Реферат по статье “Система обслуживания с ветвящимися потоками вторичных требований.”

Предмет: Теория случайных процессов

Александр Сергеевич Баклашов

Содержание

1 Введение

При проектировании различных систем — от автоматических систем управления (АСУ) до вычислительных систем, значительную роль играют приоритетные модели теории массового обслуживания. Среди таких систем можно выделить системы с различными динамическими ориентациями и режимами переключения, которые выделяются в контексте исследования моделей, включающих в себя несколько типов требований. В таких моделях разнообразие типов требований подразумевает необходимость применения соответствующих системных ориентаций и режимов переключения при их изменении.

Помимо этого, в рассматриваемых системах стоит задача оптимизации приоритетного обслуживания, заключающаяся в разработке оптимальных стратегий обслуживания, которые могли бы обеспечить эффективное управление различными типами требований в системе.

В данной работе рассматривается задача определения оптимальной дисциплины обслуживания в системе с несколькими типами требований, имеющих различные приоритеты и поступающих как извне (первичных), так и в результате обслуживания (вторичных) и ветвящимися потоком вторичных требований. Модели такого типа становятся актуальными при исследовании работы ЭВМ в различных режимах, исследовании информационно-поисковых и других различных систем.

Приведем конкретные примеры, где такие системы становятся значимыми. Предположим, мы рассматриваем работу компьютера в пакетном режиме. В этом случае после начальной обработки пакета данных определяется количество программ в нем. Каждая программа может запросить разнообразные ресурсы, такие как вызов стандартных программ, доступ к оперативной памяти, или обращение к внешним устройствам для получения дополнительной информации.

Еще один пример связан с задачей поиска информации в массивах данных. После анализа некоторого массива можно обнаружить, что необходимая инфор-

мация на самом деле содержится в одном из других массивов.

Эти примеры демонстрируют, насколько важным является эффективное управление разнообразными запросами и ресурсами в системах с ветвящимися потоками требований, исследование и оптимизация которых имеют непосредственное прикладное значение. Организация работы таких систем также включает оперативное определение приоритетов обслуживания требований. Данная работа демонстрирует, что относительно линейного функционала потерь оптимальной является именно приоритетная дисциплина, для которой приводится алгоритм её построения.

2 Постановка задачи. Предположения

В работе “Система обслуживания с ветвящимися потоками вторичных требований” рассматривается однолинейная система обслуживания (система обслуживания (СО), в которой все поступающие заявки обслуживаются на одном приборе), выполняющая r типов операций. Длительности выполнения отдельных операций являются независимыми случайными величинами (СВ) с функциями распределения (ФР) $\beta_i(\cdot)$, при этом $\beta_i(0) = 0$, имеющими первые два момента b_i, b_{i2}. Первичные требования на выполнение операции типа i образуют пуассоновский поток с интенсивностью λ_i, $\lambda_i \geq 0$, $\overline{1,r}$. Помимо этого, имеют место случаи, в которых на некоторые, но не на все операции первичных требований не поступает. В результате выполнения операции типа i вызвавшее ее требование считается обслуженным, но с вероятностью $q_i(n) = q_i(n_1, ..., n_r)$ возникает набор $n = (n_i, ..., n_r)$ вторичных требований на выполнение операций различных типов, которые мгновенно поступают в очереди (неограниченные) для требований соответствующего типа.

Непосредственно вслед за этим по набору $l = (l_1, ..., l_r)$, где l_i — число требований в i- й очереди, с помощью функции управления $u(l)$ выбирается очередное требование на обсуживание. При этом, если $u(l) = i$, то будет обслуживаться требование типа i.

Предполагается, что простои прибора при наличии требований не допускаются, т.е. любое требование, заставшее прибор свободным, немедленно начинает обслуживаться и, если $u(l) = i$, то $l_i > 0$ (начинает обслуживаться требование типа i). В момент, когда система освобождается и нельзя направить требование на обслуживание, мы доопределяем $u(0) = 0$.

Для описания поведения системы введём процесс $L(t) = (L_i(t), ..., L_i(t))$, где $L_i(t)$ — число требований типа i в момент t.

Далее зададим c_i — стоимость единицы времени пребывания в системе требования типа r. Задача заключается в определении функции управления, минимизирующей потери в единицу времени в стационарном режиме работы системы.

При сделанных далее предположениях функционал, определяющий эти потери, записывается в виде

$$J = \sum_{i=1}^{r} c_i L_i \tag{2.1}$$

где L_i — среднее число требований типа i в системе в стационарном режиме.

