АНАЛИЗ ПРОИЗВОДИТЕЛЬНОСТИ РАСПРЕДЕЛЕННЫХ ВЫЧИСЛИТЕЛЬНЫХ КОМПЛЕКСОВ НА ПРИМЕРЕ СИСТЕМЫ X-COM

А.С. Хританков

Введение

В данной работе рассматривается задача анализа производительности распределенных систем. Показывается, каким образом может быть использована предложенная в предыдущих работах модель производительности систем с расписанием для решения поставленной задачи анализа. Применение модели и метода анализа демонстрируется на примере системы метакомпьютинга X-Com в эксперименте по докингу протеинов. Целью анализа было понять, каким образом можно было бы уменьшить объем задействованных вычислительных ресурсов для решения задачи докинга. Для этого модель производительности была расширена, и в данной работе предложено понятие ресурсной эффективности, показывающее насколько меньше ресурсов использовала бы некоторая эталонная система для решения той же задачи, что и реальная система. В результате анализа было обнаружено, что количество использованных для решения задачи ресурсов можно было бы сократить примерно в полтора раза, если бы управление узлами системы было организовано по-другому.

Проблема определения эффективности параллельных и распределенных вычислений и ее увеличения была отчетливо сформулирована в докладе [1]. Предложенный ранее в [2-3] и развиваемый в данной работе подход позволяет частично решить поставленную проблему и указать для рассматриваемого эксперимента причины снижения эффективности вычислений и способы ее повышения без необходимости проведения детального анализа работы каждого узла. Для анализа производительности используется лишь небольшой набор строго определяемых характеристик, вычисляемых на основе непосредственно и достаточно просто измеримых параметров системы. Описываемый подход применим для широкого класса систем, от кластерных до полностью распределенных. Разумеется, предлагаемый здесь более «модельный» подход не столь точен, как подробный анализ, но является существенно менее трудоемким и позволяет получить предварительные результаты.

Подход к анализу производительности включает следующие основные этапы:

1. Формулировка цели анализа и построение эталонной системы, с которой производится сравнение реальной системы.
2. Проведение вычислительного эксперимента, который необходимо проанализировать, и сбор данных о процессе вычислений.
3. Обработка данных и расчет характеристик производительности.
4. Анализ полученных результатов.

В следующем разделе проводится краткое описание основных концепций модели систем с расписанием и вводится понятие ресурсной эффективности. Вторая часть работы посвящена анализу производительности системы X-Com с использованием предложенного подхода.

Автор статьи выражает благодарность Соболеву С.И. за предоставленный для анализа журнал работы системы X-Com.

Модель оценки производительности

Используемая в качестве основы, модель систем с расписанием была представлена в работах [2-3]. Здесь приводится краткое ее описание, используемое при дальнейшем изложении. По определению, под *распределенной вычислительной системой* понимается совокупность n вычислителей, объединенных вычислительной сетью, при этом набор вычислителей, выделенных для решения некоторой задачи A, может изменяться во время решения. Обозначим время решения задачи через T. Выделение i-го вычислителя для решения задачи задается *расписанием*, описываемым функцией $h_i(t)$. Функция $h_i(t)$, равна 1, если вычислитель выделен для решения задачи, и 0 в противном случае. Для каждого вычислителя полагается известным эталонное время $\overline{T}_i$ решения задачи A с помощью некоторого фиксированного эталонного алгоритма. *Эталонное время* $\overline{T}$ решения вычислительной системой задачи A определяется как минимальное время t, удовлетворяющее соотношению:

$$\int_0^t \sum_{i=0}^{n} \frac{h_i(\tau)}{\overline{T}_i} d\tau = 1 \quad .$$

Эталонной производительностью вычислителя называется величина $\pi_i = L / \overline{T}_i$, эталонной производительностью $\pi(t)$ системы будем называть сумму:

$$\pi(t) = \sum_{i=0}^{n} h_i(t)\,\pi_i \quad .$$

Для построенной модели определяются понятия *доступности* вычислителя $\rho_i(t)$, *эффективности* E и *коэффициента ускорения* (speedup) вычислителя S_i:

$$\rho_i(T)=\frac{\int_0^T h_i(\tau)\,d\tau}{T} \quad , \quad S_i=\frac{\overline{T}_i}{T} \quad , \quad E=\frac{\overline{T}}{T} \quad .$$

