УДК 519.687.4

А. Г. Феоктистов, А. С. Корсуков

Институт динамики систем и теории управления СО РАН
ул. Лермонтова, 134, Иркутск, 664033, Россия
E-mail: agf@icc.ru

РАЗРАБОТКА GRID-СИСТЕМЫ С ДЕЦЕНТРАЛИЗОВАННЫМ УПРАВЛЕНИЕМ ПОТОКАМИ ЗАДАНИЙ

В докладе рассматриваются вопросы разработки Grid-системы с применением методов и средств спецификации распределенных информационно-вычислительных ресурсов и управления ими в процессе вычислений. Обсуждаются различные способы доступа пользователей к ресурсам Grid-системы. Предлагается способ децентрализованного управления потоками заданий в этой системе.

Ключевые слова: распределенные вычисления, Grid, спецификация ресурсов, управление потоками заданий.

Введение

В настоящее время для организации распределенных вычислений, требуемых при решении фундаментальных и прикладных ресурсоемких научных задач, активно применяется технология Grid [Foster et al., 2001. P. 201], базирующаяся на интеграции распределенных информационно-вычислительных и коммуникационных ресурсах и их совместном использовании в процессе вычислений. Инфраструктура Grid включает [Коваленко и др., 2004. С. 4] средства вычислений и обработки информации (суперкомпьютеры, вычислительные кластеры, отдельные персональные компьютеры, серверы баз данных и др.), объединенные телекоммуникационной средой, и системное программное обеспечение, предназначенное для управления процессом выполнения заданий пользователей в этой среде.

Зачастую, Grid создается на базе вычислительных кластеров. Кластер представляет собой совокупность компьютеров, объединенных локальной сетью и предназначенных для решения ресурсоемких (по процессорному времени, оперативной памяти и памяти на жестких дисках) вычислительных задач, и относится к классу многопроцессорных вычислительных систем. Компьютеры (рабочие станции), входящие в состав такого кластера, называются его вычислительными узлами. Один из вычислительных узлов назначается главным (управляющим) узлом кластера. Локальная сеть, используемая в вычислительном кластере, представляет собой коммуникационную среду для взаимодействия (передачи данных и управляющих инструкций) вычислительных узлов кластера между собой. Процесс решения задач на кластере осуществляется системой управления заданиями (СУПЗ). Диспетчер СУПЗ выполняет следующие функции: регистрацию и подключение вычислительных ресурсов и пользователей; организацию работы с файлами; поддержку многозадачного режима работы, управление очередями заданий, планирование загрузки вычислительных ресурсов; интеграцию с программными средствами поддержки параллельных вычислений.

Организованная на базе кластеров Grid в большинстве случаев ее использования обеспечивает возможность удаленного доступа к ресурсам (узлам) вычислительной сети и позволяет: определить вычислительные возможности конкретного узла (количество процессоров, объем оперативной памяти и т.п.) и степень его работоспособности; выполнить на этом узле некоторое независимое задание или обработать один из блоков данных при решении больших задач, позволяющих осуществить распределение по данным (см., например, [Воеводин и др., 2007. С. 35; Демичев и др., 2007. С. 10]).

Однако существуют виды фундаментальных и прикладных научных приложений, требующих более широких возможностей:

- получения вычислительных услуг нетиражируемых программных комплексов, размещенных в узлах Grid и обладающих специфическим программным интерфейсом;

- решения мульти-дисциплинарных задач, в процессе которого необходима интеграция ряда распределенных информационно-вычислительных ресурсов на основе автоматического планирования последовательности их использования;
- выполнения ряда взаимозависимых заданий, включенных в процесс решения одной общей задачи.

Кроме того, формы ведения вычислительных работ в Grid, обусловленные сложным системным программным обеспечением, используемым для организации Grid (например, Globus Toolkit, gLite, Virtual Data Toolkit и т. д.) во многом ориентированы на специалистов с достаточно высоким уровнем квалификации в системном программировании. Это обстоятельство сдерживает широкое распространение технологий Grid во многих приложениях, важных для специалистов в своих прикладных областях, и актуализирует вопросы создания «дружественных» средств доступа к ресурсам Grid.

