параметров данной политики пополнения запасов относительно некоторого экономического (стоимостного) критерия. Эта же модель алгоритмическими методами изучена в [11]; она же в случае конечного размера орбита для абсолютно нетерпеливых p-заявок изучена в [12].

В работе [13] изучены три класса марковских моделей одноканальных СОЗ с (s, S)-политикой пополнения запасов. В модели типа I нет буфера для ожидания p-заявок, при этом если в момент поступления извне p-заявки в системе отсутствуют ресурсы или канал занят, то эта заявка либо с определенной вероятностью уходит в орбит, либо с дополнительной вероятностью окончательно покидает систему; если в момент поступления из орбита p-заявки в системе отсутствуют ресурсы или канал занят, то она опять либо с некоторой вероятностью возвращается в орбит, либо с дополнительной вероятностью окончательно уходит из системы. В моделях типа II и III имеются конечные буфера переменных размеров для ожидания p-заявок, при этом в модели типа II размер буфера равен текущему уровню запасов системы, а в модели типа III — максимальному размеру склада системы (т.е. S). При этом в моделях последнего типа потеря p-заявок или их поступление в орбит определяется состоянием буфера. В указанной работе для каждого типа моделей построены соответствующие трехмерные ЦМ и показано, что все они имеют трехдиагональную производящую матрицу. Исходя из этого, с использованием матрично-геометрического метода найдены их стационарные распределения, а также некоторые характеристики изучаемых моделей СОЗ.

В [14] изучена марковская модель одноканальной СОЗ с бесконечными очередями терпеливых p-заявок двух типов, при этом используется (s, S)-политика пополнения запасов; нагрузочные параметры входящих потоков, а также штрафы за ожидание в очереди одной заявки каждого потока за единицу времени являются различными. Найдено условие эргодичности изучаемой СОЗ, поставлена и решена задача нахождения оптимальной дисциплины обслуживания разнотипных p-заявок, где критерием служит дисконтированный средний штраф за ожидание в очереди на бесконечном горизонте планирования. Также найдено простое условие определения типа p-заявки, выбираемой для обслуживания: каждый раз из очереди необходимо выбирать заявку, для которой величина $\mu_i c_i$ наибольшая, где μ_i — интенсивность обслуживания p-заявки i-го типа и c_i — штраф за ожидание в очереди одной p-заявки i-го типа за единицу времени, $i = 1, 2$. Отметим, что это правило совпадает с известной оптимальной дисциплиной обслуживания в одноканальных СМО с бесконечной очередью. Данная работа также содержит обзор работ по этому направлению.

В [15] изучена марковская модель одноканальной СОЗ с конечной очередью терпеливых и высокоприоритетных p-заявок, где кроме этих заявок поступают и низкоприоритетные. Отличительная особенность низкоприоритетных заявок заключается в том, что их обслуживание не уменьшает уровень ресурсов системы. Обслуживание заявок последнего типа осуществляется лишь тогда, когда в моменты их поступления канал системы свободен; иначе они уходят в орбит конечного размера, при этом в орбите они становятся нетерпеливыми, т.е. после определенного времени пребывания в орбите они могут покидать систему. Найдено совместное распределение числа p-заявок в очереди, числа заявок в орбите и уровня запасов системы.

В [4, 5] получены формулы мультипликативного вида для вычисления совместного распределения длины очереди и уровня запасов в моделях СОЗ с различными схемами пополнения запасов, которые описываются марковскими моделями одноканальных СМО.

В [16, 17] изучены задачи нахождения оптимальной политики пополнения в марковских моделях СОЗ с конечной и бесконечной очередью соответственно.

Отметим, что в доступной литературе известны также работы, в которых исследованы модели СОЗ с мгновенной поставкой запасов. Однако они здесь

не анализируются, так как объектом исследования данной работы являются СОЗ с запаздывающей поставкой запасов. Модели СОЗ с мгновенной поставкой запасов изучены в работах [18, 19].

Из анализа литературы следует, что в известных моделях СОЗ, как правило, предполагают, что *p*-заявки в очереди абсолютно терпеливы, а время выполнения заказа не зависит от объема поставки. Вместе с тем в реальных СОЗ *p*-заявки в очереди являются нетерпеливыми, при этом вероятность их ухода из очереди зависит от текущего уровня запасов системы. Кроме того, параметр ф.р. времени выполнения заказа также зависит от объема поставки. Исходя из этого, в данной работе предложены новые модели, в которых учитываются указанные особенности реальных СОЗ и предложены методы их точного и асимптотического анализа.

Описание моделей СОЗ

Схема изучаемой одноканальной СОЗ с ограниченной очередью *p*-заявок показана на рис. 1. Система имеет склад ограниченного объема S. В эту систему поступает пуассоновский поток *p*-заявок с интенсивностью λ. Для простоты изложения предположим, что каждая *p*-заявка требует ресурса единичного размера. Время обслуживания *p*-заявок — случайная величина, имеющая экспоненциальную ф.р. с параметром μ. По завершении обслуживания *p*-заявки уровень ресурсов на складе системы уменьшается на единицу.

