ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 22 МАТЕМАТИКИ Выпуск 4
2015

Я. М. Агаларов, **М. Г. Коновалов** (Москва, ИПИ ФИЦ ИУ РАН)**. Об ограничении нагрузки в системе с гетерогенными ресурсами, произвольным временем обслуживания и дедлайном.**

Обслуживающая система состоит из M серверов с очередями (накопителями) бесконечной емкости. Производительности серверов, вообще говоря, неодинаковы. В систему поступает пуассоновский поток заявок на выполнение заданий. В момент поступления заявка должна быть (мгновенно) либо принята, либо отклонена. Принятая заявка в тот же момент должна быть направлена на один из серверов, и, если он свободен, то сразу же начинается выполнение задания, а если занят, то задание становится в очередь. Принятие заявки сопровождается получением платы за обслуживание в размере 1.

Выполнение заданий осуществляется на сервере в порядке поступления, без прерываний. Время выполнения задания на сервере определяется как θR^{-1}, где θ — случайная величина с фиксированным распределением и конечным первым моментом, а $R > 0$ — производительность конкретного сервера. Если оказывается, что время пребывания задания в системе, превысило установленный дедлайн, то системе начисляется штраф, равный $C > 0$.

Задача заключается в отыскании алгоритма управления заявками, при котором максимальна целевая функция, определяемая как предельный средний доход системы, складывающийся из платы за обслуживание и штрафа за превышение дедлайна. Дополнительное условие состоит в том, что время выполнения каждого задания заранее не известно.

Пусть $t = 0, 1, \ldots$ — номера последовательных моментов поступления заявок в систему и пусть $n_t^{(i)}$ означает количество заданий на сервере i в момент с номером t, $n_t = (n_t^{(1)}, \ldots, n_t^{(M)})$. В каждый момент с номером t в зависимости от значения n_t необходимо выбрать «управление» m из множества $M = \{0, 1, \ldots, M\}$, причем значение $m = 0$ соответствует отклонению заявки, а значение $m > 0$ соответствует адресации заявки на сервер m. Статические стратегии управления имеют вид $f : \mathbf{N}_+^M \to M$, где $\mathbf{N}_+^M$ — неотрицательная целочисленная решетка. Иными словами, если $n_t = n = (n^{(1)}, \ldots, n^{(M)}) \in \mathbf{N}_+^M$, то заявка обслуживается согласно значению функции $f(n)$.

В докладе рассмотрены статические стратегии, имеющие пороговую структуру, которая характерна для ряда задач об управлении ресурсами. Они определяются с помощью набора неотрицательных чисел $T = (T^{(1)}, \ldots, T^{(M)})$:

$$f(n) = f_T(n) = \begin{cases} 0, & \text{если } \max_{1 \leqslant i \leqslant M}(T^{(i)} - n^{(i)}) \leqslant 0, \\ g(n), & \text{если } \max_{1 \leqslant i \leqslant M}(T^{(i)} - n^{(i)}) > 0, \end{cases}$$

где $g(n)$ — некоторая заданная функция, отвечающая за размещение уже принятой заявки.

Таким образом, задача ограничения нагрузки сводится к отысканию оптимальных пороговых значений T. Для ее решения предлагаются два подхода: 1) численный алгоритм, опирающийся на аналитическое решение, полученное для случая $M = 1$, и 2) алгоритм, использующий имитационную модель и градиентный метод адаптивной оптимизации марковской последовательности. Обсуждаются результаты вычислительных экспериментов с предложенными алгоритмами, в которых применяются различные варианты «распределяющей» функции g.

Работа выполнена при поддержке Российского фонда фундаментальных исследований (грант 15-07-03406).