УДК 519.687.1 DOI: 10.14529/cmse190101

ОПТИМИЗАЦИЯ УТИЛИЗАЦИИ ПРИ ВЫДЕЛЕНИИ РЕСУРСОВ ДЛЯ ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ С СЕТЬЮ АНГАРА*

© 2019 А.В. Мукосей, А.С. Семенов, А.С. Симонов

АО «Научно-исследовательский центр электронной вычислительной техники»
(117587 Москва, Варшавское шоссе, д. 125, стр. 15)
E-mail: mukosey@nicevt.ru, semenov@nicevt.ru, simonov@nicevt.ru

Поступила в редакцию: 22.07.2018

В данной работе рассматривается высокоскоростная вычислительная сеть Ангара с топологией «многомерный тор». Работа посвящена оптимизации фрагментации, возникающей в результате последовательного выделения вычислительных узлов в многоузловой системе при заданном требовании о том, что сетевой трафик разных пользовательских заданий не должен пересекаться. Данная работа является продолжение работы по оптимизации фрагментации ресурсов исследуемой вычислительной системы. В данной работе к учету фрагментации при выборе узлов добавлен метод запуска пользовательских заданий, основанный на политике выбора первого подходящего задания (First-Fit) в некотором рассматриваемом окне заданий. Исследование разработанного метода проводилось с помощью симулятора работы вычислительной системы. Рассмотрен набор различных вычислительных систем с трехмерными и четырехмерными топологиями, размер минимальной системы — 32 вычислительных узла, максимальной — 144 узла. Для каждой системы задана синтетическая очередь заданий, параметры которой приближены к реально возможной и основаны на данных, полученных с вычислительного кластера Desmos на базе сети Ангара. В качестве критерия качества метода выбора узлов рассматривается средняя утилизация ресурсов вычислительной системы и среднее время ожидания заданий в очереди. Исследованы различные размеры окон заданий. Исследование показало, что увеличение утилизации ресурсов для предложенного метода выбора узлов составило в среднем 7 % и на 36,6 % сокращает значение времени ожидания задания в очереди по сравнению с базовым методом.

Ключевые слова: коммуникационная сеть Ангара, многомерный тор, планирование ресурсов, фрагментация, выбор узлов.

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Мукосей А.В., Семенов А.С., Симонов А.С. Оптимизация утилизации при выделении ресурсов для высокопроизводительных вычислительных систем с сетью Ангара // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2019. Т. 8, № 1. С. 5–19. DOI: 10.14529/cmse190101.

Введение

В АО «НИЦЭВТ» разработана высокоскоростная коммуникационная сеть Ангара [1, 2] с топологией «многомерный тор». В маршрутизаторе сети реализована бездедлоковая, адаптивная маршрутизация, основанная на правилах «пузырька» (Bubble flow control, [3]) и «порядка направлений» (Direction ordered routing, DOR, [4, 5]) с использованием битов направлений [5]. Благодаря алгоритму First Step/Last Step «нестандартного первого и последнего шага» [5] аппаратно поддерживается обход отказавших узлов или линков. Эффективность этого метода по поддержанию связности в сети с отказами была показана в статье [6].

*Работа рекомендована Программным комитетом международной конференции «Суперкомпьютерные дни в России»

В настоящий момент для сети Ангара разрабатываются алгоритмы по выделению ресурсов при условии отсутствия пересечения сетевого трафика различных заданий, при этом особое значение имеет проблема фрагментации, возникающая в результате последовательного выделения вычислительных узлов.

Для сетей с тороидальной топологией существует несколько стратегий выделения ресурсов [7]. Возможно разделение вычислительной системы на партиции, по которым размещаются задачи пользователей. Такая стратегия может снижать эффективность использования кластера из-за выделения большего числа узлов, чем требовалось, или невозможности выделить доступный набор узлов из разных партиций. Данная стратегия использовалась в суперкомпьютерах IBM Blue Gene/P, Blue Gene/Q [8, 9], в которых ее недостатки компенсировались большим числом не очень мощных по производительности вычислительных узлов и адекватным выбором размера партиции.

Вторая стратегия используется в серии суперкомпьютеров Cray XT/XE, где расположение выделенных узлов [10] не зависит от топологии. Такой способ выделения ресурсов может привести к деградации производительности ввиду наличия конкурирующего трафика.

Помимо возникающей фрагментации, на эффективное использование ресурсов влияют методы определения порядка запуска пользовательских задач. Такая задача является NP-сложной. Существует множество алгоритмов планирования, направленных на оптимизацию использования вычислительных ресурсов по разным параметрам. FCFS [11] (First Come First Served) — политика обработки очереди заданий в порядке, в котором задания поступили. Такая политика обеспечивает справедливость в отношении порядка поступления заданий, но может снижать утилизацию из-за простаивания ресурсов.

В работе [12] проводится сравнение различных вариантов алгоритмов, таких как: First Come First Served (FCFS, первым пришел — первым обслужен), Priority Queue (очередь с приоритетами), Shortest Job First (кратчайшая задача первая), Longest Job First (продолжительная задача первая) и других.

Одним из самых эффективных на данный момент алгоритмов многие авторы называют алгоритм Backfill [13] — алгоритм обратного заполнения, который является расширением политики FCFS. Для этого алгоритма необходимо наличие оценки о времени выполнения каждого задания. В алгоритме состояние системы по мере завершения работы запущенных заданий сопоставляется с очередью заданий. В статье рассматриваются консервативный вариант алгоритма, когда не допускается выполнения задания, если это повлияет на время запуска приоритетного задания, и агрессивный вариант, позволяющий запустить задание, если это не изменит времени запуска запланированного приоритетного задания.

При большом числе пользователей возникает задача справедливого распределения ресурсов, то есть избегания ситуации, когда один пользователь захватит все вычислительные ресурсы на длительное время. Различные варианты алгоритмов справедливого распределения ресурсов реализованы в программных системах управления ресурсами: PBS [14], Torque [15], MAUI [16], SGE [17], MBC-1000 [18], SLURM [19].

Данная статья является продолжением работы [20], в данной работе к методу выделения узлов с оценкой фрагментации добавлена возможность перестановки пользовательских заданий в очереди. Примененный алгоритм перестановки заданий в очереди основан на политике First-Fit (выбора первого подходящего задания); в алгоритме возможность перестановки ограничивается некоторым рассматриваемым окном заданий.