Для реализации перечисленных выше возможностей необходимы специальные методы и средства для формулирования соответствующих постановок задач и спецификации заданий по их решению, а также для управления процессом выполнения сформированных заданий в Grid.

Архитектура Grid-системы

Представленная в данной работе Grid-система состоит из трех основных частей: серверной, клиентской и исполнительной. Ниже рассматривается функционирование этой Grid-системы, организованной на базе кластеров с различными операционными системами (ОС) и диспетчерами СУПЗ (рис. 1).

Серверная часть предназначена для организации соединения пользователей с ресурсами Grid-системы. Структура серверной части является распределенной и включает Grid-шлюз и web-сервер.

Grid-шлюз представляет собой рабочую станцию с установленными пакетами Globus Toolkit [1], GridWay [2] и пакетами СУПЗ, используемыми для управления кластерами Grid-системы. Grid-шлюз является точкой входа пользователей из другой Grid или с отдельной машины в сети Интернет, на которой также установлен пакет Globus Toolkit. Grid-шлюз обеспечивает следующие стандартные [3] для Grid функции: соединение с пользователем, его сертификацию и авторизацию, получение задания пользователя, выбор необходимых ресурсов для его выполнения, пересылку задания на удаленный ресурс и получение результатов вычислений. Планирование ресурсов для выполнения поступившего задания осуществляется с помощью пакета GridWay. Остальные функции реализуются службами Globus Toolkit.

На web-сервере Grid-системы размещается Web-Interface Manager (WIM), представляющий собой скрипт языка PHP и выполняющийся на машине с установленным пакетом Apache. WIM обеспечивает пользователю возможность отправки его задания в Grid-систему и передачи этого задания на тот или иной кластер, в зависимости от системных требований, необходимых для выполнения задания. Спецификация и запуск заданий производится с помощью специального командного языка, позволяющего описать инструкции обработки задания для диспетчера СУПЗ выбранного кластера.

Клиентская часть обеспечивает средства доступа пользователей к Grid-системе. Сформировать и запустить задание в Grid-системе можно одним из двух способов, приведенных ниже.

Первый способ обеспечивает запуск заданий из другой Grid или с отдельной машины в сети Интернет, на которой также установлен пакет Globus Toolkit. Чтобы отправить задание на выполнение пользователь должен составить паспорт задания в формате языка RSL (Resource Specification Language). Отправка задания в Grid-систему происходит с помощью утилиты globusrun-ws. На машине, с которой будет производиться запуск задания, должен находиться действующий сертификат пользователя, от имени которого запускается задание.

[1] http://globus.org/toolkit/
[2] http:// www.gridway.org/
[3] С точки зрения промежуточного программного обеспечения (middleware).