Обозначим через $Q_i(z) = Q_i(z_1..., z_i)$ производящую функцию (ПФ) числа вторичных требований, возникающих в результате выполнения операции типа i:

$$Q_i(z) = \sum_{n \geq 0} q_i(n) z^n \equiv \sum_{n_1 \geq 0, \ldots, n_r \geq 0} q_i(n_1, ..., n_r) z_1^{n_1} ... z_r^{n_r}$$

2.1 Вывод формулы (2.1)

Средние потери за единицу времени на интервале $[0, T]$ равны

$$E\{\frac{1}{T} \int_0^T \sum_{i=1}^{r} c_i L_i(t) dt\}$$

а функционал J является пределом этого выражения при $T \to \infty$ который в силу эргодичности процесса $L(t)$ существует и не зависит от начального состояния, но, возможно, принимает бесконечное значение.

Пусть $L_i = \lim E\{L_i(t)\}$. Покажем, что $L_i < \infty$. Действительно, величина $V = L_1 b_1 + ... + L_r b_r$, которая имеет смысл стационарного среднего времени разгрузки системы, и величина $L = (L_i, ..., L_r)$ конечны или бесконечны одновременно. Но по своему определению V не превосходит среднего времени до окончания периода регенерации„ которое, как известно из теории восстановления [1], равно $\theta_2 / 2\theta_1$. Такие же рассуждения устанавливают конечность величин x_{ij}.

При вычислении предела удобно в качестве начального распределения процесса $L(t)$ выбрать его стационарное распределение. При таком выборе $EL_i(t) = L_i, i = \overline{1, r}$ при всех $t \geq 0$.

Так как $EL_i(t)$ положительны и конечны при $t \geq 0$, то, меняя порядок интегрирования, получаем

$$E\{\frac{1}{T}\int_0^T \sum_{i=1}^r c_i L_i(t)dt\} = \frac{1}{T}\int_0^T \sum_{i=1}^r c_i E L_i(t)dt = \sum_{i=1}^r c_i L_i$$

Формула выведена.

Далее, сделаем необходимые предположения для дальнейших действий:

2.2 Предположение 1.

Первые два момента распределения числа вторичных требований конечны при всех i:

$q_{ij} = \frac{\partial}{\partial z_j} Q_i(z)|_{z=1} < \infty$; $q^i_{jk} = \frac{\partial^2}{\partial z_j \partial z_k} Q_i(z)|_{z=1} < \infty$

Из теории Фробениуса [2] известно, что у матрицы $Q = ||q_{ij}||$ с неотрицательными элементами существует единственное собственное значение $\xi > 0$, такое, что модули всех остальных собственных значений не превосходят ξ.

2.3 Предположение 2.

$\xi < 1$; таким образом, ветвящийся процесс вторичных требований вырождается с вероятностью 1.

Условимся далее считать, что в матричных операциях символы многокомпонентных величин обозначают векторы-столбцы, а символы со штрихом (транспонирование) — векторы-строки.

Положим $\lambda = (\lambda_1, ..., \lambda_r)$, $b = (b_1, ..., b_r)$.

Докажем, что вектор $f = (I - Q)^{-1} b$ определен и имеет положительные компоненты $f > 0$, где неравенство для векторов понимается в покомпонентном смысле, а I обозначает единичную матрицу соответствующей смыслу формулы размерности.

Из приведённого предположения и сходимости ряда $1 + x + x^2 + ...$ на спектре (наборе собственных векторов) матрицы Q следует, что матрица $I - Q$ следует, что $I - Q$ обратима и справедливо представление [2]:

$$(I - Q)^{-1} = I + Q + Q^2 + ... \tag{2.2}$$

Отсюда следует, что вектор $f = (I—Q)^{-1}b$ положителен (покомпонентно), если положителен b.

Неравенство для векторов понимается в покомпонентном смысле, а I обозначает единичную матрицу соответствующей смыслу формулы размерности.

2.4 Предположение 3.

$\rho < 1$, где, как мы докажем далее $\rho = \lambda' f$ имеет смысл загрузки системы.

