УДК 681.324

СИМОНЕНКО В.П.

МАТЕМАТИЧЕСКАЯ ПОСТАНОВКА ЗАДАЧИ ДИНАМИЧЕСКОГО РАСПРЕДЕЛЕНИЯ РАБОТ В GRID СИСТЕМАХ И ОЦЕНКИ КАЧЕСТВА РЕШЕНИЯ

В статье рассматривается общая математическая модель динамического планирования в рас пределенной, неоднородной GRID системе. Показано, что назначение задачи на вычислительный ресурс сводится к проблеме поиска максимального паросочетания в двудольном графе.

This paper presents a general mathematical model of dynamic scheduler for distributed heterogeneous GRID system. It is shown that searching of computational resource for a task can be solved as maximum matching problem for bipartite graph.

Введение

Распределения задач по ресурсам в GRID [1] системе является одной из наиболее сложных задач организации распределенных вычислений. Сложность задачи распределения или динамического планирования обусловлено неоднородностью как объекта распределения, так и неоднородностью распределяемых задач. Наиболее известные планировщики или диспетчеры задач (заданий) для GRID систем Platform LSF, Windows HPC Server 2008, PBS, Condor, SGE , LoadLever , MOSIX и внешний планировщик MAUI [2-11] предназначены для оптимизации распределения потока задач (зада-ний) на ресурсы системы. Следует отметить, что если учитывать свойство неоднородности GRID системы, то такое рас пределение не всегда приводит к равномерной загрузке ресурсов [12] и требует применения нового класса пространственных планировщиков, учитывающих и приоритетность задач и неоднородность вычислительной системы.

Постановка задачи

В неоднородной системе распределенной обработки данных (GRID), состоящей из N ресурсов, на момент времени распределения τ имеются N_τ свободных ресурсов и M независимых, готовых к выполнению заданий [1].

– *Система ресурсов* задана графом системы $G_R=(V_R,E_R,W_{VR},W_{ER})$, где:

- Множество вершин $V_R=\{R_1, R_2,..., R_N\}$, каждый элемент которого представляет один из N ресурсов системы и $R_i \in N|$ (множество натуральных чисел), i=1..N.
- Множество дуг $E_R=\{E_1, E_2,.., E_d\}$, каждый элемент которого определяют связи между двумя ресурсами $E_i=\{R_i,R_j\}$, где $R_i,R_j \in V_R$ и $0 \le d \le N^2$.
- Множество весов вершин $W_{VR}=\{WVR_1,WVR_2,...,WVR_N\}$, где $WVR_i=\{RE_i, RT_i\}$. Для $\forall i=1..N$, $RE_i \in \Re^+$ (множество положительных действительных чисел) есть характеристика ресурса R_i, $RT_i \in \{0$ и $\Re^+\}$ – состояния ресурсов.
- Множество весов дуг $W_{ER}=\{WER_1, WER_2,.., WER_p\}$. Это множество можно представить в виде некоторой матрицы $RC=RC[i,j] \in \Re^+$, где i=1..N и j=1..N.

– *Поток M заданий*, задан множеством $V_J=\{Job_1, Job_2,..., Job_M\}$, каждый элемент которого представляет одно из M заданий и $Job_i=\{JN_i, JE_i, JL_i, JM_i, JP_i\}$, $\forall i=1..N$:

- $JN_i \in N|$ – номер задания;
- $JE_i \in \Re^+$ – объем работы задания i;
- $JL_i=\{(R^1, \varphi_1),.., (R^q, \varphi_q)\}$, где $R^l \in V_R$ – ресурс, с которым данное задание требует обмена данными, $\varphi_1 \in \Re^+$ – объем передачи, l=1..q, $q \in N|$;
- $JM_i=\{0$ или $R^i\}$ – маска задания, где $R^i \in V_R$ – номер ресурса, на котором возможно или желательно выполнять данное задание;
- $JP_i \in \Re^+$ – приоритет данного задания.

