МЕТОДЫ ОПТИМИЗАЦИИ СОСТАВА И СТРУКТУРЫ ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ

Фельдман Л.П., Михайлова Т.В.
Кафедра ПМиИ ДонГТУ
feldman@r5.dgtu.donetsk.ua

Abstract

Feldman L.,Michailova T. Structure optimization methods for high- performance systems. *There is approximation method for synthesis closed queuing systems without calculations error estimation. In this work structure optimization task for systems with a permanent number of tasks is put, the methods of numerical solution with Markov chains are worked.*

Введение

Основополагающие методы оптимизации разомкнутых стохастических сетей, используемых в качестве моделей систем оперативной обработки (СОО), были разработаны Л. Клейнроком [1]. Критерии эффективности ВС предлагались в статьях В.М. Глушкова [2]. Основные методы и результаты перечисленных выше работ изложены в книге С.А. Майорова [3]. В ней приведено обоснование выбора критерия сбалансированности СОО, представленной разомкнутой стохастической сетью. В настоящее время нет аналитических методов решения задачи синтеза ВС, представленной замкнутыми стохастическими сетями. В данной работе поставлена задача оптимизации состава и структуры ВС, обрабатывающих постоянное число задач, разработаны методы численного ее решения с использованием цепей Маркова.

1. Постановка задачи синтеза вычислительных систем

Для корректной постановки задачи проектирования ВС необходимо располагать сведениями о классе задач, решение которых возлагается на систему: сложности задач, характеризующих потребности задач в вычислительных ресурсах; трудоемкости задач, определяющих количество работы, выполняемой каждым из устройств системы в ходе решения. Исходя из параметров трудоемкости задач и порядка их решения, зависящего от способа планирования работ в ВС, определяются трудоемкости $\theta_1,\theta_2,...,\theta_n$ обслуживания каждого этапа задачи в каждом из n устройств. В качестве модели ВС, используется экспоненциальная стохастическая сеть. Средняя длительность обслуживания заявки в системе и при заданной трудоемкости этапа равна $\nu_i = \theta_i/V_i$, а интенсивность обслуживания $\mu_i = V_i/\theta_i$, где V_i – быстродействие i-й системы. В процессе решения одной задачи происходит в среднем α_i обращений к системе i, в которой она на каждом этапе задерживается на время u_i, следовательно, эта система задерживает получение ответа в среднем на время $\alpha_i u_i$. С учетом этого среднее время пребывания задачи в сети составит

$$U = \sum_{i=1}^{n} \alpha_i u_i. \qquad (1)$$

Если предположить линейную зависимость стоимости устройств, составляющих систему, от их быстродействия, то ВС будет стоить:

$$S = \sum_{i=1}^{n} c_i V_i , \qquad (2)$$

где c_i - стоимость единицы производительности устройства типа i.

В [3] задача оптимального проектирования формулируется следующим образом:

1.Синтезировать систему, обеспечивающую решение λ_0 задач в единицу времени при минимально возможном времени ответа, причем стоимость системы не должна превышать заданного значения S*.

2.Синтезировать систему, обеспечивающую решение λ_0 задач в единицу времени при среднем времени ответа, не превосходящим заданной величины U*, причем стоимость системы должна быть минимальной.

Результатами решения задач 1 и 2 является определение значений быстродействий V_i, $i = \overline{1,n}$.

2. Модель системы клиент-сервер с непрерывным временем

Упрощенная модель системы клиент- сервер приведена на рис. 1. В ней M рабочих станций пользователей и один сервер. Предполагается, что все рабочие станции заняты пользователями, каждый из которых может послать только один запрос на сервер. Таким образом, в системе постоянно находится M запросов, из них m находится у пользователей, остальные M-m на сервере.

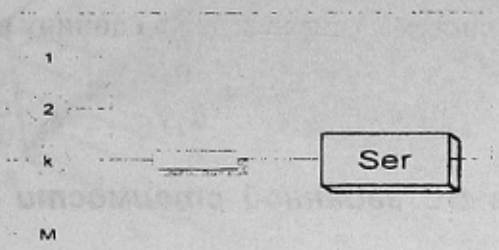

Рисунок 1. Упрощенная модель системы клиент – сервер

Функционирование рассматриваемой системы представлена замкнутой стохастической сетью, содержащей две системы массового обслуживания (СМО), в которой циркулирует M заявок. Граф передач этой сети изображен на рис. 2, на котором введены следующие обозначения: S_1, S_2 - СМО соответствующие клиентам и серверу; S_0 - фиктивная система, введенная для фиксации событий завершения задач пользователями; p_{12} - вероятность поступления заявки пользователя на сервер, p_{10} - вероятность завершения задачи пользователем.