Для того, чтобы доказать, что $\rho = \lambda' f$ имеет смысл загрузки системы, необходимо доказать эргодичность процесса $L(t)$. Докажем это:

Для доказательства эргодичности процесса $L(t)$ ограничимся пока только точками регенерации типа 0 и воспользуемся теоремой Смита [1]. Периодами регенерации в данном случае являются циклы занятости, которые, поскольку поток пуассоновский, имеют абсолютно непрерывную ФР. Чтобы установить существование некоторых моментов цикла занятости, сделаем это для периода занятости. В предположениях 1, 2, 3 справедливо следующее утверждение:

а) преобразования Лапласа—Стилтьеса функции распределения (ПЛС ФР) периода занятости, начинающегося единственным требованием типа i, удовлетворяет системе уравнений

$$\pi_i(s) = \beta_i(s + \sum_{j=1}^{r} \lambda_i[1 - \pi_i(s)])Q_i[\pi_1(s), ..., \pi_r(s)], i = \overline{1, r} \qquad (2.3)$$

б) вышеприведённая система имеет единственное решение $\pi_i(s), i = \overline{1, r}$ такое, что каждая из $\pi_i(s)$ представляет собой ПЛС собственных ФР.

Доказательство этих утверждений опускаем, так как п. а) доказывается достаточно просто, если воспользоваться методами работы [3], а доказательство п. б), напротив, громоздко.

Дифференцируя (2.3) по s и полагая $s = 0$, получим уравнение для первых моментов $\pi_i, i = \overline{1, r}$. Известно, что вероятностный смысл имеет минимальное решение этого уравнения, поэтому достаточно доказать единственность его конечного решения. В предположении конечности уравнение приводится к виду $A\pi = b$.

Отсюда, продолжая рассуждения доказательства, приведённого в предположении 2:

Матрица A также положительна (покомпонентно), если положителен b. Действительно,

$$A = I—Q - b\lambda' = (I - Q)[I—(I—Q)^{-}ib\lambda'] = (I - Q)(I—f\lambda')$$

Поскольку $(f\lambda')^n = f\lambda' f...\lambda' = \rho^{n-1} f\lambda'$ и согласно предположению 3 $\rho < 1$, то $(I - f\lambda')^{-1} = I + f\lambda' + \rho f\lambda' + ... = I + (1 - \rho)^{-1} f\lambda'$ Следовательно, решение уравнения $Ag = b$ имеет вид

$$g = A^{-1}b = (I - f\lambda')^{-1}(I - Q)^{-1}b = (1 - \rho)^{-1} f.$$

Известно [2], что наряду с (2.2) аналогичное представление имеет место и для матрицы Q_m, составленной из любых m строк и столбцов Q с одинаковыми номерами, взятых в любом порядке. Поэтому для строк и столбцов соответствующих матриц и компонент соответствующих векторов с номерами $i_1, ...i_m$ имеем

$$g_m = (1 - \rho_m) f_m,$$

где $f_m = (I - Q)^{-1} b_m, \rho_m = \lambda_{i_1} f_{i_1} + ... + \lambda_{i_m} f_{i_m}$.

В соответствии с рассуждениями выше, следует $\pi = (1 - \rho)^{-1} f$. С помощью двукратного дифференцирования системы выше и аналогичных рассуждений получаем для вектора вторых моментов $\pi_2 = (\pi_{12}, ..., \pi_{r2})$ уравнение $A\pi_2 = \varphi$, где

$$\varphi_i = (1 + \lambda'\pi)^2 b_{i2} + 2(1 + \lambda'\pi) b_i \sum_{j=1}^{r} q_{ij}\pi_j + \sum_{j,k=1}^{r} q_{kj}^i \pi_k \pi_j$$

Из доказанного вытекает существование двух первых моментов цикла занятости θ_1, θ_2 и эргодичность процесса $L(t)$. Для θ_1 имеем выражение $\theta_1 = \lambda_0^{-1}(1 + \lambda'\pi) = [\lambda_0(1 - \rho)]^{-1}$ из которого, применяя снова теорему Смита, имеем $P\{L(t) = 0\} = 1 - \rho$, т.е. ρ имеет смысл загрузки системы.

Матрица Q может быть как неразложимой, так и разложимой, что в приложениях даже более естественно. Разложимую матрицу можно привести к блочному виду [2].

2.5 Предположение 4.

В каждом диагональном блоке матрицы Q найдется такой индекс i, что $\lambda_i > 0$. Это предположение обеспечивает возможность появления в системе требований всех типов.

