На правах рукописи

Соболев Сергей Игоревич

УПРАВЛЕНИЕ ПОТОКАМИ ЗАДАНИЙ В РАСПРЕДЕЛЕННЫХ НЕОДНОРОДНЫХ ВЫЧИСЛИТЕЛЬНЫХ СРЕДАХ

Специальность 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Автореферат
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва – 2008

Работа выполнена в Научно-исследовательском вычислительном центре Московского государственного университета имени М.В. Ломоносова

Научный руководитель: доктор физико-математических наук, член-корреспондент РАН Воеводин Владимир Валентинович

Официальные оппоненты: доктор физико-математических наук, член-корреспондент РАН Абрамов Сергей Михайлович

доктор физико-математических наук, Ильин Вячеслав Анатольевич

Ведущая организация: Межведомственный суперкомпьютерный центр РАН

Защита состоится 4 июля 2008 года в 15 часов на заседании диссертационного совета Д 501.002.09 Московского государственного университета имени М.В. Ломоносова по адресу: 119992, г. Москва, Ленинские горы, д.1, стр. 4, НИВЦ МГУ, конференц-зал.

С диссертацией можно ознакомиться в библиотеке НИВЦ МГУ.

Автореферат разослан 30 мая 2008 года.

Учёный секретарь
диссертационного совета Суворов В.В

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Развитие телекоммуникационных технологий сегодня позволяет эффективно объединять распределенные вычислительные ресурсы и формировать вычислительные среды, не уступающие традиционным суперкомпьютерам в решении представительных классов больших задач. У таких сред есть безусловные достоинства: они доступны, строятся на уже существующей компьютерной инфраструктуре и потому не требуют дополнительных вложений. По уровню своей производительности такие среды приближаются к классическим суперкомпьютерам. Однако на пути их построения и использования стоят серьезные проблемы, препятствующие их массовому внедрению в вычислительную практику: большая неоднородность, значительная распределенность, непредсказуемость конфигурации, сложность администрирования.

Активные исследования в области организации распределенных вычислений и разработки метакомпьютерных систем начались не так давно, с середины 90-х годов прошлого века. Развитие исследований происходит сразу в нескольких направлениях – от адаптации конкретных задач для запуска на множестве компьютеров, доступных через сеть Интернет, до формирования общих стандартов, определяющих методы и технологии создания глобальных многофункциональных распределенных систем. Разрабатываемые программные системы, будучи ориентированными на достаточно узкий спектр решаемых задач, обладают целым рядом ограничений. Так, наиболее развитый программный комплекс Globus Toolkit имеет внушительный объем программного кода, требует значительных усилий по установке, настройке и поддержанию своей инфраструктуры в работоспособном состоянии, что фактически исключает его использование при необходимости оперативного проведения распределенных вычислительных экспериментов.

Большой разброс в направлениях исследований, отсутствие серьезной апробации и во многом экспериментальный характер работ пока не позволяет говорить о наличии в настоящее время комплексного программного обеспечения для поддержки приложений и пользователей в распределенных вычислительных средах. Необходим инструментарий, позволяющий формировать вычислительные среды из доступных компьютерных ресурсов, допускающий легкую адаптацию прикладных программ и эффективную организацию расчетов.

Целями данной диссертационной работы является разработка подходов к решению больших задач в распределенных вычислительных средах, разработка технологии управления потоками заданий в распределенных вычислительных средах, реализация программного комплекса для организации распределенных вычислительных сред на основе доступных компьютерных ресурсов и проведения масштабных расчетов в таких средах.

Основные результаты, выносимые на защиту:

1. Разработаны эффективные подходы к решению больших задач в масштабных неоднородных распределенных вычислительных средах с динамически изменяющейся конфигурацией, объединяющих все основные типы компьютерных ресурсов.
2. Разработана архитектура и технологическая основа системы управления потоками заданий в вычислительных средах. Система объединяет механизмы распределения заданий по доступным ресурсам, средства управления заданиями и мониторинга текущего состояния среды, обработки статистических данных и визуализации.