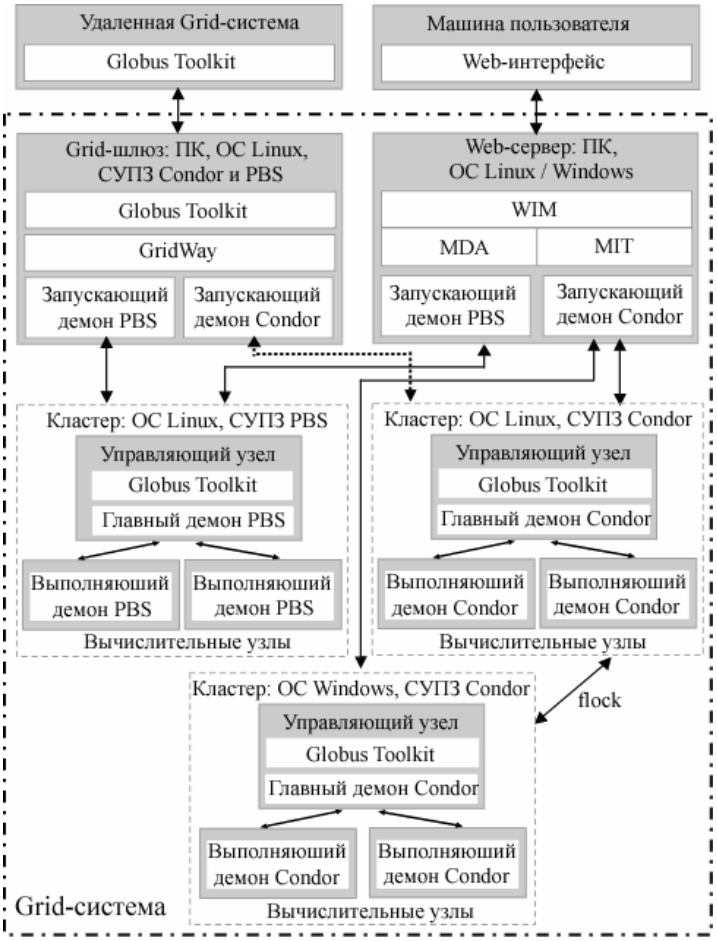

Рис. 1. Схема функционирования Grid-системы

Данный сертификат создается пользователем на запускающей машине и отправляется владельцам Grid-системы на утверждение разрешения использования их ресурсов. После прохождения проверки сертификата, составления паспорта задания и отправки задания на выполнение в Grid-систему пользователь может просматривать статус и ход выполнения задания. Пользователь имеет возможность отменить выполнение задания в процессе его выполнения, запустив утилиту globusrun-ws с соответствующими ключами. После выполнения задания на одном из кластеров результаты счета будут помещены в те папки на машине пользователя, которые были указаны им в паспорте заданий.

Второй способ базируется на запуске заданий с помощью web-интерфейса. В этом случае пользователь должен пройти процедуру регистрации на web-сервере. Зарегистрированному пользователю необходимо ввести логин и пароль для входа в систему и дальнейшего управления процессом запуска заданий. После прохождения процедуры авторизации пользователю предлагается специальная web-форма, предназначенная для создания паспорта задания. Для запуска задания пользователю необходимо указать в соответствующих полях web-формы необходимые атрибуты паспорта этого задания (исполняемую программу, файлы с исходными данными, файлы с результатами счета и т. д.). WIM автоматически транслирует данные из web-формы в паспорт задания СУПЗ выбранного кластера [4] и передает этот паспорт планировщику СУПЗ. С помощью web-интерфейса WIM позволяет пользователю приостановить выполнение задания, удалить задание и продолжить вычисление приостановленного задания. Все эти функции обеспечивают пользователю полный контроль над заданием. После выпол-

[4] Такая процедура возможна благодаря разработанному автором командному языку описания и управления заданиями.

нения задания пользователь может скачать полученные результаты и сохранить их на свой жесткий диск.

Исполнительная часть обеспечивает выполнение заданий в Grid-системе на вычислительных узлах, которые представляют собой кластеры под управлением операционных систем Windows или Linux и диспетчеров СУПЗ PBS [5] или Condor [6]. В качестве СУПЗ кластера в Grid-системе можно использовать и другие СУПЗ, совместимые с пакетом Globus Toolkit (например, Cleo [7]).

При выборе механизма обслуживания очереди заданий на выполнение в Grid-системе следует учитывать следующие факторы: задания выполняются в пакетном режиме, время выполнения задания заранее неизвестно, задания могут различаться по степени важности их выполнения, задания могут относиться к разным видам с точки зрения планирования процесса их выполнения.

В Grid-системе дисциплина обслуживания очереди заданий базируется на комбинированном применении принципов FIFO (первый пришедший обслуживается первым), переключения выполнения и учета приоритетов заданий. Так как создание Grid-системы в первую очередь обусловлено использованием ее ресурсов для решения большой задачи в рамках некоторой глобальной Grid-сети, то для заданий поступающих из такой сети назначаются приоритеты более высокие, чем для заданий поступающих через web-сервер. Приоритеты для заданий разных видов устанавливаются администратором Grid-системы.