При этом справедливо соотношение

$$E=\left(\sum_{i=1}^{n}\frac{\rho_i(\overline{T})}{S_i}\right)^{-1} \quad . \qquad (1)$$

Подход к оценке производительности состоит в определении цели анализа, построении модели эталонной системы в соответствии с этой целью и последующего сравнения реальной системы с эталонной на основе характеристик эффективности и коэффициентов ускорения. Приведенные выше соотношения соответствуют универсальной эталонной системе, построенной в предположении линейного ускорения, отсутствия накладных расходов на распараллеливание и произвольно делимой задачи. Универсальная эталонная система может быть использована для исследования произвольных систем с расписанием. В данной работе предлагается дополнительная характеристика производительности — *ресурсная эффективность*, выражающая, насколько полно используются предоставленные системе вычислительные ресурсы.

Рассмотрим простую модель учета затрат на вычисления, в которой задана только удельная стоимость $c_i>0$ использования вычислителя i в единицу времени. Будем полагать, что стоимость вычислений $\Phi(t)$ системы $\boldsymbol{R}$ к моменту времени t складывается только из стоимостей использования вычислителей $\Phi_i(t)$ и выражается формулой:

$$\Phi_i(t)=\int_0^t c_i\,h_i(\tau)\,d\tau \quad ,$$

$$\Phi(t)=\sum_{i=1}^{n}\Phi_i(t) \quad .$$

Под стоимостью использования вычислителя можно понимать, в том числе, и количество условных элементарных операций, в которых измеряется трудоемкость L при фиксированном алгоритме решения задачи. В этом случае $c_i=\pi_i$, и выполняется соотношение $\Phi(\overline{T})=L$.

Ресурсной эффективностью системы $\boldsymbol{R}$ будем называть отношение стоимости вычислений эталонной системы к реальной стоимости вычислений:

$$E_\phi=\frac{\Phi(\overline{T})}{\Phi(T)} \qquad (2).$$

Функция стоимости $\Phi(t)$ не убывает по времени и $T\le\overline{T}$, поэтому ресурсная эффективность не превосходит единицы $E_\phi\le 1$. Средняя удельная стоимость использования $\phi(t)$ системы и вычислителя $\phi_i(t)$ за время t определяются соотношениями:

$$\phi(t)=\sum_{i=1}^{n}\phi_i(t) \quad ,$$

$$\phi_i(t)=\Phi_i(t)/t=c_i\rho_i(t) \quad ,$$

Отсюда следует, что средняя стоимость вычислений $\phi_i(t)$ не может превышать удельную стоимость c_i :

$$\phi_i(t)\le c_i \quad .$$

Ресурсную эффективность E_ϕ можно выразить через эффективность системы E :

$$E_\phi=E\frac{\sum_{i=1}^{n}c_i\rho_i(\overline{T})}{\sum_{i=1}^{n}c_i\rho_i(T)}=E\frac{\phi(\overline{T})}{\phi(T)} \quad .$$

Вкладом $\kappa_i(t)$ *вычислителя* i в стоимость использования будем называть отношение средней удельной стоимости вычислителя к средней удельной стоимости системы

$$\kappa_i(t)=\frac{\phi_i(t)}{\phi(t)} \quad .$$

Введем понятие, аналогичное коэффициенту ускорения. *Удешевлением* вычислений системы по отношению к вычислителю i будем называть отношение стоимости эталонного решения на вычислителе i к действительной стоимости решения в системе:

$$Z_i=\frac{\overline{\Phi}_i}{\Phi(T)} \quad ,$$

$$\overline{\Phi}_i=c_i\overline{T}_i \quad .$$

Для ресурсной эффективности справедливо соотношение, аналогичное (1):

$$E_\phi = \left(\sum_{i=1}^{n} \frac{\kappa_i(\overline{T})}{Z_i} \right)^{-1} .$$

