Программный компонент управления платформами исполнения

В.П. Гергель, А.В. Линев, А.В. Сысоев

В рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2012 годы» разработан высокопроизводительный программный комплекс моделирования наноразмерных атомно-молекулярных систем, построенный на основе концепции iPSE (Intelligent Problem Solving Environment), ориентированный на поддержку высокопроизводительных вычислений на многокластерных системах. В настоящей работе описывается программный компонент управления платформами исполнения, реализованный в рамках указанного комплекса в качестве системного уровня, обеспечивающего выполнение задач.

1. Введение

В 2008-2009 гг. Нижегородский государственный университет им. Н.И. Лобачевского (ННГУ) участвовал в реализации проекта по разработке высокопроизводительного программного комплекса для квантово-механических расчётов и моделирования наноразмерных атомно-молекулярных систем и комплексов (головным исполнителем являлся Санкт-Петербургский государственный университет информационных технологий, механики и оптики).

Цель проекта состояла в создании программного комплекса прикладных алгоритмов, программ и математических моделей для квантово-механических расчётов: от геометрических характеристик сложной атомно-молекулярной структуры до макропараметров системы. Программный комплекс построен на основе концепции iPSE (Intelligent Problem Solving Environment) [1], в которой в дополнение к принципам PSE формализуются экспертные знания об организации процесса изучения явления средствами компьютерного моделирования и определяются принципы построения интеллектуальной системы управления параллельными вычислительными процессами в распределенной иерархической среде, включающей в себя одну или несколько суперЭВМ различной архитектуры [2].

Программный комплекс состоит из четырех базовых блоков:

- блок проблемно-ориентированных компонентов (вычислительных сервисов);
- интеллектуальная система конструирования заданий, позволяющая осуществлять интеллектуальную поддержку пользователя при формализации заданий и их эффективное исполнении в распределенных вычислительных средах;
- блок человеко-компьютерного взаимодействия, реализующий функции ввода и вывода данных;
- блок управления платформами исполнения, обеспечивающий распределенный запуск и мониторинг исполнения заданий.

В данной работе описывается программный компонент управления платформами исполнения (multi-cluster management system, MCMS).

2. Программный компонент управления платформами исполнения

2.1. Назначение MCMS

Программный компонент управления платформами исполнения является самостоятельной системой, которая обеспечивает распределенный запуск и мониторинг исполнения заданий на наборе высокопроизводительных вычислительных комплексов, объединенных общей высокоскоростной сетью. В ходе работы компонентов блока не формируется сама стратегия управления – блок лишь транслирует управляющие команды, поступающие от интеллектуальной системы конструирования заданий [2].

Основное назначение программного компонента заключается в построении дополнительного уровня абстракции над существующими системами управления кластером. Данный уровень должен предоставлять программный интерфейс, используя который можно выполнить следующие действия:

- получить список доступных кластеров;
- получить информацию о конфигурации указанного кластера и его текущей загрузке;
- зарезервировать на выбранном кластере ресурсы для выполнения задачи;
- запустить задачу на выбранном кластере, определить состояние задачи, дождаться завершения задачи;
- принудительно завершить задачу на выбранном кластере.

Программный компонент MCMS обеспечивает работу с кластерами под управлением операционных систем Linux и Windows, а так же различных систем управления: Torque [3], Microsoft HPC Server 2008 [4], Метакластер [5] (разрабатывается в ННГУ).

Множество приложений, которые возможно запустить посредством MSMC определяется отдельно. Для каждого приложения требуется индивидуальная настройка сопряжения с MCMS.

2.2. Архитектура MCMS

MCMS содержит два основных модуля: интегратор управления и трансивер, – и взаимодействует со сторонними модулями: интеллектуальная система конструирования заданий (планировщик), ftp-сервер, сервер СУБД (PostgreSQL), система управления кластером, вычислительные пакеты, установленные на конкретном кластере.

В развернутом виде MCMS имеет структуру, представленную на рис. 1.

