М. Ю. Китаев, В. В. Рыков, Система обслуживания с ветвящимися потоками вторичных требований, *Автомат. и телемех.*, 1980, выпуск 9, 52–61

УДК 65.012.122

СИСТЕМА ОБСЛУЖИВАНИЯ С ВЕТВЯЩИМИСЯ ПОТОКАМИ ВТОРИЧНЫХ ТРЕБОВАНИЙ

М. Ю. КИТАЕВ, В. В. РЫКОВ

(Москва)

Показано, что оптимальной относительно линейного функционала потерь дисциплиной обслуживания в однолинейной системе с простейшим первичным и ветвящимся вторичным потоками требований нескольких типов и произвольной длительностью обслуживания является дисциплина приоритетного типа. Приводится алгоритм определения оптимальной дисциплины.

1. Введение

При проектировании АСУ, информационно-поисковых и вычислительных систем широкое применение находят приоритетные модели теории массового обслуживания [1]. В частности, приходится проводить исследования систем с ориентацией и переключениями [2—4], а также решать задачи оптимизации приоритетного обслуживания [1, 5—8]. Оказалось, что приоритетные дисциплины являются оптимальными в более широком классе динамических дисциплин, использующих информацию о текущем состоянии системы [5—7].

В настоящей работе рассматривается предложенная в [8] задача определения оптимальной дисциплины обслуживания в системе с несколькими типами требований, которые могут поступать как извне (первичные), так и в результате обслуживания (вторичные). Такие модели возникают при исследовании работы ЭВМ в различных режимах, информационно-поисковых системах и т. п. Например, при работе ЭВМ в пакетном режиме в результате первичной обработки пакета определяется число программ в пакете, каждая программа в свою очередь может требовать различные ресурсы: вызов стандартных программ в оперативную память, обращение к внешним запоминающим устройствам за дополнительной информацией и т. п. В задаче поиска в результате просмотра некоторого массива выясняется, что необходимая информация содержится в одном из других массивов.

Организация работы систем указанного вида включает в себя оперативное определение порядка обслуживания требований. В работе показано, что относительно линейного функционала потерь оптимальной является приоритетная дисциплина и приводится алгоритм ее построения.

2. Постановка задачи. Предположения

Рассмотрим однолинейную систему обслуживания, выполняющую r типов операций. Длительности выполнения отдельных операций являются независимыми случайными величинами (СВ) с функциями распределения (ФР) $B_i(\cdot)$, $B_i(0)=0$, имеющими первые два момента b_i, b_{i2}. Первичные требования на выполнение операции типа i образуют пуассоновский поток с интенсивностью λ_i, $\lambda_i \geqslant 0$, $i=\overline{1,r}$. В частности, возможно, что 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 $q_i(n)=q_i(n_1,\ldots,n_r)$ возникает набор $n=(n_1,\ldots,n_r)$ вторичных требований на выполнение операций различных типов, которые мгновенно поступают в очереди (неограниченные) для требований соответствующего типа. Непосредственно вслед за этим по набору $l=(l_1,\ldots,l_r)$, где l_i — число требований в i-й очереди, с помощью функции управления $u(l)$ выбирается очередное требование на обслуживание. 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 $l_i>0$. В момент, когда система освобождается и нельзя направить требование на обслуживание, мы доопределяем $u(0)=0$.

Поведение системы описывается процессом $L(t)=(L_1(t),\ldots,L_r(t))$, где $L_i(t)$ — число требований типа i в момент t.

