УДК 004.4, 004.7

Отказоустойчивое децентрализованное управление ресурсами грид

В.В. Корнеев[1], Д.В. Семенов[1], П.Н. Телегин[2], Б.М. Шабанов[2]

[1]*ФГУП «НИИ «Квант» (г. Москва)*
[2]*Межведомственный суперкомпьютерный центр РАН* (*г. Москва*)

Resilient Decentralized GRID Resources Control

V.V. Korneev[1], D.V. Semenov[1], P.N. Telegin[2], B.M. Shabanov[2]

[1]*Federal state unitary enterprise «Research and development institute «Kvant», Moscow*
[2]*Joint Supercomputer Center of the Russian Academy of Sciences, Moscow*

Представлена децентрализованная система управления грид, обеспечивающая автоматическое восстановление выполнения заданий при отказах аппаратно-программных средств. Установлены отказы, которые не допускают автоматического восстановления выполнения заданий и требуют повторного их запуска пользователем с прохождением процедуры аутентификации пользователя.

Ключевые слова: грид; децентрализованное диспетчирование заданий; отказоустойчивость; защита от несанкционированного доступа.

A decentralized GRID resources control system, providing the automatic jobs recovery when the software or hardware fails, has been presented. The faults, which are not compatible with the automatic jobs recovery and require the manual rerun, have been determined.

Keywords: grid, dispatching, resilience, unauthorized access

Подход к созданию распределенной системы управления (СУ) ресурсами и заданиями грид представлен в работе [1]. СУ состоит из распределенных по грид и взаимодействующих друг с другом управляющих процессов – менеджеров СУ. Информационные связи между менеджерами СУ образуют ациклический граф (рис.1). На управляющей машине (УМ) каждой ВС обязательно запущен менеджер М1. Менеджеры М2 могут выполняться на управляющих машинах ВС или на специально выделенных компьютерах каждой ВС. Количество и распределение менеджеров М2 по физическим ресурсам определяются выполнением требований по пропускной способности системы управления грид [2] и надежности.

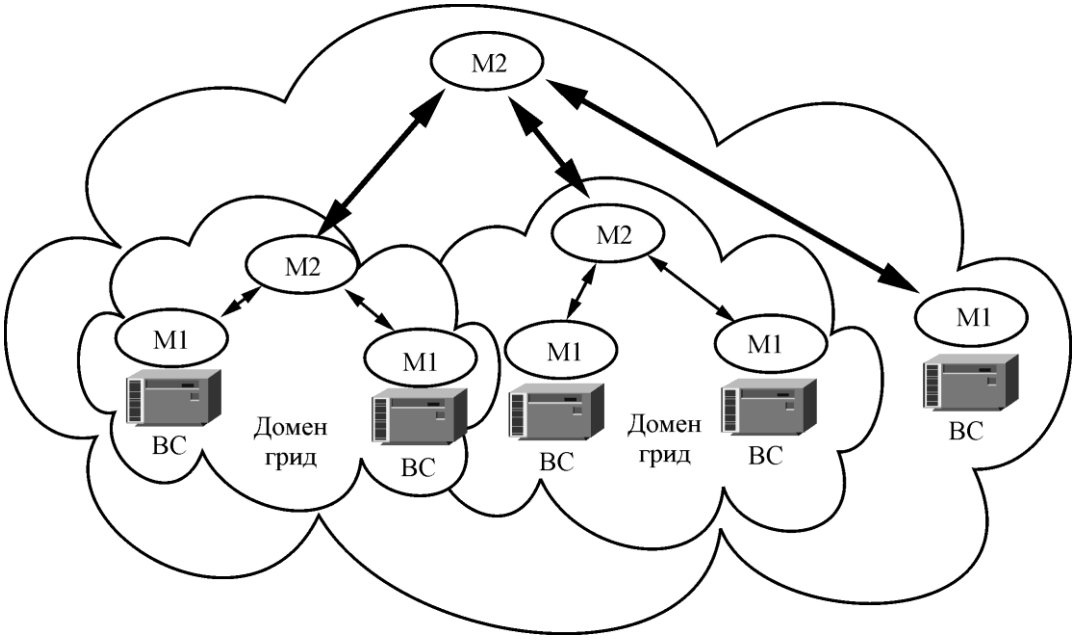

Рис.1. Структура СУ грид-среды

В [1] представлены алгоритмы децентрализованного диспетчирования заданий менеджерами СУ, применяемые в грид МСЦ РАН [3]. В настоящей работе рассматриваются подходы к реализации СУ [4], обеспечивающие отказоустойчивость и защиту от несанкционированного доступа самой СУ грид на базе Globus Toolkit (GT).