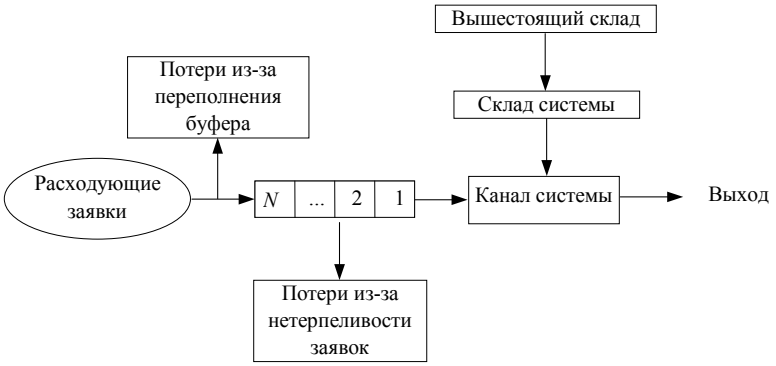

Рис. 1

Рассматривается два класса моделей СОЗ: системы с ограниченной и неограниченной очередью *p*-заявок. Предполагается, что в моменты поступления в систему *p*-заявки не имеют информации об уровне ресурсов на складе системы, т.е. в модели с ограниченной очередью максимальная длина очереди *p*-заявок может быть равна N; иными словами, если *p*-заявка поступила в момент, когда в очереди уже имеется N таких заявок, то независимо от уровня ресурсов на складе системы она теряется с вероятностью 1. Вместе с тем в модели с неограниченной очередью любая поступившая *p*-заявка независимо от уровня запасов системы принимается с вероятностью 1.

Нетерпеливость *p*-заявок проявляется в период ожидания в очереди: допустимые времена ожидания в очереди *p*-заявок, когда уровень запасов системы равен m, являются независимыми случайными величинами, имеющими экспоненциальные ф.р. со средними $\tau^{-1}(m)$. Иными словами, нетерпеливая *p*-заявка теряется из очереди, если до окончания допустимого интервала ожидания не освобождается канал обслуживания; при этом допустимое время пребывания в очереди зависит от текущего уровня запасов системы.

P-заявки не обслуживаются, если в системе отсутствуют ресурсы, т.е. отпуск ресурсов по *p*-заявкам продолжается, пока склад системы не станет пустым. Пополнения склада системы ресурсами осуществляются с помощью снабжающих заявок (*c*-заявки) согласно политике (s, S). Таким образом, когда уровень запасов системы становится меньшим некоторой пороговой (критический уровень запасов) величины s или равным ей, отправляется заказ на вышестоящий склад на поставку ресурсов объема $S - s$. При этом требуется, чтобы после выполнения заказа уровень ресурсов на складе системы был не меньше указанной величины s. Следовательно, для предотвращения случаев многократных заказов ресурсов необходимо выполнение соотношения $s < S/2$; иными словами, возможными значениями s являются числа $s = 0,1,...,\left[\frac{S}{2}\right]-1$, где $[a]$ — целая часть a.

Сделанный заказ выполняется с некоторой задержкой *c*-заявок, вызванной доставкой и выгрузкой ресурсов на склад данной системы, т.е. время выполнения заказа не может быть нулевым. Если принято, что критический уровень запасов равен s, то указанное время имеет экспоненциальную ф.р. с параметром $v(m)$, который в общем случае зависит от текущего уровня m ресурсов на складе системы, $m = 0, 1, ..., s$.

Обслуживания *c*- и *p*-заявок осуществляются на различных каналах, и эти процессы не зависят один от другого. Иными словами, допускается отпуск ресурсов *p*-заявкам во время их выгрузки на склад системы (т.е. во время обслуживания *c*-заявок).

Задача состоит в определении совместного распределения уровня запасов системы и длины очереди *p*-заявок. Решение этой задачи позволит определить также усредненные характеристики изучаемой СОЗ: средний уровень ресурсов на складе (Q_{av}), среднюю длину очереди *p*-заявок (L_{av}) и вероятность потери *p*-заявок probability of Blocking (PB).

Методы расчета характеристик СОЗ

Сначала рассмотрим модель СОЗ с ограниченной очередью. Функционирование данной СОЗ описывается двумерной ЦМ с состояниями вида (m, n), где m — уровень ресурсов на складе, n — число *p*-заявок в очереди. Эта цепь конечна, и ее фазовое пространство состояний (ФПС) определяется так:

$$E = \{(m, n) : m = 0, 1, ..., S;\ n = 0, 1, ..., N\}. \qquad (1)$$

Рассмотрим задачу определения элементов производящей матрицы (Q-матрицы) данной двумерной ЦМ. Интенсивность перехода из состояния $(m_1, n_1) \in E$ в состояние $(m_2, n_2) \in E$ обозначим $q((m_1, n_1), (m_2, n_2))$.

В общем случае переходы между состояниями ФПС (1) связаны со следующими событиями: (i) с поступлением *p*-заявок, (ii) с завершением обслуживания *p*-заявок, (iii) с уходом *p*-заявок из очереди из-за их нетерпеливости и (iv) с поступлением ресурсов из вышестоящего склада. Исходя из принятой политики пополнения запасов, необходимо различать случаи при определении исходного состояния $(m_1, n_1) \in E$: 1) $m_1 > s$; 2) $m_1 \le s$.

Сначала рассмотрим случай $m_1 > s$. Здесь выходы из данного состояния (m_1, n_1) из-за событий типа (iv) невозможны, так как в таких состояниях склад не может пополняться ресурсами. Если поступает некоторая *p*-заявка (события типа (i)), то она присоединяется к очереди при выполнении условия $n_1 < N$;

иными словами, осуществляется переход из данного состояния в состояние $(m_1,\ n_1+1)\in E$. Интенсивность такого перехода равна λ. Если в исходном состоянии выполняется условие $n_1 = N,$ то поступившая *p*-заявка теряется. По завершении обслуживания *p*-заявки (события типа (ii)) в исходном состоянии $(m_1, n_1)\in E,$ $n_1 > 0,$ осуществляется переход в состояние $(m_1-1,\ n_1-1)\in E$. Интенсивность такого перехода равна μ. Если некоторая *p*-заявка уходит из очереди необслуженной (события типа (iii)), то происходит переход из данного состояния в состояние $(m_1, n_1-1)\in E$. Интенсивность такого перехода равна $n_1\tau(m_1)$. Следовательно, для случаев $m_1 > s$ указанные выше элементы Q-матрицы определяются так:

$$q((m_1, n_1),\ (m_2, n_2)) = \begin{cases} \lambda, & \text{если } m_2 = m_1, \quad n_2 = n_1+1,\ n_1 \le N-1, \\ \mu, & \text{если } m_2 = m_1-1,\ n_2 = n_1-1,\ n_1 > 0, \\ n_1\tau(m_1), & \text{если } m_2 = m_1, \quad n_2 = n_1-1,\ n_1 > 0, \\ 0 & \text{в остальных случаях.} \end{cases} \tag{2}$$