Для прояснения смысла введения ресурсной модели эффективности рассмотрим пример системы с одним вычислителем и двумя расписаниями. Пусть имеется система с расписанием из одного вычислителя производительности $\pi_1 = 1$. Будем полагать удельную стоимость вычислений равной эталонной производительности $c_1 = \pi_1$. Предположим, что при первом запуске вычислитель сразу выделяется для решения задачи трудоемкости $L = 1$ и решает ее за время $T^{(1)} = 2$, при этом функция расписания $h_1^{(1)}(t) \equiv 1$. Эффективность системы составляет $E^{(1)} = 1/2$, ресурсная эффективность $E_\phi^{(1)} = 1/2$. Пусть при втором запуске вычислитель выделяется только через $t_0 = 1/\pi_1 = 1$, при расписании $h_1^{(2)}(t) = 1$, $t \geq t_0$, и $h_1^{(2)}(t) = 0$ в противном случае. Эталонное время решения составляет $\overline{T}^{(2)} = t_0 + L/\pi_1 = 2$. Реальное время решения задачи будет $T^{(2)} = t_0 + T^{(1)} = 3$. Эффективность системы теперь равна $E^{(2)} = 2/3$, но ресурсная остается прежней $E_\phi^{(2)} = E^{(2)} \cdot Q^{(2)} = 1/2$. Здесь $Q^{(2)}$ - *коэффициент формы расписания*, определяемый как отношение удельных стоимостей

$$Q^{(2)} = \frac{\phi(\overline{T}^{(2)})}{\phi(T^{(2)})} = \frac{1/2}{2/3} = \frac{3}{4} .$$

Форма расписания характеризует распределение стоимости вычислений относительно эталонного и реального времени решения задачи. В приведенном примере во втором случае использовалось расписание с растущей стоимостью, или, для краткости, *растущее расписание*, в котором суммарная стоимость вычислений при $t \in [\overline{T}, T]$ больше стоимости вычислений на промежутке $[0, \overline{T})$. Для растущего расписания $Q < 1$ и ресурсная эффективность E_ϕ меньше эффективности по времени E. Примером расписания с убывающей стоимостью может служить расписание для системы из двух вычислителей, которые сначала при $t \in [0, \overline{T})$ работают вместе, а потом при $t \in [\overline{T}, T]$ второй вычислитель продолжает решение задачи один. Для убывающего расписания форма расписания $Q > 1$ и ресурсная эффективность больше эффективности по времени. Для неоднородных систем $Q = 1$, поэтому ресурсная эффективность и эффективность по времени совпадают. Обратное, вообще говоря, неверно, то есть из того, что $Q = 1$, не следует, что в системе с расписанием вычислители доступны на протяжении всего времени решения. Например, для распределенной системы, работающей с эффективностью 1, ресурсная эффективность и эффективность также совпадают и коэффициент формы расписания равен 1.

Исследование производительности системы X-Com

Модель была использована при оценке производительности системы X-Com в эксперименте по докингу протеин-лиганд. Система метакомпьютинга X-Com разрабатывается в НИВЦ МГУ [4]. Система построена по иерархическому принципу и позволяет осуществлять распределенное решение полностью декомпозируемых задач, то есть задач, разбиваемых на независимые подзадачи. Рассмотренная задача докинга заключается в нахождении конфигурации молекул белка и лиганда с минимальной энергией взаимодействия. Схема решения задачи состоит в разбиении исходной задачи на множество подзадач и распределении подзадач между вычислительными узлами системы. На каждом узле системы для решения подзадач запускается клиент системы X-Com и программа SOL, непосредственно решающая подзадачу [5-6].

