Планирование заданий в территориально распределенной системе с абсолютными приоритетами

А. В. Баранов, А. И. Тихомиров*

Межведомственный суперкомпьютерный центр — филиал ФНЦ НИИ системных исследований РАН, Москва, Россия

*Контактный e-mail: TEMA4277@rambler.ru

Исследуется модель территориально распределенной вычислительной системы с абсолютными приоритетами и непрогнозируемым временем выполнения заданий. Для исследуемой модели сформулированы требования и подходы к планированию заданий, определены показатели эффективности планирования, разработан алгоритм планирования заданий.

Ключевые слова: грид, абсолютные приоритеты, управление ресурсами, время выполнения задания, планирование.

Введение

Для повышения производительности и надежности расчетов отдельные высокопроизводительные вычислительные установки (ВУ) нередко объединяются в территориально распределенные системы (ТРС). Для такого объединения чаще всего применяются грид-технологии [1–3]. Основной единицей вычислительной работы в таких системах является вычислительное задание, под которым будем понимать набор, включающий входные данные, программу их обработки и паспорт задания. Паспорт задания — специальный объект, описывающий ресурсные требования задания: количество процессоров (ядер), объем оперативной памяти и дискового пространства, заказанное время выполнения задания и др. Паспорт задания подготавливается пользователем — владельцем задания. Различные задания от разных пользователей образуют один или несколько потоков заданий.

В территориально распределенной системе всегда присутствует как минимум один, так называемый глобальный, поток заданий. Задания глобального потока могут быть обработаны на любой вычислительной установке из состава ТРС (подмножества ВУ ТРС). В то же время на каждую отдельную ВУ может поступать локальный поток, задания из которого допускают обработку только на локальных ресурсах ВУ и не могут быть распределены на другие вычислительные установки. Таким образом, на вычислительные ресурсы каждой установки поступает два потока заданий: глобальный и локальный. Жесткое разделение ресурсов ВУ между потоками приводит к фрагментации ресурсов и их простою [4]. В связи с этим задания из обоих потоков обычно могут быть распределены на любые доступные ресурсы вычислительной установки.

Очередность использования вычислительных ресурсов ВУ устанавливается с помощью приоритетов заданий. При этом могут применяться как относительные, так и абсолютные приоритеты. Использование относительных приоритетов позволяет определить место задания в очереди без вытеснения с решающего поля ВУ выполняющихся заданий. Абсолютные приоритеты допускают прерывание выполнения (вытеснение) менее приоритетных заданий.

Проведенный авторами анализ существующих подходов к построению территориальной распределенной системы показал, что в большинстве случаев исследователи рассматривают модели ТРС с равными или относительными приоритетами заданий. В отличие от распространенных подходов, авторами предпринимается попытка создания ТРС с абсолютными приоритетами. Настоящая статься посвящена начальному этапу этой работы.

1. Модель территориально распределенной системы с абсолютными приоритетами

В предложенной авторами модели ТРС важное место занимает глобальная система управления ресурсами (ГСУР). Основной функцией ГСУР является диспетчеризация глобального потока заданий [5], что в общем случае предполагает:

- глобальное планирование доступных вычислительных ресурсов ТРС;
- поддержание глобальной очереди вычислительных заданий;
- определение вычислительной установки для выполнения задания — так называемой целевой ВУ;
- доставку входных данных на целевую вычислительную установку;
- мониторинг состояния ВУ из состава ТРС и распределенных заданий.

Глобальная система управления ресурсами основана на децентрализованной схеме диспетчеризации заданий, не предусматривающей наличие единого диспетчера, осуществляющего распределение заданий. При такой схеме диспетчеры равноправны и распределены по всем вычислительным устройствам ТРС. Взаимодействие диспетчеров осуществляется с использованием общего пула [6] — информационной системы (ИС). Применение децентрализованной схемы позволяет обеспечить высокий уровень надежности и масштабируемости системы по сравнению с централизованной и иерархической схемами [4].