Стоит отметить, что для решения задачи планирования иногда используются генетические алгоритмы, например, в [21], однако предварительные результаты авторов данной статьи показывают, что использование данного механизма ведет к слишком большому времени работы процедуры выбора узлов.

Также в данной работе по сравнению с предыдущей работой авторов модифицирован принцип формирования очереди пользовательских заданий, который стал более приближенным к реальности и основан на данных полученных с вычислительного кластера Desmos [22].

Статья организована следующим образом. В разделе 1 приводятся необходимые формальные определения и постановка задачи. В разделе 2 описаны разработанные алгоритмы. В разделе 3 проведено исследование построенных алгоритмов.

Разработанный алгоритм выбора вычислительных узлов для потока пользовательских заданий, сокращающий фрагментацию вычислительной системы совместно с политикой поиска ресурсов, основанной на выборе первого подходящего задания в среднем дает прирост утилизации ресурсов 7 % и на 36,6 % сокращает значение времени ожидания задания в очереди по сравнению с базовым. Исследования проведены на симуляторе вычислительной системы на топологиях с числом узлов до 144. Разработан метод генерации синтетической очереди пользовательских заданий, параметры которой приближены к реально возможной.

1. Определения и постановка задачи

В данном разделе приводятся формальные определения, которые в дальнейшем будут использоваться в статье.

Рассмотрим вычислительную систему, узлы которой объединены в тороидальную топологию. Размерности тора обозначим $(d_1, d_2, ..., d_n)$, а множество всех узлов вычислительной системы обозначим $N = \{u | u = (u_1, ..., u_n), \forall i \ u_i \in \mathbb{Z}_{d_i}\}$, а общее число узлов — $|N|$. Расстояние на множестве N определим следующим образом: $L(u, v) = \sum\limits_{i=1}^{n} |u_i - v_i|, \ \forall u, v \in N$.

Состояние системы S можно описать множествами узлов, доступных и недоступных для выделения, обозначим эти множества N_{free} и N_{locked}, соответственно.

Будем называть *маршрутизируемым* множеством узлов в коммуникационной сети Ангара такое множество, что для каждого узла этого множества существует сетевой маршрут в любой другой узел множества, удовлетворяющий правилам маршрутизации сети Ангара, а также весь сетевой трафик узлов множества не выходит за его пределы.

Будем называть *заданием* W — число узлов W_{nodes}, запрашиваемое пользователем в момент времени W_{start} на время W_{time}, а *ресурсами для задания* — маршрутизируемое множество узлов, размер которого не меньше, чем W_{nodes}. *Потоком заданий* назовем множество различных заданий W.

Ранее в работе [23] авторами статьи решалась проблема поиска маршрутизируемого множества заданного размера в коммуникационной сети Ангара с учетом топологии и маршрутизации. Обозначим алгоритм, решающий эту проблему как $Find\_Systems(W, S)$. На вход этому алгоритму подается состояние системы S и задание W с требуемым числом вычислительных узлов W_{nodes}. Результатом работы алгоритма является набор вариантов ресурсов для задания. Необходимо заметить, что особенностью алгоритма является то, что все ресурсы для задания представляют собой многомерные прямоугольники.

Под *утилизацией* ресурсов U вычислительного кластера будем понимать среднее значение утилизации по всем вычислительным узлам:

$$U = \frac{\sum\limits_{i=1}^{|N|} U_i}{|N|}, U_i = \frac{T_i}{T},$$

где U_i — утилизация i-го вычислительного узла, T — время работы вычислительного кластера, T_i — полезное время работы i-го вычислительного узла.

Обозначим *значение времени нахождения задания в очереди относительно запрошенного времени* как $T^i_{delay} = \frac{Q^i}{W^i_{time}}$, где W^i_{time} — запрошенное время для задания W^i, Q^i — время ожидания задания W^i в очереди. За *среднее значение времени нахождения задания в очереди относительно запрошенного времени задания* примем $T_{mean} = \frac{\sum_{i=1}^{k} T^i_{delay}}{k} = \frac{1}{k}\sum_{i=1}^{k} \frac{Q^i}{W^i_{time}}$, где k — число различных заданий в потоке.

За оценку качества решения для потока пользовательских заданий возьмем утилизацию ресурсов вычислительного кластера и среднее значение времени нахождения задания в очереди относительно запрошенного времени. Эти характеристики используются по аналогии с работой [24].

Во введенных обозначениях проблема, которую решает данная статья, будет формулироваться следующим образом. Для заданного вычислительного кластера и последовательности заданий $Q = W^1, ..., W^k$ требуется найти ресурсы для всех заданий из последовательности, которые будут максимизировать утилизацию вычислительного кластера и минимизировать среднее значение времени нахождения задания в очереди относительно запрошенного времени.

2. Алгоритм выбора узлов

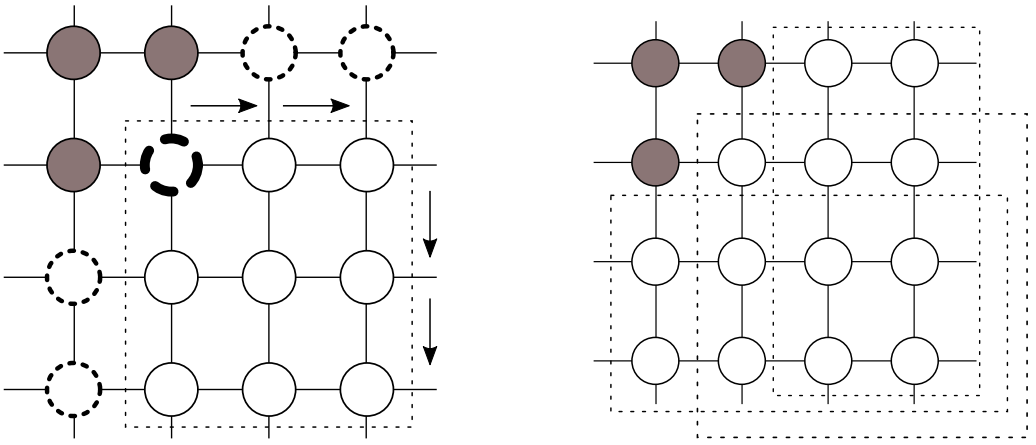
Задача упаковки контейнера является NP-полной задачей. В данной статье представлен алгоритм выбора узлов, основанный на методах, предложенных в работе [25], посвященной трехмерной упаковке контейнера. Идея алгоритма выбора узлов заключается в расположении задания таким образом, чтобы максимизировать оставшееся пространство в многомерном торе. Этот алгоритм предложен в работе авторов [20], однако для удобства восприятия приведен в данном тексте.