Для анализа был доступен журнал работы системы, любезно предоставленный разработчиками системы. Из журнала были получены данные о времени старта узлов и запроса узлами новых подзадач для решения, времени получения решения каждой подзадачи. Кроме того, из журнала были получены данные о количестве подзадач и о количестве узлов в системе. Используя доступную информацию об архитектуре системы X-Com, были сделаны следующие предположения относительно процесса решения задачи:

- Каждая подзадача имеет уникальный номер.
- Исходная задача делима на подзадачи, которые могут быть решены независимо друг от друга.
- Каждое клиентское приложение X-Com было запущено на одном вычислительном ядре.
- Узлы одного кластера имеют одинаковую производительность при решении рассматриваемой задачи докинга.
- Подзадачи распределялись между узлами без учета трудоемкости их решения.

В эксперименте было решено 5217 подзадач на более чем 1600 ядрах кластеров Скиф-МГУ, Скиф-Урал и кластера НСК-160 НГТУ. Эксперимент продолжался около 48 часов, суммарное процессорное время составило порядка 58 тысяч часов.

Можно выделить три обширных класса причин снижения эффективности:

1. *Накладные расходы.* Накладные расходы на распараллеливание, которые мы в данном расчете полагаем малыми.
2. *Распределение подзадач.* Изменение размера задачи при параллельном решении, по сравнению с последовательным. Сюда также относится повторное решение подзадач.

3. *Использование ресурсов.* Простой узлов в ожидании получения подзадачи для решения, в том числе решение задач «в холостую», когда результат решения подзадачи не был вычислен или не был передан на управляющий узел.

В данном исследовании мы проанализируем влияние третьей причины снижения эффективности. Для этого рассмотрим две эталонные системы: система $\bar{\boldsymbol{R}}_{clust}$ соответствует управлению структурой системы на уровне кластеров, система $\bar{\boldsymbol{R}}_{node}$ соответствует управлению отдельными узлами, то есть позволяет при необходимости останавливать узлы, входящие в состав системы. В действительности, узлы выделялись системе по-кластерно в монопольном режиме, что соответствует системе $\bar{\boldsymbol{R}}_{clust}$. Эталонная система $\bar{\boldsymbol{R}}_{node}$ моделирует ситуацию, когда не происходит снижения эффективности вследствие неполного использования предоставленных системе ресурсов.

Для удобства, в качестве трудоемкости решения всей задачи мы возьмем достаточно большое число $L = 10^{11}$. Согласно извлеченным из журнала работы системы данным, всего было решено $N_{unique} = 5217$ различных подзадач, общее число решенных подзадач, с повторами, составляет $N_{total} = 7525$. Полагая трудоемкость задачи неизменной при переходе к параллельному решению, нужно положить число подзадач равным $N = N_{unique}$. Для описания работы системы X-Com воспользуемся моделью интервальной системы [3]. Вычислителям модели нужно сопоставить «сессии» работы клиентского приложения X-Com с управляющим узлом. При перезапуске клиента на том же ядре, взаимодействие происходит в рамках новой сессии. Поэтому количество вычислителей в модели $n = 1664$ немного превышает число задействованных ядер. Эталонные характеристики вычислителей рассчитываются также по данным журнала следующим образом. Вычислители в рамках одного кластера будем считать одинаковыми по производительности. На каждом кластере из всех решенных на нем подзадач было выделено множество различных подзадач, получено выборочное распределение времени решения подзадачи на вычислителях кластера и рассчитано среднее время решения одной подзадачи на каждом кластере. Гистограммы выборочных распределений для кластеров Скиф-МГУ и Скиф-Урал с учетом повторно решенных подзадач приведены на рис. 1.