Кластер под управлением Linux и СУПЗ Condor включает главный (управляющий) узел и вычислительные узлы, на которых происходит выполнение пользовательских задач. Данный кластер обеспечивает выполнения заданий пользователя, реализованных для выполнения под управлением Linux. Диспетчер (главный демон) СУПЗ Condor отвечает за управление вычислительными узлами, входящими в состав этого кластера. После получения задания, которое нужно выполнить в Linux, он ищет свободные вычислительные ресурсы и направляет на них задание. Если нет свободных узлов, то задание помещается в очередь и ожидает выполнения. Результаты счета на этом кластере становятся доступными, как только задание выполнено. Далее они отправляются диспетчеру СУПЗ Condor. После этого, в зависимости от способа запуска задания, соответствующая служба Globus Toolkit или WIM отправляют результаты счета пользователю.

Кластер под управлением Windows и СУПЗ Condor служит для выполнения пользовательских заданий под управлением Windows. Данный кластер работает идентично вышеописанному кластеру, за исключением того, что задания могут поступать только от WIM.

В Grid-системе кластеры с диспетчером СУПЗ Condor могут пересылать с помощью flock-процедуры [8] друг другу задания, выполнение которых не требует определенной операционной системы (например, задания на языке java). Применение flock-процедуры может потребоваться в том случае, когда у пересылающего кластера не хватает свободных вычислительных узлов.

Кластер под управлением Linux и СУПЗ PBS ориентирован на выполнение заданий пользователей разработанных под Linux. Обработка заданий на данном кластере происходит таким же образом, что и в кластерах описанных выше, за исключением того, что взаимодействие со службами Globus Toolkit и WIM осуществляет диспетчер СУПЗ PBS.

Планирование

Поступающие в Grid-систему задания разделены условно на два потока (рис. 2). Первый поток формируется из стандартных (допускаемых пакетом Globus Toolkit) заданий, поступивших из других Grid. Во второй поток поступают задания, отправленные через web-интерфейс.

[5] http://www.clusterresources.com/pages/products/torque-resource-manager.php
[6] http://www.cs.wisc.edu/condor/
[7] http://sourceforge.net/projects/cleo-bs/
[8] Condor. Version 6.9.2. Manual (http://www.cs.wisc.edu/condor/manual/v6.9/).