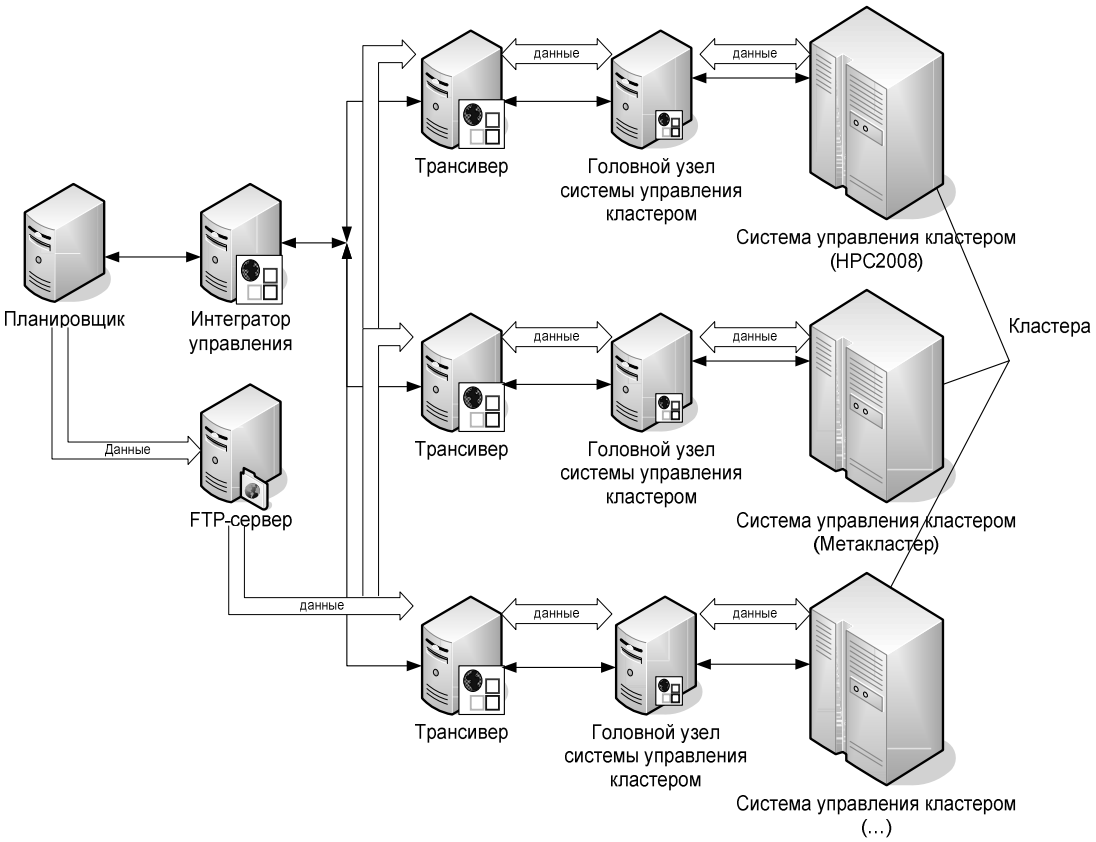

Рис. 1. Компонентный состав MCMS

Рассмотрим назначение модулей MCMS.

- Планировщик (интеллектуальная система конструирования заданий).

Планировщик является пользователем MCMS и не входит в его состав (разработан в СПбГУ ИТМО). Основное назначение – формирование стратегии выполнения множества задач

на наборе высокопроизводительных вычислительных комплексов, объединенных посредством высокоскоростной сети (мультикластере).

- Интегратор управления.

Интегратор управления обеспечивает централизованный унифицированный доступ к программно-аппаратным платформам, работающим в составе мультикластера, предоставляя соответствующий программный интерфейс.

Интегратор управления обеспечивает:

- получение информации о мультикластере;
- резервирование ресурсов под задачу на выбранном кластере;
- создание, запуск и останов задачи на выбранном кластере.

Обеспечивает хранение, обновление и выдачу информации о создаваемых задачах с использованием базы данных реляционного типа (используется СУБД PostgreSQL).

Интегратор управления реализован в виде веб-сервиса с использованием технологии ASP.NET.

- FTP-сервер.

Обеспечивает хранение входных и выходных данных заданий. Для обслуживания всех кластеров используется единый FTP-сервер. Может быть развернут на узлах под управлением операционных систем Linux (в настоящий момент используется vsftpd) и Windows (используется IIS FTP).

- Трансивер.

Трансивер обеспечивает централизованный унифицированный доступ к конкретной программно-аппаратной платформе. Трансивер обеспечивает сопряжение интегратора управления с конкретным типом системы управления кластером. Для каждого кластера используется отдельный экземпляр трансивера.

Трансивер обеспечивает для конкретного кластера:

- получение информации о кластере;
- резервирование ресурсов под задачу;
- запуск и останов задачи.

Реализован в виде Windows service и Linux daemon, является сервером распределенного приложения, реализованным с применением технологии .NET Remoting. В роли клиентского приложения для трансивера выступает интегратор управления.

- Система управления кластером.

Выполняется на головном узле кластера, обеспечивает трансивер информацией о состоянии узлов кластера и выполняемых на нем задач. По запросу трансивера выполняет запуск и останов задач. Каждый кластер обслуживается независимо работающей системой управления.

На текущий момент поддерживаются следующие системы управления кластером:

- Torque,
- HPC2008,
- Метакластер.

Для того, чтобы обеспечить поддержку дополнительных систем управления, необходимо разработать дополнительный трансивер, реализующий функциональность, объявленную в общем для всех трансиверов интерфейсе.

