РЕАЛИЗАЦИЯ УНИФИЦИРОВАННОГО ДОСТУПА К ИНФОРМАЦИИ О СОСТОЯНИИ МУЛЬТИКЛАСТЕРА В ПРОГРАММНОМ КОМПОНЕНТЕ УПРАВЛЕНИЯ ПЛАТФОРМАМИ ИСПОЛНЕНИЯ MCMS

В.Д. Кустикова

Введение

При использовании вычислительных кластеров большое значение имеет правильное распределение нагрузки по узлам, в особенности, если кластер обладает неоднородной структурой: различается производительность вычислительных узлов, объем установленной в них оперативной памяти, характеристики сетевых соединений.

Планировщик заданий - компонент, отвечающий за построение расписания запуска заданий на имеющихся вычислительных ресурсах. Как правило, он входит в состав системы управления кластером, но в ряде случаев допускается возможность подключения планировщиков сторонних разработчиков. Для принятия решений планировщику необходима информация о составе и характеристиках аппаратных средств кластера (статический компонент), а также об их текущей загруженности (динамический компонент).

Данная работа посвящена задаче обеспечения планировщика статической и динамической информацией и способе ее решения в программном компоненте управления платформами исполнения Multi-Cluster Management System (MCMS), реализованном в ННГУ в рамках выполнения ОКР по теме "Разработка высокопроизводительного программного комплекса для квантово-механических расчетов и моделирования наноразмерных атомно-молекулярных систем и комплексов".

Стратегии планирования

Основная задача планировщика - оптимальным образом определить набор узлов для запуска пользовательских приложений на основании информации о запрашиваемых и доступных ресурсах. При этом под оптимальностью решения понимается достижение в процессе функционирования многопроцессорной системы некоторой цели. В зависимости от специфики заданий, которые попадают на исполнение на кластер, и количества имеющихся вычислительных ресурсов в качестве цели часто выделяют минимизацию среднего времени ожидания задачи в очереди либо минимизацию числа незанятых ресурсов в каждый момент времени.

При решении задачи планирования используется стратегия распределения задач, обычно вводящая дополнительные характеристики задач и политики выбора узлов, а также расширяющая список требований к данным, необходимым для принятия решения. Наиболее распространены следующие стратегии:

- списочные алгоритмы (First Come First Served – FCFS и его модификации) [1];
- алгоритм обратного заполнения (Backfill) [2];
- алгоритм Gang Scheduling [3], [4];
- алгоритм, использующий множество очередей (feedback algorithm) [5];
- генетические алгоритмы;
- смешанные алгоритмы.

Модификации перечисленных алгоритмов планирования могут предоставлять пользователю возможность определять для задачи необходимый минимальный размер оперативной памяти на узле, объем свободного пространства на жестком диске и другие параметры и ограничения, что требует наличия дополнительной информации об имеющихся вычислительных ресурсах.

В реальных системах планирования кластерных ресурсов наибольшее распространение получили алгоритмы FCFS (или его модификация – Priority FCFS) и Backfill.

Средства мониторинга вычислительных ресурсов

Планировщику для выполнения итерации планирования необходима актуальная информация о состоянии вычислительных ресурсов. Эти данные предоставляются планировщику средствами мониторинга вычислительных ресурсов. Рис. 1 представляет частичный список характеристик, которые предоставляются рядом существующих средств мониторинга, входящих в состав систем управления кластером или работающими независимо.

Имя характеристики	Ganglia [6]	Cacti [7]	Nagios [8]	Hyperic [9] (доступные метрики под Linux)	Средства Windows HPC Server 2008 [10], [11]	Средства Torque [12]
Имя/ip-адрес узла	+	+	+	+	+	+
Состояние узла	-	+	+	+	+	+
Количество ядер	+	-	?	?	+	+
Количество принятых и отправленных бит/байт по сети за некоторый промежуток времени	+	+	+	+	+	+
Размер свободного пространства на жестком диске	+	-	+	-	+	+
Общий размер жесткого диска	+	-	+	-	- (?)	+
Размер оперативной памяти	+	-	?	-	- (?)	+
Размер доступной оперативной памяти	+	-	?		+	+
Количество работающих процессов в системе	+	-	-	-	-	-
Общее количество процессов в системе	+	+	-	+	-	-
Проверка активности процесса с указанным именем	-	-	+	-	-	-
Средний процент загрузки процессора в течение фиксированного времени	+	+	+	+	+	+
Время последней загрузки системы	+	-	-	-	+	-
Процент простоя процессора или процессоров	+	-	?	+	-	-
Тактовая частота процессора	+	-	?		+	-
Максимальный размер файла подкачки	+	-	+	+	-	+
Размер свободного свопа	+	-	+	+	-	-

