Разработка автоматизированного программного комплекса кластерных расчетов

Г.А. Макеев, А.Л. Штангеев, А.В. Юлдашев

В статье представлен программный комплекс, позволяющий автоматизировать расчеты задач на кластерных системах. Описывается структура программного комплекса и технологии, использованные при его реализации. Приводится пример применения программного комплекса для автоматизированной оптимизации гидродинамических моделей.

1. Введение

В настоящее время кластерные системы получили широкое распространение для решения научных, технических и промышленных задач, требующих значительных вычислительных ресурсов. Данная работа была инициирована в результате сотрудничества УГАТУ и ООО «РН-УфаНИПИнефть» в области организации кластерных расчетов для решения задач моделирования фильтрационных течений углеводородов в пористой среде. Начиная с 2000 года, в УГАТУ накапливался опыт эксплуатации и администрирования кластерных систем. В то же время в «РН-УфаНИПИнефть» имелась традиционная потребность в вычислительных ресурсах, вылившаяся в 2006 году в приобретение нового 8-узлового кластера на базе двухпроцессорных серверов HP с вычислительной коммуникационной средой Myrinet для решения задач гидродинамического моделирования. Тогда и началась разработка описываемого программного комплекса, который изначально строился на основе существовавшей в «РН-УфаНИПИнефть» клиент-серверной системы управления расчетами на вычислительном кластере Cluster Client-Server. Особенностями данной системы являлись автоматизация действий пользователя на всех этапах расчета гидродинамической модели на кластере и наличие клиентской программы, реализованной для работы в операционной системе Windows. В то же время серверная часть, полностью прозрачная с точки зрения пользователей кластера, обладала лишь базовой функциональностью менеджера ресурсов, характеризовалась политикой обслуживания заданий First Come First Served, назначала задания на узлы методом First Fit, не предусматривала наличие уникальных учетных записей пользователей и в целом имела низкие показатели масштабируемости и надежности.

В связи с этим в УГАТУ ведется разработка автоматизированного программного комплекса кластерных расчетов, в котором произведена попытка объединить наработки в области автоматизации действий пользователя, создания проблемно-ориентированного дружественного клиентского интерфейса и современные достижения в сферах управления и планирования в кластерных системах. Разработка собственного программного комплекса также связана с желанием учесть специфику решаемых задач, а именно, показатели эффективности выполнения их параллельных реализаций на кластерных системах, в целях оптимизации алгоритмов планирования и балансировки нагрузки для данного класса задач.

Сейчас рассматриваемый комплекс, в первую очередь, включает в себя полностью переработанную многопоточную реализацию серверного компонента для ОС Linux. Функционально сервер представляет собой сетевое приложение, принимающее запросы от клиентов, обрабатывающее и перенаправляющее их во внешнюю систему пакетной обработки заданий (СПОЗ). На данный момент реализована поддержка системы TORQUE. Кроме этого основными функциями сервера являются планирование ресурсов и создание контрольных точек заданий. Сервер совместим с разработанными ранее графическими клиентами, хотя в настоящее время ведется разработка интерфейса нового поколения, после чего планируется создать клиент, ориентированный на работу через веб-интерфейс. Также роль клиента может играть система автоматизированной оптимизации гидродинамических моделей на основе вычислительной среды Matlab. Кроме того, программный комплекс включает в себя модуль ведения статистики расчетов на

кластере, который позволяет с помощью веб-интерфейса получить детальную информацию обо всех выполненных заданиях, а также ресурсах, затраченных пользователями кластера.

За время своего существования программный комплекс активно использовался для организации и автоматизации кластерных расчетов задач гидродинамического моделирования на нескольких кластерных системах с применением многопоточных и MPI-версий гидродинамических симуляторов BOS, More и Eclipse. Однако, заложенные в него возможности, позволяют также поддерживать расчеты с применением решателей, входящих в пакеты ANSYS, STAR-CD и др.