2.1. Алгоритм построения прямоугольников максимального размера

Назовем *прямоугольником максимально возможного размера* MSS ($MaxSpaceSize$) многомерный прямоугольник, состоящий только из узлов N_{free}, который нельзя расширить ни в одну из его сторон. Расширить прямоугольник может быть невозможно по двум причинам — либо по соответствующему измерению тора достигнуто максимальное количество узлов в кольце (расширение невозможно), либо сторона прямоугольника граничит с узлом из множества N_{locked}. Множество различных прямоугольников MSS характеризуют меру фрагментированности системы.

Алгоритм поиска различных прямоугольников MSS ($Find\_MSSs(S)$) реализован следующим образом. Из множества N_{free} выбирается узел $u_1 \in N_{free}$. Выбранный узел последовательно расширяется во все стороны, пока это возможно. Полученное множество узлов обозначим MSS_1. На следующем этапе выбирается узел $u_2 \in N_{free} \backslash MSS_1$ и аналогичным образом строится множество MSS_2. Алгоритм продолжается до тех пор,

пока множество $N_{free} \setminus \bigcup_{iter=1}^{Iters} MSS_{iter}$ не пусто, где $Iters$ — число итераций алгоритма. Псевдокод алгоритма представлен на рис. 2а. Обозначим множество различных MSS_{iter}, как $MSSs$. Важно отметить, что каждый прямоугольник строится независимо от остальных прямоугольников, в предположении доступности всех изначально свободных узлов N_{free}.



а) Построение прямоугольника максимально возможного размера

б) Все прямоугольники максимально возможного размера, построенные алгоритмом

Рис. 1. Выделение прямоугольников максимального размера

```
Input:
S -- массив, характеризующий состояние системы
S[u], может принимать следующие значения: 0 - free, 1 - locked, 2 - discovered
Output:
MSSs -- массив прямоугольников максимального размера
Find_MSS(S)
{
        Iter = 1
        dirs -- массив доступных направлений в торе, например, +x,-x,+y,-y...
        MSSs -- результирующий массив, изначально пуст
        for u in S {
                if S[u] == free {
                        MSSs[Iter].push_back(u)
                        for dir in dirs {
                                MSSs[Iter].extend(dir)
                        }
                        for v in MSSs[Iter] {
                                S[v] = discovered
                        }
                        Iter++
                }
        }
        return MSSs
}
```

Рис. 2. Псевдокод алгоритма Find_MSS поиска прямоугольников максимального размера

Иллюстрация работы алгоритма приведена на рис. 2а и рис. 2б, на которых в двумерной решетке узлы множества N_{locked} закрашены, а свободные узлы N_{free} — нет. Жирным контуром на рис. 2а выделен узел, из которого поочередным расширением построен двумерный прямоугольник, который обозначен пунктирной линией. Узлы, выделенные полужирным пунктиром, соответствуют множеству узлов N_{free}, не входящих в построенный прямоугольник. Из этих узлов будут строиться последующие прямоугольники. Все построенные прямоугольники MSS показаны на рис. 2б.

2.2. Оценка состояния вычислительной системы на основе прямоугольников максимального размера

Для оценки состояния вычислительной системы предложена функция φ, которая тем больше, чем большее число прямоугольников максимального размера имеется в системе:

$$\varphi(S) = N * MSS_{max}^{nnodes} + |MSS_{max}|,$$

где MSS_{max} — множество прямоугольников максимального размера, имеющих наибольшее число узлов MSS_{max}^{nnodes}, $|MSS_{max}|$ — число таких прямоугольников, S — текущее состояние системы.

Эта метрика была добавлена в алгоритм $Find\_Systems(W, S)$ поиска маршрутизируемого множества заданного размера. Для каждого найденного маршрутизируемого множества оценивается значение функции $\varphi(S')$, где S' – состояние вычислительной системы S после выделения узлов. Для увеличения утилизации вычислительного кластера требуется выбирать решения с наибольшим значением функции φ. Модифицированный алгоритм $Find\_Systems(W, S)$, в котором возможные варианты систем отсортированы с учетом значения функции φ, в дальнейшем будем обозначать $Find\_Systems_{MSS}(W, S)$.

2.3. Первоначальный алгоритм выбора узлов для кластеров с сетью Ангара

Алгоритм, который изначально работал на кластерах с сетью Ангара, устроен следующим образом. Для всего кластера строится таблица маршрутизации [23]. Для требуемого числа узлов W_{nodes} и допустимого числа транзитных узлов $N_{transit}$ строятся всевозможные разложения чисел $W_{nodes}, W_{nodes} + 1, ..., W_{nodes} + N_{transit}$ на n множителей, таких что $1 \leqslant p_i \leqslant d_i, \forall i \in 1..n$, где p_i — множитель разложения. Все такие разложения обозначим D. Эти разложения описывают всевозможные размеры прямоугольников, подходящих под решение задачи W. Средним диаметром прямоугольника, соответствующего разложению $D_j \in D$, назовем среднее арифметическое всех расстояний между узлами прямоугольника: $\frac{\sum_{u,v \in D_j, u!=v} L(u,v)}{|D_j|}$.

Следующий этап выбора узлов — поиск множества узлов вычислительного кластера, которое можно покрыть одним из найденных прямоугольников таким образом, чтобы в покрытии присутствовали только узлы из множества N_{free}, то есть доступные для выделения. Поиск покрытия начинается с разложений с наименьшим средним диаметром. При первом найденном решении алгоритм заканчивает свою работу.

3. Экспериментальное исследование

3.1. Симулятор вычислительного кластера

Для оценки утилизации ресурсов вычислительного кластера разработан симулятор очереди задач (заданий) и модель состояния кластера. На вход симулятору подается поток пользовательских задач $Q = W^1, ..., W^k$. На выходе выдается полное время работы всего кластера T, время работы каждого узла T_i и время предоставления ресурсов для каждого задания. Используя эти данные, можно вычислить утилизацию ресурсов вычислительного кластера U и среднее значение времени нахождения задания в очереди T_{mean}.