В качестве ВУ ТРС рассматриваются вычислительные кластеры. Под вычислительным кластером понимается параллельная масштабируемая вычислительная система, включающая набор высокопроизводительных вычислительных модулей, объединенных коммуникационными сетями, и находящаяся под управлением единой локальной системы управления ресурсами (ЛСУР) [7–9]. В качестве ЛСУР может выступать одна из распространенных в настоящее время систем пакетной обработки — PBS, Load Leveler, SLURM, Moab, или система управления прохождением параллельных заданий (СУППЗ). Локальная система управления ресурсами выполняет следующие функции:

- прием локального потока заданий;
- ведение локальной очереди заданий;
- выделение локальных ресурсов для выполнения вычислительных заданий;
- освобождение выделенных ресурсов после завершения задания.

Таким образом, исследуемая модель ТРС представляет собой грид-систему с многоуровневым планированием вычислительных ресурсов (рис. 1): на уровне локальных

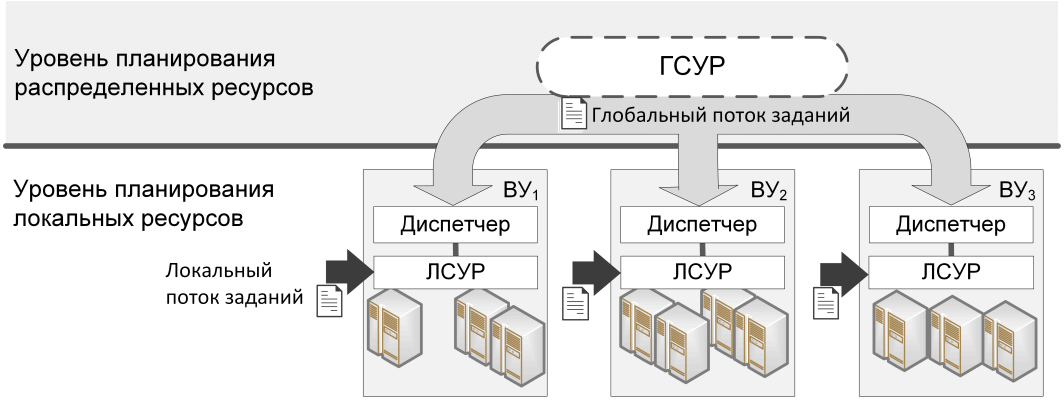

Рис. 1. Модель территориально распределенной системы с абсолютными приоритетами

ресурсов планирование осуществляется ЛСУР, на уровне распределенных ресурсов — системой диспетчеров, размещенных на каждой вычислительной установке.

В процессе исследования в качестве примера локальной системы управления ресурсами авторами была применена СУППЗ — отечественная система пакетной обработки заданий, разработанная в ИПМ им. М.В. Келдыша РАН и МСЦ РАН. Одной из отличительных черт СУППЗ является поддержка механизма фоновых заданий, допускающих в процессе своего выполнения многократные прерывания с возвращением в очередь и последующим новым запуском с восстановлением состояния выполнения.

Статистика работы СУППЗ в МСЦ РАН за 2016 г. показала, что фоновые задания расходуют до 20 % высокопроизводительных вычислительных ресурсов. При этом механизм фоновых заданий применяет сравнительно небольшое число пользователей, но, несмотря на это, фоновые задания расходуют существенную долю суперкомпьютерных ресурсов и для расширения применения механизма фоновых заданий существует значительный потенциал. Можно сделать вывод, что применение механизма фоновых заданий позволит рассмотреть модель ТРС с абсолютными приоритетами и организовать планирование заданий с вытеснением.