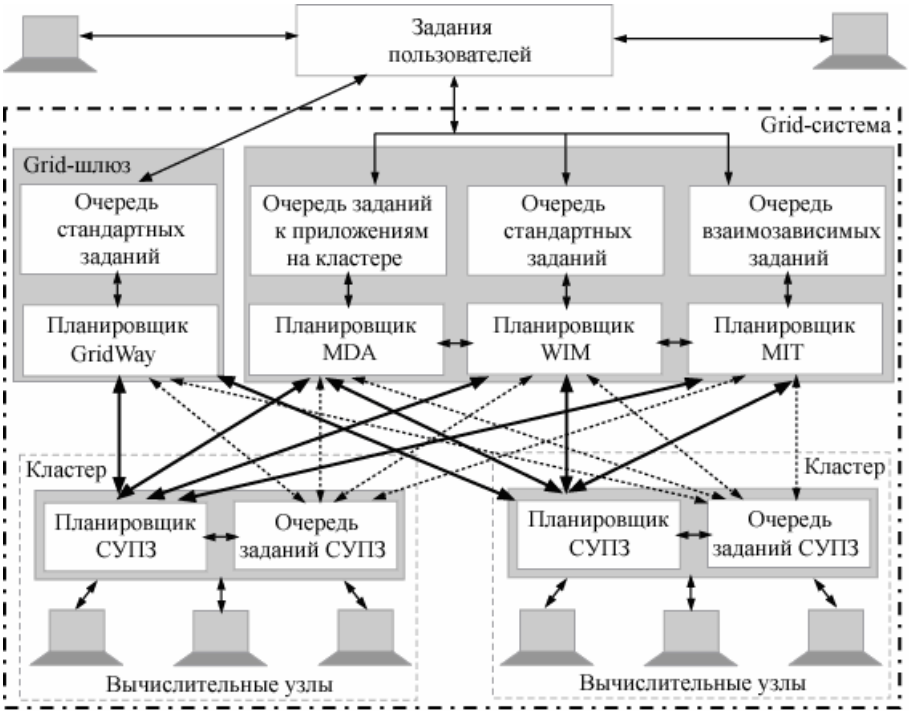

Рис. 2. Управление потоками заданий

Распределение заданий из первого потока. Поступившее на Grid-шлюз стандартное задание ставится в очередь и ждет назначения на один из свободных вычислительных ресурсов. Перед тем как запустить задание пакету Globus Toolkit необходимо знать состояние вычислительных узлов кластеров Grid-системы. Поиск свободных ресурсов производит пакет интеллектуального распределения заданий по кластерам GridWay. Пакет GridWay периодически опрашивает состояние кластеров и, как только их вычислительные узлы освобождаются, передает эту информацию пакету Globus Toolkit. Последний направляет находящееся на данный момент в голове очереди задание на выполнение главному демону диспетчера СУПЗ освободившегося кластера. Диспетчер СУПЗ формирует локальную очередь заданий на кластере и передает задание на выполнение вычислительным узлам кластера.

Распределение заданий из второго потока. С помощью WIM, помимо стандартных заданий, можно запускать так же следующие виды заданий:

- задания, использующие программное обеспечение, размещенное в узлах Grid-системы;
- взаимосвязанные задания, требующие частично упорядоченного выполнения входящих в них подзадач.

Для первого вида создан менеджер распределенных приложений – MDA (Manager of Distributed Applications), который распределяет задания на нужные вычислительные узлы кластеров, где установлены необходимые приложения. Планировщик MDA взаимодействует с планировщиком WIM с целью закрепления за заданием выбранных ресурсов перед передачей этого задания диспетчеру СУПЗ кластера.

Менеджер взаимосвязанных заданий MIT (Manager of Interrelated Tasks) занимается распределением заданий второго вида по кластерам и отслеживает ход их выполнения. Подход к планированию работ MIT основан на двух важных решениях. Во-первых, для Grid-системы построена модель, включающая спецификации аппаратных компонентов (компьютеров и коммуникационных интерфейсов), спецификации системного и прикладного программного обеспечения, а так же множество зависимостей между программно-аппаратными ресурсами этой системы. Во-вторых, для управления кластерами использованы диспетчеры СУПЗ, совместимые с инструментарием пакета Globus Toolkit.

На верхнем уровне абстракции модель ресурсов Grid-системы может быть представлена в виде структуры R=<HW, CND, SW, P, TP, PT, NL, NC, NH, NS, JOB, JS, JP, JSM, JM>,

где HW (HardWare) – множество значимых характеристик аппаратных средств узлов Grid-системы, элементы которого представляют собой тройки <PP, RAM, DM>, где PP (Peak Performance) – пиковая производительность узла, RAM – объем оперативной памяти узла и DM (Disk Memory) – объем дисковой памяти узла;

CND (Channels and Network Devices) – технические средства коммуникаций Grid-системы;

SW (SoftWare) – множество программных средств, размещенных в узлах Grid-системы; подмножества SA, UA $\subseteq$ SW (System Applications and Users Applications) представляют соответственно совокупности системных и пользовательских приложений;

P (Parameters) – множество параметров (концептов) приложений из SA и UA; подмножества SYS, IN, OUT $\subseteq$ P определяют соответственно параметры системных приложений из SA, входные и выходные параметры пользовательских приложений из UA;

TP (Types of Parameters) – множество допустимых типов параметров в Grid-системе (простые и структурированные типы параметров, в том числе, файлы);

PT $\subseteq$ P $\times$ TP (Parameters and types) – отношение, определяющее множество типизированных параметров в Grid-системе;

NL (Nodes Locations) - множество адресов узлов Grid-системы;

NC $\subseteq$ NL $\times$ CND (Nodes and Channels) – отношение, определяющее топологию сети передачи данных Grid-системы;

NH $\subseteq$ NL $\times$ HW (Nodes and Hardware) – отношение, определяющее множество аппаратных ресурсов в узлах Grid-системы;