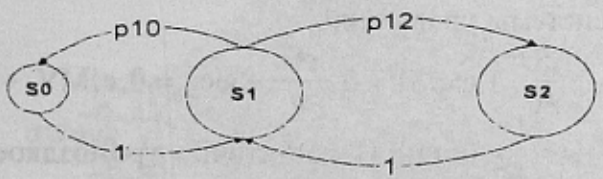

Рисунок 2. Граф передач сети клиент- сервер

Поскольку в ВС находится постоянное число задач, то предполагается, что после завершения очередной задачи пользователь приступает к следующей. На рис. 2 этому соответствует передача через систему S_0. По графу передач определяются соотношения интенсивностей потоков заявок поступающих в каждую из систем

$$\lambda_0 = p_{10}\lambda_1 , \lambda_2 = p_{12}\lambda_1 .$$

Выражая λ_1, λ_2 через λ_0, получены: $\lambda_1 = \alpha_1\lambda_0, \lambda_2 = \alpha_2\lambda_0$, где $\alpha_1 = 1/p_{10}, \alpha_2 = p_{12}/p_{10}$, $\alpha_1 - \alpha_2 = 1$, т.к. $p_{10}+p_{12}=1$.

Если за состояние системы принять распределение заявок по СМО (m,M-m), $m=\overline{0,M}$, то можно вычислить стационарные вероятности состояний, используя теорему Джексона:

$$\pi(m, M-m) = \frac{(\alpha_1\nu_1)^m(\alpha_2\nu_2)^{M-m}}{m!\sum_{i=1}^{M}\frac{(\alpha_1\nu_1)^i(\alpha_2\nu_2)^{M-i}}{i!}}, \quad m = \overline{0,M}, \tag{3}$$

Основные характеристики сети, выраженные через стационарные вероятности:

- загрузка сервера

$$\rho_2 = 1 - \pi(M,0), \tag{4}$$

- среднее число задач, находящихся у пользователей и на сервере

$$m_1^s = \frac{\alpha_1\nu_1\rho_2}{\alpha_2\nu_2}, m_2^s = M - m_1^s, \tag{5}$$

- среднее время ответа на запрос пользователя

$$u = \frac{\nu_2 M}{\rho_2} - \frac{\alpha_1}{\alpha_2}\nu_1, \tag{6}$$

- общее время решения задачи

$$U = \alpha_2 u + \alpha_1\nu_1, \tag{7}$$

- производительность системы (число задач в единицу времени)

$$\lambda_0 = \frac{\rho_2}{\alpha_2\nu_2}. \tag{8}$$

3. Задача синтеза ВС заданной стоимости с минимальным временем ответа

Рассматривается решение следующей задачи: определить быстродействие рабочих станций V_1 и сервера V_2, обеспечивающих минимальное время решения задачи U так, чтобы стоимость системы с M рабочими станциями не превышала заданного значения S*. Таким образом, необходимо найти минимум функции U (7) при условии

$$c_1MV_1+c_2V_2 \le S^*, \quad V_1>0, \quad V_2>0. \tag{9}$$

Применяя метод множителей Лагранжа, определяется минимум функции

$$G=U+\omega(c_1MV_1+c_2V_2-S^*), \tag{10}$$

где ω - неопределенный постоянный множитель. В этом случае V_1 и V_2 и ω определяются как решение системы уравнений

$$\frac{\partial U}{\partial V_1} + \omega c_1 M = 0, \frac{\partial U}{\partial V_2} + \omega c_2 = 0, c_1MV_1 + c_2V_2 \le S^*, \tag{11}$$

Выражение для функции U достаточно громоздкое, поэтому для выполнения преобразований, необходимых для получения системы (11), используется пакет *Mathematica* [6]. Ниже приводится программа формирования уравнений (11).

Все вычисления получены для следующих значений исходных данных: трудоемкость одного этапа решения задачи в миллионах операций $\Theta_1=1$, $\Theta_2=2$; количество этапов обслуживания $\alpha_1=11$, $\alpha_2=10$; c1= 2 тыс.грв/ Мфлоп, c2= 4 тыс.грв/ Мфлоп, S = 90 тыс.грв., M=8 ; U=90 с.

Выражение для функции U (7) определятся следующим образом:

```
U t1_, t2_, v1_, v2_, a_, M_ = a * t2 * M r2 t1, t2, v1, v2, a, M  v2;
```

$$U(t1,t2,v1,v2,a,M) = \frac{a\,M\,t2}{v2\left|1 - \frac{E^{\frac{(1-a)\,t1\,v2}{a\,t2\,v1}}\left(\frac{(1-a)\,t1\,v2}{a\,t2\,v1}\right)^{M}\,\text{Gamma}\,2+M}{1+M\;M!\;\text{Gamma}\,1+M,\frac{1-a\;t1\,v2}{a\,t2\,v1}}\right|} \qquad (12)$$

Частные производные по v1 и по v2 находятся операторами:

```
dU1 t1_, t2_, v1_, v2_, a_, M_ = D U t1, t2, v1, v2, a, M , v1
dU2 t1_, t2_, v1_, v2_, a_, M_ = D U t1, t2, v1, v2, a, M , v2
```

Формулы для частных производных достаточно сложны, поэтому аналитическое решение трансцендентной системы уравнений (11) едва ли возможно. Анализируя графики частных производных по v1 и по v2 (рис. 3, рис. 4), можно сделать вывод: производные отрицательны в рассматриваемой области.