3. Реализован программный комплекс для организации распределенных вычислительных сред и проведения расчетов на доступных компьютерных ресурсах. Разработанный программный комплекс может быть использован как для оперативного развертывания вычислительных экспериментов различного масштаба, так и для создания на его основе постоянно действующих вычислительных сервисов.
4. Разработанный программный комплекс успешно прошел апробацию в ходе решения большого числа вычислительно сложных задач биоинженерии, биоинформатики, проектирования лекарственных препаратов, электродинамики и ряда других с использованием множества географически распределенных компьютеров с различными архитектурами, режимами работы и административной принадлежностью.

Научная новизна диссертации состоит в разработке принципов построения системы для управления потоками заданий в распределенных вычислительных средах. Предложенная архитектура ориентирована на многопользовательскую работу в средах, характеризующихся большими количественными характеристиками, динамичностью, неоднородностью входящих в нее узлов, большой латентностью во взаимодействии параллельных процессов. Архитектура обладает свойствами распределенности, переносимости, масштабируемости.

Все исследования, выполненные в рамках данной работы, ориентированы на применение технологии распределенных вычислений для решения реальных вычислительно сложных задач из различных областей науки.

Практическая значимость. Представленная в работе технология ориентирована на максимальное отражение особенностей вычислительных сред, вычислительных и

организационных реалий: неоднородность, динамичность и априорную неопределенность конфигурации среды, простоту организации и проведения расчетов, использование всех основных административно определенных режимов работы вычислительных ресурсов.

Применимость и востребованность результатов диссертационной работы определяется двумя факторами. Во-первых, распространение компьютерных методов исследований в таких областях, как биоинженерия и биоинформатика, фармацевтика, машиностроение, энергетика, проектирование новых материалов и других ведет ко все более и более широкому использованию высокопроизводительных вычислительных систем, а в этой нише для многих задач предлагаемый подход имеет наилучшие экономические показатели. Во-вторых, развитие телекоммуникационных технологий ведет к неизбежному сближению параметров традиционных суперкомпьютеров и вычислительных сред, поэтому спектр расчетов, эффективно проводимых по технологиям данной работы, будет неуклонно расширяться.

Апробация работы. Основные положения работы прошли обсуждение на научных семинарах НИВЦ МГУ, на совещаниях по проблемам развития высокопроизводительных вычислений в УГАТУ (г. Уфа). Результаты работы представлялись на всероссийской научной конференции "Научный сервис в сети Интернет: технологии распределенных вычислений" (г. Новороссийск, 19-24 сентября 2005 г.), на всероссийской научной конференции "Научный сервис в сети Интернет: технологии параллельного программирования" (г. Новороссийск, 18-23 сентября 2006 г.), на международных конференциях "Распределенные вычисления и Грид-технологии в науке и образовании" (г. Дубна, 29 июня - 2 июля 2004 г., 26-30 июня 2006 г.), на научной конференции "Ломоносовские чтения" (г. Москва, 21 апреля 2005 г.).

Разработанная система была включена в состав экспозиции Роснауки и демонстрировалась на выставке высоких технологий в рамках XI Петербургского международного экономического форума (г. Санкт-Петербург, 8-10 июня 2007 г.).

Публикации. По теме диссертации опубликовано 6 научных работ. На систему X-Com, лежащую в основе разработанного программного комплекса, получено свидетельство Роспатента о регистрации программ для ЭВМ № 2006611361 от 12.05.2006.

Структура и объем работы. Диссертация состоит из введения, 3-х глав, заключения, приложения и списка литературы. Общий объем диссертации – 98 страниц.

СОДЕРЖАНИЕ РАБОТЫ

Введение носит постановочный характер, содержит обоснование актуальности и необходимости создания программных средств для управления потоками заданий в распределенных неоднородных вычислительных средах.

Первая глава посвящена обзору и анализу наиболее известных программных средств для организации распределенных вычислений. На основе этого анализа выбирается направление исследований, уточняется постановка задачи и требования к результатам исследований.

При организации вычислений в распределенной среде центральную роль играет программное обеспечение, позволяющее всем компонентам среды работать над единой задачей, распределяющее задания и собирающее результаты, обеспечивающее взаимодействие пользователей со средой и выполняющимися в ней процессами. В настоящее время

исследования и разработка таких программных средств ведется в нескольких направлениях.