NS $\subseteq$ NL $\times$ SW (Nodes and Software) – отношение, определяющее множество программных ресурсов в узлах Grid-системы;

JOB – множество заданий пользователей Grid-системы;

JS $\subseteq$ JOB $\times$ SW (Jobs and Software) – отношение, определяющее множество приложений в заданиях;

JP $\subseteq$ JOB $\times$ P (Jobs and Parameters) – отношение, определяющее множество параметров приложений в заданиях;

JSM (Job Start Modes) – множество допустимых режимов запуска задания в Grid-системе;

JM $\subseteq$ JOB $\times$ JSM (Jobs and Modes) – отношение, определяющее возможность запуска задания в допустимых режимах Grid-системы.

Данная модель лежит в основе базы знаний, используемой планировщиком MIT для планирования последовательности выполнения взаимозависимых заданий. Планировщик MIT, используя информационно-логические связи между объектами модели ресурсов Grid-системы, выполняет частичное упорядочение подзадач общего задания, находит необходимые для выполнения задания вычислительные ресурсы, взаимодействует с планировщиком WIM с целью закрепления за заданием выбранных ресурсов, передает задание диспетчеру СУПЗ и осуществляет дальнейший контроль выполнения этого задания.

Таким образом, децентрализованный способ управления потоками заданий заключается, во-первых, в распределении заданий пользователей по различным направлениям их обработки с различными приоритетами обслуживания в очереди и, во-вторых, в применении ряда планировщиков, предназначенных для работы со специальными видами заданий.

Командный язык заданий

Данный язык представляет совокупность команд, предназначенных для управления пользовательскими заданиями на узлах кластера. С помощью командного языка WIM может быть легко и гибко настроен для взаимодействие с используемым диспетчером СУПЗ за счет применения унифицированных шаблонов описания для команд СУПЗ. Этот подход основан на схожести форматов вызовов основных команд различных СУПЗ и их опций (параметров). Основными функциями командного языка являются: перехват и трансляция сообщений диспетчеров СУПЗ, вывод различного рода статистики (информации о свободных и занятых вычислительных узлах кластера, количестве выполняющихся, ожидающих выполнения и выполненных заданий и т. д.). Командный язык также позволяет управлять учетными записями пользователей.

Командный язык включает три типа команд.

1. Команды управления заданиями. Основное назначение рассматриваемой группы команд – это обеспечение запуска пользовательских приложений на узлах кластера, приостановка выполнения заданий, продолжение выполнения временно приостановленных заданий и отмена выполнения задания на кластере (удаление пользовательского задания).

2. Команды просмотра статистики процесса выполнения заданий. Данная группа команд служит для контроля пользователями процесса выполнения заданий на кластере. Эти команды позволяют просматривать различного рода информацию о ходе выполнения задания, статистику по заданиям определенного пользователя и по всем пользовательским заданиям, выполнявшимся на вычислительных узлах кластера за выбранный промежуток времени.

3. Команды управления учетными записями пользователей. Данная группа команд ориентирована в большей степени на их использование администратором кластера или Grid-системы, так как основным назначением этих команд является редактирование, удаление, блокировка или разблокирование уже созданных и регистрация новых учетных записей пользователей.

Пользователю предлагается набор web-форм с элементами управления, каждый из которых обрабатывается PHP скриптом в соответствии с используемым шаблоном описания команды СУПЗ для данного элемента управления. Таким образом, настройка PHP скрипта на ту или иную СУПЗ заключается в замене в шаблоне имени команды и используемых ей опций (параметров). Шаблон команды включает следующие поля:

<имя команды> <опция_1> <опция_2> … <опция_n>.

Ниже приведены примеры описания команд для СПУЗ Condor и PBS.

Команда запуска задания:
СУПЗ Condor: condor_submit <паспорт задания>;
СУПЗ PBS: qsub < паспорт задания>.
Команда приостановки выполнения задания:
СУПЗ Condor: condor_hold <номер задания>;
СУПЗ PBS: qhold <номер задания>.
Команда удаления задания:
СУПЗ Condor: condor_rm <номер задания>;
СУПЗ PBS: qdel <номер задания>.

