УДК 004.75

В.Г. Хорошевский, М.Г. Курносов, С.Н. Мамойленко

ПРОСТРАНСТВЕННО-РАСПРЕДЕЛЕННАЯ МУЛЬТИКЛАСТЕРНАЯ ВЫЧИСЛИТЕЛЬНАЯ СИСТЕМА: АРХИТЕКТУРА И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ[1]

Представлены архитектура и функциональная структура пространственно-распределенной мультикластерной вычислительной системы, созданной и развиваемой Центром параллельных вычислительных технологий Сибирского государственного университета телекоммуникаций и информатики совместно с Институтом физики полупроводников им. А.В. Ржанова СО РАН. Приведено описание программного обеспечения системы, включающее средства, созданные коллективом ведущей научной школы по распределенным вычислительным системам (НШ 5176.2010.9).

Ключевые слова: *распределенные вычислительные системы*, *GRID*, *параллельное мультипрограммирование*, *эффективное выполнение параллельных программ.*

Современный этап развития вычислительной техники и телекоммуникационных технологий характеризуется построением пространственно-распределенных мультикластерных вычислительных систем (ВС) [1, 2]. В архитектурном плане такая ВС представляется как множество кластеров, взаимодействие между которыми осуществляется через телекоммуникационную сеть (в общем случае – сеть Интернет). Каждый кластер, в свою очередь, является пространственно-сосредоточенной распределённой ВС, состоящей из множества вычислительных узлов, взаимодействующих через свою телекоммуникационную подсистему. Конфигурация вычислительного узла допускает варьирование в широких пределах – от однопроцессорного до композиции из многоядерных процессоров и специализированных ускорителей (например, GPGPU).

Центром параллельных вычислительных технологий (ЦПВТ) Государственного образовательного учреждения высшего профессионального образования «Сибирский государственный университет телекоммуникаций и информатики» (ГОУ ВПО «СибГУТИ») совместно с Лабораторией вычислительных систем учреждения Российской академии наук Института физики полупроводников им. А.В. Ржанова Сибирского отделения РАН (ИФП СО РАН) создана и развивается пространственно-распределенная мультикластерная вычислительная система [3].

1. Архитектура пространственно-распределенной мультикластерной ВС

Действующая конфигурация пространственно-распределённой мультикластерной ВС (GRID-модель) в свой состав включает более 120 процессорных ядер и имеет пиковую производительность несколько TFLOPS. Система (рис. 1) объединяет 9

[1] Работа выполнена в рамках интеграционного проекта № 113 СО РАН, при поддержке РФФИ (гранты № 09-07-90403, 10-07-05005, 08-07-00022, 08-08-00300), Совета по грантам Президента РФ для поддержки ведущих научных школ (грант НШ-5176.2010.9) и в рамках государственного контракта № 02.740.11.0006 с Минобрнауки РФ.

пространственно-распределенных кластеров, причем кластеры A-G расположены в ЦПВТ ГОУ ВПО «СибГУТИ» (центр г. Новосибирска), а кластеры H, I – в Лаборатории вычислительных систем ИФП СО РАН (Академгородок, СО РАН).

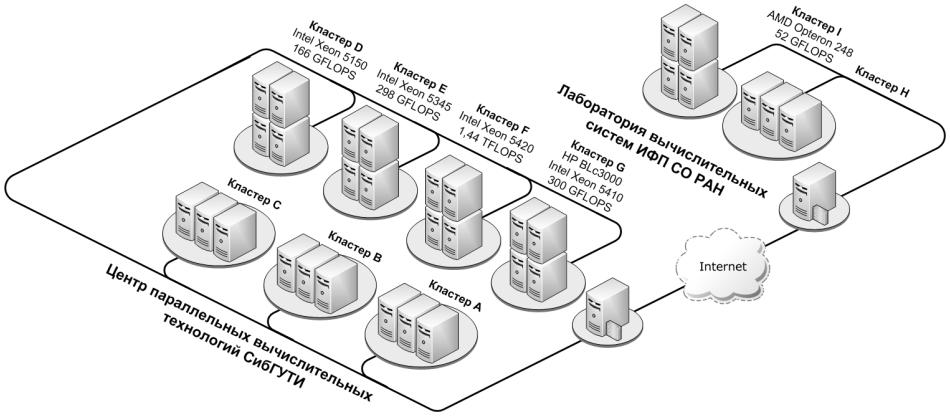

Рис. 1. Конфигурация пространственно-распределенной мультикластерной вычислительной системы

Кластеры A, B, C и H являются кластерами рабочих станций. Причем кластера A, B и C представляют собой компьютерные классы, располагающиеся в учебных лабораториях кафедры вычислительных систем ГОУ ВПО «СибГУТИ», а H – совокупность персональных компьютеров лаборатории вычислительных систем ИФП СО РАН. Кластеры D, E, F, G и I работают в круглосуточном режиме.