Рис. 1. Некоторые характеристики узлов, предоставляемые различными средствами мониторинга

Из приведенных данных видно, что независимые средства мониторинга предоставляют значительно более широкий спектр метрик по сравнению со средствами, встроенными в системы управления. Необходимо отметить, что в таблице приведен не полный перечень предоставляемых характеристик, в частности, некоторые средства ориентированы на мониторинг различных сетевых интерфейсов и предоставляют набор соответствующих метрик.

Все рассматриваемые средства мониторинга, за исключением средств Microsoft Windows HPC Server 2008, являются свободно распространяемыми, либо для них существует Community-версия, что позволяет попробовать их в практическом использовании. В частности, нас интересуют предоставляемые ими интерфейсы для получения данных.

Рассматриваемые средства мониторинга предоставляют различные механизмы взаимодействия для получения информации о вычислительных ресурсах:

- Ganglia позволяет использовать TCP/IP или UDP протокол для получения информации в формате XML, а также включает Web-интерфейс.
- Nagios имеет командный интерфейс и Web-интерфейс.
- Cacti предоставляет доступ по протоколу SNMP для выдачи сообщений фиксированного формата, а также реализует графический интерфейс для отображения результатов мониторинга.
- Enterprise-версия Hyperic предоставляет Java-ориентированный программный интерфейс, а также Web-интерфейс.
- Microsoft Windows HPC Server 2008 поддерживает графический, командный, программный (COM для приложений на C++, .NET APIs для приложений на .NET) интерфейсы и Microsoft Powershell.
- Torque поддерживает командный интерфейс.

Необходимо отметить, что в рамках одного средства различные интерфейсы могут предоставлять различный набор метрик, например, Microsoft PowerShell позволяет получить средний процент загрузки сетевых интерфейсов в то время, как командный интерфейс такой информации не предоставляет.

Характеристики узлов, используемые планировщиками

Набор метрик, учитываемых планировщиком, определяется стратегией планирования. На рис. 2 представлены характеристики, используемые некоторыми компонентами планирования.

Имя характеристики	Maui [15]	Планировщик Windows HPC Server 2008		Планировщик Torque[1]
	Backfilling	Priority FCFS	Backfilling	FCFS
Имя/ip-адрес узла	+	+	+	+
Название операционной системы	+	-	-	-
Состояние узла	+	+	+	+
Архитектура процессора	+	-	-	-
Количество ядер	+	+	+	+
Текущая загрузка процессора	+	-	-	-
Размер свободного пространства на жестком диске	+	-	-	-
Размер доступной оперативной памяти	+	-	-	+
Общий размер жесткого диска	+	-	-	-
Максимальное количество задач, которые могут одновременно работать на узле	+	-	+	-
Количество задач, которые одновременно работают на узле	+	-	+	-
Тактовая частота процессора	+	-	-	-

[1] Поддерживает возможность интеграции с некоторыми сторонними компонентами планирования. Например, планировщик может быть реализован разработчиком на процедурном языке BaSL или в скрипте, написанном на Tcl.

Рис. 2. Характеристики узлов, используемые компонентами планирования

Необходимо отметить, что совокупность используемых характеристик зависит от фактически используемой стратегии, в частности, настройки системы управления кластером определяют используемые планировщик (встроенный в систему или внешний) и алгоритм планирования, что и определяет совокупность необходимых характеристик.