Пусть теперь в исходном состоянии $(m_1, n_1)\in E$ выполняется условие $m_1 \le s$. В этом состоянии интенсивности переходов для указанных выше событий типа (i)–(iii) определяются аналогично соотношениям (2). Вместе с тем в момент поступления заказа из вышестоящего склада объемом $S-s$ происходит переход из этого состояния в состояние (m_1+S-s, n_1); интенсивность такого перехода равна $v(m_1)$. Следовательно, для случаев $m_1 \le s$ указанные выше элементы Q-матрицы определяются так:

$$q((m_1, n_1),\ (m_2, n_2)) = \begin{cases} \lambda, & \text{если } m_2 = m_1, \quad n_2 = n_1+1, & n_1 \le N-1, \\ \mu, & \text{если } m_2 = m_1-1,\ n_2 = n_1-1, & n_1 > 0, \\ n_1\tau(m_1), & \text{если } m_2 = m_1, \quad n_2 = n_1-1, & n_1 > 0, \\ v(m_1), & \text{если } m_1 \le s, & m_2 = m_1+S-s,\ n_2 = n_1, \\ 0 & \text{в остальных случаях.} \end{cases} \tag{3}$$

С учетом соотношений (2), (3) заключаем, что все состояния этой конечной двумерной ЦМ являются сообщающимися, следовательно, в этой системе существует стационарный режим.

Пусть $p(m, n)$ — стационарная вероятность состояния $(m, n)\in E$. Эти вероятности удовлетворяют системе уравнений равновесия (СУР), которая составляется на основе соотношений (2), (3). В матричной форме она записывается так:

$$\boldsymbol{p}Q = \mathbf{0}, \tag{4}$$

где $\boldsymbol{p}$ — вектор-строка $\boldsymbol{p} = (p(m, n) : (m, n)\in E)$ размерности $|E|$ ($|E|$ — количество элементов множества E), а $\mathbf{0}$ — нулевой вектор-столбец такой же размерности. К этой СУР добавляется условие нормировки:

$$\sum_{(m,\,n)\in E} p(m, n) = 1. \tag{5}$$

После нахождения совместного распределения уровня ресурсов на складе системы и длины очереди *p*-заявок можно вычислить усредненные характеристики

исследуемой СОЗ. Так, средняя длина очереди p-заявок (L_{av}) и средний уровень ресурсов на складе (S_{av}) определяются как математические ожидания соответствующих случайных величин. Иными словами, эти параметры определяются из следующих формул:

$$L_{av} = \sum_{n=1}^{N} n \sum_{m=0}^{S} p(m, n); \tag{6}$$

$$S_{av} = \sum_{m=1}^{S} m \sum_{n=0}^{N} p(m, n). \tag{7}$$

Потери p-заявок происходят либо из-за переполненности буфера для ожидания p-заявок, либо из-за их нетерпеливости. Иными словами, p-заявки теряются в следующих случаях: в момент поступления такой заявки в очереди отсутствует свободное место либо до окончания допустимого интервала ожидания не освобождается канал обслуживания. Следовательно, для вычисления вероятности потери p-заявок (PB) получим следующую формулу:

$$PB = \sum_{m=0}^{S} p(m, N) + \sum_{m=0}^{S} \sum_{n=1}^{N} p(m, n) P_l(m, n), \tag{8}$$

где $P_l(m, n)$ означает вероятность того, что в состоянии (m, n) p-заявка теряется из-за нетерпеливости. Эта величина определяется так:

$$P_l(m, n) = \frac{n\tau(m)}{n\tau(m) + \mu I(m \geq 1) + \lambda I(n < N)}, \tag{9}$$

где $I(A)$ — индикаторная функция события A.

Относительно решения СУР (4), (5), отметим, что, к сожалению, из-за сложной структуры ее матрицы не удается найти ее аналитическое решение. Более того, изучаемая двумерная ЦМ не является хотя бы квазипроцессом размножения и гибели, для расчета которой существуют эффективные численные методы [20, 21] (напомним, что двумерная ЦМ называется квазипроцессом размножения и гибели, если Q-матрица одномерной цепи, которая получается в результате соответствующей перенумерации состояний исходной двумерной цепи, трехдиагональная). Поэтому для ее расчета приходится использовать известные (стандартные) численные методы теории марковских процессов [22, 23]. Однако известные методы работоспособны лишь для моделей малой и умеренной размерности и бесполезны для моделей большой и сверхбольшой размерности. Заметим также, что реальные СОЗ являются именно большими системами.

Здесь используется приближенный метод для расчета стационарного распределения двумерных ЦМ [24], позволяющий осуществить асимптотический анализ характеристик данной модели СОЗ при больших размерностях склада системы и объема буферного накопителя для ожидания p-заявок. При этом принимается, что СОЗ работает в условиях большой нагрузки, т.е. интенсивность поступления p-заявок намного превосходит интенсивность их обслуживания. Ранее этот метод успешно применялся для анализа различных моделей СМО [24–27].

При выполнении этого допущения рассмотрим следующее расщепление ФПС (1):

$$E = \bigcup_{m=0}^{S} E_m, \;\; E_{m_1} \bigcap E_{m_2} = 0, \;\; m_1 \neq m_2, \tag{10}$$

где $E_m = \{(m, n) \in E : n = 0, 1, ..., N\}$.