3 Основные соотношения

Для вычисления вероятностных характеристик процесса $L(t)$ при всех: допустимых функциях управления удобно пользоваться аппаратом регенерирующих процессов с несколькими типами точек регенерации. Действительно, если рассматривать процесс $L(t)$ в последовательные моменты $t_n, n = 1, 2, ...$ окончания обслуживания требований, то его поведение после t_n не зависит от его предшествующей траектории, а зависит лишь от состояния $L(t_n)$ в момент tn и значения функции переключения. Назовем t_n моментом регенерации типа i, $i = \overline{0, r}$, если $u(L(t_n)) = i$. В предложении 3 при доказательстве эргодичности процесса $L(t)$ показано, что при сделанных предположениях процесс является эргодическим при любом допустимом управлении. Это значит, что его стационарные характеристики не зависят от начального состояния. Положим $t_0 = 0$, $L(0) = 0$. Определим функции $H_0(t)$, $H_i(t, z)$ рядами:

$$H_0(t) = \sum_{n=0}^{\infty} P\{t_n < t, L(t_n) = 0\} \tag{3.1a}$$

$$H_i(t, z) = z_i^{-1} \sum_{l:u(l)=i} z^l \sum_{0}^{\infty} P\{t_n \leq t, L(t_n) = l\}, i = \overline{1, r} \tag{3.1b}$$

которые, очевидно, сходятся при всех t>0. Следующая теорема устанавливает связь между введенными функциями и между ними и производящей функцией $L(t)$ в произвольный момент t:

$$P(t, z) = E\{z^{L(t)}\} = \sum_{l \geq 0} z^l P\{L(t) = l\}$$

Обозначим через $p(s, z)$ преобразование Лапласа (ПЛ) функции $P(t, z)$ в точке s, а через $\chi_0(s)$ и $\chi_i(s, z)$ — преобразования Лапласа—Стилтьеса (ПЛС) соответственно функцией $H_0(t)$ и $H_i(t, z)$, $i = \overline{1, r}$.

Теорема 1. Справедливы соотношения

$$p(s,z) = \sum_{i=1}^{r} \left(\frac{\lambda_i}{s+\lambda_0} \chi_0(s) + \chi_i(s,z) \right) z_i \frac{1-\beta_i(s+\lambda_0-\lambda' z)}{s+\lambda_0-\lambda' z} \tag{3.2a}$$

$$\chi_0(s) + \sum_{i=1}^{r} z_i \chi_i(s,z) = \sum_{i=1}^{r} \left(\frac{\lambda_i}{s+\lambda_0} \chi_0(s) + \chi_i(s,z) \right) \times \beta_i(s+\lambda_0-\lambda' z) Q_i(z) \tag{3.2b}$$

Здесь $\beta_i(\cdot)$ — ПЛС ФР $B_i(\cdot)$, $\lambda_0 = \lambda_1 + \ldots + \lambda_r$. Существуют ненулевые пределы

$$\lim_{s\to 0} sp(s,z) = \lim_{t\to\infty} P(t,z) = P(z)$$

$$\lim_{s\to 0} s\chi_i(s,z) = \lim_{t\to\infty} \frac{1}{t} H_i(t,z) = \chi_i(z) = \sum_{k\geq 0} h_i(k) z^k, i = \overline{0,r}$$

Доказательство:

Вывод (3.2a) и (3.2b) производится на основе вероятностной интерпретации производящих функций [3]. При таком подходе каждое поступающее в систему требование i-го типа независимо от остальных требований и от состояния системы называется красным с вероятностью z_i и синим с вероятностью $1 - z_i$. Для доказательства (3.2b) выпишем равенство:

$$dH_0(t) + \sum_{i=1}^{r} z_i d_i H_i(t,z) = \sum_{i=1}^{r} \int_0^T d_x H_i(x,z) dB_i(t-x) e^{-(t-x)(\lambda_0-\lambda' z)} +$$

$$+ \int_0^t dH_0(x) \sum_{i=1}^{r} \int_0^{t-x} \lambda_i e^{-\lambda_0 y} dy dB_i(t-x-y) Q_i(z) e^{-(t-x-y)(\lambda_0-\lambda' z)}$$

Слева в этом равенстве стоит вероятность того, что в момент регенерации, находящийся в бесконечно малой окрестности точки t, все требования в системе красные или их вовсе нет, справа — вероятность того же события, найденная по формуле полной вероятности через вероятности событий, которые могли иметь

место в предшествующий t момент регенерации x и на периоде регенерации.

Переходя к ПЛС, получаем (3.2b). Аналогично получается (3.2a). Существование пределов есть следствие эргодичности процесса $L(t)$, доказанного в утверждении 3, и вложенной по моментам регенерации марковской цепи (доказательство аналогично [4]), а также абелевой теоремы [5], причём

$$\chi_i(z) = \lim_{T\to\infty} T^{-1} H_i(T, z), i = \overline{1, r}$$

$$P(z) = \lim_{T\to\infty} T^{-1} \int_0^T P(t, z)dt = \lim_{T\to\infty} P(t, z)$$

Доказательство завершено.