Архитектура СУ грид. С точки зрения пользователя, неадекватное функционирование управления компьютерной сети, вызванное несанкционированным доступом к разделяемым ресурсам пользователей и запущенных ими процессов, рассматривается как отказ. При построении сложных систем на базе традиционных сетевых технологий борьба с отказами, возникающими из-за несанкционированного доступа к ресурсам, выносится на прикладной уровень, на котором возможность найти процессы, злонамеренно использовавшие последствия нарушения политики безопасности, весьма ограничена. Трудность построения эффективно функционирующих сложных систем на базе традиционных сетевых технологий привела к возникновению грид.

Свойства грид [5] нашли практическую реализацию в пакете GT. В грид на базе GT протокол грид, выделяющий грид среди компьютерных сетей, обеспечивает защиту от несанкционированного доступа к ресурсам путем взаимной аутентификации сторон при доступе к ресурсу. Любой процесс, требующий, как клиент, доступа к ресурсу, получает доступ посредством обращения к модулю управления ресурсами GRAM (Globus Resource Allocation Manager), контролирующему этот ресурс (рис.2).

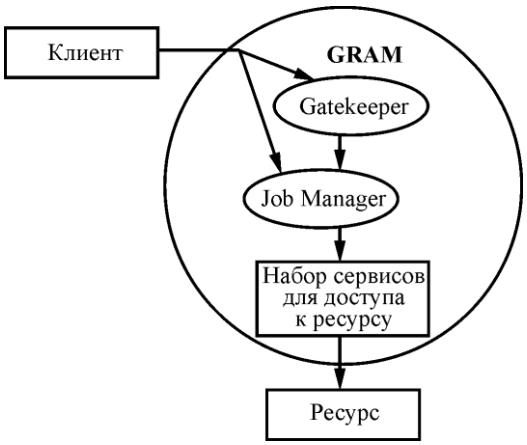

Рис.2. Доступ «клиента» к «ресурсу» посредством GRAM

При этом GRAM запускает с правами суперпользователя процесс Gatekeeper, который выполняет взаимную аутентификацию клиента и вычислительного ресурса с использованием сертификатов стандарта

X-509 и, возможно, протокола SSL. Это обеспечивает надлежащее разграничение доступа к ресурсу и исключение несанкционированного доступа. Собственно в этом заключается существенная особенность GT: каждый ресурс имеет свой сертификат и список сертификатов клиентов, которым разрешен доступ к рассматриваемому ресурсу. Сертификат считается достоверным (а аутентификация успешной), если принимающей стороне известен подписавший его центр сертификации и сертификат подписан правильно с учетом срока его действия.

После успешной аутентификации клиента процесс Gatekeeper ограничивает свои права правами этого клиента (непривилегированного пользователя) и запускает экземпляр процесса Job Manager Instance (JMI), осуществляющий дальнейшее управление доступом к ресурсу. Устанавливаемая по умолчанию конфигурация GRAM порождает для доступа к ресурсу процесс Unix. Предусмотрена возможность расширения функциональности процесса JobManager для реализации требуемого механизма доступа к ресурсу в виде набора скриптов.

Возможность создать собственный набор скриптов широко используется разработчиками грид, например, для контроля выполнения политик доступа к ресурсу, а также биллинга ресурсов и применения экономических моделей предоставления ресурсов на базе соглашений между поставщиками и потребителями ресурсов.

При построении грид на базе GT локальные системы пакетной обработки (СПО) заданий каждой ВС грид-среды выступают в качестве ресурса, доступ к которому контролируется экземплярами GRAM. В пакет GT включен набор скриптов для ряда распространенных СПО: LSF, PBS, PRUN, CONDOR и др. При запуске задания служба GRAM определяет используемую СПО и в дальнейшем применяет соответствующий скрипт. В ходе создания СУ грид добавлен скрипт, выполняющий запуск заданий под управлением системы очередей СУПЗ МВС [6], широко используемой на отечественных суперкомпьютерах семейства МВС. Данная СПО также обеспечивает возможность синхронного запуска ветвей одной параллельной программы на вычислительных модулях разных ВС грид.