Программный компонент управления платформами исполнения

Цель создания программного компонента управления платформами исполнения MCMS состояла в обеспечении унифицированного доступа к совместно используемому множеству кластеров, обслуживаемых различными системами управления (мультикластеру). Архитектура MCMS представлена на рис. 3.

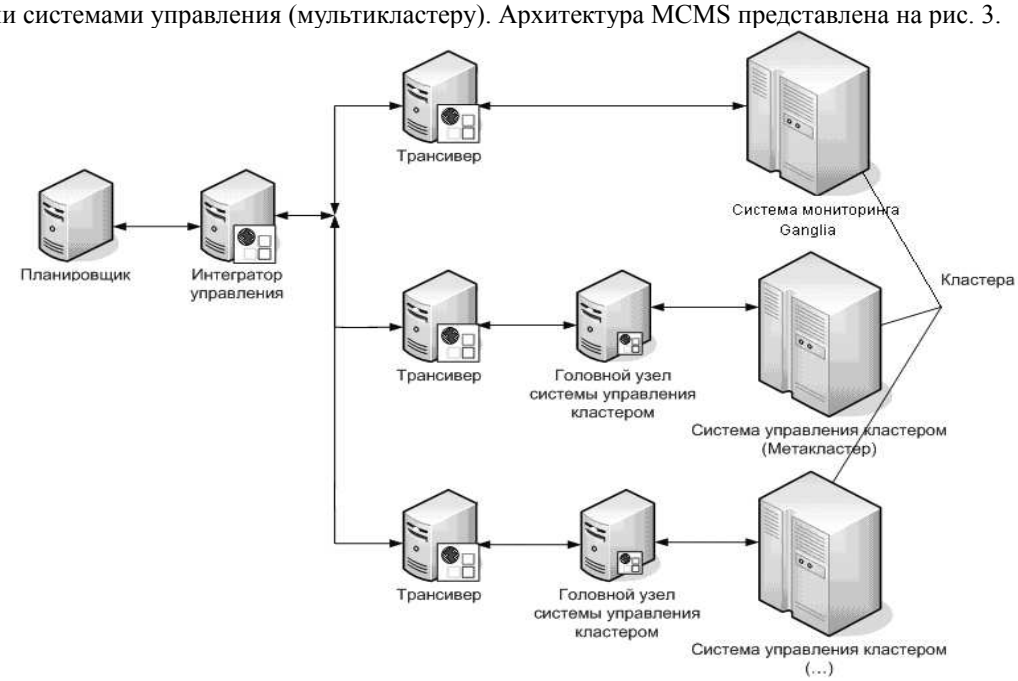

Рис. 3. Архитектура программного компонента управления платформами исполнения MCMS

MCMS состоит из двух основных модулей.

- Интегратор управления – web-сервис, предоставляющий унифицированный программный интерфейс управления мультикластером.

- Трансивер – модуль, который обеспечивает непосредственное взаимодействие с конкретной системой управления кластером (Microsoft Windows HPC Server 2008 [10], «Метакластер» [13], Torque [14]) или системой мониторинга ресурсов (Ganglia [6]).

Планировщик использует возможности MCMS и не входит в его состав.

Планировщик получает данные об аппаратных характеристиках и текущем состоянии узлов мультикластера в результате обращения к Интегратору управления, который ретранслирует запросы доступным трансиверам. Трансиверы получают значения характеристик для всех доступных узлов каждого кластера, используя соответствующий интерфейс взаимодействия с системой управления кластером или с системой мониторинга. В настоящий момент используются следующие механизмы доступа:

- трансивер Torque использует удаленный вызов команд интерфейса системы управления средствами удаленного терминала (ssh, [17]);
- трансивер Windows HPC Server 2008 использует программный интерфейс .NET API и командный интерфейс Windows PowerShell;
- трансивер Метакластера использует программный интерфейс .NET Remoting;
- трансивер Ganglia использует непосредственное TCP-соединение.