Каждый кластер укомплектован: вычислительными узлами, управляющим узлом, двумя телекоммуникационными системами (вычислительной и сервисной), а также средствами бесперебойного электропитания. Кластер D (Xeon16) объединяет 4 вычислительных узла, на каждом из которых размещено по два двухъядерных процессора Intel Xeon 5150 (Woodcrest) с тактовой частотой 2,66 GHz. Пиковая производительность кластера D – 166 GFLOPS. Кластер E (Xeon32) состоит из 4-х вычислительных узлов на базе двух процессоров Intel Quad Xeon E5345 с тактовой частотой 2,33 GHz. Пиковая производительность кластера E – 298 GFLOPS. Кластер F (Xeon80) объединяет 10 вычислительных узлов с двумя процессорами Intel Quad Xeon E5420 с тактовой частотой 2.5 GHz. Пиковая производительность кластера E – 800 GFLOPS. Кластер I – 5 вычислительных узлов с двумя процессорами AMD Opteron 248 – 252 (Sledgehammer) с тактовой частотой 2,2 – 2,6 GHz. Пиковая производительность кластера – 47 GFLOPS.

В 2009 году ЦПВТ ГОУ ВПО «СибГУТИ» стал участником проекта «Университетский кластер», в рамках которого компанией Hewlett Packard предоставлен кластер (G), укомплектованный 4 вычислительными узлами HP ProLiant BL220c, на которых размещено два четырехъядерных процессора Intel Xeon E5410 с тактовой частотой 2,33 GHz. Пиковая производительность кластера G – 298 GFLOPS.

Любой из кластеров способен функционировать как автономно, так и в составе пространственно-распределённой мультикластерной распределенной ВС. Телекоммуникационные системы кластеров построены на базе технологий Gigabit и Fast Ethernet. Для объединения кластеров используется сеть Internet (технология VPN). Мультикластерная ВС допускает масштабирование путем организации взаимодействия с множеством других кластеров.

2. Программное обеспечение

Пространственно-распределенная мультикластерная ВС укомплектована системным программным обеспечением, включающим инструментарий параллельного мультипрограммирования (рис. 2). Инструментарий – это модели, методы и программное обеспечение организации функционирования распределенных ВС, при решении множества задач, представленных параллельными программами.

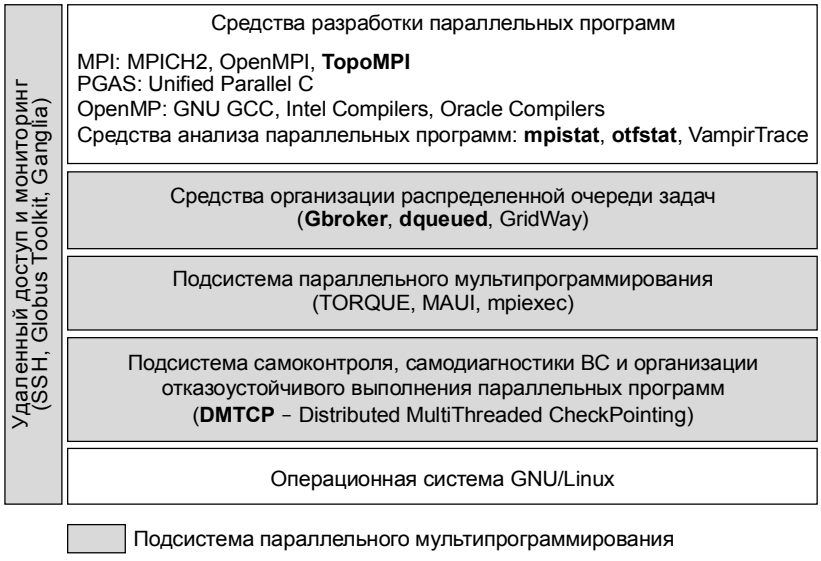

Рис. 2. Программное обеспечение пространственно-распределенной мультикластерной вычислительной системы

Стандартные компоненты системного программного обеспечения представлены:

• сетевой операционной системой GNU/Linux (дистрибутив CentOS 5.5, версия ядра 2.6.18);

• средствами разработки, отладки и анализа последовательных и параллельных программ:

- компиляторы: Gcc, Sun, Intel;

- математические библиотеки: GNU Scientific Library, AMD Core Math Library, Intel Math Kernel Library;

- библиотеки передачи сообщений между ветвями параллельных программ:

· распределенные приложения – MPI: MPICH2, OpenMPI, IntelMPI;

· параллельные программы – OpenMP: gcc, sun, intel;

- средствами отладки и анализа программ: gdb, vampire, gprof и т.д.