Предел $s\chi_0(s)$ при $s \to 0$ вычисляется непосредственно. Как известно из [1], $\chi_0(s) = [1-\theta(s)]^{-1}$ где $\theta(s)$ — ПЛС ФР периода регенерации, который выражается через $\pi_i(s)$ — ПЛС ФР периодов занятости, открывающихся требованием типа i, $i = \overline{1, r}$, в виде $\theta(s) = (s + \lambda_0)^{-1}[\lambda_1\pi_1(s) + ... + \lambda_r\pi_r]$. Отсюда (см. предположение 3) интересующий нас предел $\chi_0 = \lambda_0(1 - \rho)$. С учетом этого видно, что (3.2a), (3.2b) в пределе переходят соответственно в

$$P(z) = \sum_{i=1}^{r} [\lambda_i(1 - \rho) + \chi_i(z)] \frac{1 - \beta_i(\lambda_0 - \lambda' z)}{\lambda_0 - \lambda' z} \qquad (3.3a)$$

$$(1 - \rho) \sum_{i=1}^{r} \lambda_i(z_i - 1) = \sum_{i=1}^{r} [z_i - b_i(z)][\lambda_i(1 - \rho) + \chi_i(z)] \qquad (3.3b)$$

, где $b_i(z) = \beta_i(\lambda_0 - \lambda' z)Q_z(z)$

Соотношения (3.3a), (3.3b) позволяют получить все необходимые характеристики процесса $L(t)$ и сформулировать проблему оптимизации в виде задачи линейного программирования.

4 Задача линейного программирования

Связь между функционалом (2.1) и функцией управления $u(\cdot)$ осуществляется путем введения переменных

$$x_{ij} = \frac{\partial}{\partial z_j}\chi_i(z)|_{z=1} = \sum_{l:u(l)=i} l_i h_i(l)$$

обладающих важным свойством: $x_{ij} = 0$ тогда и только тогда, когда требования типа j имеют приоритет перед требованиями типа i.

С помощью этих величин и соотношений (3.3a), (3.3b) задача минимизации функционала (2.1) сводится к задаче линейного программирования

$$J = \sum_{i,j=1}^{r} b_i c_i x_{ij} \Rightarrow min \tag{4.1}$$

при ограничениях

$$\sum_{i=1}^{r}(a_{ij}x_{ik} + a_{ik}x_{ij}) = \gamma_{ik}, \quad j,k = 1,r \tag{4.2}$$

Выведем формулы (4.1) и (4.2).

Обозначим интенсивность выполнения операций типа i $R_i = \lambda_i(1-\rho) + \chi_i(1)$. Учитывая, что $\frac{\partial}{\partial z_j}[z_i - b_i(z)]|_{z=1} = a_{ij}$, продифференцируем (3.3b) по z_j в точке $z = 1$. Получим $A'R = (1—\rho)\lambda$. Повторяя рассуждения доказательства из предложения 2 для A', находим $R = A'^{-1} \times (1-\rho)\lambda = (I - Q')\lambda$, который, как и следовало ожидать, не зависит от управления.

Путем двукратного дифференцирования равенства (3.3b) по z_j и z_k в точке $z = 1$, замечая, что

$$\frac{\partial^2}{\partial z_j \partial z_k}[z_i - b_i(z)]|_{z=1} = -(\lambda_j\lambda_k b_{i2} + \lambda_j b_i q_{ik} + \lambda_k b_i q_{ij} + q^i_{jk})$$

получаем

$$\sum_{i=1}^{r}(a_{ij}x_{ik}+a_{ik}x_{ij})=\gamma_{jk},\quad 1\le j,\quad k\le r,$$

где $\gamma_{jk}=\gamma_{kj}>0$,

$$\gamma_{jk}=\sum_{i=1}^{r}(\lambda_j\lambda_k b_{i2}+\lambda_j b_i q_{ik}+\lambda_k b_i q_{ij}+q^i_{jk})R_i$$

Формула (3.3а) позволяет выразить L_i через переменные x_{ij}:

$$L_j=\frac{\partial}{\partial z_j}P(z)|_{z=1}=\sum_{i=1}^{r}b_i x_{ij}+\sum_{i=1}^{r}R_i(\lambda_j b_{i2})+b_i$$

Так как от управления зависит только первая сумма, то задача минимизации функционала (2.1) эквивалентна задаче минимизации функционала (который обозначим той же буквой J) (4.1) при ограничениях (4.2).

Крайними точками множества, задаваемого ограничениями (4.2), являются матрицы $X=||x_{ij}||$, такие, что $x_{ii}>0$, $i=\overline{1,r}$ и одно из чисел x_{ij}, x_{ji} при $i\neq j$ равно нулю, а другое — положительно.