- Агенты системы управления кластером.

Выполняются на узлах кластера, обеспечивают выполнение заданий системы управления кластером на конкретных узлах.

- Вычислительные сервисы.

Вычислительные сервисы представляют собой прикладные пакеты, установленные на узлах кластеров и готовые к использованию для решения прикладных задач. Для каждого прикладного пакета требуется индивидуальная организация сопряжения с трансивером и системой управления кластером. В настоящий момент выполнена интеграция ~20 пакетов, среди которых имеются как разработанные в рамках проекта, так и сторонние открытые и коммерческие пакеты.

2.3. Взаимодействие MCMS с системами управления кластерами

Основная задача MCMS состоит в предоставлении программного интерфейса управления кластерами, независимого от типа используемых систем управления. Непосредственное взаимодействие является задачей программного модуля «Трансивер», который обеспечивает сопряжение со следующими системы управления: Torque [3], Microsoft Windows HPC Server 2008 [4], «Метакластер» [5] (разработка ННГУ).

Трансивер обеспечивает выполнение следующих операций:

- предоставление корректной информации о составе (имена узлов кластера) и аппаратных и программных характеристиках (размер жесткого диска, количество ядер, частота процессора, список установленного программного обеспечения и т.д.) кластера;
- предоставление корректной информации о текущей загруженности кластера (текущая загруженность процессора каждого узла, загруженность сетевого интерфейса, идентификаторы задач, исполняющихся на каждом узле, и т.п.);
- запуск и остановка заданий;
- резервирование ресурсов для запуска задачи и отмена резервирования.

Разработчики систем управления кластерами предоставляют различные наборы средств взаимодействия, посредством которых можно реализовать перечисленную выше функциональность. Трансивер использует соответствующие возможности интерфейсов для сопряжения с системами управления:

- трансивер Torque использует удаленный вызов команд интерфейса системы управления средствами удаленного терминала;
- трансивер Windows HPC Server 2008 использует программный интерфейс .NET API и командный интерфейс Windows PowerShell;
- трансивер Метакластера использует интерфейс .NET Remoting, позволяющий получить удалённый доступ к системе управления кластером «Метакластер».

Дополнительно к перечисленным реализован трансивер Ganglia, который не является трансивером в общем смысле, то есть не обеспечивает запуск и выполнение заданий, а используется для получения информации о статических и динамических характеристиках кластера от распределенной системы мониторинга высокопроизводительных вычислительных систем Ganglia. Остальные трансиверы используют значения статических и динамических характеристик, полученные либо от системы управления кластером, либо от трансивера Ganglia.

Архитектура трансивера допускает расширение его функциональности для обеспечения сопряжения с другими системами управления кластерами.

2.4. Развертывание MCMS

Минимальный набор модулей и сопутствующего ПО, пригодный для независимой демонстрации и тестирования MCMS, содержит следующие программные средства.

- Операционная система: Linux (используется OpenSUSE 11.1 x64, также поддерживаются RedHat Enterprise Linux и Gentoo Linux) или Windows (используется Windows Server 2008 x64).
- Система управления кластером: HPC2008, Метакластер или Torque; возможно совместное использование системы управления кластером и системы сбора динамических характеристик (загруженность кластера) Ganglia.
- Трансивер.
- Интегратор управления.
- Сервер ftp.
- Сервер базы данных – хранит информацию о всех когда-либо исполнявшихся задачах (историю задач), в том числе – состояниях активных задач.
- Планировщик – web-сервис, «заглушка», эмулирующая работу реального планировщика.
- Тестовая консоль – консольное приложение, выполняет последовательность вызовов методов планировщика с целью проверки правильности функционирования системы в целом.

Пример диаграммы развертывания программного комплекса с использованием одного кластера приведен на рис. 2.