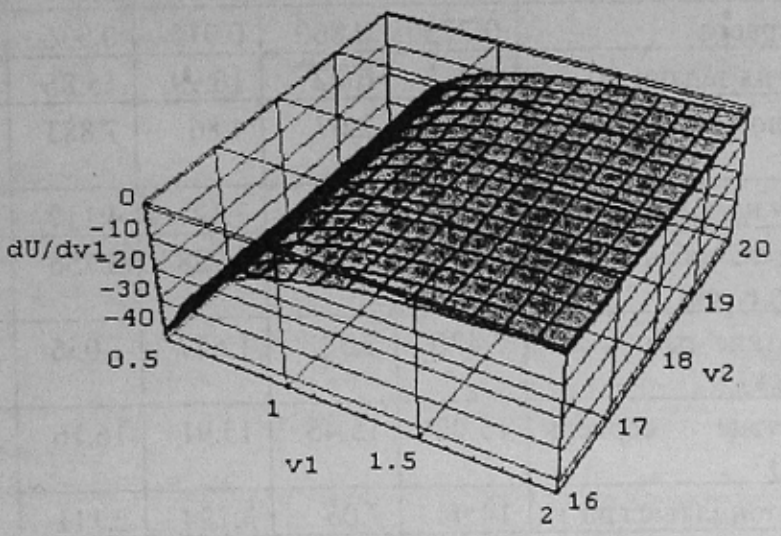

Рисунок 3. График зависимости производной функции U(v1,v2) по v1

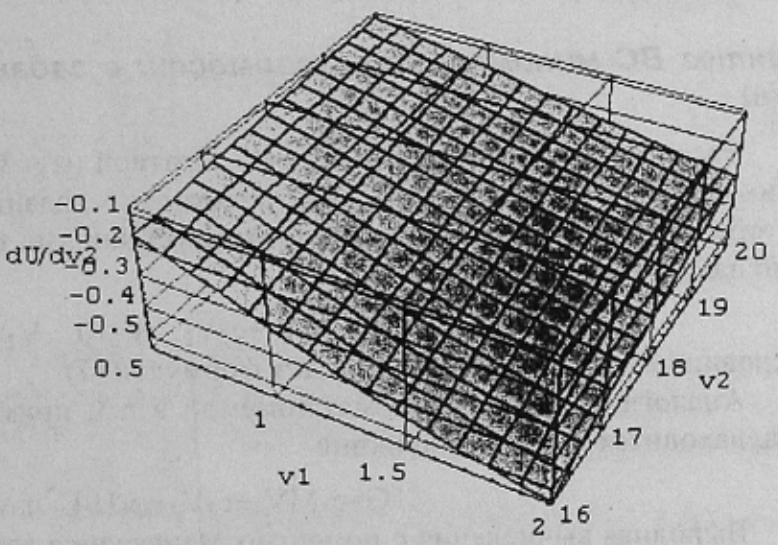

Рисунок 4. График зависимости производной функции U(v1,v2) по v2

Система (11) определяется следующим образом:

```
Q1[t1_, t2_, v1_, v2_, a_, M_, c1_, c2_] =
 dU1[t1, t2, v1, v2, a, M] - c1 / c2 * M * dU2[t1, t2, v1, v2, a, M];
 Q2[v1_, v2_, M_, c1_, c2_, s_] = s - (c1 * v1 * M + c2 * v2);

 q1 = Q1[1, 2, v1, v2, 10, 10, 2, 4] == 0;
 q2 = Q2[v1, v2, 10, 2, 4, 90] == 0;
 x = FindRoot[{q1, q2}, {v1, 1}, {v2, 4}]
```

Вычислены оптимальные значения быстродействий рабочих станций и сервера и другие характеристики ВС в зависимости от количества клиентов M (табл. 1).