Одно из направлений заключается в разработке методов организации глобальных вычислительных экспериментов, использующих ресурсы компьютеров, подключенных к сети Интернет. Среди наиболее известных проектов такого рода стоит отметить научный проект SETI@home, анализирующий полученные от радиотелескопа данные с целью поиска в них фрагментов сигналов искусственного происхождения. Существует значительное число аналогичных проектов, решающих конкретные задачи из других областей науки. Схожесть подобных проектов и их основное ограничение – возможность работы в рамках одного проекта только над одной задачей – привело к созданию ряда платформ, унифицирующих разработку подобных проектов за счет использования единых клиентских модулей, интерфейсов и протоколов. В диссертации обсуждается одна из таких платформ – BOINC.

Другое направление развития метакомпьютерных вычислений опирается на использование процессорного времени простаивающих компьютеров в пределах некоторой организации. Изначально на такой режим работы ориентировалась система распределения последовательных задач Condor. Со временем данная система эволюционировала, и в настоящее время она представляет собой систему управления очередями заданий, которые могут запускаться на доступных распределенных ресурсах. Система Condor поддерживает большинство современных программно-аппаратных платформ, она способна работать как в рамках локальной сети, так и через сеть Интернет, и ориентирована на управление прохождением множества независимых приложений.

Следующее направление развития распределенных компьютерных технологий состоит в разработке общих стандартов и проектирование на их основе многофункционального программного обеспечения,

поддерживающего значительный набор распределенных сервисов. Инициатива Grid-технологий направлена на создание согласованной открытой стандартизованной среды, обеспечивающей скоординированное разделение ресурсов различного плана. Базовым программным обеспечением Grid де-факто стал программный комплекс Globus Toolkit. Безусловно, это направление работы является крайне перспективным, однако в настоящий момент оно во многом носит экспериментальный характер. Globus Toolkit тяжел в установке и сложен в использовании, к тому же от версии к версии внутренние соглашения и протоколы системы могут меняться без учета обратной совместимости. Полноценная поддержка этого комплекса обеспечивается только для среды UNIX. ПО gLite, взятое за основу в европейском проекте EGEE, является несколько более проблемно-ориентированным, однако перечисленные особенности характерны и для него

Еще одним направлением развития является разработка программных инструментариев для адаптации прикладных программ, обладающих значительным ресурсом внутреннего параллелизма, для их выполнения в распределенных неоднородных компьютерных средах. Примером такого инструментария служит базовый уровень системы X-Com.

Проведенные исследования программных средств для организации метакомпьютерных сред позволяют выделить целый ряд общих моментов. Системы, ориентированные непосредственно на организацию распределенных вычислений, имеют сходную архитектуру, они базируются на клиент-серверных технологиях, осуществляют распределение заданий по инициативе клиента и обладают множеством других общих черт. Основные отличия проявляются в списке поддерживаемых платформ, интерфейсах программирования (API), возможности оперативного развертывания метакомпьютерных сред, трудозатратах на адаптацию

прикладных приложений для выполнения в распределенной среде, уровне предоставляемых сервисов для пользователей.

Однако для комплексной поддержки пользователей и прикладных программ в распределенной вычислительной среде полноценного решения пока не существует. Необходимо прежде всего сформулировать набор требований для программного комплекса, поддерживающего наиболее распространенные программно-аппаратные платформы, позволяющего оперативно проводить масштабные распределенные расчеты и создавать на своей основе постоянно действующие "метакомпьютерные вычислительные центры", предоставляющие удобный функциональный пользовательский интерфейс для работы с приложениями и обеспечивающие возможность подключения дополнительных доступных ресурсов. Анализ структуры прикладных задач и существующих систем для организации распределенных вычислений показал, что проектируемый комплекс должен соответствовать следующим требованиям.

Масштабируемость. Для достижения максимальной производительности программный комплекс должен поддерживать среды со значительным числом (порядка десятков тысяч) подключенных узлов. Комплекс должен обеспечивать надежную и эффективную работу в рамках масштабных распределенных сред.