Определение 1: Г есть отображение множества заданий $V_J=\{Job_1, Job_2,.., Job_M\}$ на множество ресурсов $V_R=\{R_1, R_2,.., R_N\}$ графа системы $G_R=(V_R,E_R,W_{VR},W_{ER})$, если результат отображения $\Gamma(V_J,V_R)$ есть некоторое множество A: $A=\{a_1, a_2,.., a_n\}$, где $a_i=(R^i, J^i)$, $R^i \in V_R$, $J^i \in V_J$, i=1..n, $n \in N|$.

Обозначим $AR=\{R^1, R^2,.., R^n\}$, $AJ=\{J^1, J^2,..,$

$J^n\}$. Таким образом, $|A|=|AR|\cap|AJ|$, $AR\subseteq V_R$, $AJ\subseteq V_J$.

Определение 2: отображение Г есть ***распределение*** заданий V_J на ресурсы V_R, если его результат $\Gamma(V_J,V_R)=A$, где $A=\{(R^1, J^1), (R^2, J^2),.., (R^n,J^n)\}$ удовлетворяет следующему условию: для $\forall i=1..n$, $R^i\notin AR\backslash R^i$, $J^i\in AJ\backslash J^i$, где $AR=\{R^1, R^2,.., R^n\}$, $AJ=\{J^1, J^2,.., J^n\}$. ***Размером*** данного распределения $\Gamma(V_J,V_R)$ является число n. Тогда $\Gamma(V_J,V_R)\rightarrow A$, $n=|A|$.

Определение 3: результат распределения заданий на ресурсы $A=\Gamma(V_J,V_R)$ называем ***расписанием*** для данного распределения Г. Пара $a_i=(R^i, J^i)$, i=1..n, называется ***назначением*** задания $J^i\in V_J$ на ресурс $R^i\in V_R$.

Определение 4: пусть $X=\{A^1, A^2,.., A^z\}$, $z\in N|$ — множество результатов всех возможных распределений для множества заданий V_J и для множества ресурсов V_R. Тогда $\Gamma(V_J,V_R)\equiv X$. Распределение заданий на ресурсы $\Gamma(V_J,V_R)\rightarrow A^*$ есть ***максимальное распределение*** для данных множества заданий V_J и множества ресурсов V_R если:

1) $n^*=|A^*|$;
2) $n^*=\max\{|A^1|, |A^2|,.., |A^z|\}$.

Определение 5: пусть Δ есть некоторая функция от назначения $a_s=(R^s, J^s)$ (то есть назначения задания J^s на ресурс R^s, $R^s\in V_R$ и $J^s\in V_J$). Тогда $\Delta(a_s)=\Phi$ или $\Phi=\Delta(R^s, J^s)$ и $\Phi_i=\Delta(a_i)=\Delta(R^i, J^i)$, где i=1..n, назовем ***весом назначения*** $a_i=(R^i, J^i)$ по Δ.

Определение 6: сумму весов всех назначений $\{a_1, a_2,.., a_n\}$ назовем ***весом D(A) расписания*** A.

То есть: $D(A)=\sum_{i=1}^{n}\Delta(a_i)$.

Определение 7: пусть $Xm=\{A_1, A_2,.., A_m\}$, $m\in N|$, есть множество всех максимальных распределений для множества заданий V_J и для множества ресурсов V_R. Тогда расписание $A^*=\Gamma(V_J,V_R)$ – ***оптимальное расписание*** распределения (заданий V_J на ресурсы V_R) Г по измерению Δ, если A* удовлетворяет следующим условиям:

1) $A^*=\{(R^1, J^1), (R^2, J^2),.., (R^n, J^n)\}$ является результатом максимального распределения для данных множества заданий V_J и множества ресурсов V_R , то есть $|A^*|=\max\{|A^1|, |A^2|,.., |A^z|\}$ (определение 5);