• программным обеспечением организации взаимодействия пространственно-распределенных кластерных ВС и диспетчеризации пользовательских заданий: Globus Toolkit, GridWay.

Инструментарий параллельного мультипрограммирования включает:

• средства самоконтроля и самодиагностики;

• подсистему организации функционирования ВС в мультипрограммных режимах, включающую средства вложения параллельных программ и реализации эффективных групповых обменов между ветвями параллельных программ;

• средства организации распределенной очереди задач и диспетчер пользовательских запросов;
• подсистему анализа параллельных программ;
• средства мониторинга и организации удаленного доступа к ресурсам ВС.

2.1. Вложение параллельных программ в иерархических системах

Коммуникационные среды современных распределенных ВС используют неоднородные каналы связи между ресурсами. Более того, в (мульти)кластерных и GRID-системах коммуникационные среды имеют иерархическую организацию, в которых первым уровнем является сеть связи между кластерами, вторым – сеть связи внутри кластеров, третьим – среда доступа процессоров (или ядер) вычислительного узла к общей памяти. Скорости передачи информации на различных уровнях таких коммуникационных сред существенно различны. По этой причине время выполнения параллельных программ на ВС в немалой степени зависит от того, как они «вложены» в систему (на какие ядра назначены ветви и через какие каналы связи они взаимодействуют).

На основе метаэвристики имитации отжига (Simulated Annealing) созданы [4, 5] стохастические последовательный и параллельный алгоритмы субоптимального вложения в распределенные ВС параллельных программ с целью минимизации времени их выполнения. В ходе выполнения алгоритмов ветви, обменивающиеся большими объемами данных, распределяются на один вычислительный узел, где они взаимодействуют через его общую память. Последнее обеспечивает сокращение времени реализации межпроцессорных обменов и, как следствие, сокращение времени выполнения параллельной программы. Алгоритмы реализованы как дополнительный функциональный модуль к существующим средствам запуска параллельных программ.

Проведены эксперименты по вложению параллельных MPI-программ из пакетов NAS Parallel Benchmarks и High-Performance Linpack в действующие вычислительные кластеры на базе многоядерных процессоров компаний Intel и AMD. В среднем, время выполнения тестовых MPI-программ с вложением предложенными алгоритмами на 30 – 40 % меньше времени выполнения программ с вложением средствами библиотек MPI (MPICH2 и OpenMPI). Предложенные алгоритмы характеризуются полиномиальной трудоемкостью и поддерживают вложение параллельных программ с количеством ветвей до 10^6.

2.2. Диспетчеризация задач в пространственно-распределенных ВС

Для организации функционирования пространственно-распределенных ВС и GRID-систем в мультипрограммном режиме обслуживания потока задач разработаны децентрализованные алгоритмы и средства диспетчеризации заданий. На каждой подсистеме пространственно-распределенной ВС функционирует диспетчер, который поддерживает локальную (для кластера) очередь задач. При поступлении задачи диспетчер запрашивает у диспетчеров из локальной окрестности количество задач в их очередях и оценку времени, через которое поступившая задача может начать решаться при передаче её в соответствующую очередь. Далее, диспетчер, используя систему мониторинга, определяет пропускную способность каналов связи между кластерами. После этого он выбирает кластер, в котором (с

учетом передачи самой задачи в очередь и её данных в кластер) быстрее всего начнет выполняться задача. Важно отметить, что при передаче задачи в очередь выбранной подсистемы, для неё, аналогичным образом, периодически будет осуществляться поиск ресурсов. Это обеспечивает адаптацию диспетчеров под динамически изменяющуюся загрузку ресурсов пространственно-распределенных ВС.

Предложенный алгоритм реализован в программном пакете GBroker [6]. Пакет имеет расширяемую архитектуру и допускает интеграцию с системами пакетной обработки заданий. В пакет (рис. 3) входят диспетчер gbroker, клиентское приложение gclient и средство мониторинга производительности каналов связи netmon на уровне стека протоколов TCP/IP.

Модуль gbroker устанавливается в каждом кластере и обеспечивает на интерфейс с локальной системой пакетной обработки заданий (на данный момент – TORQUE). Модуль netmon устанавливается вместе с gbroker. Сервисы netmon собирают информацию о производительности каналов связи между кластерами. Модуль gclient обеспечивает интерфейс между пользователем и системой.

Администратор настраивает локальные окрестности диспетчеров gbroker, указывая, какие диспетчеры с какими могут обмениваться программами из своих очередей, и настраивает сервис netmon.