Введем некоторые формальные определения, необходимые для описания работы симулятора. *Очередью* симулятора Q_{now} назовем набор заданий из потока, для которых не выделялись ресурсы и время их запуска T_{start} меньше текущего симулируемого времени t. *Окном заданий* Q_{window} *размера* w назовем некоторое множество заданий таких, что $\forall W^i \in Q_{window}, i - i_{min} < w$, где i_{min} — минимальный индекс задания из множества Q_{now}. *Временем занятости узла* u системы S, назовем время, на которое узел u был выделен для некоторого задания W. В начальный момент времени $t = 0$ время занятости всех узлов равно 0. Операцией *выделения набора узлов* на время T_{alloc} назовем увеличение времени занятости для этих узлов на время T_{alloc}. *Временем изменения системы* T_S назовем время, через которое освободится хотя бы один из выделенных узлов. *Временем изменения очереди* T_{queue} назовем время, через которое хотя бы одно задание перейдет из потока заданий в очередь симулятора. Тогда *временем ожидания симулятора* T_{sleep} назовем минимальное время до изменения состояния симулятора: $T_{sleep} = min(T_S, T_{queue})$.

Алгоритм работы симулятора устроен следующим образом. Если окно заданий не пусто, симулятор выполняет процедуру поиска маршрутизируемого множества для каждого задания из окна по очереди. Если удалось найти решение, то симулятор выделяет найденные ресурсы на необходимое время, а также удаляет это задание из очереди. Если решение не было найдено, то симулятор выполняет процедуру поиска для следующего задания из окна. Если ни одно решение ни для одного задания из окна не было найдено, время симулятора сдвигается на время ожидания T_{sleep}, а время занятости каждого занятого узла u системы S уменьшается на T_{sleep}. Если очередь заданий пуста, а все узлы перешли в состояние свободных, то симулятор завершает свою работу.

3.2. Результаты исследования

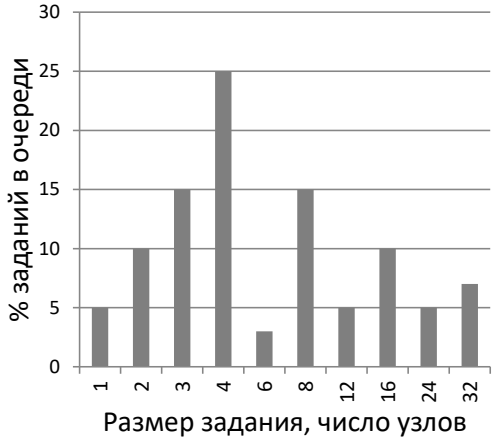
Исследование разработанного алгоритма проводилось на симуляторе для вычислительных систем, представленных в таблице.

Таблица

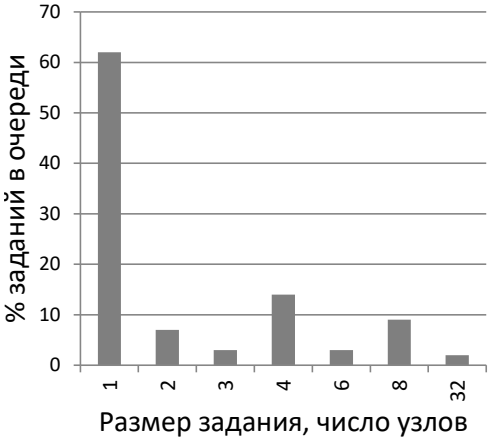
Моделируемые системы

Количество узлов	**3х-мерный тор**	**4х-мерный тор**
32	4x4x2	4x2x2x2
36	4x3x3	3x3x2x2
64	4x4x4	4x4x2x2
96	6x4x4	4x4x3x2
144	8x6x3	4x4x3x3

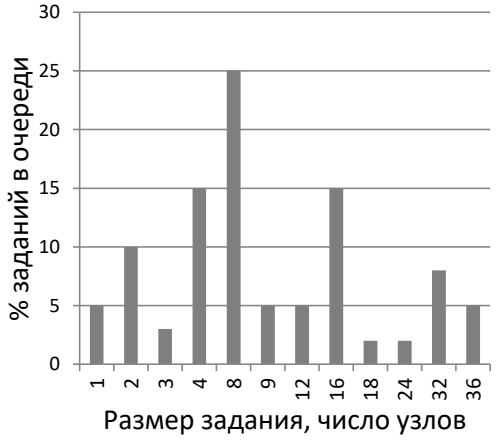
Поток заданий для каждой из систем характеризуется вероятностью появления задания для каждого числа узлов от 1 до максимального. Распределение таких вероятностей представлено на рис. 3. На рис. 2б представлено реальное распределение пользовательских задач (заданий) по количеству узлов, полученное с гибридного суперкомпьютера Desmos на базе сети Ангара, имеющую топологию 4х-мерный тор 4×2×2×2 [22]. Остальные распределения долей заданий по количеству узлов — синтетические, основанные на предположении о том, что чаще всего встречаются заданий с требуемым числом узлов, равным степеням двойки. Вероятности для остальных чисел узлов равны 0.



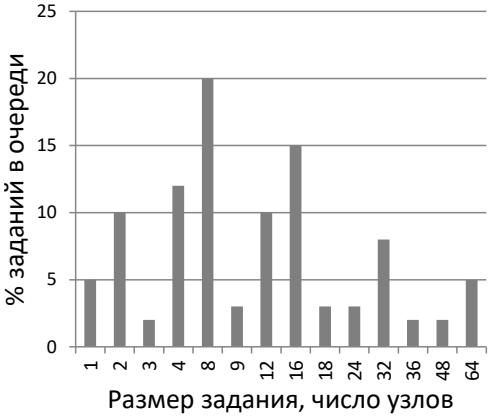
а) Для систем из 32 узлов



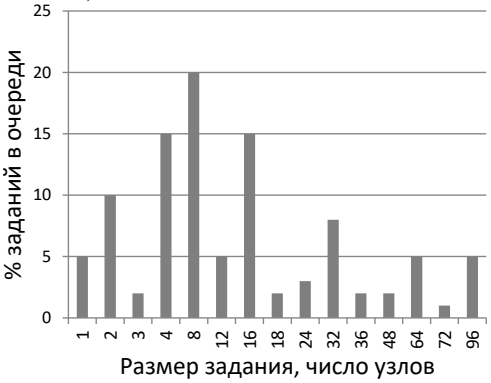
б) Для систем из 32 узлов, кластер Desmos