Таким образом, отметим две главных отличительных особенности исследуемой модели. Первая — планирование заданий осуществляется с использованием единой для всех вычислительных устройств территориально распределенной системы абсолютных приоритетов. Вторая особенность, логически следующая из первой, заключается в том, что использование абсолютных приоритетов делает затруднительным составление расписания запусков заданий [4, 6]. Составленное расписание придется перестраивать каждый раз при поступлении в систему высокоприоритетного задания, причем с возможным вытеснением выполняющихся заданий с более низким приоритетом. Возможность прерывания выполнения задания делает непрогнозируемым время его выполнения, и планирование следует осуществлять с учетом этого факта. Необходимость при планировании учитывать непрогнозируемое время также еще обосновывается тем, что по статистике МСЦ РАН более 60 % заданий завершаются раньше заказанного времени не менее чем на час.

Система абсолютных приоритетов предполагает, что основным показателем эффективности планирования является минимизация среднего времени обработки высокоприоритетных заданий. Под временем обработки понимается временной интервал от момента поступления задания в ТРС до момента окончания выполнения на вычисли-

тельных ресурсах ВУ. Время обработки задания может быть поделено на несколько этапов:

- время пребывания задания в глобальной очереди ($T_{\text{г.о}}$);
- время доставки задания (включая передачу входных данных) в целевую ВУ ($T_{\text{д}}$);
- время пребывания задания в локальной очереди целевой ВУ ($T_{\text{л.о}}$);
- время выполнения задания на вычислительных ресурсах целевой ВУ ($T_{\text{в}}$).

Время обработки задания рассчитывается по формуле

$$T_{\text{о.з}} = T_{\text{г.о}} + T_{\text{д}} + T_{\text{л.о}} + T_{\text{в}}. \tag{1}$$

Время пребывания высокоприоритетного задания в глобальной и локальных очередях может быть сокращено за счет распределения высокоприоритетного задания в ВУ, вычислительные ресурсы которого заняты обработкой менее приоритетных заданий. Так, при поступлении в систему задания с приоритетом, превосходящим приоритеты остальных заданий, часть заданий с меньшим приоритетом прерывает свое выполнение и освобождает занимаемые вычислительные ресурсы под высокоприоритетное задание. Приостановленные задания смогут продолжить свое выполнение после завершения высокоприоритетного задания или на других подходящих вычислительных ресурсах, если те станут свободными.

Для сокращения времени доставки задания на целевую вычислительную установку необходимо при распределении высокоприоритетных заданий выбирать установку, время передачи задания в которую будет минимальным. Среди эквивалентных по времени доставки задания ВУ следует выбирать вычислительную установку с наивысшей производительностью.

Таким образом, критерий эффективности планирования в ТРС с абсолютными приоритетами может быть определен следующим образом:

$$T_{\text{о.з}}(T_{\text{г.о}}, T_{\text{д}}, T_{\text{л.о}}, T_{\text{в}}) \longrightarrow \min. \tag{2}$$

2. Алгоритм планирования заданий в ТРС с абсолютными приоритетами

При осуществлении планирования заданий из глобального потока ТРС часто применяются следующие методы и алгоритмы.

- Алгоритмы, разработанные для планирования территориально распределенных ресурсов: алгоритм на основе экономических принципов [10, 11], алгоритм опережающего планирования.
- Алгоритмы, адаптированные под планирование в ТРС: алгоритм равномерного планирования Round-Robin, алгоритм обратного заполнения Backfill [12], алгоритм справедливого распределения ресурсов Fairshare [13] и алгоритм FCFS.
- Методы планирования территориально распределенных ресурсов с использованием стратегий планирования. Использование стратегий планирования предполагает выбор алгоритма планирования в зависимости от наступивших в системе событий. Примером стратегии планирования ТРС может служить циклическая схема планирования (ЦСП) [7].

Стоит отметить, что ни один из указанных алгоритмов не может быть непосредственно применен для построения ТРС с абсолютными приоритетами, в связи с этим

авторами предложен собственный алгоритм планирования, при разработке которого использованы следующие подходы.