2) Вес расписания $A^*=\{a_1, a_2,.., a_n\}$ максимален из $Xm=\{A_1, A_2,.., A_m\}$, то есть:

$$D(A^*)=\sum_{i=1}^{n}\Delta(a_i^*)=\max\{D(A_1),$$

$$D(A_2),...,D(A_m)\}=\max_{j=1}^{m}\{D(A_j)\}$$

Требование: нужно найти ***оптимальное*** (максимальное по весу заданной функции Δ) ***расписание*** $A=\{(R^1, J^1), (R^2, J^2),.., (R^n, J^n)\}$, $n\in N|$ максимального распределения Г (по определению 7) для N_τ свободных ресурсов (V_R) и M готовых к выполнению заданий (V_J).

– Общая схема решения

Определим *модель* решения для задачи оптимизации и распределения (математическая постановка которой представлена в [2,3]) на основе модели оптимизации и распределения, представленной в [4,5].

Решение данной задачи для N_τ ресурсов $V_R=\{R_1, R_2,.., R_N\}$ и M заданий $V_J=\{J_1, J_2,.., J_M\}$ состоит из следующих этапов:

1 Определение функции Δ весов назначения. Определяются веса $\delta_{i,j}$ ($i=1..N_\tau$, $j=1..M$) всех возможных назначений по функции Δ.

2 Поиск оптимального расписания $A=\{a_1, a_2,.., a_n\}$, где $a_i=(R^i, J^i)$, $R^i\in V_R$, $J^i\in V_J$, $i=1..n$, $n\in N|$, которое удовлетворяет условиям определения 7 и весовым значениям, определенным на первом этапе.

– Определение функции измерения качества решения

При оптимизации и распределении, функцию Δ для измерения веса назначения задания J_j на ресурс R_i ($R_i\in V_R$ и $J_j\in V_J$), можно определить следующим образом:

$$\Delta(R_i,J_j)=\delta_{i,j}=$$

$$\prod_{k=1}^{K}P_k^{i,j}\times\prod_{x=1}^{H}C_x^{i,j}\times\sum_{y=1}^{G}L_yO_y^{i,j} \quad (1)$$

Где,

- $\prod_{k=1}^{K}P_k^{i,j}$ – величина приоритета назначения (R_i, J_j). Она вычисляется путем умножения величин всех K приоритетов $P_k^{i,j}\in\Re^+$ не только заданий, но и ресурсов. В приоритете учитываются разные факторы: время ожидания заданий, работоспособность ресурсов и т.д.).
- $\prod_{x=1}^{H}C_x^{i,j}$ – результат анализа H обязательных требований, $C_x^{i,j}$ определяет степень выполнения обязательного требования x для назна-

чения задания J_j на ресурс R_i, $C_x^{i,j} \in \{0,1\}$. Например, требования наличия каналов передачи, объема требуемой памяти, наличия программ, данных и т.д. $C_x^{i,j}=1$, если ресурс R_i полностью удовлетворяет требованиям задания J_j, $C_x^{i,j}=0$ в противном случае.

- $\sum_{y=1}^{G} L_y O_y^{i,j}$ – результат анализа G оптимизирующих требований, где $O_y^{i,j} \in \Re^+$ и $\mathbf{0} \le O_y^{i,j} \le \mathbf{1}$ – степень выполнения оптимизирующего требования **y** назначения задания J_j на ресурс R_i; $L_y \in \Re^+$ и $L^d \le L_y \le L^u$ – весовой коэффициент оптимизирующего требования y.

В предложенной системе представлений исходной информации имеем :

- $\prod_{k=1}^{K} P_k^{i,j}$ вычисляется с помощью приоритета JP_j задания J_j из выражения:

$$\prod_{k=1}^{K} P_k^{i,j} = \mu_i \times \rho_j,$$

где $\rho_j = JP_j = 1/Tw_j$ (Tw_j – время ожидания задания J_j в системе),

$$\mu_i = \begin{cases} M^o, если R_i = J, M_j = R^* \\ 1, если R_i = J, M_j \in \{0, R^*\} \end{cases}$$

(μ_i – маска задания).