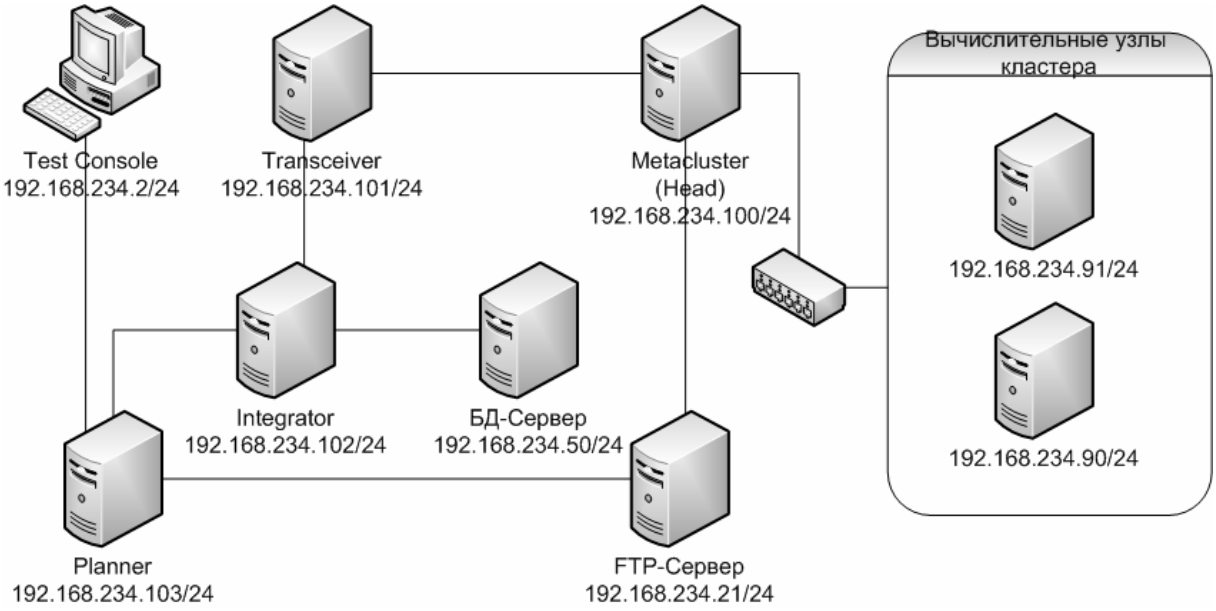

Рис. 2. Пример диаграммы развертывания MCMS

Для обеспечения быстрой и корректной установки и настройки работоспособного комплекса была разработана схема полуавтоматического развертывания основных программных средств.

- Развертывание операционной системы.

Система полуавтоматического развертывания использует OpenSUSE 11.1 x64 и обеспечивает ее установку на «чистую» аппаратную платформу, настройку сетевой подсистемы и необходимых сервисов (NFS-сервер, NFS клиент, DNS-сервер, FTP-сервер, SSH-сервер, RSH-сервер и т.д.), установку и настройку вычислительных пакетов. Состав программного обеспечения и настройки управляющего узла системы управления кластером отличаются от ПО и настроек вычислительных узлов.

При выполнении установки группы узлов необходимо устанавливать их имена и IP-адреса в настройках системы развертывания.

- Система управления кластером.

Используется система управления кластером Torque 2.3.6. Обеспечивается установка требуемых пакетов и формирование конфигурации.

- Компоненты MCMS.

Для обеспечения функционирования MCMS выполняется установка и настройка следующих программных средств: mono с пакетом xsp, web-сервер (используется apache), ftp-сервер (используется vsftpd), сервер СУБД (используется PostgreSQL), интегратор управления, трансивер.

- Тестовая консоль.

Тестовая консоль не требует установки и может быть запущена на любом узле с установленным .NET framework или mono с пакетом xsp, и имеющим возможность установить сетевое соединение с узлом, на котором запущен планировщик.

В ходе выполнения проекта был построен территориально распределенный мультикластер, включающий 96 узлов, объединенных в 5 кластеров.

3. Заключение

Используемая архитектура программного компонента управления платформами исполнения (MCMS) позволяет создать дополнительный уровень абстракции над системами управления кластером, что обеспечивает централизованный унифицированный доступ к программно-аппаратным платформам, работающим в составе мультикластера, и эффективно решает задачу управления несколькими кластерами. Данная архитектура также позволяет расширять список поддерживаемых систем управления кластером без изменения системы в целом.

Разработанная схема развертывания позволяет в сжатые сроки обеспечить установку и настройку нескольких высокопроизводительных вычислительных комплексов для совместной работы в качестве одного мультикластера либо настройку работающего кластера для включения в состав уже функционирующего мультикластера.

Исследования выполнены в рамках ОКР по теме «Разработка высокопроизводительного программного комплекса для квантово-механических расчетов и моделирования наноразмерных атомно-молекулярных систем и комплексов».

Литература

1. В.Н. Васильев и др. Высокопроизводительный программный комплекс моделирования атомно-молекулярных наноразмерных систем // Научно-технический вестник СПбГУ ИТМО. Выпуск 54. Технологии высокопроизводительных вычислений и компьютерного моделирования. – СПб.: СПбГУ ИТМО, 2008. – С. 3-12.
2. Ковальчук С.В. и др. Особенности проектирования высокопроизводительных программных комплексов для моделирования сложных систем // Информационно-управляющие системы, 3(34), 2008. - С. 10-18.
3. Официальная страница системы управления Torque – [http://www.clusterresources.com/products/torque-resource-manager.php].
4. Официальная страница Microsoft HPC Server 2008 – [http://www.microsoft.com/hpc/en/us/].
5. Система управления «Метакластер» – [www.cluster.software.unn.ru].