ДИНАМИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ РЕСУРСОВ ДЛЯ ПРИЛОЖЕНИЙ

Матвеев Г. А., Первин А.Ю., Трушкова Е. А.

ИПС РАН, г. Переславль-Залесский, Россия

С помощью математической теории оптимизации динамических систем проводится исследование автоматического управления аппаратными ресурсами, которое способно учитывать ценность выделенных приложению ресурсов (процессорной мощности, памяти и т. п.) при текущей пользовательской нагрузке [1]. Это позволит оптимально использовать ресурсы в процессе эксплуатации системы приложений, а именно, динамически добавлять или убирать количество каждого ресурса во время работы приложения с учетом потребностей и приоритетов других приложений системы.

Построим математическую модель рассматриваемой задачи для системы, состоящей из n приложений, использующих m различных категорий ресурсов.

Обозначим через $r_{ij}(t)$, $i=1,\ldots,n$, $j=1,\ldots,m$, - количество ресурса j, выделенного приложению i в момент времени t. Пусть ограничения на использование ресурсов записываются в виде

$$r_j^- \le \sum_{i=1}^{n} r_{ij}(t) \le r_j^+. \qquad (1)$$

Предполагается, что каждое приложение имеет индивидуальный набор характеристик, которые однозначно описывают его текущее состояние с точки зрения пользователя. Характеристика зависит от объема ресурсов, предоставленных приложению, и от оказываемой на приложение с течением времени неуправляемой пользовательской нагрузки $L_i(t)$, $i=1,\ldots,n$. При этом характеристики могут быть составлены либо на испытательном стенде до запуска приложения в эксплуатацию, либо непосредственно в процессе работы приложения. Обозначим через $p_{ij} = p_{ij}(r_{i1}(t), r_{i2}(t), \ldots, r_{im}(t), L_i(t))$, $i=1,\ldots,n$, $j=1,\ldots,k_i$, - характеристику j приложения i, через $\tilde{p}_{ij}$ - целевой уровень соответствующей характеристики (желаемое значение характеристики, ниже которого она не должна опускаться во время эксплуатации приложения).

С помощью функций

$$\sigma_i(t) = \sum_{j=1}^{k_i} \alpha_{ij} p_{ij}\left(r_{i1}(t),\ldots,r_{im}(t),L_i(t)\right),\ i=1,\ldots,n, \qquad (2)$$

можно получить некоторую суммарную оценку характеристик каждого из приложений в момент времени t. Здесь через α_{ij} обозначены весовые коэффициенты, выбираемые согласно значимости каждой характеристики приложения. Будем называть функции $\sigma_i(t)$ уровнями сервиса приложений.

Задача состоит в поддержании значения уровня сервиса (2) не ниже некоторого целевого уровня сервиса (поддержание каждой характеристики не ниже ее целевого уровня) на дискретном наборе моментов времени $T = \{t_0, t_0 + h, \ldots, t_0 + qh\}$. Данная задача равносильна минимизации отклонения уровня сервиса

$$\delta_i = \sum_{j=1}^{k_i} \alpha_{ij} \sum_{t=0}^{q} \max\left\{0, \tilde{p}_{ij} - p_{ij}\left(r_{i1}(t_0 + th), r_{i2}(t_0 + th), \ldots, r_{im}(t_0 + th), L_i(t_0 + th)\right)\right\} \to \min$$

для системы приложений в совокупности, с учетом приоритета каждого из приложений. Обозначим через δ_i^+ максимум δ_i при всех допустимых значениях ресурсов (минимум δ_i, очевидно, равен нулю).

Тогда исходную задачу динамического распределения ресурсов можно поставить в виде дискретной задачи оптимального управления следующего вида:

$$\begin{aligned}
&y_i(t+h) = y_i(t) + \sum_{j=1}^{k_i} \max\left\{0, \tilde{p}_{ij} - p_{ij}\left(r_{i1}(t), \ldots r_{im}(t), L_i(t)\right)\right\}, \quad t \in T, \\
&y_i(t_0) = 0, \quad i = 1, \ldots, n, \quad r_j^- \le \sum_{i=1}^{n} r_{ij}(t) \le r_j^+, \\
&F(y(t_0 + qh)) = \sum_{i=1}^{n} \beta_i \frac{y_i(t_0 + qh)}{\delta_i^+} \to \min,
\end{aligned} \tag{3}$$

где через β_i обозначены весовые коэффициенты, выбираемые согласно приоритету каждого приложения (большее значение весового коэффициента соответствует более высокому приоритету соответствующего приложения). Здесь роль управлений играют функции $r_{ij}(t)$, $i = 1, \ldots, n$, $j = 1, \ldots, m$.

Рассмотрим один из самых простых вариантов поставленной задачи оптимального управления (3), а именно, предположим, что на рассматриваемом промежутке времени управления постоянны $r_{ij}(t) = r_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$. Тогда, согласно всем введенным обозначениям, задача динамического управления ресурсами для системы приложений может быть сформулирована в виде задачи поиска минимума функции $i \times j$ переменных $F\left(y(t_0 + qh)\right) = G(r_{ij})$ при условии (1), т. е.

$$F\big(y(t_0+qh)\big)=G(r_{ij})\to\min_{r},\ r_j^-\le\sum_{i=1}^{n}r_{ij}\le r_j^+.$$

Спецификой рассматриваемой задачи является зависимость вида динамической системы (3) от неизвестных функций $L_i(t), i=1,\ldots,n$. При этом предполагается, что известны значения этих функций в некоторые предшествующие рассматриваемому отрезку времени моменты $t=t_0-sh,t_0-(s-1)h,\ldots,t_0-h$. Этот факт позволяет осуществлять прогноз изменения функций $L_i(t)$ на рассматриваемом отрезке времени (например, с помощью численных методов аппроксимации функций одной переменной).

Предлагается производить в режиме параллельного счета (на кластерном вычислительном устройстве) условную минимизацию функции $G(r_{ij})$ для различных прогнозов изменения функций $L_i(t)$, что позволит получить закон оптимального управления для динамической системы (3) с учетом различного поведения неуправляемой нагрузки $L_i(t)$. Следовательно, в процессе эксплуатации системы приложений можно будет динамически корректировать распределение ресурсов между приложениями с учетом изменения неуправляемой нагрузки, выбирая наиболее подходящий прогноз. Полученное таким образом управление (количество выделяемых ресурсов) будет носить кусочно-постоянный характер.

Литература

1. Московский А.А., Первин А.Ю., Walker B. Оптимальное управление ресурсами виртуальных инструментов на вычислительном кластере // Тр. четвертой межд. конф. "Параллельные вычисления и задачи управления" (PACO'2008), Москва, 2008. ИПУ им. В. А. Трапезникова РАН. С. 968-978. ISBN 978-5-91450-016-7.