Расщепление (10) означает, что класс состояний E_m содержит те состояния (m, n) из исходного ФПС (1), в которых уровень ресурсов равен m независимо от длины очереди p-заявок. Далее в исходном ФПС (1) определяется следующая функция укрупнения:

$$U(m, n) = \langle m \rangle, \tag{11}$$

где $\langle m \rangle$ — укрупненное состояние, которое объединяет в себе класс состояний E_m, $m = 0, 1, ..., S$. Обозначим $\Omega = \{\langle m \rangle : m = 0, 1, ..., S\}$.

Стационарное распределение исходной модели приближенно определяется следующим образом (см. [24]):

$$p(m, n) \approx \rho_m(n)\,\pi(\langle m \rangle), \tag{12}$$

где $\rho_m(n)$ — вероятность состояния (m, n) внутри расщепленной модели с пространством состояний E_m, а $\pi(\langle m \rangle)$ — вероятность укрупненного состояния $\langle m \rangle \in \Omega$.

Из расщепления (10) видно, что во всех состояниях внутри расщепленной модели с ФПС E_m первая компонента постоянная, и потому все состояния из этого класса определяются лишь второй компонентой. При этом из соотношений (2), (3) получаем, что стационарные вероятности состояний внутри расщепленной модели с ФПС E_m вычисляются как вероятности состояний классической модели Эрланга *M/M/N*/0 с нагрузкой $\sigma_m = \lambda / \tau(m)$, т.е.

$$\rho_m(n) = \frac{\sigma_m^n}{n!}\rho_m(0), \quad m = 0, 1, ..., N, \tag{13}$$

где $\rho_m(0) = \left(\sum_{n=0}^{N} \frac{\sigma_m^n}{n!} \right)^{-1}$.

Интенсивность перехода из укрупненного состояния $\langle m_1 \rangle$ в другое укрупненное состояние $\langle m_2 \rangle$ обозначим $q(\langle m_1 \rangle, \langle m_2 \rangle)$, $\langle m_1 \rangle, \langle m_2 \rangle \in \Omega$. С учетом (2), (3) и (13) после определенных математических преобразований получаем граф переходов между состояниями укрупненной модели (рис. 2):

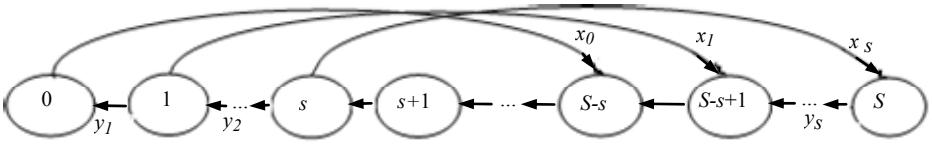

Рис. 2

$$q(\langle m_1 \rangle, \langle m_2 \rangle) = \begin{cases} x_{m_1}, & \text{если } 0 \le m_1 \le s, \quad m_2 = m_1 + S - s, \\ y_{m_1}, & \text{если } 1 \le m_1 \le S, \quad m_2 = m_1 - 1, \\ 0 & \text{в остальных случаях,} \end{cases} \tag{14}$$

где $x_{m_1} = v(m_1)$, $y_{m_1} = \mu(1 - \rho_{m_1}(0))$.

Из соотношений (14) удается выразить все вероятности состояний через вероятность $\pi(s+1)$ следующим образом (здесь промежуточные преобразования также опускаются):

$$\pi(< m >) = \begin{cases} \alpha_m \pi(\langle s+1 \rangle), & \text{если } 0 \le m \le s, \\ \beta_m \pi(\langle s+1 \rangle), & \text{если } s+1 \le m \le S-s, \\ \gamma_m \pi(\langle s+1 \rangle), & \text{если } S-s+1 \le m \le S. \end{cases} \tag{15}$$

Здесь и далее приняты следующие обозначения:

$$\alpha_m = \prod_{i=m+1}^{s+1} \frac{y_i}{x_{i-1} + y_{i-1}},\ y_0 := 0;\ \beta_m = \frac{y_{s+1}}{y_m};\ \gamma_m = \frac{1}{y_m} \sum_{i=m-S+s}^{s} \alpha_i x_i . \quad (16)$$

Вероятность $\pi(s+1)$ находится из условия нормировки, т.е.

$$\pi(s+1) = \left(\sum_{m=0}^{s} \alpha_m + \sum_{m=s+1}^{S-s} \beta_m + \sum_{m=S-s+1}^{S} \gamma_m \right)^{-1}. \quad (17)$$

Тогда с учетом соотношений (12)–(17) находится совместное распределение уровня ресурсов на складе системы и длины очереди *p*-заявок. Далее с использованием (6)–(8) получаем следующие формулы для приближенного расчета характеристик системы обслуживания-запасания с ограниченной очередью:

$$L_{av} = \sum_{n=1}^{N} n \sum_{m=0}^{S} \rho_m(n)\, \pi(\langle m \rangle); \quad (18)$$

$$S_{av} = \sum_{m=1}^{S} m\, \pi(\langle m \rangle); \quad (19)$$

$$PB = \sum_{m=0}^{S} \rho_m(N)\, \pi(\langle m \rangle) + \sum_{m=0}^{S} \sum_{n=1}^{N} \rho_m(n)\, \pi(\langle m \rangle) P_l(m, n). \quad (20)$$

Теперь рассмотрим модель СОЗ с неограниченной очередью. Функционирование данной системы также описывается двумерной, в данном случае бесконечной ЦМ с состояниями вида (m, n), где m — уровень ресурсов в складе, n — число *p*-заявок в очереди, т.е. ФПС этой модели бесконечномерно и задается так:

$$E = \{(m, n) : m = 0, 1, ..., S;\ n = 0, 1, ...\}. \quad (21)$$

Замечание 1. Здесь и далее в целях упрощения изложения для обоих типов моделей используются одинаковые обозначения для ФПС, стационарных распределений и характеристик системы. Однако из контекста ясно, о каких именно моделях будет идти речь.