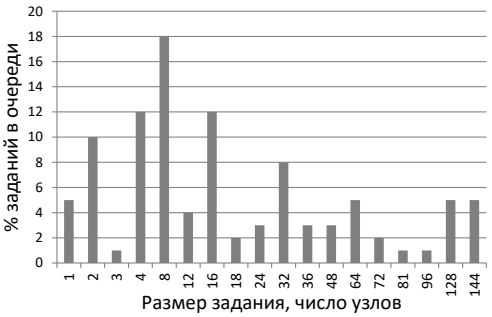
в) Для систем из 36 узлов



г) Для систем из 64 узлов



д) Для систем из 96 узлов



е) Для систем из 144 узлов

Рис. 3. Распределение заданий для систем с разным числом узлов

Задания равномерно случайно размещаются на временной шкале в диапазоне $\left[0; 60^2 * 24 * 30\right]$ для распределений полученных на кластере Desmos и на диапазоне $\left[0; 4 * 60^2 * 24 * 30\right]$ для остальных распределений вне зависимости от числа требуемых узлов.

Время продолжительности заданий соответствует распределению представленному на рис. 4, полученному в результате анализа запускаемых заданий на кластере Desmos. По оси y представлено отношение требуемого времени для задания к максимальному времени задания. На кластере Desmos максимальное время задания ограничено одними сутками. По оси x представлено процентное отношение заданий. Распределение разбито на две составляющие: на интервале $[0; 90]$ представляет из себя 10 в степени линейной функции, такой что в точке 0 оно принимает значение $0,01$, а в точке 90 — значение 99; на интервале $[90; 100]$ представляет линейную функцию и принимает значения от 99 до 100.

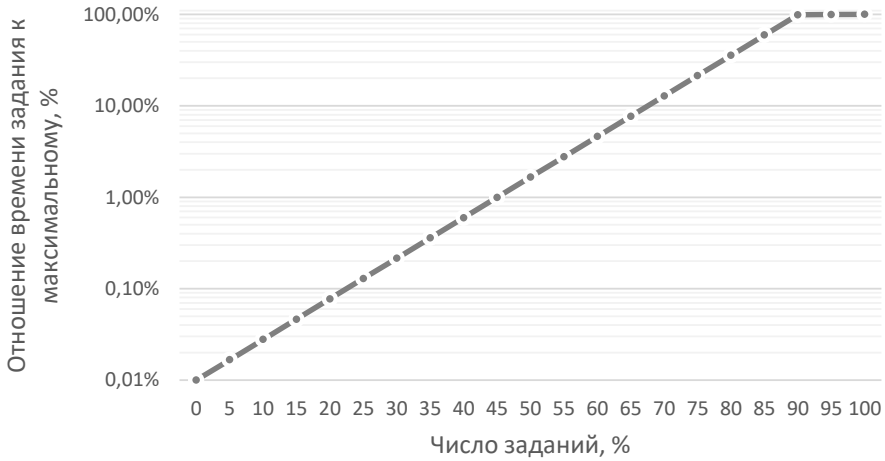


Рис. 4. Распределение времен заданий

Для исследования разработанного алгоритма поиск ресурсов для заданий производился тремя различными способами: методом $Find\_Systems$ без применения разработанной метрики ($Find\_Systems$); методом $Find\_Systems_{MSS}$ с применением разработанной метрики ($Find\_Systems + MSS$); методом, который изначально функционировал на кластерах с сетью Ангара ($base$).

Исследования проводились на окнах заданий размера $1, 2, 4, 8, 16, 32, 64, 128$.

На рис. 5 представлено среднее значение утилизации вычислительного кластера по всем системам в зависимости от размера окна. Разработанный алгоритм с применение разработанной метрикой оценки фрагментированности в данных условиях в среднем дает увеличение на $0,5$ % относительно алгоритма без учета фрагментированности системы и в среднем на 7 % относительно базового. С ростом размера окна утилизация во всех экспериментах увеличивается.

На рис. 6 представлено среднее значение времени нахождения задания в очереди по всем системам в зависимости от размера окна. Метод $Find\_Systems + MSS$ в среднем дал прирост на 2 % относительное метода $Find\_Systems$ и на $36,6$ % относительно $base$. С ростом размера окна значение времени ожидания задания в очереди в среднем значительно сокращается.

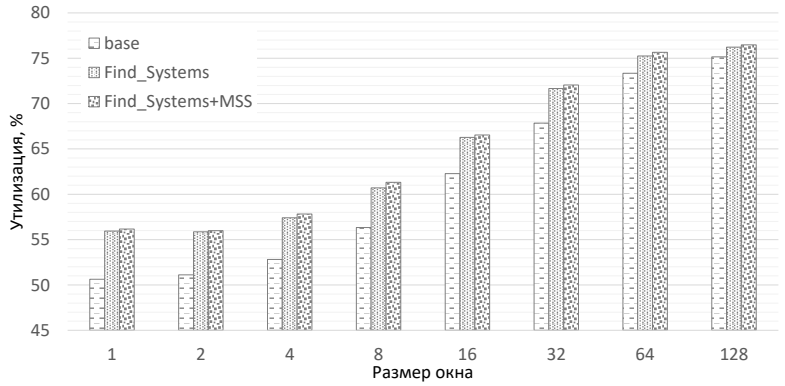


Рис. 5. Сравнение утилизации вычислительного кластера для различных систем, методов поиска и размеров окон