Распределенность. Это требование отражает одно из базовых свойств комплекса. Комплекс должен поддерживать работу с распределенными ресурсами через сеть Интернет, при этом удаление и географическое местонахождения подключаемых узлов не должно влиять на работоспособность этих узлов и комплекса в целом.

Переносимость. Необходимо обеспечить поддержку максимального числа современных программно-аппаратных платформ как для клиентской, так и для серверной части комплекса. Для каждой клиентской платформы должна быть реализована возможность исполнения программного кода

прикладной задачи, созданного и оптимизированного именно для этой платформы.

Неоднородность. Программный комплекс должен быть ориентирован на совместную работу узлов на основе различных платформ, с различными конфигурациями и политиками использования ресурсов. Должен быть обеспечен широкий спектр режимов работы клиентской части программного комплекса, обеспечивающий его корректную совместную работу со штатным программным обеспечением узлов, позволяющий в каждом конкретном случае выбрать оптимальный способ запуска распределенных приложений.

Оперативность. Комплекс должен поддерживать максимально простую и быструю процедуру установки и настройки. Это требование наиболее актуально для клиентской части. Использование комплекса не должно предусматривать наличие административного доступа к ресурсам.

Адаптивность прикладных задач. Должна быть обеспечена легкая адаптация прикладных программ для работы в распределенных вычислительных средах.

Пользовательская функциональность. Пользователи программного комплекса должны работать с привычным окружением, аналогичным окружению традиционной высокопроизводительной вычислительной системы. В частности, должны быть реализованы средства управления заданиями пользователей в распределенной среде.

Информативность. Программный комплекс должен предоставлять наглядную и подробную информацию о ходе вычислений, о состоянии распределенной среды. Комплекс должен иметь механизмы для сбора и анализа статистики по завершенным расчетам.

В качестве технологической основы разрабатываемого программного комплекса был выбран базовый уровень системы метакомпьютинга X-Com. Система X-Com, ориентированная на адаптацию и запуск приложений в распределенной среде,

является инструментарием низкого уровня, однако ее архитектура хорошо соответствует ряду перечисленных выше требований, а именно масштабируемости, распределенности и переносимости. Разрабатываемый программный комплекс, используя базовый уровень X-Com, предоставляет набор сервисов для комплексной поддержки пользователей и приложений при работе в распределенной вычислительной среде.

Вторая глава посвящена исследованию методов управления потоками заданий в метакомпьютерных средах, создаваемых на базе технологии X-Com. В таких средах можно выделить три независимых уровня управления заданиями: серверный, сетевой и клиентский. Предлагается архитектура программного комплекса X-Com/VMC (Рис. 1), реализующего распределение заданий на всех этих уровнях.

Центральной частью X-Com/VMC является подсистема, работающая на серверном уровне и позволяющая организовывать потоки заданий по схеме, привычной для пользователей традиционных высокопроизводительных систем. Обсуждаются два подхода к организации очередей заданий:

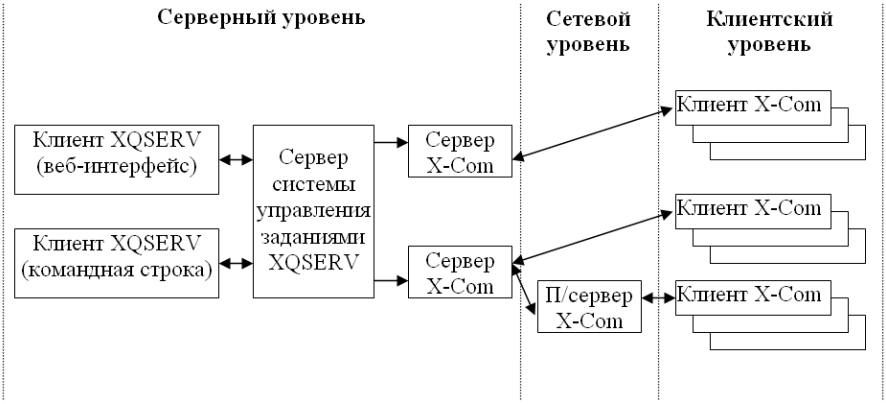

Рис. 1. Архитектура программного комплекса X-Com/VMC

однопоточный, обеспечивающий последовательное выполнение каждого из заданий и задействующий для их решения все доступные компьютерные ресурсы, и многопоточный, позволяющий разделить всю среду на классы по заданным признакам и осуществить одновременное выполнение нескольких приложений в распределенной среде.