Действительно, из (4.2) следует, что $x_{ij}+x_{ji}\ge\gamma_{ij}>0$, $i\neq j$, $x_{ii}\ge\gamma_{ii}/2>0$, а подсчет числа линейно-независимых ограничений показывает, что невырожденное базисное решение должно содержать $\frac{r(r-1)}{2}$ нулей. Среди матриц указанного вида выделяются такие, которые одновременной перестановкой строк и столбцов приводятся к треугольной форме. Эти матрицы соответствуют приоритетным дисциплинам обслуживания. У матриц, которые не обладают этим свойством, найдутся такие индексы $i_1,i_2,...,i_k$, что $x_{i_2i_1}=0,...,x_{i_ki_{k-1}}=0,x_{i_1i_k}=0$. С учетом смысла переменных x_{ij} ясно, что эти крайние точки вообще не соответствуют какой бы то ни было допустимой функции управления, так как с ненулевой вероятностью в системе в момент переключения могут оказаться требования типов $i_1,...,i_k$ (и только они), и в этом случае никакое требование не может быть поставлено на обслуживание. Это значит, что множество (4.2) содержит планы, не отвечающие допустимым функциям управления. Исключать недопустимые планы удобнее, обращаясь к двойственной задаче линейного программирования. Одновременное исследование прямой и двойственной задач позволяет найти область значений

параметров системы, в которой оптимальна заданная приоритетная дисциплина. Исследование задачи линейного программирования, приведенное далее, показывает, что оптимальными планами задачи (4.1), (4.2) могут быть лишь матрицы, приводимые к треугольной форме, соответствующие приоритетным дисциплинам обслуживания. Таким образом, справедлива.

Теорема 2. Существует оптимальное управление системой, которое осуществляется с помощью приоритетной дисциплины обслуживания.

4.1 Исследование задачи линейного программирования

Введем, как предписывает алгоритм вывод формул (4.1) и (4.2), приоритетную дисциплину. Перенумеруем индексы (чтобы избежать субиндексов): $i_k \to k$, так что $1 \succ \ldots \succ r$.Соответствующий план задачи (4.1) и (4.2) X^* будет верхней треугольной матрицей. Двойственные (4.2) ограничения с помощью симметричной матрицы Y запишутся в виде $AY \leq bc'$, в чем нетрудно убедиться, выписывая функцию Лагранжа. Для того чтобы план X^*. был оптимальным в задаче (4.1), (4.2), в силу основной теоремы линейного программирования [6] необходимо и достаточно выполнения условий дополняющей нежёсткости. В данном случае это эквивалентно существованию решения системы уравнений - неравенств $AU \leq bc'$, $AU = bc'$, где равенства имеют место на главной диагонали и выше, а ниже главной диагонали - неравенства. Следовательно, задача сводится к тому, чтобы установить необходимые и достоточные условия разрешимости указанной системы. Запишем ее в развернутом виде:

$$\sum_{i=1}^{k} a_i^k u_{ik} + \sum_{i=k+1}^{r} a_i^k u_{ki} = c_k b^k \tag{4.3a}$$

$$\sum_{i=1}^{k} \overline{a}_i^k u_{ik} + \sum_{i=k+1}^{r} \overline{a}_i^k u_{ki} \leq c_k b^k \tag{4.3b}$$

Здесь a_i - i-й, столбец матрицы A, вектор ω^k $(\overline{\omega}^k)$ размерности k $(r-k)$ получается из r-мерного вектора ω отбрасыванием $r-k$ последних (k первых) компонент.

Покажем, что необходимым и достаточным условием разрешимости (4.3) яв-

ляется система неравенств $c^m_{r-m} \geq 0, m = \overline{0, r-1}$. Легко видеть, что если приоритеты назначены в соответствии с алгоритмом вывода формул (4.1) и (4.2), то это условие выполняется автоматически. Доказательство производится по индукции. По условию $c^0_k > 0, k = \overline{1, r}$. Система (4.3) при $k = r$ решается относительно вектора (столбца) $u_r = (u_{1r}, ..., u_{rr})$, $u_r = (1-\rho)^{-1} c_r f^r$. Выразим a_r через компоненты вектора f^r (см. доказательство из предположения 2):

$$a_r = (f^r_r)^{-1}[c^0_r b - (f^r_1 a_1 + ... + f^r_{r-1} a_{r-1})]$$

и подставим это выражение в (4.3), $k = \overline{1, r-1}$. Введем новые переменные $u^1_{ij} = u_{ij} - (f^r_j / f^r_r) u_{ir}$, $i = \overline{1, r-1}$. Легко видеть, что $||u^1_{ij}||$ опять симметричная матрица. В результате проведенного преобразования получим эквивалентную форму (4.3):