Элементы Q-матрицы данной двумерной ЦМ определяются аналогично (2) и (3). Средняя длина очереди *p*-заявок и средний уровень ресурсов на складе также определяются аналогично (6) и (7), но при этом следует иметь в виду, что в указанных формулах параметр N принимается равным бесконечности, т.е. в них необходимо положить $N = \infty$. Относительно определения вероятности потери *p*-заявок отметим, что в данной модели отсутствует первое слагаемое в формуле (10), так как очередь для *p*-заявок имеет бесконечную длину. Следовательно, в данной модели для вычисления вероятности потери *p*-заявок получим следующую формулу:

$$PB = \sum_{m=0}^{s} \sum_{n=1}^{\infty} p(m, n)\, P_l(m, n). \quad (22)$$

Замечание 2. Здесь следует иметь в виду, что при определении функции $P_l(m, n)$ в правой части формулы (9) $I(n < N) = 1$ для любого $n = 1, 2, ...$.

В данной модели нахождение стационарного распределения соответствующей бесконечномерной ЦМ становится еще более сложной проблемой, чем для модели с ограниченной очередью. Здесь для этой цели не удается использовать соответствующую СУР для стационарных вероятностей состояний, поскольку невозможно найти аналитические выражения для их вычисления или разработать другие эффективные численные процедуры. Использование для этой цели метода двумерных производящих функций сопровождается известными вычислительными и методологическими трудностями. Поэтому для нахождения стационарного

распределения этой бесконечномерной ЦМ используется достаточно подробно описанный выше приближенный метод. Специфика его применения для данной модели в краткой форме излагается ниже.

Рассматривается аналогичное (10) расщепление ФПС (21) и соответствующим образом строится функция укрупнения (11).

В данном случае стационарные вероятности состояний внутри расщепленной модели с ФПС E_m вычисляются как вероятности состояний модели $M/M/\infty$ с нагрузкой σ_m, т.е.

$$\rho_m(n) = \frac{\sigma_m^n}{n!} e^{-\sigma_m}, \quad n = 0, 1, 2, \ldots \ . \tag{23}$$

Интенсивности переходов между состояниями укрупненной модели определяются аналогично (14), но следует иметь в виду, что в этом случае $y_{m_1} = \mu(1 - e^{-\sigma_{m1}})$.

Далее вероятности состояний укрупненной модели вычисляются в соответствии с формулами (15)–(17).

Средний уровень ресурсов также определяется с помощью формулы (19). С учетом (23) из (18) при $N = \infty$ получим следующую простую формулу для вычисления средней длины очереди в модели СОЗ с неограниченной очередью:

$$L_{av} = \sum_{m=0}^{S} \pi(\langle m \rangle)\sigma_m . \tag{24}$$

Вероятность потери p-заявок из-за нетерпеливости в данной модели определяется так:

$$PB = \sum_{m=0}^{S} e^{-\sigma_m} \pi\big(\langle m \rangle\big) \sum_{n=1}^{\infty} \frac{\sigma_m^n}{n!} P_l(m, n) . \tag{25}$$

Численные результаты

Разработанные алгоритмы позволяют изучить поведение характеристик (6)–(8) исследуемых моделей СОЗ относительно изменения как структурных параметров системы (т.е. S и N), так и критического уровня запасов (т.е. s), при котором делается заказ на поставку ресурсов. Для конкретности изложения предположим, что нагрузочные параметры p-заявок (т.е. λ и μ) фиксированны, а критический уровень запасов поддается управлению. Исходя из последних допущений, здесь изучается поведение характеристик (6)–(8) относительно изменения критического уровня запасов.

Следует ожидать, что функции $\tau(m),\ m = 0, 1, \ldots, S,$ и $v(m),\ m = 0, 1, \ldots, s,$ убывающие. Действительно, логично предположить, что если p-заявки имеют информацию о текущем уровне запасов системы, то с уменьшением этого уровня интенсивность потери p-заявок из-за нетерпеливости будет расти, а интенсивность пополнения запасами должна увеличиваться (т.е. чем меньше запасов системы, тем быстрее они должны пополняться). Вместе с тем для общности анализа здесь рассматривается два варианта изменения функций $\tau(m),\ m = 0, 1, \ldots, S,$ а именно, в первом (оптимистическом) варианте предполагается, что эти функции убывающие, а во втором (пессимистическом) они возрастают относительно уровня запасов. Для конкретности изложения принимается, что в первом варианте $\tau(m) = 1/(m+1)$, а во втором $\tau(m) = (m+1)/(m+2)$, $m = 0, 1, \ldots, S$. Для простоты изложения принимается, что во всех вариантах $v(m) = 1$ для любого $m = 0, 1, \ldots, s$.

Исходные данные гипотетической модели с ограниченной очередью *p*-заявок выбирались таким образом: $S = 50,\ N = 100,\ \lambda = 5,\ \mu = 0{,}5$. Приведенный ниже анализ основан исключительно на этих данных.