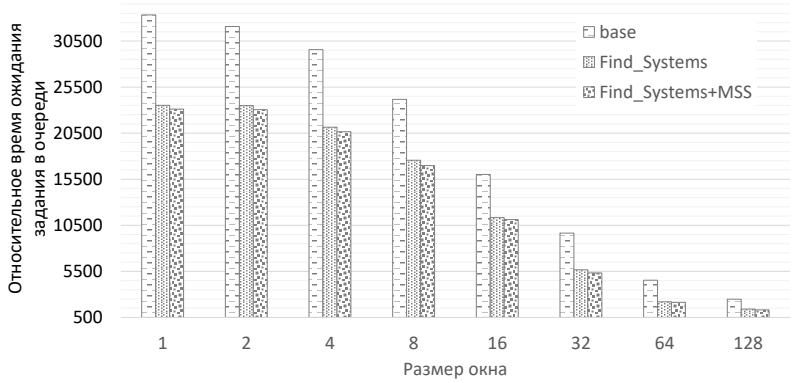


Рис. 6. Сравнение средних значений времени нахождения задания в очереди относительно запрошенного времени для различных систем, методов поиска и размеров окон

Заключение

В данной работе представлен метод выбора вычислительных узлов для потока пользовательских заданий, сокращающий фрагментацию вычислительной системы в сети Ангара с топологией многомерный тор. Метод основан на алгоритме выделения ресурсов для задания таким образом, чтобы максимизировать оставшееся пространство в многомерном торе. Это достигается построением прямоугольников максимального размера, которые возможно вписать в систему после размещения очередного пользовательского задания. Каждое множество узлов, подходящее для размещения задания, оценивается предложенной функцией, учитывающей размер и количество найденных прямоугольников максимального размера.

Проведено исследование применимости данного метода в политике поиска ресурсов для пользовательских заданий, основанной на выборе первого подходящего задания (First-Fit) в некотором рассматриваемом окне заданий.

Разработан метод генерации синтетической очереди пользовательских заданий, параметры которой приближены к реально возможной и основанны на данных, полученных с вычислительного кластера Desmos на базе сети Ангара.

Проведено экспериментальное исследование разработанного алгоритма для различных конфигураций вычислительных систем с топологией многомерный тор с общим числом узлов 32, 36, 64, 96 и 144, при этом рассмотрены 3х-мерные и 4х-мерные конфигурации топологий. Были рассмотрены различные варианты размера окна заданий.

Показана эффективность использования изменения порядка заданий для вычислительных систем с топологией многомерный тор с применением метода по оптимизации фрагментации. Разработанный метод с учет фрагментированности системы в среднем дает прирост утилизации 7 % и на 36, 6 % сокращает значение время ожидания задания в очереди по сравнению с базовым методом.

Добавление возможности изменения порядка заданий увеличивает утилизацию вычислительных ресурсов, время ожидания задания в очереди сокращается. Однако условия применения этого механизма требуют дальнейшего исследования.

Литература

1. Агарков А.А., Исмагилов Т.Ф., Макагон Д.В., Семенов А.С., Симонов А.С. Результаты оценочного тестирования отечественной высокоскоростной коммуникационной сети Ангара // Суперкомпьютерные дни в России: Труды международной конференции (Москва, 26–27 сентября 2016 г.). М.: Изд-во МГУ, 2016. С. 626–639.
2. Симонов А.С., Макагон Д.В., Жабин И.А., Щербак А.Н., Сыромятников Е.Л., Поляков Д.А. Первое поколение высокоскоростной коммуникационной сети «Ангара» // Наукоемкие технологии. 2014. Т. 15. № 1. С. 21–28.
3. Puente V., Beivide R., Gregorio J.A., Prellezo J.M., Duato J., Izu C. Adaptive Bubble Router: a Design to Improve Performance in Torus Networks // Proceedings of the International Conference Parallel Processing (ICPP). 1999. P. 58–67. DOI: 10.1109/ICPP.1999.797388.
4. Adiga N.R., Blumrich M., Chen D. Blue Gene/L Torus Interconnection Network // IBM Journal of Research and Development. 2005. Vol. 49. No. 2. P. 265–276. DOI: 10.1147/rd.492.0265.
5. Scott S.L. The Cray T3E Network: Adaptive Routing in a High Performance 3D Torus. 1996.
6. Пожилов И.А., Семенов А.С., Макагон Д.В. Алгоритм определения связности сети с топологией «многомерный тор» с отказами для детерминированной маршрутизации // Программная инженерия. 2015. № 3. С. 13–19.
7. Lan Z., Tang W., Wang J., Yang X., Zhou Z., Zheng X. Balancing Job Performance with System Performance via Locality-aware Scheduling on Torus-connected Systems // 2014 IEEE International Conference on Cluster Computing (CLUSTER). 2014. P. 140–148. DOI: 10.1109/CLUSTER.2014.6968751.
8. IBM Redbooks Publication: IBM System Blue Gene Solution: Blue Gene/Q System Administration. 2013. 282 p.
9. Tang W., Lan Z., Desai N., Buettner D., Yu Y. Reducing Fragmentation on Torus-Connected Supercomputers // Proceedings of the 2011 IEEE International Parallel Distributed Processing Symposium (IPDPS'11). IEEE Computer Society, Washington, DC, USA. 2011. P. 828–839 DOI: 10.1109/IPDPS.2011.82.
10. Cray Document: Managing System Software for Cray XE and Cray XT Systems. 2010.
11. Schwiegelshohn U., Yahyapour R. Analysis of First-Come-First-Serve Parallel Job Scheduling // SODA. 1998. Vol. 98. P. 629–638.
12. Полежаев П.Н. Исследование алгоритмов планирования параллельных задач для кластерных вычислительных систем с помощью симулятора // Параллельные

вычислительные технологии (ПаВТ'2010): Труды международной конференции (Уфа, 29 марта–2 апреля 2010 г.). Челябинск: Издательский центр ЮУрГУ, 2010. С. 287–298.

13. Mu'alem A.W., Feitelson D.G. Utilization, Predictability, Workloads, and User Runtime Estimates in Scheduling the IBM SP2 with Backfilling // IEEE Transactions on Parallel and Distributed Systems. 2001. Vol. 12. No. 6. P. 529–543. DOI: 10.1109/71.932708.

14. Henderson R.L. Job Scheduling Under the Portable Batch System // Workshop on Job Scheduling Strategies for Parallel Processing. Springer, Berlin, Heidelberg, 1995. P. 279–294.

15. Staples G. TORQUE Resource Manager // Proceedings of the 2006 ACM/IEEE Conf. on Supercomputing. ACM, 2006. P. 8.

16. Jackson D., Snell Q., Clement M. Core Algorithms of the Maui Scheduler // Workshop on Job Scheduling Strategies for Parallel Processing. Springer, Berlin, Heidelberg, 2001. P. 87–102.