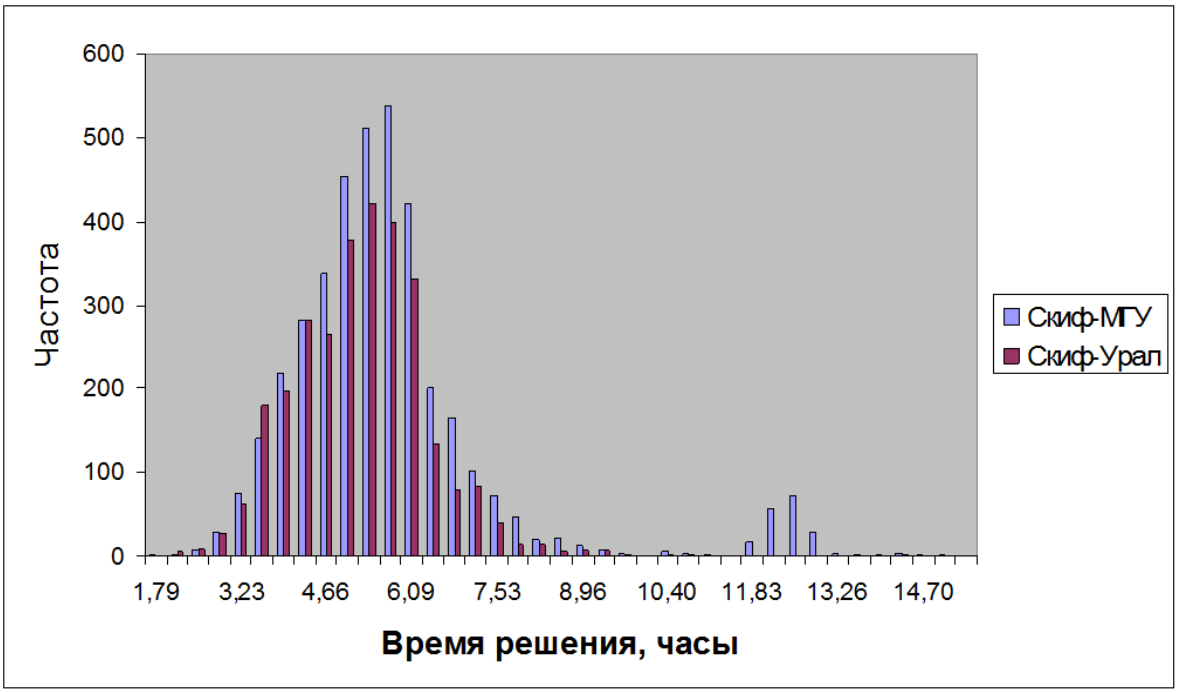

Рис. 1. Распределение длительности решения подзадач для кластеров Скиф-Урал и Скиф-МГУ.

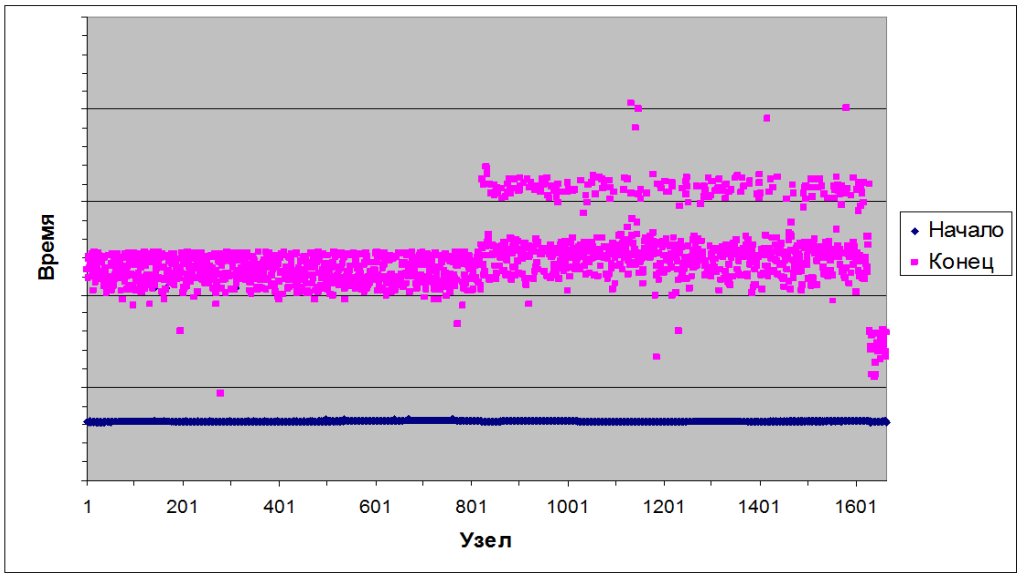

Рис. 2. Карта использования вычислителей.