$$\sum_{i=1}^{k} a^k_i u^1_{ik} + \sum_{i=k+1}^{r-1} a^k_i u^1_{ki} = c^1_k b^k,$$

$$\sum_{i=1}^{k} \overline{a}^k_i u^1_{ik} + \sum_{i=k+1}^{r-1} \overline{a}^k_i u^1_{ki} \leq c^1_k b^k, \quad k = \overline{1, r-1},$$

$$u_r = c^0_r (1-\rho_r)^{-1} f^r, \quad c^0_r > 0, \; c^1_j = c^0_j - (f^r_j / f^r_r) c^0_r, \; j = \overline{1, r-1}.$$

Предположим, что проделано m аналогичных шагов, в результате которых система приняла вид

$$\sum_{i=1}^{k} a^k_i u^m_{ik} + \sum_{i=k+1}^{r-m} a^k_i u^m_{ki} = c^m_k b^k \tag{4.4a}$$

$$\sum_{i=1}^{k} \overline{a}^k_i u^m_{ik} + \sum_{i=k+1}^{r-m} \overline{a}^k_i u^m_{ki} \leq c_k \overline{b}^k, \; k = \overline{1, r-m}, \tag{4.4b}$$

$$u_{r-j} = c^0_{r-j} (1-\rho_{r-j})^{-1} f^{r-j}, \quad c^j_{r-j} \geq 0, \; j = \overline{0, m-1}. \tag{4.4c}$$

(Смысл символа A_j здесь тот же самый, что и Q_j). Требуется доказать, что $c^m_{r-m} \geq 0$ и указать переход к эквивалентной форме (4.4), в которой в соотношения не входят элементы столбца a_{r-m}. В (4.4) рассмотрим систему равенств

при $k = r - m$. Очевидно $u_{r-m} = c^m_{r-m} A^{-1}_{r-m} b^{r-m}$. Так как $A^{-1}_{r-m} b^{r-m} > 0$, а соответствующие элементы матрицы A отрицательны, то подстановка u_{r-m} в неравенства (4.4) при $k = r-m$ показывает, что указанные неравенства выполняются тогда и только тогда, когда $c^m_{r-m} \geq 0$. Если $c^m_{r-m} = 0$, то a_{r-m} входит в систему (ибо $u_{r-m} = 0$). Если же $c^m_{r-m} > 0$, то из (4.4) при $k = r - m$ имеем

$$\overline{a}^m_{r-m} \leq (u^m_{r-m\ r-m})^{-1}(c^m_{r-m}\overline{b}^m - \sum_{i=1}^{r-m-1} \overline{a}^m_i u^n_{im}).$$

$$\overline{a}^m_{r-m} \leq (u^m_{r-m\ r-m})^{-1}(c^m_{r-m}\overline{b}^m - \sum_{i=1}^{r-m-1} \overline{a}^m_i u^m_{im}).$$

Подставляя эти выражения в (4.4) и полагая

$$u^{m+1}_{ij} = u^m_{ij} - (f^{r-m}_j / f^{r-m}_{r-m}) u^m_{ir-m}, i = \overline{1, r-m-1}$$

преобразуем систему (4.4) к виду

$$\sum_{i=1}^{k} a^k_i u^{m+1}_{ik} + \sum_{i=k+1}^{r-m-1} a^k_i u^{m+1}_{ki} = c^{m+1}_k b^k \tag{4.5a}$$

$$\sum_{i=1}^{k} \overline{a}^k_i u^{m+1}_{ik} + \sum_{i=k+1}^{r-m-1} \overline{a}^k_i u^{m+1}_{ki} \leq c^{m+1}_k \overline{b}^k,\ k = \overline{1, r-m}, \tag{4.5b}$$

$$u_{r-j} = c^j_{r-j}(1 - \rho_{r-j})^{-1} f^{r-j},\ \ c^j_{r-j} \geq 0,\ j = \overline{0, m}. \tag{4.5c}$$

$$c^{m+1}_j = c^m_j - (f^{r-m}_j / f^{r-m}_{r-m}) c^m_{r-m},\ j = \overline{1, r-m-1} \tag{4.5d}$$

где $||u^{m+1}_{ij}||$ - снова симметричная матрица, если такой же была по индуктивному предположению $||u^m_{ij}||$. Тем самым индуктивный шаг завершен и требуемое утверждение доказано. Очевидно, что последовательность $1 \succ \ldots \succ r$ обеспечивает выполнение условия $c^m_{r-m} \geq 0,\ m = \overline{0, r-1}$.