Пусть далее c_i — стоимость единицы времени пребывания в системе требования типа i. Задача заключается в определении функции управления, минимизирующей потери в единицу времени в стационарном режиме работы системы. В приложении 3 показано, что при сделанных далее предположениях функционал, определяющий эти потери, записывается в виде

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

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

Сделаем необходимые для дальнейшего предположения. Обозначим через $Q_i(z)=Q_i(z_1,\ldots,z_r)$ производящую функцию (ПФ) числа вторичных требований, возникающих в результате выполнения операции типа i:

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

Предположение 1. Первые два момента распределения числа вторичных требований конечны при всех i:

$$q_{ij}=\frac{\partial}{\partial^7} Q_i(z)\,|_{z=1}<\infty; \quad q_{jk}{}^i=\frac{\partial^2}{\partial z_j\,\partial z_k} Q_i(z)\,|_{z=1}<\infty.$$

Из теории Фробениуса известно [9], 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.

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 $\lambda=(\lambda_1,\ldots,\lambda_r)$, $b=(b_1,\ldots,b_r)$. В приложении 1 показано, 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.

Предположение 3. $\rho<1$, где $\rho=\lambda' f$ имеет, как показано в приложении 2, 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 [9].

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 [8]. Действительно, если рассматривать процесс $L(t)$ в последовательные моменты t_n, $n=1,2,\ldots$ 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 t_n и значения функции переключения. Назовем t_n моментом регенерации типа i, $i=\overline{0,r}$, если $u(L(t_n))=i$. В приложении 2 показано, что при сделанных предположениях процесс $L(t)$ 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

$$H_0(t)=\sum_{n=0}^{\infty}\mathbf{P}\{t_n\leqslant t,\ L(t_n)=0\},\tag{2}$$

$$H_i(t,z)=z_i^{-1}\sum_{l:u(l)=i}z^l\sum_{n=0}^{\infty}\mathbf{P}\{t_n\leqslant t,\ L(t_n)=l\},\quad i=\overline{1,r},$$

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:

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

Обозначим через $p(s,z)$ преобразование Лапласа (ПЛ) функции $P(t,z)$ в точке s, а через $\varkappa_0(s)$ и $\varkappa_i(s,z)$ — преобразования Лапласа — Стилтьеса (ПЛС) соответственно функций $H_0(t)$ и $H_i(t,\ z)$, $i=\overline{1,r}$.

Теорема 1. Справедливы соотношения

$$p(s,z)=\sum_{i=1}^{r}\left\{\frac{\lambda_i}{s+\lambda_0}\varkappa_0(s)+\varkappa_i(s,z)\right\}z_i\frac{1-\beta_i(s+\lambda_0-\lambda'z)}{s+\lambda_0-\lambda'z},\tag{3}$$

$$\varkappa_0(s)+\sum_{i=1}^{r}z_i\varkappa_i(s,z)=\sum_{i=1}^{r}\left\{\frac{\lambda_i}{s+\lambda_0}\varkappa_0(s)+\varkappa_i(s,z)\right\}\times$$
$$\times\beta_i(s+\lambda_0-\lambda'z)Q_i(z).\tag{4}$$

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

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

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

Доказательство см. в приложении 4.

Предел $s\varkappa_0(s)$ при $s\to 0$ вычисляется непосредственно. Как известно [10], $\varkappa_0(s)=[1-\tau(s)]^{-1}$, где $\tau(s)$ — ПЛС ФР периода регенерации, который выражается через $\pi_i(s)$ — ПЛС ФР периодов занятости, открывающихся требованием типа i, $i=\overline{1,r}$, в виде $\tau(s)=(s+\lambda_0)^{-1}[\lambda_1\pi_1(s)+\ldots$ $\ldots+\lambda_r\pi_r]$. Отсюда (см. приложение 2) интересующий нас предел $\varkappa_0=$ $=\lambda_0(1-\rho)$. С учетом этого видно, что (3), (4) в пределе переходят соответственно в

$$P(z)=\sum_{i=1}^{r}[\lambda_i(1-\rho)+\varkappa_i(z)]\frac{1-\beta_i(\lambda_0-\lambda'z)}{\lambda_0-\lambda'z}, \tag{5}$$

$$(1-\rho)\sum_{i=1}^{r}\lambda_i(z_i-1)=\sum_{i=1}^{r}[z_i-b_i(z)][\lambda_i(1-\rho)+\varkappa_i(z)], \tag{6}$$

где $b_i(z)=\beta_i(\lambda_0-\lambda'z)Q_i(z)$.