Результаты выполненных численных экспериментов показаны на рис. 3–5, где *o* — $\tau(m) = (m+1)/(m+2)$; *x* — $\tau(m) = 1/(m+1)$. На рис. 3 отображена зависимость вероятности потери *p*-заявок от параметра *s* в модели с ограниченной очередью, на рис. 4 — зависимость среднего числа *p*-заявок в очереди от параметра *s* в модели с ограниченной очередью, на рис. 5 — зависимость среднего уровня запасов от параметра *s* в модели с ограниченной очередью. Из рис. 3 видно, что в оптимистическом варианте изменения функций $\tau(m),\ m = 0, 1, ..., S,$ вероятность потери *p*-заявок больше, чем в пессимистическом варианте. На первый взгляд эти результаты кажутся нелогичными, так как в оптимистическом варианте *p*-заявки покидают очередь с меньшей интенсивностью, чем в пессимистическом варианте, и потому следовало ожидать, что в первом варианте вероятность потери *p*-заявок также должна быть меньше по сравнению со вторым вариантом. Однако в первом варианте *p*-заявки покидают очередь с меньшей интенсивностью, что приводит к увеличению длины очереди, и тем самым растет вероятность потери *p*-заявок из-за переполнения буфера; во втором варианте *p*-заявки покидают очередь с большой интенсивностью, что приводит к уменьшению длины очереди, и тем самым уменьшаются потери *p*-заявок из-за переполнения буфера. Поскольку вероятность потери *p*-заявок определяется как сумма двух величин, определяющих вероятности потери от переполненности буфера для ожидания p-заявок и от их нетерпеливости (см. формулы (8)), то конечное значение функции *PB* определяется скоростями изменения этих составляющих. Так, для выбранных исходных данных вероятность потери из-за переполнения буфера в пессимистическом варианте практически равна нулю, а в оптимистическом эта величина плавно меняется в пределах 0,042 и 0,2; вместе с тем вероятность потери из-за нетерпеливости *p*-заявок в пессимистическом варианте практически постоянна и приближенно равна 0,4, а в оптимистическом варианте эта величина плавно меняется в пределах 0,3198 и 0,4225. В результате этого графики функции *PB* имеют вид, показанный на рис. 3. Отметим, что в обоих вариантах функция *PB* изменяется с очень малыми скоростями, а именно, в пессимистическом варианте она почти постоянна относительно изменения критического уровня ресурсов (приблизительно равна 0,4085), в оптимистическом же варианте ее минимальное значение равно 0,4645 (достигается при $s = 0$), а максимальное — 0,5184 (достигается при s = 7). Иными словами, значения функции *PB* в различных вариантах изменения функций $\tau(m),\ m = 0, 1, ..., S,$ почти не отличаются.

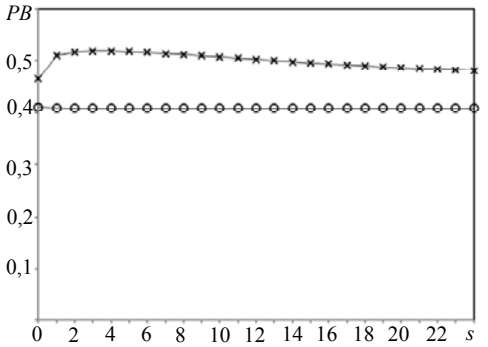

Рис. 3

Вместе с тем значения функции L_{av} в различных вариантах изменения функций $\tau(m),\ m = 0, 1, ..., S,$ существенно отличаются (рис. 4). При этом в оптимистическом варианте ее значения становятся заметно большими, чем при использовании пессимистического варианта для поведения *p*-заявок в очереди. Этого

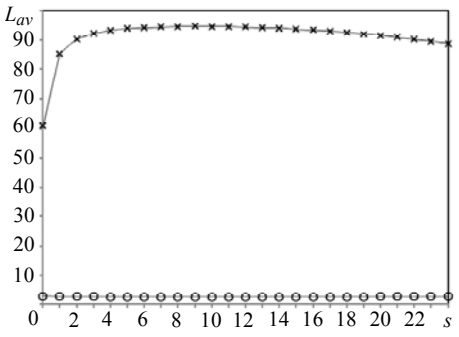

Рис. 4

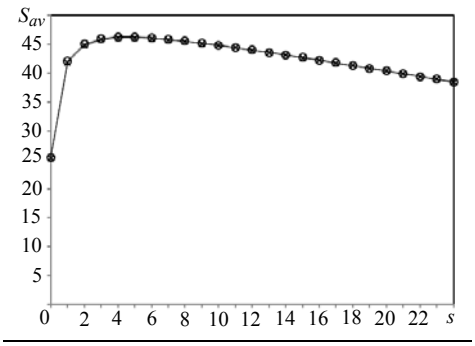

Рис. 5

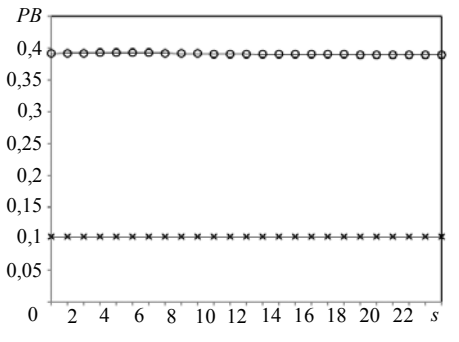

Рис. 6

следовало ожидать, так как в оптимистическом варианте *p*-заявки покидают очередь с меньшей интенсивностью по сравнению с пессимистическим вариантом. Отметим, что максимальная средняя длина очереди (примерно 95 заявок) в оптимистическом варианте наблюдается при $s = 13$, а минимальная средняя длина очереди (примерно 61 заявка) в этом варианте имеет место при $s = 0$; в пессимистическом варианте эта функция почти постоянна и ее значение примерно равно 3.

Из рис. 5 видно, что средний уровень ресурсов в системе почти не зависит от характера поведения *p*-заявок в очереди, так как в обоих вариантах его значения почти совпадают. При этом интересным является резкое увеличение значения этой функции при начальных значениях параметра *s*; так, при $s = 0$ имеем $S_{av} = 25$, а уже при $s = 1$ значение этой функции $S_{av} = 41$. Далее эта функция растет медленно, достигает максимального значения 46,13 в точке $s = 5$ и потом плавно убывает.

Теперь рассмотрим результаты численных экспериментов для той же гипотетической модели с неограниченной очередью *p*-заявок, т.е. принимается, что $S = 50$, $N = \infty$, $\lambda = 5$, $\mu = 0{,}5$.