Второй пик для кластера Скиф-МГУ связан с повторным решением трудоемких подзадач, при исключении последних, выборочное распределение имеет один пик для обоих кластеров, как распределение для кластера Скиф-Урал на приведенном графике.

Так как информация о действительном расписании выделения узлов недоступна, будем использовать имеющиеся данные журнала работы для составления оценочного расписания для систем $\bar{R}_{clust}$ и $\bar{R}_{node}$. Начало и окончание работы вычислителей, соответствующих сессиям, удобно представить графически. Пронумеруем вычислители сначала кластера Скиф-Урал, затем Скиф-МГУ и НСК-160. Карта использования вычислителей представлена на рис. 2.

Расписание системы $\bar{R}_{clust}$ будем считать от первого начавшего свою работу вычислителя до последнего завершившего. При этом, мы не учитываем возможность того, что вычислитель в момент завершения выделения кластера решал задачу. Тем не менее, учитывая случайный характер длительности решения подзадачи и большое количество вычислителей в кластере, максимальное по всем вычислителям время окончания дает неплохую оценку снизу времени завершения выделения кластера. Расписание системы $\bar{R}_{node}$ включает только промежутки действительного решения вычислителями подзадач и непосредственно изображено на рис. 2. Зная расписание работы вычислителей и эталонные их производительности несложно рассчитать эталонное время решения для каждой из двух систем по определению.

Положим удельную стоимость равной эталонной производительности $c_i = \pi_i$, тогда $\Phi(\bar{T}) = L$. Результаты расчета характеристик производительности по отношению к системе $\bar{\boldsymbol{R}}_{clust}$ приведены в табл. 1.

Таблица 1. Характеристики производительности по отношению к системе $\boldsymbol{R}_{clust}$.

	Скиф-Урал	Скиф-МГУ	НСК-160
$\bar{T}$	5,05 часа x 5217 порций	5,16 часа x 5217 порций	5,72 часа x 5217 порций
π_i	1,053	1,033	0,931
$h_i(t)$	с 17:20 06.11 до 19:00 07.11	с 17:25 06.11 до 17:10 08.11	с 17:20 06.11 до 07:05 07.11
S_i	~550	~560	~620
	0,535	0,998	0,287

Значения эффективностей реальной системы по отношению к эталонным системам $\bar{\boldsymbol{R}}_{clust}$ и $\bar{\boldsymbol{R}}_{node}$ приведены в табл. 2. Системы используют разные расписания, поэтому значения ресурсных эффективностей E_ϕ отличаются при тех же значениях эффективности по времени E.

Таблица 2. Производительность по отношению к различным эталонным системам

	$\vec{h}(t)$	$\bar{T}$	T	E	E_ϕ
	по кластерам	~16,3 часа	47,8 часов	0,35	0,45
	по узлам	~16,3 часа	47,8 часов	0,35	0,66

Ресурсную эффективность работы реальной системы, работающей по расписанию системы $\bar{\boldsymbol{R}}_{clust}$, по отношению к системе $\bar{\boldsymbol{R}}_{node}$ можно вычислить как отношение ресурсных эффективностей:

$$E_\phi^{node} = \frac{L/\Phi_{clust}(T)}{L/\Phi_{node}(T)} = \frac{\Phi_{node}(T)}{\Phi_{clust}(T)} = 0{,}68 \quad .$$

Данное значение можно интерпретировать следующим образом. Допустим, мы выбрали систему $\bar{\boldsymbol{R}}_{node}$ в качестве эталонной для сравнения. Система $\bar{\boldsymbol{R}}_{clust}$ отличается от нее только расписанием выделения вычислителей. Ресурсная эффективность системы $\bar{\boldsymbol{R}}_{clust}$ по отношению к системе $\bar{\boldsymbol{R}}_{node}$ показывает насколько меньше вычислительных ресурсов использует система $\bar{\boldsymbol{R}}_{node}$, чем система $\bar{\boldsymbol{R}}_{node}$, то есть на 68% в данном случае. Значит, если бы в эксперименте использовался алгоритм выделения вычислителей, дающий расписание системы $\bar{\boldsymbol{R}}_{node}$, то на решение задачи было бы затрачено в полтора раза меньше ресурсов.