1. *Событийно-ориентированный подход.* Алгоритм начинает выполнение при возникновении в ТРС новых событий, наиболее важными из которых являются освобождение ресурсов и появление новых заданий. Запуски алгоритма независимы друг от друга. Подобный подход применим в алгоритмах планирования ЛСУР.
2. *Двухэтапное планирование.* Алгоритм осуществляет планирование в два этапа [6]. Первый этап возникает в момент добавления задания в ГСУР. На этом этапе производится анализ, какие вычислительные установки ТРС могут выполнить глобальное задание, т. е. формируется список альтернатив ВУ, удовлетворяющих ресурсным требованиям задания. При отсутствии подходящих вычислительных ресурсов поступившее задание помещается в глобальную очередь [4]. Второй этап начинается в момент распределения задания в целевую ВУ, на котором рассматривается каждая вычислительная установка из списка альтернатив и осуществляется анализ целесообразности размещения задания в рассматриваемую ВУ.
3. *Контроль загруженности ВУ ТРС.* При распределении задания важно учитывать загруженность вычислительных установок ТРС. В общем случае загруженность ВУ ТРС оценивается как наличие свободных вычислительных ресурсов. В системе с абсолютными приоритетами для оценивания загруженности ВУ такой информации недостаточно. Дополнительно требуется анализировать максимальный приоритет заданий, находящихся в процессе выполнения на ВУ.
4. *Высокоприоритетное планирование.* Задание может быть распределено на ВУ ТРС только в том случае, если его приоритет строго больше максимального приоритета заданий этой установки. В случае если на всех вычислительных установках ТРС происходит обработка заданий, приоритет которых выше планируемого, планируемое задание помещается в глобальную очередь, при этом считается, что все ВУ ТРС заняты обработкой более приоритетных заданий.
5. *Эвристический коэффициент совместимости.* В [5] показано, что принятие решения о распределении задания, основанного только на приоритете, может негативно сказаться на эффективности планирования всего потока заданий. Другими словами, для повышения эффективности планирования потока заданий в целом необходимо рассмотреть альтернативные способы ранжирования системы заданий. Так, в работе [5] рассматриваются альтернативные способы ранжирования системы заданий с использованием эвристического коэффициента “совместимости” характеристик заданий и целевой ВУ для принятия решений о распределении задания. При учете свойств исследуемой ТРС коэффициент “совместимости” основывается на приоритете задания и времени передачи входных данных задания между ВУ ТРС. Важность учета времени передачи входных данных задания заключается в том, что для некоторых заданий это время может превосходить время основных вычислительных операций.
6. *Горизонт планирования.* Планирование высокоприоритетных заданий на ВУ с заданиями с меньшим приоритетом может привести к тому, что ранее распределенные в ВУ задания не будут успевать обрабатываться, так как постоянно будут вытесняться поступающими более приоритетными. Для предотвращения такой ситуации локальная очередь каждого узла ограничивается горизонтом планирования [4]. В общем случае горизонт планирования может быть как количественной (по числу заданий в очереди), так и временной характеристикой локальной оче-

реди. Однако непредсказуемость времени выполнения каждого задания оставляет возможность использования только количественного горизонта планирования. При достижении горизонта планирования (т. е. определенного числа распределенных глобальных заданий) направление заданий на эту вычислительную установку прекращается, пока не будет выполнено хотя бы одно ранее распределенное задание. При достижении горизонта планирования на всех ВУ планируемое в ГСУР задание помещается в глобальную очередь и ожидает освобождения вычислительных ресурсов.

Разработанный авторами алгоритм планирования заданий для модели ТРС с абсолютными приоритетами состоит из следующих шагов.