Совокупность предоставляемых характеристик фактически определяется суммарными ограничениями систем управления кластерами и мониторинга и в настоящий момент включает:

- характеристики кластера: имя кластера, количество узлов, топология сети, пропускная способность и латентность сетевых соединений (сетевые параметры не определяются автоматически и должны статически задаваться администратором);
- характеристики узла: DNS-имя узла, число ядер, частота и производительность ядер, объем оперативной памяти, список установленного программного обеспечения;
- состояние узла: текущий процент загрузки ядер, объем свободной физической памяти, текущий процент загрузки сетевых интерфейсов.

Перечисленный набор характеристик достаточен для реализации большинства стратегий планирования.

Заключение

Разработанная компонента управления платформами исполнения MCMS ориентирована на взаимодействие с различными системами управления кластерами, системами мониторинга ресурсов и планировщиками. Компонента позволяет получать актуальную информацию об аппаратном составе и текущей загруженности узлов кластера, используя различные механизмы сопряжения с системами мониторинга ресурсов, и предоставляет планировщику совокупность информации, достаточную для работы большинства стратегий планирования.

Исследования выполнены в рамках ОКР по теме «Разработка высокопроизводительного программного комплекса для квантово-механических расчетов и моделирования наноразмерных атомно-молекулярных систем и комплексов» [16].

ЛИТЕРАТУРА:

1. Saeed Iqbal, Rinku Gupta, Yung-Chin Fang. Planning Considerations for Job Scheduling in HPC Clusters // Dell Power Solutions, c. 133-136, 2005.
2. David Jackson, Quinn Snell, Mark Clement. Core Algorithms of the Maui Scheduler // Lecture Notes in Computer Science Job Scheduling Strategies for Parallel Processing: 7th International Workshop, JSSPP 2001, Cambridge, MA, USA, June 16, 2001. Revised Paper, Volume 2221, pp. 87-102, 2001, ISSN: 0302- 9743.
3. Uwe Schwiegelshohn, Ramin Yahyapour Improving First Come First Served job scheduilung by gang scheduling // Computer Engineering Institute, University Dortmund, Germany: JSSPP'98, LNCS 1459, p. 180-198, 1998;
4. Julita Corbalan Gonzalez. Coordinated Scheduling and Dynamic Performance Analysis in Multiprocessor Systems - [http://www.tdr.cesca.es/TESIS_UPC/AVAILABLE/TDX-0723102-094622//07Jcg07de08.pdf].
5. O. Sename, D. Simon, D. Robert. Feedback scheduling for real-time control of systems with communication delays // ETFA'03 9th IEEE International Conference on Emerging Technologies and Factory Automation, Lisbonne. Volume 2, 16-19. pp. 454 - 461 vol.2 , 2003.
6. Ganglia Monitoring System - [http://ganglia.sourceforge.net/].
7. Cacti: The Complete RRDTool-based Graphing Solution - [http://www.cacti.net].
8. Nagios - The Industry Standard in IT Infrastructure Monitoring - [http://www.nagios.org].
9. Systems Monitoring, Server Monitoring & Systems Management Software | Hyperic - [http://www.hyperic.com].
10. Windows HPC Server 2008 | Microsoft Supercomputing | Supercomputers - [http://www.microsoft.com/hpc/en/us/].
11. Understanding Job Scheduling Policies - [http://technet.microsoft.com/en-us/library/dd197402(WS.10).aspx].
12. Portable Batch System Administrators Guide – [http://lsec.cc.ac.cn/chinese/lsec/doc/v2.3_admin.pdf].
13. Система управления «Метакластер» – [www.cluster.software.unn.ru].

14. TORQUE Resource Manager – [http://www.clusterresources.com/products/torque].
15. Maui Cluster Scheduler – [http://www.clusterresources.com/products/maui].
16. В.Н. Васильев, А.В. Бухановский, С.А. Козлов, В.Г. Маслов, Н.Н. Розанов. Высокопроизводительный программный комплекс моделирования наноразмерных атомно-молекулярных систем // Научно-технический вестник. СПбГУ ИТМО «Технологии высокопроизводительных вычислений и компьютерного моделирования», № 54. СПб: Университетские телекоммуникации, 2008. С.3-13.17. SharpSSH - A Secure Shell (SSH) library for .NET – [http://www.tamirgal.com/blog/page/SharpSSH.aspx].