Соотношения (5), (6) позволяют получить все необходимые характеристики процесса $L(t)$ и сформулировать проблему оптимизации в виде задачи линейного программирования.

4. Задача линейного программирования

Связь между функционалом (1) и функцией управления $u(\cdot)$ осуществляется путем введения переменных

$$x_{ij}=\frac{\partial}{\partial z_j}\varkappa_i(z)\big|_{z=1}=\sum_{l:u(l)=i} l_j h_i(l),$$

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 (5), (6) задача минимизации функционала (1) сводится к задаче линейного программирования

$$J=\sum_{i,j=1}^{r}b_ic_jx_{ij}\Rightarrow\min \tag{7}$$

при ограничениях

$$\sum_{i=1}^{r}(a_{ij}x_{ik}+a_{ik}x_{ij})=\gamma_{jk},\qquad j,k=\overline{1,r}. \tag{8}$$

Здесь $a_{ij}=\delta_{ij}-b_i\lambda_j-q_{ij}$, $\gamma_{jk}=\gamma_{kj}>0$ не зависят от управления. Вывод (7) и (8) приведен в приложении 5.

Крайними точками множества, задаваемого ограничениями (8), 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 (8) следует, что $x_{ij}+x_{ji}\geqslant\gamma_{ij}>0$, $i\neq j$, $x_{ii}\geqslant\gamma_{ii}/2>0$, а подсчет числа линейно-независимых ограничений показывает, что невырожденное базисное решение должно содержать $^1/_2r(r-1)$ 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 $i_1, i_2, \ldots, i_k$, что $x_{i_2i_1}=0,\ldots,x_{i_ki_{k-1}}=0$, $x_{i_1i_k}=0$. 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 $i_1, \ldots, i_k$ (и только они), и в этом случае никакое требование не может быть поставлено на обслуживание. Это значит, что множество (8) содержит планы, не отвечающие допустимым функциям управления. Исключать недопустимые планы удобнее, обращаясь к двойственной задаче линейного программирования [6]. Одновременное исследование прямой и двойственной задач позволяет найти область значений параметров системы, в которой оптимальна заданная приоритетная дисциплина [11]. Исследование задачи линейного программирования, приведенное в приложении 6, показывает, что оптимальными планами задачи (7), (8) 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.

5. Алгоритм назначения приоритетов

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

$$(9) \qquad (I-Q_{r-m}) f^{r-m} = b^{r-m},$$

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

$$(10) \qquad c_i^m = c_i^{m-1} - \frac{f_i^{r-m+1}}{f_{i_{r-m+1}}^{r-m+1}} c_{i_{r-m+1}}^{m-1}, \qquad i \neq i_r, \ldots, i_{r-m+1} \qquad (c_i^0 = c_i)$$

и выбрать индекс i_{r-m} (если таких несколько, то любой из них) из условия $c_{i_{r-m}}^m / f_{i_{r-m}}^{r-m} \leqslant c_i^m / f_i^{r-m}$, $i \neq i_r, \ldots, i_{r-m+1}$. 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. $i_1 \succ \ldots \succ i_r$. Заметим, что определение $i_1, \ldots, i_r$ корректно, так как уравнения (9) разрешимы и $f^m > 0$, $m=\overline{0, r-2}$ (см. приложение 1).

6. Обсуждение и примеры

В зависимости от вида матрицы Q рассмотренная модель включает в себя многофазные системы и системы с обратной связью [7], в частности решение задачи, поставленной в [12]. 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 $b_{i_1}/c_{i_1} \geqslant \ldots \geqslant b_{i_r}/c_{i_r}$, что является хорошо известным результатом [5].

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, $i=1, 2, 3$, 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 $3>2>1$.