1. Задание поступает в локальную очередь диспетчера $i \in L$, где L — множество всех диспетчеров ВУ ТРС.
2. Диспетчер i опрашивает диспетчеров $j \in L$ и получает от них информацию о текущей загруженности и максимальном приоритете заданий локальной очереди.
3. На основе собранной на шаге 2 информации осуществляется поиск вычислительных установок, способных обработать поступившее задание, т. е. формируется список альтернатив ВУ.
4. Если список альтернатив пуст, то задание помещается в глобальную очередь в ожидании освобождения подходящих ресурсов, иначе для задания осуществляется выбор целевой вычислительной установки (переход к п. 7).
5. Производится оценка времени передачи входных данных задания до каждого из ВУ из списка альтернатив.
6. На основе информации о максимальном приоритете задания в локальной очереди и оцененном времени передачи для каждой ВУ из списка альтернатив выбирается целевая ВУ. Целевой вычислительной установкой i^* считается подсистема, в которой, во-первых, максимальный приоритет заданий локальной очереди меньше приоритета поступившего задания (данное условие позволяет минимизировать время, которое задание затратит на ожидание выполнения в локальной очереди ВУ), а, во-вторых, для такой вычислительной установки время передачи исходных данных минимально. Другими словами, целевой ВУ назначается система с максимальным коэффициентом совместимости. Таким образом, подлежащий минимизации целевой функционал может быть записан в следующем виде:

$$i^* = \arg\min_{j \in L}\{T_{\text{о.з}}(j)\}, \tag{3}$$

где $T_{\text{о.з}}(j)$ — время обработки задания на ВУ, $j \in L$, рассчитанное в соответствии с (1).
7. Задание распределяется в целевую вычислительную установку.

3. Свойства разработанного алгоритма планирования

В работе [4] для планирования территориально распределенных ресурсов не рекомендуется применять событийно-ориентированный подход, вместо этого предлагается осуществлять планирование с использованием очередей. Причина заключается в том, что ГСУР не может моментально отреагировать на возникшее событие, т. е. шаг диспетчеризации [12] требует определенного времени.

В общем случае шаг диспетчеризации включает:

- выбор задания из глобальной очереди;
- пересылку входных данных задания на целевую ВУ ТРС;
- добавление задания в ЛСУР.

Временные затраты на выбор задания и добавление его в ЛСУР в большинстве случаев можно считать пренебрежимо малыми, чего нельзя сказать о времени доставки входных данных задания на целевую ВУ.

Однако мы считаем, что в ТРС с абсолютными приоритетами событийно-ориентированный подход вполне применим благодаря использованию высокоприоритетного планирования, позволяющего минимизировать как время ожидания глобального задания в локальной очереди, так и время простоя вычислительных ресурсов.

В самом деле, распределение глобального задания, приоритет которого больше приоритетов всех заданий на целевой ВУ, позволяет при готовности глобального задания к запуску, т. е. по окончании передачи входных данных, вытеснить задания с меньшим приоритетом и приступить к выполнению. Рассмотрение ТРС с неотчуждаемыми ресурсами подразумевает, что при подготовке глобального задания вычислительные ресурсы могут быть заняты обработкой заданий из локальной очереди. Приоритет заданий из локальной очереди строго ниже приоритета подготавливаемого задания, поэтому они не помешают запуску глобального задания.

4. Макет территориально распределенной системы

В соответствии с рассмотренной моделью ТРС авторами подготовлен макет подобной ТРС, изображенный на рис. 2. Глобальная очередь заданий размещается в специальной распределенной информационной системе. Пунктирные стрелки указывают пересылку паспортов заданий, сплошные линии — пересылку входных данных. Планирование вычислительных ресурсов осуществляется с использованием приведенного в настоящей статье алгоритма. В процессе опытной эксплуатации в систему введен модельный поток заданий, все задания которого успешно обработаны системой, опытная эксплуатация