Подсистема управления заданиями реализуется на основе клиент-серверной технологии. Сервер подсистемы, работая совместно с серверными компонентами X-Com, осуществляет координацию всех расчетов в распределенной среде. Сервер поддерживает структуры данных для хранения очереди заданий, осуществляет запуск и останов заданий, ведет статистику по использованию ресурсов вычислительной среды. Клиентская часть подсистемы предоставляет интерфейс для работы пользователей в интерактивном режиме либо с помощью специализированных веб-сервисов. Интерфейс включает в себя набор команд, позволяющих ставить задания в очередь, наблюдать за состоянием очереди и заданий, останавливать задания, осуществлять административные функции.

Для оптимизации сетевых обменов и увеличения масштабируемости распределенных вычислительных сред в архитектуру X-Com вводится дополнительный сетевой уровень промежуточных серверов. Промежуточные серверы X-Com позволяют организовать сетевую структуру распределенной среды в виде произвольного дерева. Они берут на себя функцию буферизации входящих и исходящих данных между центральным сервером X-Com и заданными подмножествами вычислительных узлов. С точки зрения центрального сервера X-Com промежуточный сервер представляет собой обычный узел с высокой производительностью, а с точки зрения нижележащих узлов промежуточный сервер является центральным сервером системы. Такая организация позволяет в ходе расчета снизить нагрузку на центральный сервер системы, а также подключать к

расчету ресурсы с нестабильными коммуникационными каналами.

Одним из важнейших моментов обеспечения эффективной обработки потоков заданий в распределенной среде является выбор режима запуска клиентской части X-Com на вычислительных узлах. Во второй главе диссертационной работы исследуются различные способы запуска клиентов X-Com: в монопольном режиме, в моменты простоя, в фоновом режиме с пониженным приоритетом, с использованием штатных систем очередей вычислительных систем. Описываются механизмы, реализующие указанные способы запуска и составляющие основу клиентского уровня управления прохождением заданий.

В любой момент времени пользователи распределенных вычислительных сред должны иметь возможность получить актуальную информацию о состоянии среды в целом и отдельных ее компонентов, а также о ходе текущих расчетов в ней. Программный комплекс X-Com/VMC поддерживает несколько способов визуализации процесса вычислений: выдача подробной технической информации в формате HTML, удобном для просмотра в любом браузере, и модульный механизм, анализирующий вывод сервера X-Com в формате XML и позволяющий отобразить различные аспекты вычислительного процесса в наглядной графической форме. Для предоставления статистики по завершившимся расчетам в состав комплекса X-Com/VMC включена программа для анализа лог-файлов сервера X-Com. Эта программа позволяет оценить затраченные на вычисления ресурсы (общее время, процессорное время, объем переданных данных), а также предоставляет сведения об эффективности использования ресурсов. Эффективность вычислений в распределенной среде может оцениваться с различных точек зрения: минимизация накладных расходов (коммуникационная эффективность), отсутствие избыточности вычислений и потерь порций данных (комплексная

эффективность), минимизация сетевых обменов. В конце второй главы обсуждаются факторы, влияющие на производительность среды и эффективность ее использования для прикладных расчетов.

Третья глава диссертационной работы посвящена опыту практического применения разработанного программного комплекса X-Com/VMC. Глава содержит рекомендации по установке, настройке и использованию комплекса.

С использованием программного комплекса X-Com/VMC был выполнен ряд расчетов для решения реальных прикладных задач. Расчеты для задачи дифракции электромагнитного поля проводились на фоне работы суперкомпьютерного комплекса НИВЦ МГУ, причем для решения использовались только простаивающие ресурсы центра. Несмотря на это, комплексная эффективность расчета составляла около 82%. В разное время к расчету было подключено от 78 до 144 процессоров комплекса. Задача целиком была решена за двое суток, при этом суммарное процессорное время составило 196 процессоро-дней.