- $\prod_{x=1}^{H} C_x^{i,j}$ вычисляется с помощью сравнения требований по коммуникациям задания $JL_j=\{(R^1, \varphi_1),..,(R^q,\varphi_q)\}$ с множеством дуг графа системы ресурсов $E_R=\{E_1, E_2,.., E_d\}$:
для $\forall l=1..q$: $CC_l^{i,j} = 1$, если $(R_i, R^l) \in E_R$;

$CC_l^{i,j} = 0$, если $(R_i, R^l) \notin E_R$;

T.e. $\prod_{x=1}^{H} C_x^{i,j} = C^{i,j} = \prod_{l=1}^{q} CC_l^{i,j}$.

$\sum_{y=1}^{G} L_y O_y^{i,j}$ вычисляется как сумма обратных величин времени выполнения $Te_{i,j}$ и времени, затрачиваемом на коммуникации $Tc_{i,j}$.

Коэффициент производительности ресурса $RE_i = k_i$ определяется из WVR_i, объем работы задания $JE_j=\varepsilon_j$ – из матрицы весов дуг графа системы ресурсов $RC[k,l]=\beta_{k,l}$, где k=1..N и l=1..N, объемы требований заданий по коммутациям из $JL_i=\{(R^1, \varphi_1),..,(R^q, \varphi_q)\}$.

Тогда $Te_{i,j}$ и $Tc_{i,j}$. вычисляются из следующих выражений:

$$Te_{i,j} = \varepsilon_j * k_i \; ; \; Tc_{i,j} = \sum_{l=1}^{q} (\varphi_l * \beta_{i,l}).$$

Таким образом, имеем:

$$\sum_{y=1}^{G} L_y O_y^{i,j} = 1/Te_{i,j} + 1/Tc_{i,j} =$$

$$1 / (\varepsilon_j * k_i) + 1 / \sum_{l=1}^{q} (\varphi_l * \beta_{i,l})$$

Из выражения (1) имеем:

$$\Delta(R_i,J_j)=\delta_{i,j}=(\mu_i \times \rho_j)*C^{i,j} *$$

$$(1 / (\varepsilon_j * k_i) + 1 / \sum_{l=1}^{q} (\varphi_l * \beta_{i,l})).$$

Очевидно, что $\delta_{i,j} \ge 0$ для $\forall i=1..N_\tau$, $j=1..M_\tau$. Поэтому $\inf(\Delta(R_i, J_j)) = 0$.

В случае отсутствия связи между ресурсами i и l, RC[i,l] получает такое значение $\beta_{i,l}$, что $Tc_{i,,j} = \sum_{l=p}^{q} (\varphi_l * \beta_{i,l}) > \lambda_0$, где λ_0 некоторое заданное число. Число λ_0 есть порог для определения существования связи между двумя ресурсами. Время выполнения $Te_{i,j}$ имеет некоторую нижнюю границу $\mathbf{T^0}$.

Верхняя граница диапазона изменения $\delta_{i,j}$ определяется следующим образом:

$$\sup(\Delta(R_i,J_j)) =$$

$$(\mu_{Oj} \times \rho_{Oj}) \times [1/T^O + 1/\lambda_O] = \delta_{max}.$$

Определенные значения $\delta_{i,j}$ для $i=1..N_\tau$, $j=1..M_\tau$ хранятся в матрице $JR[1..N_\tau, 1..M_\tau]$.