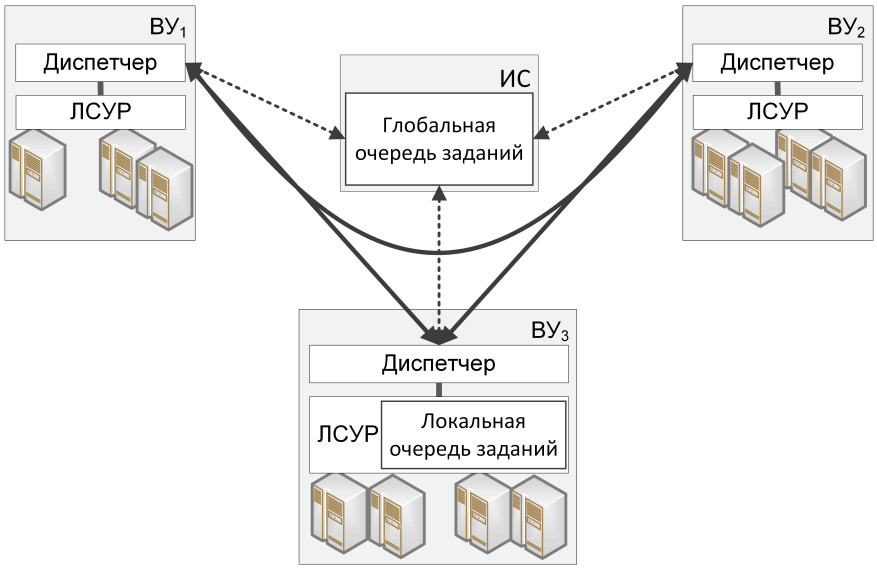

Рис. 2. Макет децентрализованной территориально распределенной системы с абсолютными приоритетами

подтвердила работоспособность макета и жизнеспособность предложенного авторами подхода. Ближайшей перспективой работы являются экспериментальные исследования по оценке эффективности и ресурсоемкости разработанного алгоритма.

Таким образом, в данной работе новыми являются следующие положения и результаты:

- модель TPC с абсолютными приоритетами;
- децентрализованная схема диспетчеризации и алгоритм планирования заданий в TPC с абсолютными приоритетами.

Список литературы / References

[1] **Foster, I.** The anatomy of the Grid: enabling scalable virtual organizations // Intern. J. of High Performance Computing Applications. 2001. Vol. 15, No. 3. P. 200–222.

[2] **Foster, I.** The physiology of the Grid: an open grid services architecture for distributed systems integration // Computer Networks: The Intern. J. of Computer and Telecommunications Networking. 2002. Vol. 40, No. 1. P. 5–17.

[3] **Коваленко В.Н.** Эволюция и проблемы Grid // Открытые системы. 2003. № 1. С. 23–33.
Kovalenko, V.N. Evolution and problems of Grid // Open systems. 2003. No. 1. P. 23–33. (In Russ.)

[4] **Коваленко В.Н., Коваленко Е.И., Корягин Д.А.** Планирование ресурсов Grid на основе локальных расписаний // Тр. Первой Всерос. научн. конф. “Методы и средства обработки информации”. М.: МГУ им. М.В. Ломоносова, 2003. С. 205–210
Kovalenko, V.N., Kovalenko, E.I., Koryagin, D.A. Resource manager for GRID with global job queue and planning based on local schedules // Proc. of the First All-Russ. Sci. Conf. “Methods and Means of Information Processing”. Moscow: MGU im. M.V. Lomonosova, 2003. P. 205–210. (In Russ.)

[5] **Топорков В.В., Емельянов Д.М., Потехин П.А.** Формирование и планирование пакетов заданий в распределенных вычислительных средах // Вестн. ЮУрГУ. Сер.: Вычисл. математика и информатика. 2015. Т. 4, № 2. С. 44–57.
Toporkov, V.V., Emelyanov, D.M., Potekhin, P.A. Job batch generation and scheduling in distributed computing environment // Bulletin of the South Ural State Univ. Ser.: Comput. Math. and Software Eng. 2015. Vol. 4, No. 2. P. 44–57. (In Russ.)

[6] **Коваленко В.Н., Коваленко Е.И., Шорин О.Н.** Разработка диспетчера заданий грид, основанного на опережающем планирвании. М.: Ин-т прикл. математики им. М.В. Келдыша, 2005. 28 с.
Kovalenko, V.N., Kovalenko, E.I., Shorin, O.N. Development of grid job dispatcher based on lookahead scheduling. Moscow: In-t Prikl. Matematiki im. M.V. Keldysha, 2005. 28 p. (In Russ.)