17. Gentzsch W. Sun Grid Engine: Towards Creating a Compute Power Grid // Cluster Computing and the Grid, 2001. Proceedings. First IEEE/ACM International Symposium on. IEEE, 2001. P. 35–36.

18. Баранов А.В., Смирнов С.В., Храмцов М.Ю., Шарф С.В. Модернизация СУПЗ МВС-1000 // Материалы Всероссийской научной конференции «Научный сервис в сети Интернет». Новороссийск, 2008.

19. SchedMD L. L. C. SLURM Workload Manager. 2018. https://slurm.schedmd.com/overview.html (дата обращения: 20.09.2018)

20. Мукосей А.В., Семенов А.С. Оптимизация фрагментации при выделении ресурсов для высокопроизводительных вычислительных систем с сетью Ангара // Параллельные вычислительные технологии (ПаВТ'2018): Труды международной научной конференции (Ростов-на-Дону, 2–6 апреля 2018 г.). Челябинск: Издательский центр ЮУрГУ, 2018. С. 310–318.

21. Woo S.H. Task Scheduling in Distributed Computing Systems with a Genetic Algorithm // High Performance Computing on the Information Superhighway. 1997. HPC Asia'97. IEEE. 1997. P. 301–305.

22. Вечер В.С., Кондратюк Н.Д., Смирнов Г.С., Стегайлов В.В. Гибридный суперкомпьютер на базе сети Ангара для задач вычислительного материаловедения // Суперкомпьютерные дни в России: Труды международной конференции (Москва, 25–26 сентября 2017 г.). М.: Изд-во МГУ, 2017. С. 557–571.

23. Мукосей А.В., Семенов А.С., Приближенный алгоритм выбора оптимального подмножества узлов в коммуникационной сети Ангара с отказами // Вычислительные методы и программирование. 2017. Т. 18. С. 53–64.

24. Баранов А.В., Киселёв Е.А., Ляховец Д.С. Квазипланировщик для использования простаивающих вычислительных модулей многопроцессорной вычислительной системы под управлением СУППЗ // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2014. Т. 3. № 4. С. 75–84. DOI: 10.14529/cmse140405.

25. Gonçalves J.F., Resende M.G.C. A Parallel Multi-population Biased Random-key Genetic Algorithm for a Container Loading Problem // Computers & Operations Research. February 2012. Vol. 39. No. 2. P. 179–190. DOI: 10.1016/j.cor.2011.03.009.

Мукосей Анатолий Викторович, научный сотрудник, сектор управления разработки вычислительной техники, акционерное общество «Научно-исследовательский центр электронной вычислительной техники» (Москва, Российская Федерация)

Семенов Александр Сергеевич, к.т.н., зам. начальника отдела архитектуры и программного обеспечения суперкомпьютеров, акционерное общество «Научно-исследовательский центр электронной вычислительной техники» (Москва, Российская Федерация)

Симонов Алексей Сергеевич, к.т.н., первый заместитель генерального директора АО «Научно-исследовательский центр электронной вычислительной техники» (Москва, Российская Федерация)

DOI: 10.14529/cmse190101

ALLOCATION OPTIMIZATION FOR REDUCING RESOURCE UTILIZATION IN ANGARA HIGH-SPEED INTERCONNECT

© 2019 A.V. Mukosey, A.S. Semenov, A.S. Simonov

JSC "NICEVT"

(Varshavskoye shosse 125, building 15, Moscow, 117587 Russia)

E-mail: mukosey@nicevt.ru, semenov@nicevt.ru, simonov@nicevt.ru

Received: 22.07.2018

This paper considers a high-speed interconnect with a multidimensional topology. The paper is devoted to the optimization of fragmentation resulting from sequential allocation of computing nodes in a supercomputer provided that network traffic from different user's tasks should not overlap. This paper is the continuation of resources fragmentation optimization work. In this work, the method for scheduling tasks based on the policy of choosing the first suitable task (First-Fit) in a certain task window has been added to the accounting for fragmentation when choosing nodes. A set of different computer systems with three-dimensional and four-dimensional topologies was considered. The minimum system size is 32 computing nodes, and the maximum is 144. A synthetic queue of tasks is set for each system. The parameters of the synthetic queues are close to real ones and are based on data received from the Desmos cluster equipped with Angara interconnect. The average utilization of the resources of the computer system and the average waiting time for the tasks in the queue is chosen as a method quality criterion. Various sizes of task windows have been evalauated. The study showed that the increase of the resources utilization for the proposed method averaged 7 % compared to the base method, and the average time spent in queue was reduced by 36.6 %.

Keywords: Angara interconnect, multidimensional torus, deterministic routing, direction ordered routing, fragmentation, allocation.

FOR CITATION

Mukosey A.V., Semenov A.S., Simonov A.S. Allocation Optimization for Reducing Resource Utilization in Angara High-speed Interconnect. *Bulletin of the South Ural State University. Series: Computational Mathematics and Software Engineering.* 2019. vol. 8, no. 1. pp. 5–19. (in Russian) DOI: 10.14529/cmse190101.

References

1. Agarkov A.A., Ismagilov T.F., Makagon D.V. Performance Evaluation of the Angara Interconnect. *Superkomp'yuternye dni v Rossii: Trudy mezhdunarodnoi konferentsii (Moskva, 26–27 sentyabrya 2016)* [Russian Supercomputing Days: Proceedings of the International Scientific Conference (Moscow, Russia, September, 26–27, 2016)]. Moscow, Publishing of Moscow State University, 2016. pp. 626–639. (in Russian)

2. Simonov A.S., Makagon D.V., Zhabin I.A., Shcherbak A.N., Syromyatnikov E.L., Polyakov D.A. The First Generation of Angara High-Speed Interconnect. *Naukoemkie tekhnologii* [Science Intensive Technologies]. 2014. vol. 15. no. 1. pp. 21–28. (in Russian)

3. Puente V., Beivide R., Gregorio J.A., Prellezo J.M., Duato J., Izu C. Adaptive Bubble Router: a Design to Improve Performance in Torus Networks. Proceedings of the International Conference Parallel Processing (ICPP). 1999. pp. 58–67. DOI: 10.1109/ICPP.1999.797388.