Средствами нашего программного комплекса с февраля 2008 года ведется решение задач гидродинамического моделирования на 84 узлах (672 ядрах) кластера УГАТУ. Рассчитано более сорока тысяч моделей с использованием различных гидродинамических симуляторов. Большая часть расчетов осуществляется при оптимизации гидродинамических моделей с помощью системы автоматизированной оптимизации.

2. Структура программного комплекса

К программному комплексу (ПК) предъявлялись следующие основные требования:

- поддержка запуска многопоточных и MPI-программ;
- наличие дружественного проблемно-ориентированного клиентского интерфейса;
- автоматизированное выполнение «рутинных» действий пользователей;
- обеспечение отказоустойчивости;
- поддержка произвольного набора плавающих лицензий и различных политик лицензирования приложений;
- учет специфики обрабатываемых заданий;
- наличие средств для представления и анализа статистики расчетов.

Для обеспечения предъявленных требований структурно в ПК было выделено три основных программных компонента: клиент, сервер и модуль статистики. Ниже подробно рассматриваются их назначение и возможности.

2.1 Клиент

Клиентская часть предназначена для управления заданиями в системе со стороны пользователей посредством взаимодействия с сервером. На данный момент существует ряд клиентов с графическим интерфейсом, взаимодействующих с сервером с помощью специализированного сетевого протокола: Windows-клиент (Delphi), кроссплатформенный клиент (Python), система автоматизированной оптимизации гидродинамических моделей на основе вычислительной среды Matlab и др. Первые два клиента ориентированы на индивидуальную постановку заданий, как правило, небольшого числа для каждого пользователя, в то время как в системе автоматизированной оптимизации гидродинамических моделей при решении одной оптимизационной задачи на практике генерируются десятки заданий.

Рассмотрим процесс индивидуального расчета гидродинамической модели на кластерной системе с помощью одного из первых двух клиентов. При помощи проблемно-ориентированного диалогового окна постановки задания выбираются все необходимые параметры запуска и указываются входные файлы модели. Далее происходит автоматический анализ главного входного файла модели на предмет подключения дополнительных файлов, все необходимые для расчета файлы архивируются и копируются на кластер (сейчас используется протокол FTP, в разработке SFTP), где затем будут распакованы сервером. Причем на кластер копируются только те файлы, которые были изменены, если производится повторный расчет. После того, как освобождаются необходимые для запуска ресурсы, сервер запускает задание. Клиент производит периодический опрос сервера на предмет изменения статуса всех запущенных заданий. Также клиент получает дополнительную информацию о состоянии расчета, которая извлекается сервером из выходных файлов задания. По завершению расчета пользователь может получить результаты моделирования в виде заархивированных выходных файлов.

Таким образом достигается автоматизация действий пользователя на всех этапах расчета гидродинамической модели на кластере, причем от пользователя требуется лишь поставить задание в систему и забрать результаты с помощью дружественного графического интерфейса клиентской программы.

2.2 Сервер

Сервер представляет собой непривилегированный демон ОС Linux, расположенный на управляющем узле кластера, и выполняет следующие основные функции:

- обслуживание клиентских запросов;
- планирование ресурсов кластерной системы;
- управление заданиями средствами внешней СПОЗ (постановка, запуск и удаление);
- мониторинг состояния запущенных заданий по выходным файлам;
- создание контрольных точек заданий.

Серверный компонент, входящий в Cluster Client-Server, не был рассчитан на взаимодействие с внешней СПОЗ. В связи с этим сервер должен был выполнять функции полноценного мониторинга и управления заданиями, для чего использовались оболочки rsh/ssh. Однако такой подход оказался неудовлетворительным с точки зрения масштабируемости и надежности, что, в свою очередь, привело к необходимости интеграции с внешней системой пакетной обработки. В настоящий момент сервер, входящий в рассматриваемый программный комплекс, поддерживает интеграцию с широко распространенным в мире высокопроизводительных вычислений менеджером ресурсов TORQUE [1]. В перспективе планируется поддерживать и другие системы: LoadLeveler и Sun Grid Engine (SGE).