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

ПРИЛОЖЕНИЕ 1

Свойства матриц $I-Q$ и $A=I-Q-b\lambda'$

Из предположения 2 и сходимости ряда $1+x+x^2+\ldots$ на спектре матрицы Q следует, что матрица $I-Q$ обратима и справедливо представление [9]

$$(\text{П.1}) \qquad (I-Q)^{-1}=I+Q+Q^2+\ldots$$

Отсюда следует, что вектор $f=(I-Q)^{-1}b$ положителен (покомпонентно), если положителен b, то же самое имеет место для матрицы A. Действительно,

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

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

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

Известно [9], что наряду с (П.1) 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 $i_1,\ldots i_m$ имеем

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

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

ПРИЛОЖЕНИЕ 2

Эргодичность процесса $L(t)$

Для доказательства эргодичности процесса $L(t)$ ограничимся пока только точками регенерации типа 0 и воспользуемся теоремой Смита [10]. Периодами регенерации в данном случае являются циклы занятости, которые, поскольку поток пуассоновский, имеют абсолютно непрерывную ФР. Очевидно, чтобы установить существование некоторых моментов цикла занятости, достаточно сделать это для периода занятости. В предположениях 1, 2, 3 справедливо следующее утверждение:

а) ПЛС ФР периода занятости, начинающегося единственным требованием типа i, удовлетворяет системе уравнений

$$(\text{П.2}) \qquad \pi_i(s)=\beta_i\left(s+\sum_{j=1}^{r}\lambda_i[1-\pi_i(s)]\right)Q_i[\pi_1(s),\ldots,\pi_r(s)),\quad i=\overline{1,r};$$

б) система (П.2) имеет единственное решение $\pi_i(s)$, $i=\overline{1,r}$, такое, что каждая из $\pi_i(s)$ представляет собой ПЛС собственной ФР.

Доказательство этих утверждений опускаем, так как п. а) доказывается достаточно просто, если воспользоваться методами работы [13], напротив, доказательство п. б) громоздко.

Дифференцируя (П.2) по s и полагая $s=0$, получим уравнение для первых моментов π_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 в соответствии с приложением 1 следует $\pi=(1-\rho)^{-1}f$. С помощью двукратного дифференцирования (П.2) и аналогичных рассуждений получаем для вектора вторых моментов $\pi_2=(\pi_{12},\ldots,\pi_{r2})$ 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 τ_1, τ_2 и эргодичность процесса $L(t)$. Для τ_1 имеем выражение $\tau_1=\lambda_0^{-1}(1+\lambda'\boldsymbol{\pi})=[\lambda_0(1-\rho)]^{-1}$, из которого, применяя снова теорему Смита, имеем $\mathbf{P}\{L(t)=0\}=1-\rho$, т. е. ρ имеет смысл загрузки системы.

ПРИЛОЖЕНИЕ 3

Вывод формулы (1)

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

$$\mathrm{E}\left\{\frac{1}{T}\int_0^T\sum_{i=1}^{r} c_iL_i(t)\,dt\right\},$$

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 $L_i=\lim\mathrm{E}\{L_i(t)\}$. Покажем, что $L_i<\infty$. Действительно, величина $V=L_1b_1+\ldots+L_rb_r$, которая имеет смысл стационарного среднего времени разгрузки системы, и величина $L=(L_1,\ldots,L_r)$ конечны или бесконечны одновременно. Но по своему определению V не превосходит среднего времени до окончания периода регенерации, которое, как известно из теории восстановления [10], равно $\tau_2/2\tau_1$. 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 $\mathrm{E}\{L_i(t)\}=L_i$, $i=\overline{1,\ r}$ при всех $t\geqslant0$.