4. Adiga N.R., Blumrich M., Chen D.. Blue Gene/L Torus Interconnection Network. IBM Journal of Research and Development. 2005. vol. 49. no. 2. pp. 265–276. DOI: 10.1147/rd.492.0265.

5. Scott S.L., et al. The Cray T3E Network: Adaptive Routing in a High. Performance 3D Torus. 1996.

6. Pozhilov I.A., Semenov A.S., Makagon D.V., Connectivity Problem Solution for Direction Ordered Deterministic Routing in nD Torus. Software Engineering. 2015. no. 3. pp. 13–19. (in Russian)

7. Lan Z., Tang W., Wang J., Yang X., Zhou Z., Zheng X. Balancing job Performance with System Performance via Locality-aware Scheduling on Torus-connected Systems. 2014 IEEE International Conference on Cluster Computing (CLUSTER). 2014. pp. 140–148. DOI: 10.1109/CLUSTER.2014.6968751.

8. IBM Redbooks Publication: IBM System Blue Gene Solution: Blue Gene/Q system administration. 2013. 282 p.

9. Tang W., Lan Z., Desai N., Buettner D., Yu Y. Reducing Fragmentation on Torus-Connected Supercomputers. In Proceedings of the 2011 IEEE International Parallel Distributed Processing Symposium (IPDPS'11). IEEE Computer Society, Washington, DC, USA. 2011. pp. 828–839 DOI: 10.1109/IPDPS.2011.82.

10. Cray Document: Managing System Software for Cray XE and Cray XT Systems. 2010.

11. Schwiegelshohn U., Yahyapour R. Analysis of First-Come-First-Serve Parallel Job Scheduling. SODA. 1998. vol. 98. pp. 629–638.

12. Polezhaev P.N. The Study of Parallel Job Scheduling Algorithms for Cluster Computing Systems Using a Simulator. *Parallelnye vychislitelnye tekhnologii (PaVT'2010): Trudy mezhdunarodnoj nauchnoj konferentsii (Ufa, 29 marta–2 aprelya 2010)* [Parallel Computational Technologies (PCT'2010): Proceedings of the International Scientific Conference (Ufa, Russia, March, 29–April, 2, 2010)]. Chelyabinsk, Publishing of the South Ural State University, 2010. pp. 287—298. (in Russian)

13. Mu'alem A.W., Feitelson D.G. Utilization, Predictability, Workloads, and User Runtime Estimates in Scheduling the IBM SP2 with Backfilling. IEEE Transactions on Parallel and Distributed Systems. 2001. vol. 12. no. 6. pp. 529–543. DOI: 10.1109/71.932708.

14. Henderson R.L. Job Scheduling Under the Portable Batch System. Workshop on Job Scheduling Strategies for Parallel Processing. Springer, Berlin, Heidelberg, 1995. pp. 279–294.

15. Staples G. TORQUE Resource Manager. Proceedings of the 2006 ACM/IEEE conference on Supercomputing. ACM, 2006. pp. 8.

16. Jackson D., Snell Q., Clement M. Core Algorithms of the Maui Scheduler. Workshop on Job Scheduling Strategies for Parallel Processing. Springer, Berlin, Heidelberg, 2001. pp. 87–102.

17. Gentzsch W. Sun Grid Engine: Towards Creating a Compute Power Grid. Cluster Computing and the Grid, 2001. Proceedings. First IEEE/ACM International Symposium on. IEEE, 2001. pp. 35–36.

18. Baranov A.V., Smirnov S.V., Khramtsov M.Yu., Sharf S.V. *Modernizatsiya SUPZ MVS-1000* [Modernization of the SUPZ MBS-1000]. *Materialy Vserossiiskoi nauchnoi konferentsii "Nauchnyi servis v seti Internet"* [Materials of the All-Russian Scientific Conference "Scientific Service on the Internet"]. Novorossiysk. 2008.

19. SchedMD L. L. C. SLURM Workload Manager. 2018. https://slurm.schedmd.com/overview.html (accessed: 20.09.2018)

20. Mukosey A.V., Semenov A.S. Allocation Optimization for Reducing Resource Fragmentation in Angara High-speed Interconnect. *Parallelnye vychislitelnye tekhnologii (PaVT'2010): Trudy mezhdunarodnoj nauchnoj konferentsii (Rostov-na-Donu, aprel' 2–6 2018)* [Parallel Computational Technologies (PCT'2018): Proceedings of the International Scientific Conference (Rostov-na-Donu, Russia, April, 2–6, 2018)]. Chelyabinsk, Publishing of the South Ural State University, 2018. pp. 310–318. (in Russian)

21. Woo S.H. Task Scheduling in Distributed Computing Systems with a Genetic Algorithm. High Performance Computing on the Information Superhighway. 1997. HPC Asia'97. IEEE. 1997. pp. 301–305.

22. Vecher V.S., Kondratyuk N.D., Smirnov G.S., Stegailov V.V. Angara-based hybrid supercomputer for efficient acceleration of computational materials science studies. *Superkomp'yuternye dni v Rossii: Trudy mezhdunarodnoj konferentsii (Moskva, sentyabr' 25–26 2017)* [Russian Supercomputing Days: Proceedings of the International Conference (Moscow, Russia, September, 25–26, 2017)]. Moscow, Publishing of Moscow State University, 2017. pp. 557–571. (in Russian)

23. Mukosey A.V., Semenov A.S. An Approximate Algorithm for Choosing the Optimal Subset of Nodes in the Angara Interconnect with Failures. Numerical methods and Programming. 2017. vol. 18. pp. 53–64. (in Russian)

24. Baranov A.V., Kiselev E.A., Lyakhovets D.S. The Quasi Scheduler for Utilization of Multiprocessing Computing System's Idle Resources Under Control of the Management System of the Parallel Jobs. *Vestnik Yuzho-Uralskogo gosudarstvennogo universiteta. Seriya "Matematicheskoe modelirovanie i programmirovanie"* [Bulletin of South Ural State University. Series: Mathematical Modeling, Programming & Computer Software]. 2014. vol. 3. no. 4. pp. 75–84. (in Russian) DOI: 10.14529/cmse140405.

25. Gonçalves J.F., Resende M.G.C. A Parallel Multi-Population Based Random-key Genetic Algorithm for a Container Loading Problem. Computers & Operations Research. February 2012. vol. 39. no. 2. pp. 179–190. DOI: 10.1016/j.cor.2011.03.009.