– Определение оптимального распределения

Множество N_τ ресурсов $V_R=\{R_1, R_2,.., R_{N\tau}\}$ и M заданий $V_J=\{J_1, J_2,..,J_M\}$ можно представлять как множество вершин некоторого графа G. Тогда множество неориентированных дуг $E =\{E_1, E_2,.., E_d\}$ между вершинами графа G соответствует множеству возможных назначений заданий J* на ресурсы R*. Исходная информация при такой постановке представляется в виде матрицы связности или двудольного графа (пример графа для 6-и ресурсов и 6-и заданий приведен на рис. 1). Дуга $E_k=\{R_i, J_j\}$, где $R_i \in V_R$ и $J_j \in V_J$, k=1..d, $0 \le d \le N_\tau \times M$, между вершинами J_j и R_i отсутствует только тогда, когда назначение задания J_j на ресурсе R_i является *"невозможным"*, то есть когда $\delta_{i,j} \le \delta_0$, где δ_0 есть некоторое заданное число (в данном примере δ_0=1).

ЗАДАНИЯ

РЕСУРСЫ	№	1	2	3	4	5	6
	1	3.5	0.2	7.1	0.6	6.1	0.1
	2	0	4.1	2.1	5.3	8.7	3.3
	3	2.1	5.9	0.3	0	7.2	0.7
	4	4.5	3.6	5.1	0.8	3.1	0
	5	8.1	7.1	0	0	4.9	0.4
	6	0.9	0	2.1	6.8	0	8.1

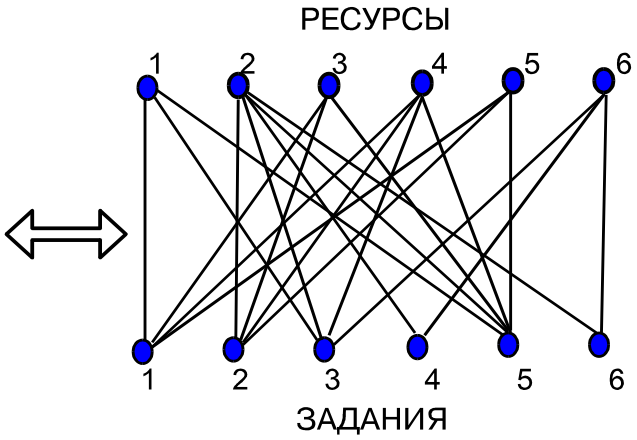

Рис.1.

Выполнение второго этапа распределения представляет собой *задачу назначения*. Существует несколько методов для решения задачи назначения для взвешенного двудольного графа

$G=(V_R,V_J, E, WE)$:

Где: $V_R=\{R_1, R_2,.., R_{N\tau}\}$ и $V_J=\{J_1, J_2,.., J_M\}$,

$E =\{E_1, E_2,.., E_d\}$, $E_k=\{R^*,J^*\}$, где $R^*\in V_R$ и $J^*\in V_J$,

$k=1..d$, $0\le d\le N_\tau\times M$.

$WE=\{WE_1, WE_2,.., WE_d\}$, $WE_k=\Delta(E_k)$,

Где: $k=1..d$, $0\le d\le N_\tau\times M$.

Решение задачи назначения для графа размером $N_\tau\times M$, где $N_\tau\ne M$ приводится к решению задачи назначения для графа размером $N\times N$, где: $N=\max\{N_\tau; M\}$

В случае планирования для однородной GRID решение задачи назначения для взвешенного графа G приводится к решению задачи назначения для невзвешенного графа G', полученного из графа G снятием весов всех дуг.

Задача назначения в такой постановке решается во многих приложениях [4,5,8,10,11]. На выбор метода и алгоритма решения влияет временная сложность, т.к. время решения задач планирования, особенно при динамическом планировании, является основным критерием.

Выделим два наиболее часто используемых подхода к решению данной задачи.

- поиск максимального потока в сети[6,7];
- поиск максимального паросочетания методом увеличивающего, чередующегося пути [7,9].