Практическое использование Grid-системы

В Grid-системе был решен ряд практических задач по исследованию систем булевых уравнений [Опарин и др., 2006. С. 884], поиску оптимального управления [Сидоров и др., 2006. С. 519], поиску генераторов двоичных последовательностей [Заикин и др., 2007. С. 83], складской логистики [Башарина и др., 2006. С. 30] и др. Так при решении задачи моделирования схем погрузочно-разгрузочных работ складского комплекса были использованы: кластер на базе учебного класса вуза (16 узлов, AMD Athlon 1,7 GHz); кластеры ИДСТУ СО РАН (8 узлов, Pentium 4 2,8 GHz с технологией гипертрейдинг; 2 узла, Pentium Dual core 2,8 GHz). Причем, эффективность расчетов в такой распределенной вычислительной среде для этой задачи составила 91-94% по сравнению со временем решения этой задачи на локальных ПК, производительность которых была сопоставима производительности различных узлов кластеров.

Заключение

Отметим следующие основные результаты, представленные в данной статье:

- выполнена реализация Grid-системы на базе пакета Globus Toolkit, включающая кластеры, функционирующие под управлением разных операционных систем и использующие различные диспетчеры СУПЗ;

- разработаны менеджеры потоков заданий WIM, MDA и MIT, обеспечивающие альтернативный пакету Globus Toolkit способ доступа пользователей к Grid-системе через web-интерфейс;
- разработан унифицированный командный язык, обеспечивающий быструю и гибкую настройку на взаимодействие с различными СУПЗ кластеров;
- предложена новая схема децентрализованного управления потоками заданий в Grid-системе.

Список литературы

Башарина О. Ю., Ларина А. В., Суханова Н. Г., Феоктистов А. Г. Моделирование торгово-складского комплекса в распределенной вычислительной среде // Моделирование. Теория, методы и средства: Материалы VI Междунар. науч.-практ. конф. Новочеркасск: ЮРГТУ, 2006. Ч. 3. С. 28–32.

Воеводин В. В. Решение больших задач в распределенных вычислительных средах // Автоматика и телемеханика. 2007. № 5. С. 32–45.

Демичев А. П., Ильин В. А., Крюков А. П. Введение в грид-технологии. Препринт. М.: НИИЯФ МГУ, 2007. 87 с.

Заикин О. С., Семенов А. А., Сидоров И. А., Феоктистов А. Г. Параллельная технология решения SAT-задач с применением пакета прикладных программ D-SAT // Вестник ТГУ. Приложение. 2007. № 23. С. 83–95.

Коваленко В. Н., Корягин Д. А. Организация ресурсов ГРИД. Препринт. М.: ИПМ им. Келдыша РАН, 2004. 25 с.

Опарин Г. А., Богданова В. Г., Феоктистов А. Г. Технология булева моделирования и решения дискретных задач в распределенной вычислительной среде // Параллельные вычисления и задачи управления: Труды III Междунар. конф. PACO'2006. М.: ИПУ РАН, 2006. С. 883–891.

Сидоров И. А., Тятюшкин А. И., Феоктистов А. Г. Распределенная информационно-вычислительная среда модульного программирования // Параллельные вычисления и задачи управления: Труды III Междунар. конф. PACO'2006. М.: ИПУ РАН, 2006. С. 505–521.

Foster I., Kesselman C., Tuecke S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations // International Journal of High Performance Computing Applications. 2001. Vol. 15. No. 3. P. 200–222. http://www.globus.org/alliance/publications/papers/anatomy.pdf

Материал поступил в редколлегию 20.05.2008

A. G. Feoktistov, A. S. Korsukov

The Development of The Grid-System with Decentralized Job Flow Control

The issue of the development of the Grid-system is considered. This process is carried out with use of methods and tools of specification of the distributed information and computing resources and their control in the process of computing. The various methods of access of users to resources of Grid-system are discussed. The new decentralized method of the job flow control is represented.

Keywords: distributed computing, Grid, specification of computing resources, job flow control.