Так как $\mathrm{E}\{L_i(t)\}$ положительны и конечны при $t\geqslant0$, то, меняя порядок интегрирования, получаем

$$\mathrm{E}\left\{\frac{1}{T}\int_0^T\sum_{i=1}^{r} c_iL_i(t)\,dt\right\}=\frac{1}{T}\int_0^T\sum_{i=1}^{r} c_i\mathrm{E}\{L_i(t)\}\,dt=\sum_{i=1}^{r} c_iL_i.$$

ПРИЛОЖЕНИЕ 4

Доказательство теоремы 1

Особенно прост вывод (3), (4) на основе вероятностной интерпретации производящих функций [13]. 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 (4) выпишем равенство

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

$$+\int_0^t dH_0(x)\sum_{i=1}^{r}\int_0^{t-x}\lambda_ie^{-\lambda_0y}\,dy\,dB_i(t-x-y)\,Q_i(z)\,e^{-(t-x-y)(\lambda_0-\lambda'z)}.$$

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 (4). Аналогично получается (3). Существование пределов есть следствие эргодичности процесса $L(t)$ и вложенной по моментам регенерации марковской цепи (этот факт доказывается аналогично тому, как это сделано в [7]), а также абелевой теоремы [14], причем

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

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

ПРИЛОЖЕНИЕ 5

Вывод соотношений (7), (8)

Обозначим интенсивность выполнения операций типа i $R_i=\lambda_i(1-\rho)+\varkappa_i(1)$. Учитывая, что $\frac{\partial}{\partial z_j}[z_i-b_i(z)]|_{z=1}=a_{ij}$, продифференцируем (6) по z_j в точке $z=1$. Получим $A'R=(1-\rho)\lambda$. Повторяя рассуждения приложения 1 для A', находим $R=A'^{-1}\times\times(1-\rho)\lambda=(I-Q')\lambda$, который, как и следовало ожидать, не зависит от управления.

Путем двукратного дифференцирования равенства (6) по z_j и z_k в точке $z=1$, замечая, что

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

получаем

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

где $\gamma_{jk}=\gamma_{kj}>0$,

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

Формула (5) 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 (1) эквивалентна задаче минимизации функционала (который обозначим той же буквой J) (7) при ограничениях (8).

ПРИЛОЖЕНИЕ 6

Исследование задачи линейного программирования

Введем, как предписывает алгоритм приложения 5, приоритетную дисциплину. Перенумеруем индексы (чтобы избежать субиндексов): $i_k\to k$, так что $1\succ\ldots\succ r$. Соответствующий план задачи (7), (8) X^* будет верхней треугольной матрицей. Двойственные (8) ограничения с помощью симметричной матрицы Y запишутся в виде $AY\leqslant bc'$, в чем нетрудно убедиться, выписывая функцию Лагранжа. Для того чтобы план X^* был оптимальным в задаче (7), (8), в силу основной теоремы линейного программирования [15] необходимо и достаточно выполнения условий дополняющей нежесткости. В данном случае это эквивалентно существованию решения системы уравнений – неравенств $AU\overset{=}{\leqslant}bc'$, 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:

$$\text{(П.3)}\qquad \sum_{i=1}^{k} a_i{}^k u_{ik} + \sum_{i=k+1}^{r} a_i{}^k u_{ki} = c_k b^k,$$

$$\sum_{i=1}^{k} \bar{a}_i{}^k u_{ik} + \sum_{i=k+1}^{r} \bar{a}_i{}^k u_{ki} \leqslant c_k \bar{b}^k.$$

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

Покажем, что необходимым и достаточным условием разрешимости (П.3) является система неравенств $c^m_{r-m}\geqslant 0$, $m=\overline{0,\ r-1}$. Легко видеть, что если приоритеты назначены в соответствии с алгоритмом приложения 5, то это условие выполняется автоматически. Доказательство по индукции. По условию $c_k{}^0>0$, $k=\overline{1,\ r}$. Система (П.3) при $k=r$ решается относительно вектора (столбца) $u_r=(u_{1r},\ldots,u_{rr})$, $u_r=(1-\rho)^{-1}c_rf^r$. Выразим a_r через компоненты вектора f^r (см. приложение 1):