5 Алгоритм назначения приоритетов

В результате осуществления $r-1$ шага строится последовательность индексов $i_1, ..., i_r$ следующим образом. На шаге $m, m = \overline{0, r—2}$ вычисляется вектор f^{r-m} из уравнения

$$(I - Q_{r-m})f^{r-m} = b^{r-m} \tag{5.1}$$

где Q_{r-m}, b_{r-m} составлены из строк и столбцов (соответственно компонент) Q, b, индексы которых не совпадают с $i_r, ..., i_{r-m+1}$ — индексами, определенными на предыдущих шагах. Далее следует положить

$$c_i^m = c_i^{m-1} - \frac{f_i^{r-m+1}}{f_{i_{r-m+1}}^{r-m+1}} c_{i_{r-m+1}}^{m-1},\ i \neq i_r, ..., i_{r-m+1}\ (c_i^0 = c_i) \tag{5.2}$$

и выбрать индекс i_{r-m} (если таких несколько, то любой из них) из условия $c_{i_{r-m}}^m / f_{i_{r-m}}^{r-m} \leq c_i^m / f_i^{r-m},\ i \neq i_r, ..., i_{r-m+1}$. Перейти на шаг $m+1$. Оптимальной будет дисциплина, при которой операция типа i_j имеет преимущество перед; операцией типа i_k, - если $j < k$, т.е. $i_1 \succ ... \succ i_r$. Заметим, что определение $i_1, ..., i_r$ корректно, так как уравнения (5.1) разрешимы и $f^m > 0$, $m = \overline{0, r-2}$ (см. предположение 2).

6 Обсуждение и примеры

В зависимости от вида матрицы Q рассмотренная модель включает в себя многофазные системы и системы с обратной связью [4], в частности решение задачи, поставленной в [7]. Рассмотрим примеры.

1. Пусть $Q = 0$. На каждом шаге алгоритма имеем $f^m = b^m$ и условие оптимальности дисциплины $i_1 \succ i_2 \succ \ldots \succ i_r$ имеет вид $b_{i1}/c_{i1} \geq \ldots \geq b_{ir/c_{ir}}$.

2. Проиллюстрируем действие алгоритма на модельном примере. Пусть однопроцессорная система работает в пакетном режиме. Выделим 3 операции: ввод пакета и его обработка (трансляция программ и подготовка к счету); счет по каждой программе; выдача результатов счета. Допустим, что средние продолжительности выполнения операций в условных единицах $b_1 = 600, b_2 = 10, b_3 = 60$, среднее число программ в пакете — 10, и каждая программа обращается к устройству ввода — вывода в среднем 1 раз, так что

$$Q = \begin{bmatrix} 0 & 10 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Ясно, что только $\lambda_1 > 0$ и должно выполняться условие $1300\lambda_1 < 1 (\rho < 1)$.

Зададим весовые коэффициенты $c_i, i = l, 2, 3$, учитывающие значимость различных операций. В данном примере можно руководствоваться соображениями экономии оперативной памяти, считая, что задержка в обработке программ, находящихся в оперативной памяти, приносит большие потери, чем те, которые еще не введены, т.е. $c_1 < c_2 = c_3$.Примем $c_2 = c_3 = 10, c_1 = 1$. В рассматриваемом примере расчет по приведенному алгоритму потребует всего 2 шага. Оптимальной является приоритетная дисциплина $3 \succ 2 \succ 1$.

Следует учитывать, что, хотя условия оптимальности зависят только от b_i, c_i и Q, необходимо существование вторых моментов b_{i2}, q_{ij}^k и выполнение условия $\rho < 1$.

По предложенному в статье алгоритму разработана программа, которая подготовлена для передачи в фонд алгоритмов и программ по ТМО.

7 Библиография

1. Кокс Д., Смит У. Теория восстановления. «Наука», 1967
2. Гантмахер Ф. Р. Теория матриц. «Наука», 1967.
3. Климов Г. П. Стохастические системы. «Наука», 1966.
4. Климов Г. Я. Системы обслуживания с разделением времени. I. Теория вероятностей и ее применения, № 3, стр. 558—576, 1974.
5. Widder D. V. The Laplace Transform. Princeton, 1946.
6. Васильев Ф. П. Лекции по методам решения экстремальных задач. Изд-во МГУ, 1974.
7. Kleinrock L., Coffman E. Distribution of attained service in time-shared systems. J. Computer and System. Sciences, v. 1, pp. 287-298, 1967.