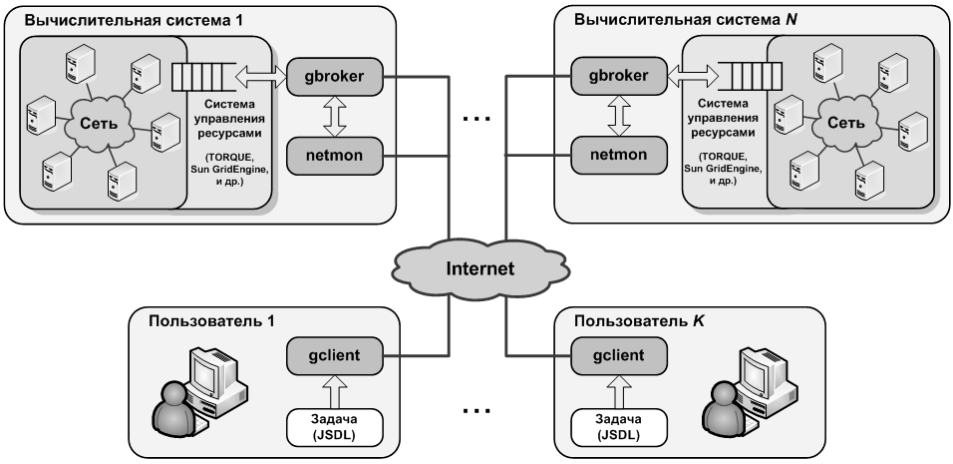

Рис. 3. Функциональная структура пакета GBroker

Пользователь формирует задание, состоящее из параллельной MPI-программы и паспорта на языке ресурсных запросов JSDL, и отправляет его средствами gclient любому из диспетчеров gbroker. Диспетчер в соответствии с описанным выше алгоритмом выбирает подсистему, на которой затем выполняется программа.

На ресурсах пространственно-распределенной мультикластерной ВС проведено исследование созданных алгоритмов и пакета GBroker. В качестве тестовых задач использовались MPI-программы из пакета NAS Parallel Benchmarks, а также программы, реализующие параллельные версии различных численных методов. Моделирование показало, что среднее время обслуживания задач при децентрализованной диспетчеризации сопоставимо с централизованной диспетчеризацией. Вместе с тем обеспечивается отказоустойчивость пространственно-распределенной мультикластерной ВС в случае выхода из строя отдельных кластеров. Время диспетчеризации достаточно мало по сравнению со временем выполнения программ.

Заключение

Созданная пространственно-распределенная мультикластерная вычислительная система используется как инструментальное средство для проведения исследований и подготовки специалистов в области параллельных вычислительных технологий. Перспективы использования созданного инструментария параллельного мультипрограммирования в промышленности подтверждаются растущими потребностями в применении распределенных вычислительных и GRID-систем в областях, где требуются высоконадежные (живучие) вычислительные средства, где решаются суперсложные задачи и моделируются современные технологические процессы и природные явления.

ЛИТЕРАТУРА

1. *Хорошевский В.Г.* Архитектура вычислительных систем. М.: МГТУ им. Н.Э. Баумана, 2008. 520 с.
2. *Хорошевский В.Г. Мамойленко С.Н. Курносов М.Г.* Архитектурные концепции, анализ и организация функционирования вычислительных систем с программируемой структурой // Информационные технологии и математическое моделирование систем: Труды Международной научной конференции. М.: РАН, 2008.
3. *Вычислительные* ресурсы Центра параллельных вычислительных технологий ГОУ ВПО «СибГУТИ». URL: http://cpct.sibsutis.ru/index.php/Main/Resources (дата обращения: 25.11.2010).
4. *Khoroshevsky V.*, *Kurnosov M.* Mapping parallel programs into hierarchical distributed computer systems // Proc. of 4th Intern. Conf. “Software and Data Technologies (ICSOFT 2009)”. Sofia: INSTICC, 2009. V. 2. P. 123–128.
5. *Курносов М.Г.* Алгоритмы вложения параллельных программ в иерархические распределённые вычислительные системы // Вестник СибГУТИ. 2009. № 2 (6). С. 20–45.
6. *Курносов М.Г.*, *Пазников А.А.* Децентрализованное обслуживание потоков параллельных задач в пространственно-распределенных вычислительных системах // Вестник СибГУТИ. 2010. № 2 (10). С. 79–86.

Хорошевский Виктор Гаврилович
Курносов Михаил Георгиевич
Мамойленко Сергей Николаевич
ГОУ ВПО «Сибирский государственный университет телекоммуникаций и информатики».
E-mail: khor@cpct.sibsutis.ru; mkurnosov@gmail.com; sergey@cpct.sibsutis.ru

Поступила в редакцию 3 декабря 2010 г.