$$a_r=(f_r{}^r)^{-1}[c_r{}^0b-(f_1{}^ra_1+\ldots+f^r_{r-1}a_{r-1})]$$

и подставим это выражение в (П.3), $k=\overline{1,\ r-1}$. Введем новые переменные $u_{ij}{}^1=u_{ij}-(f_j{}^r/f_r{}^r)u_{ir}$, $i=\overline{1,\ r-1}$. Легко видеть, что $\|u_{ij}{}^1\|$ опять симметричная матрица. В результате проведенного преобразования получим эквивалентную форму (П.3):

$$\sum_{i=1}^{k}a_i{}^ku_{ik}{}^1+\sum_{i=k+1}^{r-1}a_i{}^ku_{ki}{}^1=c_k{}^1b^k,$$

$$\sum_{i=1}^{k}\bar{a}_i{}^ku_{ik}{}^1+\sum_{i=k+1}^{r-1}\bar{a}_i{}^ku_{ki}{}^1\leqslant\bar{c}_k{}^1\bar{b}^k,\qquad k=\overline{1,r-1},$$

$$u_r=c_r{}^0(1-\rho_r)^{-1}f^r,\qquad c_r{}^0>0,\qquad c_j{}^1=c_j{}^0-(f_j{}^r/f_r{}^r)c_r{}^0,\qquad j=\overline{1,\ r-1}.$$

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

$$\sum_{i=1}^{k}a_i{}^ku_{ik}{}^m+\sum_{i=k+1}^{r-m}a_i{}^ku_{ki}{}^m=c_k{}^mb^k,$$

(П.4)

$$\sum_{i=1}^{k}\bar{a}_i{}^ku_{ik}{}^m+\sum_{i=k+1}^{r-m}\bar{a}_i{}^ku_{ki}{}^m\leqslant c_k{}^m\bar{b}^k,\qquad k=\overline{1,r-m},$$

$$u_{r-j}=\overset{j}{c}_{r-j}(1-\rho_{r-j})^{-1}f^{r-j},\qquad \overset{j}{c}_{r-j}\geqslant 0,\qquad j=\overline{0,m-1}.$$

(Смысл символа A_j здесь тот же самый, что и Q_j.) Требуется доказать, что $c^m_{r-m}\geqslant 0$, и указать переход к эквивалентной форме (П.4), в которой в соотношения не входят элементы столбца a_{r-m}. В (П.4) рассмотрим систему равенств при $k=r-m$. Очевидно $u_{r-m}=\overset{m}{c}_{r-m}\overset{-1}{A}_{r-m}b^{r-m}$. Так как $A^{-1}_{r-m}\,b^{r-m}>0$, а соответствующие элементы матрицы A отрицательны, то подстановка u_{r-m} в неравенства (П.4) при $k=r-m$ показывает, что указанные неравенства выполняются тогда и только тогда, когда $c^m_{r-m}\geqslant 0$. Если $c^m_{r-m}=0$, то a_{r-m} не входит в систему (ибо $u_{r-m}=0$). Если же $c^m_{r-m}>0$, то из (П.4) при $k=r-m$ имеем

$$\overset{m}{\bar{a}}_{r-m}\leqslant(\overset{m}{u}_{r-m\ r-m})^{-1}\left(\overset{m}{c}_{r-m}\bar{b}^m-\sum_{i=1}^{r-m-1}\bar{a}_i{}^mu_{im}{}^n\right).$$

$$\overset{m}{\bar{a}}_{r-m}\leqslant(\overset{m}{u}_{r-m\ r-m})^{-1}\left(\overset{m}{c}_{r-m}\bar{b}^m-\sum_{i=1}^{r-m-1}\bar{a}_i{}^mu_{im}{}^m\right).$$

Подставляя эти выражения в (П.4) и полагая

$$\overset{m+1}{u_{ij}}=u_{ij}{}^m-(\overset{r-m}{f_j}/\overset{r-m}{f}_{r-m})\overset{m}{u}_{ir-m},\qquad i=\overline{1,r-m-1}$$