Заключение

В работе рассмотрена проблема анализа производительности распределенных систем на примере работы системы метакомпьютинга X-Com в эксперименте по докингу протеинов. Для анализа производительности была использована модель систем с расписанием, предложенная ранее и расширенная в данной работе для анализа ресурсной эффективности. В результате анализа было обнаружено, что количество использованных для решения задачи ресурсов можно было бы сократить примерно в полтора раза, если бы управление узлами системы было организовано по-другому. Указанного сокращения использованных ресурсов удалось бы достичь, или, по крайней мере, приблизится к нему, если бы узлы системе предоставлялись согласно следующим правилам:

1. Выделение узлов производить через систему пакетной обработки.
2. Заказываемый интервал выделения узла следует выбирать так, чтобы большая часть подзадач могла быть решена. В данном случае, как следует из распределения длительности решения подзадач, интервал можно выбрать примерно 10 часов.
3. Если длительность интервала недостаточна для решения задачи, то заказать для данной задачи в два раза больший интервал.
4. По завершении решения задачи до окончания интервала, возвращать узел в пул системы пакетной обработки.
5. Не использовать принудительное исключение вычислителей из системы до завершения решения подзадачи, назначенной вычислителю.

При этом время решения задачи может возрасти. Однако, при сохранении такой же средней удельной стоимости вычислений $\phi(t)$ на протяжении всего процесса решения, как и при выделении узлов в монопольном режиме, время решения будет меньше за счет более эффективного использования ресурсов. По

сути, потребуется меньшее количество ресурсов и при том же их объеме, выделяемом в единицу времени, время решения будет меньше. Данное утверждение является следствием того, что $\Phi(t)$ не убывает по t и определения ресурсной эффективности (2).

Работа выполнена при поддержке гранта РФФИ № 09-07-00352-а

ЛИТЕРАТУРА:

1. Вл.В. Воеводин «Суперкомпьютерные технологии решения больших задач». // Труды IV Международной Конференции "Параллельные Вычисления и Задачи Управления", Москва, 27-29 октября 2008 г.
2. М.А. Посыпкин, А.С. Хританков «О понятии ускорения и эффективности в распределенных системах ». Материалы Всероссийской научной конференции "Научный сервис в сети Интернет", Новороссийск, 22-27 сентября 2008 г., Изд-во МГУ, с.149-155.
3. А.С. Хританков «Один алгоритм балансировки вычислительной нагрузки в распределенных системах» Материалы конференции «Параллельные Вычислительные Технологии», Нижний Новгород, 30 марта — 3 апреля 2009, Челябинск: Изд. ЮурГУ, 2009, с. 783-789.
4. Вл.В. Воеводин, Ю.А. Жолудев, С.И. Соболев, К.С. Стефанов. Эволюция системы метакомпьютинга X-Com. Материалы конференции Параллельные вычислительные технологии, Нижний Новгород, 30 марта - 3 апреля 2009, Челябинск: Изд. ЮУрГУ, 2009. с. 82-91.
5. В.Б. Сулимов, А.Н. Романов, Ф.В. Григорьев, О.А. Кондакова, С.В. Лущекина, С.И. Соболев Оценка энергии связывания белок-лиганд с учетом растворителя и программа докинга SOL: принцип работы и характеристики, Сборник материалов XIII российского национального конгресса "Человек и лекарство" 3 апреля 2006 г., 2006, с. 37.
6. С.И. Соболев. Использование распределенных компьютерных ресурсов для решения вычислительно сложных задач // Системы управления и информационные технологии. 2007, №1.3 (27). с. 391-395.