При динамическом планировании в GRID системах задача планирования сводится к поиску максимального паросочетания во взвешенном двудольном графе. Анализ методов ее решения показывает, что при решении задачи поиска максимального паросочетания в взвешенном двудольном графе используется поиск максимального паросочетания в невзвешенном двудольном графе. Поэтому задача поиска максимального паросочетания в невзвешенном двудольном графе. является ключевой и требует специального изучения, т.к. наиболее известные алгоритмы имеют временную сложность, ограничивающую их практическое использование.

Литература

1. Метод опережающего планирования для грид [Электронный ресурс] / В. Н. Коваленко, Е. И. Коваленко, Д. А. Корягин, Э. З. Любимский // Препринт ИПМ.2005. – № 112. – Режим доступа: http://www.keldysh.ru/ papers/2005/prep112/prep2005_112.html.
2. 5. Platform LSF 7 Update 6. An Overview of New Features for Platform LSF Administrators [Электронный ресурс] / Официальный сайт компании Platform Computing Corporation – 2009. – Режим доступа: http://www.platform.com/workload-management/whatsnew_lsf7u6.pdf.
3. Microsoft Windows Compute Cluster Server 2003 [Электронный ресурс] // Руководство пользователя – 2006. Режим доступа: https://msdb.ru/Downloads/ WindowsServer2003/CCS/CCS2003 Guide_Rus.pdf.
4. TORQUE Resource Manager Guide [Электронный ресурс] // Официальный сайт компании Cluster Resources Inc. – 2009 Режим доступа: http://www.clusterresources.com/products/torque-resource-manager.php.

5. PBS Works [Электронный ресурс] // Официальный сайт компании Altair Engineering, Inc. – Режим доступа: – 2006 http://www.pbsworks.com/.
6. Commercial-grade HPC workload and resource management [Электронный ресурс] // Официальный сайт компании Altair Engineering, Inc. – Режим доступа: – 2008 http:/ ww.pbsgridworks. com/Product.aspx? id=1.
7. Resource Library [Электронный ресурс] // Официальный сайт компании Altair Engineering, Inc. – Режим доступа: – 2007 http://www.pbsgridworks.com/ResLibSearchResult.aspx? Keywords = openpbs&industry=All&product_service=All&category=Free%20Software%20Downloads&order_by=date_ submitted.
8. What is Condor? [Электронный ресурс] // Официальный сайт продукта Condor - Режим доступа– 2006: http://www.cs.wisc.edu/condor/description.html.
9. Sun Grid Engine 6.2 Update 5 [Электронный ресурс] //Официальный сайт компании Oracle Corp. – Режим доступа: – 2009 http://www.sun.com/software/sge/index.xml.
10. IBM Tivoli Workload Scheduler LoadLeveler [Электронный ресурс] // Официальный сайт компании «Интерфейс» – 2007. – Режим доступа: http://www.interface.ru/home.asp?artId=6283. Maui Scheduler Administrator's Guide [Электронный ресурс] // Официальный сайт компании Cluster ResourcesInc. – Режим доступа: – 2008 http:// www. clusterresources. com/products /maui/ docs/index.shtml. – Загл. с экрана.
11. Moab Workload Manager [Электронный ресурс] // Официальный сайт компании Cluster Resources Inc. – Режим доступа: – 2008 http://www.clusterresources.com/products/moab-cluster-suite/ workload-manager.php.
12 Симоненко А.В. Выбор стратегии пространственного планирования в параллельных вычислительных системах. Вісник НТУУ «КПІ» Інформатика, управління та обчислювальна техника, Київ 2001 р. № 35.- с. 104-108.
13 Kaufmann A., Introduction a la combinatorique en vue des aplications, Dunod, Paris.(1968) .
14 Papadimitry X., Stayglitsh K., Combinatory optimization, algorithm and complexity, Moscow: Mir (1985).
15 Berge C. , Theorie des graphes et ses application. Dunod, Paris (1958).