преобразуем систему (П.4) к виду

$$\sum_{i=1}^{k} a_i^k u_{ik}^{m+1} + \sum_{i=k+1}^{r-m-1} a_i^k u_{ki}^{m+1} = c_k^{m+1} b^k,$$

$$\sum_{i=1}^{k} \bar{a}_i^k u_{ik}^{m+1} + \sum_{i=k+1}^{r-m-1} \bar{a}_i^k u_{ki}^{m+1} \leqslant c_k^{m+1} \bar{b}^k,$$

$$u_{r-j} = c_{r-j}^{j}(1-\rho_{r-j})^{-1} f^{r-j}, \qquad c_{r-j}^{j} \geqslant 0, \qquad j=\overline{0, m},$$

$$c_j^{m+1} = c_j^m - (f_j^{r-m} / f_{r-m}^{r-m}) c_{r-m}^{m}, \qquad j=\overline{1, r-m-1},$$

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

Поступила в редакцию
1 августа 1979 г.

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

1. *Рыков В. В.* Управляемые системы массового обслуживания (обзор). Итоги науки. Теория вероятностей. Математическая статистика. Теоретическая кибернетика, т. 12, стр. 43–153. ВИНИТИ, М., 1975.
2. *Волковинский М. И., Кабалевский А. Н.* Обслуживание со смешанным приоритетом в системе с потерями на переключение. Автоматика и телемеханика, № 11, стр. 16–22, 1975.
3. *Волковинский М. И., Кабалевский А. Н.* Анализ многоуровневых приоритетных систем с подготовкой и перенастройкой прибора. Техническая кибернетика, ч. 1, № 4, стр. 115–119, 1977, ч. 2, № 3, стр. 209–214, 1978.
4. *Мишкой Г. К.* Обслуживание с ориентацией и двумя потоками требований относительного приоритета. Техническая кибернетика, № 4, стр. 110–115, 1977.
5. *Рыков В. В., Лемберг Э. К.* Об оптимальных динамических приоритетах в однолинейных системах массового обслуживания. Техническая кибернетика, № 1, стр. 25–34, 1967.
6. *Питтель Б. Г.* Оптимальное управление в системе массового обслуживания с несколькими потоками требований. Техническая кибернетика, № 6, стр. 101–116, 1972.
7. *Климов Г. П.* Системы обслуживания с разделением времени. I. Теория вероятностей и ее применения, № 3, стр. 558–576, 1974.
8. *Рыков В. В.* Применение регенерирующих процессов при исследовании управляемых систем. Теория массового обслуживания. Тр. III Всесоюз. школы-совещания по теории массового обслуживания. Т. 1, стр. 135–148. Изд-во МГУ, 1975.
9. *Гантмахер Ф. Р.* Теория матриц. «Наука», 1967.
10. *Кокс Д., Смит У.* Теория восстановления. «Наука», 1967.
11. *Шварц Л. Б.* Об условиях реализуемости заданных стационарных режимов в приоритетных системах массового обслуживания. Автоматика и телемеханика, № 10, стр. 55–64, 1975.
12. *Kleinrock L., Coffman E.* Distribution of attained service in time-shared systems. J. Computer and System Sciences, v. 1, pp. 287–298, 1967.
13. *Климов Г. П.* Стохастические системы. «Наука», 1966.
14. *Widder D. V.* The Laplace Transform. Princeton, 1946.
15. *Васильев Ф. П.* Лекции по методам решения экстремальных задач. Изд-во МГУ, 1974.

A SERVICE SYSTEM WITH A BRANCHING FLOW OF SECONDARY CUSTOMERS

A. Yu. KITAEV, V. V. RYKOV

A discipline of the priority type is shown to be optimal with respect to a linear loss functional in a single-line system with a simple primary and a branching secondary flows of customers of several types and an arbitrary service duration. An algorithm for determining the optimal discipline is described.