Другой масштабный вычислительный эксперимент, иллюстрирующий возможность эффективного использования программного комплекса X-Com/VMC в распределенной неоднородной среде, проводился с задачей, являющейся частью процесса компьютерного проектирования лекарств. В расчете участвовало более 150 компьютеров НИВЦ МГУ и ЮУрГУ (г. Челябинск) различных конфигураций, работающих под управлением ОС Linux и MS Windows. Расчет в общей сложности продолжался около 10 дней, при этом суммарное процессорное время расчета составило 4,8 процессоро-лет, а суммарная производительность распределенной вычислительной среды в основное время расчета превысила 1 Tflops.

В этой же главе описываются методы, позволяющие проводить исследования самих распределенных сред, выявлять особенности их компонентов и на основе полученных данных

повышать эффективность их использования. На примере задачи электромагнитной динамики исследуется зависимость эффективности расчета от размера вычислительных порций, от использования различных режимов компиляции вычислительного модуля, от аппаратных характеристик узлов.

В **заключении** сформулированы основные результаты диссертационной работы:

1. Разработаны эффективные подходы к решению больших задач в масштабных неоднородных распределенных вычислительных средах с динамически изменяющейся конфигурацией, объединяющих все основные типы компьютерных ресурсов.
2. Разработана архитектура и технологическая основа системы управления потоками заданий в вычислительных средах. Система объединяет механизмы распределения заданий по доступным ресурсам, средства управления заданиями и мониторинга текущего состояния среды, обработки статистических данных и визуализации.
3. Реализован программный комплекс для организации распределенных вычислительных сред и проведения расчетов на доступных компьютерных ресурсах. Разработанный программный комплекс может быть использован как для оперативного развертывания вычислительных экспериментов различного масштаба, так и для создания на его основе постоянно действующих вычислительных сервисов.
4. Разработанный программный комплекс успешно прошел апробацию в ходе решения большого числа вычислительно сложных задач биоинженерии, биоинформатики, проектирования лекарственных препаратов, электродинамики и ряда других с использованием множества географически распределенных компьютеров с различными

архитектурами, режимами работы и административной принадлежностью.

ПУБЛИКАЦИИ

Основные результаты диссертации отражены в следующих работах:

1. Соболев С.И. Программные решения для организации метакомпьютерных вычислительных сред. Распределенные вычисления и грид-технологии в науке и образовании. Труды международной конференции - Дубна: ОИЯИ, 2004 - С. 190-193.
2. М.Ю. Медведик, Ю.Г. Смирнов, С.И. Соболев. Параллельный алгоритм расчета поверхностных токов в электромагнитной задаче дифракции на экране // Вычислительные методы и программирование. 2005. Том 6, №1. 86-95
3. С.И. Соболев. Решение прикладных задач в распределенной среде на основе технологий X-Com. Труды Всероссийской научной конференции «Научный сервис в сети Интернет. Технологии распределенных вычислений» (г. Новороссийск, 19-24 сентября 2005 г.), Изд-во МГУ, 2005, с. 9-10.
4. В.Б. Сулимов, А.Н. Романов, Ф.В. Григорьев, О.А. Кондакова, А.В. Сулимов, С.Н. Жабин, С.И. Соболев; Веб-ориентированная система молекулярного моделирования Keenbase для разработки новых лекарств, Труды Всероссийской научной конференции "Научный сервис в сети Интернет: Технологии параллельного программирования" (г. Новороссийск, 18-23 сентября 2006 г.), 2006, с. 170-172.

5. С.И. Соболев. Управление заданиями в Виртуальном метакомпьютерном центре на основе технологий X-Com. Распределенные вычисления и Грид-технологии в науке и образовании. Труды второй международной конференции (Дубна, 26 – 30 июня 2006 г.), Дубна: ОИЯИ, 2007, с. 401-404.
6. С.И. Соболев. Использование распределенных компьютерных ресурсов для решения вычислительно сложных задач // Системы управления и информационные технологии, № 1.3(27), 2007.
7. Свидетельство Роспатента об официальной регистрации программы для ЭВМ "Система метакомпьютинга X-Com" / Воеводин Вл. В, Соболев С.И., Филамофитский М.П.- № 2006611361 от 12.05.2006.