Сервер реализован как многопоточное приложение, и каждая открытая клиентом сессия на стороне сервера обслуживается отдельным потоком. При поступлении запроса на подключение клиент проходит процедуру аутентификации. Индивидуальные учетные данные (имя пользователя и пароль) пользователи получают после регистрации с помощью соответствующей вебстраницы. При проверке аутентификационных данных сервер использует информацию о пользователях, расположенную в специальном файле. Пользователь программного комплекса не является системным пользователем кластера, тем не менее, каждому пользователю ставится в соответствие некоторый системный пользователь, от имени которого посредством внешней СПОЗ будут запускаться задания.

Текущая реализация сетевого протокола не является безопасной с точки зрения удаленного доступа, так как пароль пользователя при инициализации очередной сессии передается в открытом виде. Однако безопасное удаленное использование возможно, например, с помощью организации VPN-туннеля или туннелирования через SSH. Для устранения данного недостатка программного комплекса в настоящее время разрабатывается клиент-серверный кроссплатформенный интерфейс (API) нового поколения на языке C, ориентированный на использование в приложениях, написанных на языках C/C++, Delphi, PHP, Python и Matlab.

При постановке задания в очередь сервер находит соответствующий ему класс заданий, заранее описанный в конфигурационном файле. Затем на основании параметров конфигурации (команды запуска, используемого командного интерпретатора, команд выполняемых до запуска и по окончании расчета, и т.д.) и набора макросов (%u — имя владельца, %j — имя задания, %np — число параллельных ветвей и т.д.) для данного класса заданий формируется скрипт, который будет выполняться внешней СПОЗ. Таким образом, класс заданий — набор общих параметров и действий вместе с макросами, а само задание можно назвать реализацией (объектом) данного класса со значениями макросов, передаваемыми программой-клиентом или назначаемыми конфигурацией. Далее задание с помощью функций библиотеки (API) системы пакетной обработки помещается в очередь.

Помимо стандартного мониторинга задач как набора процессов, осуществляемого средствами внешней СПОЗ, в серверной части программного комплекса предусмотрена дополнительная функциональная возможность расширенного слежения за задачами. В частности, имеется возможность текстового анализа выходного файла выполняющейся задачи, позволяющая например, определить текущий временной шаг либо возникновение ошибки. Это позволяет поль-

зователю, не выходя на кластер при помощи стандартных средств, отслеживать процесс выполнения задания.

В целях обеспечения надежности, а также в перспективе осуществления динамической балансировки нагрузки в вычислительной системе, сервер поддерживает периодическое создание контрольных точек выполняющихся заданий. На кластере УГАТУ данная функциональность обеспечивается как для многопоточных, так и для MPI-программ. Для создания контрольных точек используется библиотека Berkeley Lab Checkpoint/Restart (BLCR) [2], причем создание контрольных точек MPI-программ осуществляется при использовании MPI-реализации MVAPICH2 [3], которая в свою очередь поддерживает BLCR.

Планировщик в текущей реализации программного комплекса представляет собой самостоятельный поток серверного приложения, осуществляющий поиск и выделение необходимых для запуска задачи ресурсов в соответствии с политикой Least Utilized Node First. Предусмотрен также механизм резервации ресурсов на основании прогнозируемого времени пребывания задачи в состоянии исполнения, необходимый при использовании алгоритма обратного заполнения Backfill [4,5] выбора задания для назначения. Однако пока отсутствуют механизмы прогнозирования времени на стороне сервера или его явного указания на стороне клиента.