В случае успешной аутентификации СПО и процесса, ставящего от имени пользователя задание в очередь СПО, происходит попытка авторизации пользователя. Основная схема авторизации в GT заключается в том, что пользователь авторизуется в ВС, если в файле авторизации (grid-mapfile), извлекаемом из сертификата, идентификатору пользователя грид поставлено в соответствие имя локального пользователя данной ВС. Если идентификатор пользователя грид в файле авторизации отсутствует, то пользователь признается неавторизованным и запрос на постановку задания в очередь СПО отвергается.

Брокеры заданий. Требования к вычислительным ресурсам (количество и тип процессоров, прогнозируемое время выполнения), размещение файлов и другие сведения о своем параллельном задании пользователь оформляет в виде паспорта задания [4, 6]. Ввод паспорта задания в грид (рис.3) осуществляется удаленно посредством протокола SSH через УМ любой из входящих в ее состав ВС. При этом на этой УМ запускается специальный сервис – экземпляр брокера задания (БЗ). Экземпляр БЗ, или брокер, создается для любого задания и уничтожается при полном его завершении. БЗ производит предобработку полученных от пользователя содержащихся в паспорте задания данных о задании, а именно добавляет те характеристики, которые могли быть не указаны пользователем, но должны учитываться при выделении ресурсов, например идентификатор пользователя в грид, архитектура ВС и т.д. На основе переработанных данных БЗ формирует паспорт задания для передачи менеджеру М1 СУ, содержащий параметры задания, необходимые для принятия решения о выделении менеджерами СУ ресурсов для задания. Экземпляры брокера запускаются и функционируют от имени и с правами пользователя грид, запустившего задание.

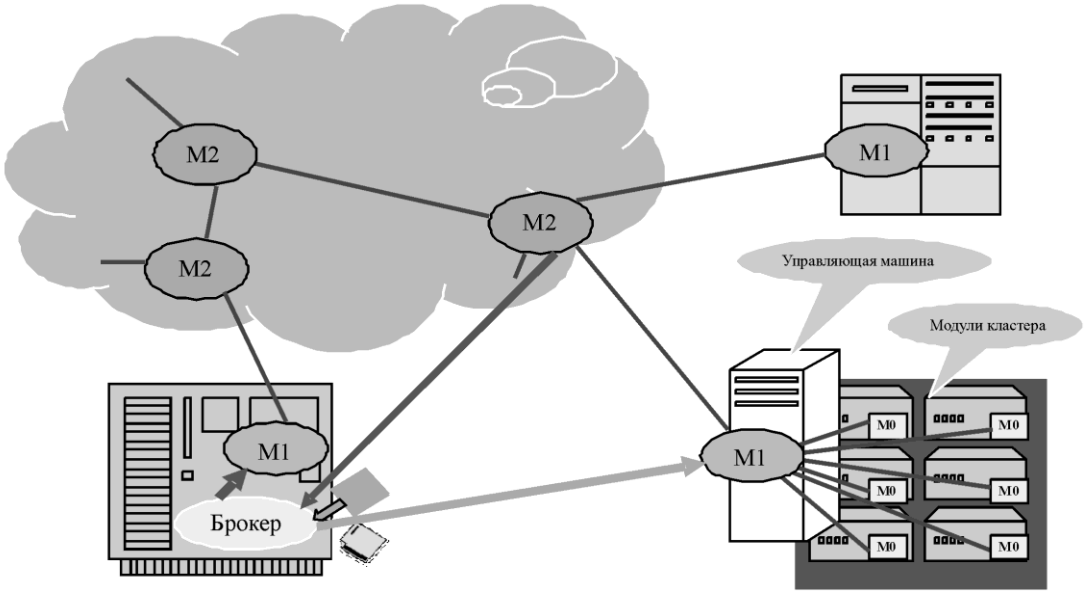

Рис.3. Взаимодействие брокера задания и менеджеров СУ
(M1 – менеджеры ВС, M2 – менеджеры доменов ССРВ)

Основное назначение БЗ:

- передача системе менеджеров паспорта задания;
- получение в ответ от системы менеджеров сетевого IP-адреса ВС, предлагаемой для запуска задания, или сетевых IP-адресов ВС, если предлагается запуск задания на нескольких ВС;
- запуск пользовательского задания на выполнение на ресурсах ВС с этим IP-адресом или этими IP-адресами посредством передачи паспорта задания в очередь их локальной СПО заданий;
- предоставление пользователю интерфейса для управления заданиями.