Из рис. 6, где показана зависимость вероятности потери *p*-заявок от параметра *s* в модели с неограниченной очередью; *o* — $\tau\ (m) = (m+1)/(m+2)$; *x* — $\tau = 1/(m+1)$, видно, что вероятности потери *p*-заявок в обоих вариантах почти постоянны. При этом, в отличие от модели с ограниченной очередью, здесь в оптимистическом варианте изменения функций $\tau(m)$, $m = 0, 1, ..., S$, вероятности потери *p*-заявок оказываются существенно меньшими, чем в пессимистическом варианте. Этого также следовало ожидать, так как в этой модели потери *p*-заявок происходят только из-за их нетерпеливости, и поэтому с ростом терпеливости *p*-заявок уменьшается вероятность их потери из очереди.

Замечание 3. В этом случае для расчета вероятности потери *p*-заявок не удается получить удобную для вычисления формулу (формулу вида (25)), так как сюда входит бесконечная сумма. Для выхода из такой ситуации здесь использован метод отсечения хвоста распределения, т.е. верхняя (бесконечная) граница суммы заменяется достаточно большой (конечной) величиной, далее она постепенно увеличивается, и эта процедура продолжается до тех пор, пока значения функции PB практически перестают изменяться.

Интересно отметить, что поведение функции L_{av} в данной модели идентично ее поведению в модели с ограниченной очередью, при этом ее значения в пессимистическом случае почти совпадают с ее значениями в модели с ограниченной очередью (рис. 7, где показана зависимость среднего числа p-заявок в очереди от параметра s в модели с ограниченной очередью; o — $\tau\,(m) = (m+1)/(m+2)$; x — $\tau\,(m) = 1/(m+1)$); максимальная средняя длина очереди (примерно 75 заявок) в оптимистическом варианте наблюдается при $s = 16$, а минимальная средняя длина очереди (примерно 52 заявки) имеет место при $s = 0$.

Как и в случае модели с ограниченной очередью, в данной модели средний уровень ресурсов в системе почти не зависит от характера поведения p-заявок в очереди (рис. 8, где отображена зависимость среднего уровня запасов от параметра s в модели с неограниченной очередью; o — $\tau\,(m) = (m+1)/(m+2)$; x — $\tau\,(m) = 1/(m+1)$). При этом здесь значения функции S_{av} при начальных значениях параметра s растут достаточно плавно и достигают максимального значения $S_{av} = 36{,}51$ в точке $s = 12$; далее эта функция медленно убывает.

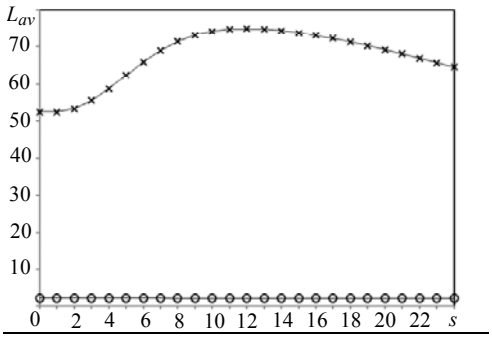

Рис. 7

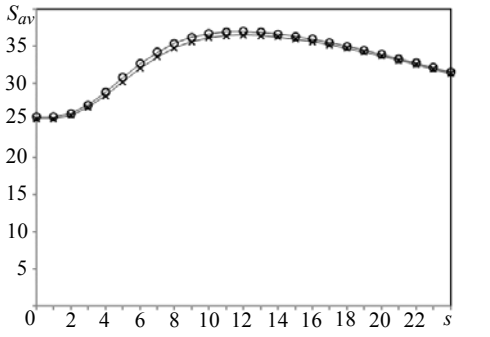

Рис. 8

Численные эксперименты показали, что характеристики изучаемой модели СОЗ существенным образом зависят от значений ее структурных (объем склада системы и размер буфера для ожидания p-заявок в очереди) и нагрузочных (интенсивности поступления и обслуживания p-заявок и задержки выполнения заказа) параметров. В каждом конкретном случае следует проводить вычисления на основе предложенных соответствующих алгоритмов.

Заключение

В данной работе предложены новые модели систем обслуживания-запасания с нетерпеливыми расходующими заявками, которые могут образовывать очередь конечной или бесконечной длины. В них интенсивность потери расходующих заявок из очереди из-за нетерпеливости, а также время выполнения заказа зависят от текущего уровня запасов системы. Политика пополнения запасов относится к классу политик двух уровней. Разработаны точный и приближенный методы для определения их характеристик. Точный метод эффективен для систем с умеренными значениями объема склада системы и длины очереди расходующих заявок. Вместе с тем в реальных системах указанные величины принимают достаточно большие значения, и в связи с этим размерности изучаемых моделей чрезмерно большие. В результате точное вычисление стационарных распределений соответствующих двумерных ЦМ оказывается сложной проблемой. Поэтому здесь предложен приближенный метод для асимптотического анализа характеристик изучаемых моделей СОЗ при высоких нагрузках расходующих заявок. Он основан на алгоритмах фазового укрупнения состояний двумерных ЦМ.

Предложенные формулы позволяют проанализировать характеристики предложенных моделей СОЗ любой размерности, а также решить практически важные задачи по оптимизации их характеристик. Последние задачи являются предметами специальных исследований.

А.З. Меліков, Л.А. Пономаренко, С.А. Багірова

АНАЛІЗ СИСТЕМ ОБСЛУГОВУВАННЯ-ЗАПАСАННЯ З НЕТЕРПЛЯЧИМИ ВИТРАЧАЛЬНИМИ ВИМОГАМИ

Запропоновано нові моделі систем обслуговування-запасання з обмеженими і необмеженими чергами нетерплячих витрачальних вимог. У них рівень ресурсів системи зменшується лише в моменти завершення обслуговування витрачальних вимог. Розроблено точний та наближений методи розрахунку характеристик запропонованих моделей. Наведено результати числових експериментів.