Таблица 1

	Число Задач M						
	2	4	8	12	16	20	24
Загрузка раб. станции	0.429	0.518	0.608	0.657	0.690	0.715	0.734
Загрузка сервера	0.775	0.860	0.915	0.936	0.948	0.955	0.960
Время реш-ия задачи (с)	3.435	6.02	10.99	15.86	20.68	25.46	30.21
Сред. число работающих станций	0.857	2.07	4.86	7.883	11.04	14.29	17.61
Число задач на сервере	1.143	1.93	3.14	4.117	4.96	5.71	6.39
Производительность системы (задач в сек)	0.582	0.664	0.728	0.756	0.774	0.786	0.794
Быстродействие раб. станции (Мфл/сек)	7.470	3.53	1.647	1.056	0.771	0.605	0.496
Быстродействие сервера (Мфлоп/сек)	15.02	15.45	15.91	16.16	16.332	16.453	16.54
Стоим.раб.станц.(тыс.грн.)	14.94	7.05	3.294	2.111	1.542	1.209	0.993
Стоим. сервера (тыс. грн)	60.11	61.80	63.64	64.66	65.33	65.81	66.17

4. Синтез ВС минимальной стоимости с заданным временем решения задачи

Рассматривается решение задачи, обратной из п.3: определить быстродействие рабочих станций V_1 и сервера V_2, обеспечивающих заданное время решения задачи U^* так, чтобы стоимость системы с M рабочими станциями была минимальна. Т.о., необходимо найти минимум функции S:

$$S= c_1MV_1+c_2V_2, \quad V_1>0, \quad V_2>0, \tag{13}$$

при условии $U < U^*$, где U определяется формулой (7).

Аналогично задаче, рассматриваемой в п.3, применяя метод множителей Лагранжа, находится минимум функции

$$G=c_1MV_1+c_2V_2+\omega(U-U^*). \tag{14}$$

Выполняя вычисления с помощью *Mathematica* для тех же данных, что и в п.3, построив систему трансцендентных уравнений для определения оптимальных быстродействий рабочих станций и сервера , с помощью которых вычислены и другие характеристики ВС (табл. 2).

Таблица 2

	Число Задач					
	2	4	8	6	10	24
Загрузка рабочей станции	0.57	0.48	0.408	0.3	0.28	0.26
Загрузка сервера	0.77	0.860	0.915	0.95	0.955	0.960
Время решения задачи (с)	90	90	90	90	90	90
Сред.число раб. станций	0.85	2.07	6.33	11.0	14.29	17.61
Число задач на сервере	1.143	1.92	3.14	4.96	5.71	6.39
Производительность системы – задач в сек	0.022	0.044	0.08	0.17	0.22	0.26
Быстродействие рабочей станции (Мфл/сек)	0.29	0.24	1.19	0.17	0.17	0.16
Быстродейст. серв.(Мфл/сек)	0.57	1.03	1.591	3.75	4.65	5.55
Стоим. раб. станции (тыс.грн.)	1.14	1.89	3.294	5.67	6.84	7.998
Стоим. сервера (тыс. грн)	2.29	4.13	7.86	15	18.61	22.22

Выводы

Получены основные характеристики (4)-(8) для системы клиент – сервер (таблица 1, таблица 2) в зависимости от количества клиентов. Анализируя результаты решения задачи синтеза с минимальным временем ответа, можно сделать вывод, что с ростом количества клиентов растет загрузка рабочих станций и сервера, быстродействие сервера также растет, а быстродействие рабочих станций уменьшается, что вполне естественно при ограниченной стоимости и заданном классе задач. В задаче с ограниченным временем ответа резко возрастает быстродействие и стоимость сервера, чтобы удовлетворить ограничение на время нахождения клиента в системе.

Для сравнения приведены некоторые характеристики, полученные по методике в [3] и с помощью марковской модели (таблица 3, где М1 - результаты марковской модели, М2 - результаты, полученные с помощью метода в [3]). Видно, что результаты при М=1 совпадают и расходятся при М>1, следовательно, предположения в [3] о зависимости интенсивности обслуживания μ_i от количества задач М $\mu_i \sim \alpha^{M/(M+1)}$ не оправданы.

Таблица 3

	М=1		М=2		М=4		М=8		М=20	
	М1	М2	М1	М2	М1	М2	М1	М2	М1	М2
Быстрод. раб. станции	15.48	15.4	7.47	13.37	3.53	11.81	1.65	10.84	0.6	10.18
Быстрод. сервера	14.7	14.75	15.02	15.81	15.42	16.59	15.9	17.07	16.3	17.4

Литература

1. Клейнрок Л. Вычислительные системы с очередями. – М.: Мир, 1979 - 600с. Последовательно - параллельные вычисления: Пер. с англ. - М.: Мир, 1985. - 456 с.
2. Глушков В.М. Два универсальных критерия эффективности вычислительных машин. // Доклады АН УССР, 1960, № 4 ,с.36 – 42.
3. Основы теории вычислительных систем/С.А.Майоров, Г.И.Новиков, Т.И.Алиев и др.
4. Wolfram S. Matematika. Ein Sistem fur Matematik auf dem Computer.-Addison – Wesley, 1994. - 993 p.

Поступила в редакційну колегію 1.03.2000 р.