Пользователь посредством не зависящего от архитектурно-технической организации ВС интерфейса управления заданием может добавить задание в грид, снять с выполнения, извлечь из очереди, отслеживать состояние задания на всех стадиях его продвижения по системе и осуществлять другие управляющие воздействия.

Менеджеры. После получения менеджером паспорта задания (от брокера или другого менеджера) менеджеры СУ согласованно вырабатывают решение о предоставлении заданию требуемых ресурсов [1]. В результате один из менеджеров, обычно это менеджер типа M1, передает БЗ, запросившему ресурсы для выполнения задания, решение о выделенных ресурсах. Вариантов ответа менеджера на запрос БЗ может быть три:

- запрошенного количества ресурсов в грид-среде нет;
- необходимого количества ресурсов в текущий момент времени в системе нет, паспорт помещен в очередь одного из менеджеров системы управления грид-среды и ответ с указанием сетевых адресов предоставляемых ресурсов придет позднее;
- возможен запуск задания на одной (или на совокупности) ВС из состава грид-среды с указанием сетевых IP-адресов предоставляемых ресурсов.

Если запуск задания возможен, то брокер, используя средства модуля управления ресурсами пакета GT, осуществляет постановку задания в очередь локальной СПО ВС, IP-адрес которой указан менеджером СУ. При передаче задания в локальную очередь СПО брокер устанавливает взаимодействие с запускаемым для управления заданием

процессом jobmanager пакета GT. Данный процесс отслеживает состояние задания в локальной очереди СПО и служит шлюзом брокеру, позволяющим удаленно получать информацию о состоянии задания, а также удалить задание из очереди или снять с выполнения. В ходе постановки задания в очередь СПО в планируемую для исполнения задания ВС доставляются файлы с необходимыми заданию данными. Для этого используются утилиты взаимодействия с файловым сервером из пакета GT, позволяющие безопасно переслать файлы в удаленную ВС и обратно в сервер.

Отказоустойчивость взаимодействия брокеров и менеджеров. При приеме паспорта задания менеджер, принявший паспорт в свою очередь, извещает брокера об этом и своем сетевом адресе. От момента отправки паспорта задания менеджеру до момента постановки задания в очередь локальной СПО брокер периодически обращается к менеджеру СУ (heartbeat), в очереди которого находится паспорт его задания, для подтверждения того, что задание еще актуально, и для получения управляющих воздействий.

В менеджерах СУ реализована функция определения живости брокеров заданий, паспорта которых находятся у них в очереди. В случае если брокер в течение определенного времени не подтверждает наличие задания в системе, то паспорт этого задания удаляется менеджером СУ из своей очереди. В случае если с менеджером, с которым взаимодействовал БЗ, пропадает связь, то после нескольких неудачных попыток эту связь восстановить, брокер инициирует очередной раунд планирования этого задания, передавая его краткий паспорт в очередь локального менеджера M1. После нескольких неудачных попыток установить связь с менеджером M1 брокер задания завершает свою работу с извещением пользователя о причине отказа. Также брокер завершит свою работу в случае, если возникнет ошибка в процедуре взаимной аутентификации между менеджером СУ и брокером.

Аналогично обрабатываются следующие сбои в ходе взаимодействия между брокером и процессом jobmanager подсистемы управления ресурсами пакета GT:

- сбои из-за ошибок в процедуре взаимной аутентификации между jobmanager и брокером;
- нарушение связи с удаленной ВС и, соответственно, со службой jobmanager;
- сбои службы jobmanager;
- неправильное функционирование СПО.

Восстановление взаимодействия со службой jobmanager возложено на брокер задания. БЗ определенное количество раз пытается повторить процедуру аутентификации или установить соединение со службой jobmanager и в случае неудачи принимает решение о необходимости выполнить новую процедуру планирования этого задания, передавая его паспорт в очередь локального менеджера M1.

Возможна ситуация, когда после принятия решения о выделении ресурсов и извещении брокера задания о предоставлении ресурсов состояние этих ресурсов изменилось и БЗ при попытке запуска задания получит отказ. Такая ситуация возможна из-за отказа оборудования или занятия ресурсов локальным пользователем, ставящим свои задания в СПО, минуя СУ грид. В этом случае брокер должен повторить запрос менеджеру M1 и начать новый раунд планирования задания.