Помимо традиционных ресурсов (процессоры, память, и т.п.) существует возможность конфигурации дополнительных ресурсов, например, лицензий на используемое на кластере программное обеспечение и правил их вычитания. Данные ресурсы представляют собой обобщенный аналог GRES-ресурсов (generic consumable resources), поддерживаемых, в частности, планировщиком Maui [6]. Произвольные правила вычитания ресурсов используются в том случае, когда условия лицензирования программного обеспечения являются не совсем стандартными. Зачастую количество лицензий, требуемое на запуск параллельной версии программы, равно количеству запускаемых ветвей параллельной программы. Однако параллельная версия гидродинамического симулятора More, к примеру, при наличии N лицензий может быть запущена на 2^N ядрах в силу особенностей ее масштабируемости.

Рассмотрим пример конфигурационного файла серверной части ПК:

```
[ .global ]       # глобальные параметры
      root_dir = /path/to/root/directory
      nodelist_file = /path/to/nodefile
      port = 12345
      ...
[ .resources ]    # ресурсы, определяемые администратором
      resource 'lic-ex1', float:10, dedicate: 1
      resource 'lic-ex2', float:20, dedicate: parallel = %np
      resource 'lic_ex3', float:20, dedicate: parallel = log(%np)/log(2)+1
      ...
[ class1 ]        # первый класс заданий
      resources = 'lic-ex1', 'lic-ex2'
      stagein_cmd = unzip /%u/%j.zip
      multithreaded = true
      cmd = /path/to/executable1 -nt %np /%u/%j.DATA
      stageout_cmd = zip -1 %j.out %j.log
      ckpt_cmd = /path/to/ckpt_script %job_master_node %jpid
      ckpt_interval = 28800
      ...
[ class2 ]        # второй класс заданий
      multithreaded = false
      resources = 'lic-ex1', 'lic-ex3'
      cmd = /path/to/mpi/bin/mpirun -machinefile %mf -n %np \
            /path/to/executable2 /%u/%j.DATA
      ...
```

Рис. 1. Пример конфигурационного файла серверной части ПК.

Конфигурационный файл разделен на несколько секций. Две из них являются служебными — секция глобальных настроек ([.global]) и секция описания дополнительных ресурсов ([.resources]).

В секции глобальных настроек следует отметить наличие параметра root_dir, значением которого является корневая директория, расположенная, как правило, на сетевом хранилище и за пределами которой сервер не имеет права изменять файлы из соображений безопасности. В секции дополнительных ресурсов определяются ресурсы, не поддерживаемые внешней СПОЗ и правила их вычитания. Например, ресурс lic-ex3 есть плавающие лицензии с правилом вычитания $\lfloor \log_2 np \rfloor + 1$, где *np* — число параллельных ветвей запускаемой программы. Ключевое слово parallel указывает на то, что лицензия будет вычитаться, только если $np > 1$. Аналогичным образом можно ввести и привязанные к узлам лицензии, и какие-либо другие специфические ресурсы.

В оставшихся секциях приведен пример конфигурации двух разных классов заданий. Первый класс ([class1]) описывает параметры многопоточной программы executable1, использующей лицензии lic-ex1 и lic-ex3, для которой с периодом 28 800 секунд (8 часов) будут создаваться контрольные точки. Во втором классе показано, как могут быть настроены параметры запуска MPI-программы. Также в приведенном примере продемонстрирована возможность совместного использования лицензий lic-ex1 заданиями обоих классов.

2.3 Модуль статистики

Модуль ведения статистики расчетов на кластере позволяет с помощью веб-интерфейса, реализованного на языке PHP, получить доступ к информации о расчетах на кластерной системе, которая помещается сервером в MySQL базу данных. Имеются возможности представления детальной информацию обо всех выполненных заданиях в виде таблиц, а также графического представления информации о ресурсах, затраченных пользователями кластера, и использовавшихся программах в виде гистограмм и диаграмм.