[7] **Топорков В.В** Модели распределенных вычислений. М.: Физматлит, 2004. 320 с.
Toporkov, V.V. Distributed Computing Models. Moscow: Fizmatlit, 2005. 320 p. (In Russ.)

[8] **Барский А.Б.** Параллельные процессы в вычислительных системах. Планирование и организация М.: Радио и связь, 1990. 256 с.
Barskiy, A.B. Parallel Processes in Computing Systems. Planning and organization. Moscow: Radio i Svyaz’, 1990. 256 p. (In Russ.)

[9] Для начинающих пользователей вычислительных кластеров. Адрес доступа: https://parallel.ru/cluster/beginnerguide.html (дата обращения 23.08.2016).
For novice users of computing clusters. Available at: https://parallel.ru/cluster/beginnerguide.html (accessed 23.08.2016). (In Russ.)

[10] **Mutz, A., Wolski, R., Brevik, J.** Eliciting honest value information in a batch-queue environment // Proc. of the "8th IEEE/ACM Intern. Conf. on Grid Computing". Austin: IEEE Computer Society, 2007. P. 291–297

[11] **Ernemann, C.** Economic scheduling in Grid computing // Job Scheduling Strategies for Parallel Proc. 2002. No. 2537. P. 129–152.

[12] **Коваленко В.Н., Семячкин Д.А.** Использование алгоритма Backfill в ГРИД // Тр. Междунар. конф. "Распределенные вычисления и Грид-технологии в науке и образовании". Дубна: Объединенный ин-т ядерных исследований, 2004. С. 139–144.
Kovalenko, V.N., Semyachkin, D.A. Using BackFill in GRID system // Proc. of the "Intern. Conf. on Distributed Computing and Grid-technologies in Science and Education". Dubna: Ob'edinennyy in-t Yadernykh Issledovaniy, 2004. P. 139–144. (In Russ.)

[13] **Towsley, D.** Analysis of fork-join program response times on multiprocessors // IEEE Trans. Parallel and Distributed System. 1990. Vol. 1, No. 3. P. 286–303.

Поступила в редакцию 20 января 2017 г.

Scheduling of jobs in a territorially distributed computing system with absolute priorities

Baranov, Anton V., Tikhomirov, Artem I.*

Joint Supercomputer Center — Branch FSC SRI for System Researches RAS, Moscow, 119334, Russia

*Corresponding author: Tikhomirov, Artem I., e-mail: TEMA4277@rambler.ru

This study addresses research and design of approaches and algorithms for scheduling of jobs with absolute priorities and unpredictable run time in a territorially distributed computing system (TDCS). To this purpose, authors designed and researched the TDCS model comprising of several high-performance computing (HPC) clusters united by communication channels with variable bandwidth. Both local and global levels of management are reviewed in the model. On a local level, jobs go through a local cluster queue to be run on a single HPC cluster. On a global level, jobs go through the global TDCS queue with global scheduler submitting jobs to one of local cluster queues. Jobs have absolute priorities. High priority job is able to interrupt execution of a low priority job and return it to the queue. Minimizing of the staying time for high priority jobs is the goal of the global scheduler. The researched model was implemented as a prototype of the TDCS. The decentralized dispatching scheme and the scheduling algorithm were designed for the prototype. Algorithm distributes jobs to the HPC clusters that consider performance, workload of clusters and bandwidth of cluster communication channels were presented. The pilot operation of the prototype was done successfully.

In authors opinion following provisions and results were obtained for the model of TDCS with absolute prioritization and for the decentralized dispatching scheme and the algorithm of scheduling jobs with absolute priorities.

Keywords: grid, absolute priorities, resource management, runnin time, scheduling.

Received 20 January 2017