A.Z. Melikov, L.A. Ponomarenko, S.A. Bagirova

ANALYSIS OF QUEUEING-INVENTORY SYSTEMS WITH IMPATIENT CONSUME CUSTOMERS

New models of queueing-inventory systems with finite and infinite queues of impatient consume customers are proposed. Here level of inventory is reduced after servicing of customers only. Both exact and approximate methods to calculate the characteristics of the proposed models are developed. The results of numerical experiments are shown.

1. *Прабху Н.* Методы теории массового обслуживания и управления запасами (изучение основных случайных процессов). — М. : Машиностроение, 1969. — 356 с.
2. *Прабху Н.* Стохастические процессы теории запасов. — М. : Мир, 1984. — 184 с.
3. *Математические* модели управления запасами / В.А. Шуенкин, В.С. Донченко, С.Н. Константинов, В.Ю. Шапировский. — Киев : ООО «Международное финансовое агентство», 1997. — 302 с.
4. *Schwarz M., Daduna H.* Queuing systems with inventory management with random lead times and with backordering // Mathematical Methods of Operations Research. — 2006. — **64**, N 3. — P. 383–414.
5. *Schwarz M., Sauer C., Daduna H., Kulik R., Szekli R.* M/M/1 queuing systems with inventory // Queuing systems. Theory and applications. — 2006. — **54**, N 1. — P. 55–78.
6. *Постан М.Я.* Применение марковских процессов для моделирования систем обслуживания встречных транспортных потоков. — Киев, 1989. — 14 с. (Препр. / НАН Украины. Ин-т кибернетики им. В.М. Глушкова; 89–6).
7. *Melikov A.Z., Molchanov A.A.* Stock optimization in transport/storage systems // Cybernetics and Systems Analysis. — 1992. — **28**, N 3. — P. 484–487.
8. *Меликов А.З.* Марковская модель процесса накопления в системах транспортно-складского типа // Электронное моделирование. — 1996. — **18**, № 6. — С. 79–83.
9. *Меликов А.З., Фаталиева М.Р.* Управление запасами систем обслуживания разнотипных встречных потоков с учетом текущей ситуации // Там же. — 1997. — **19**, № 6. — С. 106–115.
10. *Ushakumari P.V.* On (*s*, *S*) inventory system with random lead time and repeated demands // Journal of Applied Mathematics and Stochastic Analysis. — 2006. — Article ID 81508. — 22 p.
11. *Artalejo J.R., Krishnamoorthy A., Lopez-Herrero M.J.* Numerical analysis of (*s*, *S*) inventory system with repeated attempts // Annals of Operations Research. — 2006. — **141**. — P. 67–83.

12. *Lopez-Herrero M.J.* Waiting time and other first-pasage time measures in an (*s*, *S*) inventory system with repeated attempts and finite retrial group // Computers & Operations Research. — 2010. — **37**. — P. 1256–1261.
13. *Krishnamoorthy A., Jose K.P.* Comparision of inventory systems with service, positive lead-time, loss, and retrial of customers // Journal of Applied Mathematics and Stochastic Analysis. — 2007. — Article ID 37848. — 23 p.
14. *Zhao N., Lian Z.* A queueing-inventory system with two classes of customers // Journal of Production Economics. — 2011. — **129**. — P. 225–231.
15. *Yadavalli V.S.S., Anbazhaga N., Jeganathan K.* A retrial inventory system with impatient customers // Applied Mathematics and Information Science. — 2015. — **9**, N 2. — P. 637–650.
16. *Berman O., Supna K.P.* Optimal control of service for facilities holding inventory // Computers & Operations Research. — 2001. — **28**, N 3. — P. 429–441.
17. *Berman O., Kim E.* Dynamic inventory strategies for profit maximization in a service facilities with stochastic service, demand and lead time // Mathematical Methods of Operations Research. — 2004. — **60**, N 3. — P. 497–521.
18. *Berman O., Kim E.* Stochastic models for inventory management at service facilities // Stochastic Models. — 1999. — **15**, N 4. — P. 695–718.
19. *Berman O., Supna K.P.* Inventory management at service facilities for systems with arbitrary distributed service times // Ibid. — 2000. — **16**, N 3-4. — P. 343–360.
20. *Servi L.D.* Algorithmic solutions to two-dimensional birth-death processes with applications to capacity planning // Telecommunication Systems. — 2001. — **21**, N 2–4. — P.205–212.
21. *Baumann H., Sandmann W.* Numerical solution of level dependent quasi-birth-and-death processes // Proceeding Computer Science. — 2010. — **1**. — P. 1555–1563.
22. *Philippe B., Saad Y., Stewart W.J.* Numerical methods in Markov chains modelling // Operations Research. — 1992. — **40**, N 6. — P. 1156–1179.
23. *Stewart W.J.* Introduction to the numerical solution of Markov chains. — Princeton: University Press, 1994. — 539 p.
24. *Ponomarenko L., Kim C.S., Melikov A.* Performance analysis and optimization of multi-traffic on communication networks. — Heidelberg; Dortrecht; London; New York: Springer, 2010. — 208 p.
25. *Liang C., Luh H.* Cost estimation queuing model for large-scale file delivery service // International Journal of Electronic Commerce Studies. — 2011. — **2**, N 1. — P. 19–34.
26. *Liang C., Luh H.* Optimal services for content delivery based on business priority // Journal of the Chinese Institute of Engineers. — 2013. — **36**, N 4. — P. 422–440.
27. *Liang C., Luh H.* Efficient method for solving a two-dimensional Markov chain model for call centers // Industrial Management & Data Systems. — 2015. — **115**, N 5. — P. 901–922.

Получено 27.10.2015

Статья представлена к публикации членом редколлегии чл.-корр. НАН Украины Чикрием А.А.