3. Использование ПК для автоматизированной оптимизации гидродинамических моделей

Одной из целей разработки программного комплекса кластерных расчетов являлось предоставление программного интерфейса для массовых расчетов гидродинамических моделей, который позволил бы избежать системного программирования при разработке интеллектуальных клиентских приложений. Так, например, хотелось избавить клиентские приложения от необходимости поддержки различных СПОЗ, учета политик лицензирования используемого программного обеспечения и т.д.

Решение задач оптимизации и анализа чувствительности геолого-гидродинамических моделей принципиально требует проведения массовых расчетов моделей с привлечением большого количества вычислительных ресурсов, что сопряжено с проведением большого объема рутинной работы специалистов по моделированию. Общий поток работ включает в себя, в частности:

- генерацию множества вариантов некоторой модели;
- распределение задач по доступным вычислительным мощностям;
- сбор и обработку результатов выполнения моделей.

В настоящее время в УГАТУ ведется разработка системы автоматизированной оптимизации гидродинамических моделей на основе вычислительной среды Matlab, которая в рамках программного комплекса играет роль интеллектуального клиента. Типичный цикл оптимизации модели с помощью этой системы включает следующие этапы:

- постановка цели оптимизации: какие целевые функции мы хотим оптимизировать;
- определение параметров оптимизации: какие параметры модели мы можем менять, чтобы добиться цели;
- определение области определения и области поиска всех параметров;

- анализ чувствительности целевых функций к искомым параметрам;
- автоматическая оптимизация параметров с помощью алгоритмов оптимизации общего назначения (генетические алгоритмы, градиентные алгоритмы, нейросети);
- анализ полученных решений.

Выполнение этих этапов подразумевает расчет большого количества вариантов гидродинамических моделей на всех доступных вычислительных ресурсах. Система автоматизированной оптимизации в рамках программного комплекса кластерных расчетов дает возможность пользователю сконцентрироваться на решении оптимизационной задачи, снижая затраты на управление расчетами.

4. Заключение

Средствами разрабатываемого автоматизированного программного комплекса кластерных расчетов с февраля 2008 года организован доступ ООО «РН-УфаНИПИнефть» к кластеру УГАТУ для решения задач гидродинамического моделирования. Серверный компонент программного комплекса является масштабируемым приложением благодаря многопоточной реализации, а также использованию внешней системы пакетной обработки заданий TORQUE. Надежность выполнения программ на кластере поддерживается за счет создания контрольных точек заданий средствами библиотеки BLCR. Использование системы автоматизированной оптимизации в качестве клиента делает возможным проведение массовых кластерных расчетов при решении задач оптимизации и анализа чувствительности геолого-гидродинамических моделей с минимумом затрат на управление расчетами.

В дальнейшем планируется уделить особенное внимание развитию существующих алгоритмов планирования и балансировки нагрузки с учетом специфики выполняемых заданий. Также перспективной является разработка клиентского компонента на основе веб-интерфейса.

Литература

1. Tera-scale Open-source Resource and QUEue manager (TORQUE): [http://www.clusterresources.com/pages/products/torque-resource-manager.php]
2. Berkeley Lab Checkpoint/Restart: [http://ftg.lbl.gov/CheckpointRestart/CheckpointRestart.shtml].
3. Q. Gao, W. Yu, W. Huang and D. K. Panda Application-Transparent Checkpoint/Restart for MPI Programs over InfiniBand // Int'l Conference on Parallel Processing (ICPP), 2006.
4. Топорков В.В. Модели распределенных вычислений. — М.: ФИЗМАТЛИТ, 2004. - 320 с.
5. D. Jackson, Q. Snell, and M. Clement Core algorithms of the Maui scheduler // Proceedings of 7th Workshop on Job Scheduling Strategies for Parallel Processing: Lecture Notes Computer Science Volume 2221, Cambridge, MA, USA. –2001. –P. 87-102.
6. Maui Cluster Scheduler: [http://www.clusterresources.com/pages/products/maui-cluster-scheduler.php].