Перезапуск задания пользователем. Отдельно необходимо рассмотреть ситуацию, когда происходит отказ или перезагрузка управляющей машины ВС, на которой функционирует брокер задания (ВС, с которой задание было запущено). При разработке децентрализованной СУ грид с возможностью динамического перераспределения заданий проанализированы следующие три варианта реализации:

- задание запускается брокером на ВС выбранной СУ грид, средствами пакета GT;
- задание передается менеджеру выбранной ВС и запускается им средствами пакета GT;
- задание передается менеджеру выбранной ВС и запускается средствами локальной системы планирования заданий данной ВС.

В последних двух вариантах менеджеры ВС должны активно взаимодействовать по сети. Для взаимной аутентификации менеджеры должны иметь соответствующие сертификаты. Для запуска заданий от имени другого пользователя менеджеры должны запускаться с правами суперпользователя. С учетом первых трех положений обеспечения безопасности второй и третий варианты неприемлемы.

Первый вариант запуска заданий предполагает в случае перезагрузки управляющей машины ВС, с которой осуществлялся запуск задания, необходимость ручного перезапуска брокеров всех заданий, введенных в грид с данной ВС, и установления связей с этими заданиями. Для этого предлагается регистрировать в ВС все запускающиеся брокеры и использовать для перезапуска зарегистрированных брокеров данного пользователя специальный системный процесс – менеджер брокеров (МБ), который запускается на управляющей машине ВС и функционирует с правами суперпользователя.

Однако с учетом первых трех положений обеспечения безопасности, операцию восстановления брокера нельзя выполнить без участия пользователя, который должен вручную зарегистрироваться в системе. Таким образом, восстановление брокеров автоматически, сразу после перезагрузки управляющей машины, невозможно без ослабления защиты от несанкционированного доступа. Перезапуск брокеров осуществляется после того, как пользователь вновь зарегистрируется в системе и известит об этом менеджер брокеров.

Таким образом, выполнение задания может быть прервано только при крахе управляющей машины ВС, с которой запускалось задание, или отказе управляющей машины, на которой запущена СПО, принявшая задание. При восстановлении работоспособности УМ и поддержке СПО восстановления очередей заданий после сбоя выполнение задания будет продолжено средствами СПО. Считается, что перезапущенный отказавший менеджер будет при инициации иметь пустую очередь заданий.

Отказоустойчивость системы менеджеров. В СУ грид должно быть заложено решение задач масштабирования, повышения надежности и восстановления после сбоев. Эти задачи имеют эффективное централизованное решение на кластерном уровне, однако при выходе на уровень грид-среды требуются существенно другие решения, учитывающие распределенность ресурсов.

Для эффективного функционирования СУ необходимо удовлетворить некоторые требования, налагаемые на выбранную модель управления вычислительными ресурсами:

- неожиданные изменения в связности иерархии менеджеров СУ рассматриваются как обычные явления – ставшие несвязными участки иерархии должны продолжать работать автономно;
- при обнаружении менеджером отсутствия соединения со смежным менеджером он устанавливает соединение с другим менеджером, выбирая его из числа не связанных с ним на момент обнаружения отсутствия соединения и приводящим к построению графа с минимальным диаметром. Для предотвращения возможности при одновременном конфигурировании связей несколькими менеджерами образования циклов и эффекта «пинг-понг» предусмотрен механизм построения остовного дерева, аналогичный протоколу IEEE 802.1d STP (Spanning Tree Protocol), применяемому в коммутаторах.

Использование ациклических графов в системе управления грид-средой позволяет удовлетворить перечисленные требования. Каждый менеджер СУ управляет своей очередью паспортов заданий, имеет обобщенное представление о состоянии грид-среды в целом [1]. Этой информации достаточно для принятия менеджером управленческого решения о выделении заданиям контролируемых вычислительных ресурсов в случае нарушения связности или сбоев в работе менеджеров удаленных участков иерархии СУ.

Восстановление связности происходит автоматически в процессе периодического обмена менеджерами информацией о состоянии своих очередей и вычислительных ресурсов. Обновление информации происходит в момент изменения состояния системы: освобождение/занятие вычислительных ресурсов, добавление нового паспорта задания или удаление паспорта задания из очередей СУ, изменение конфигурации СУ. Частота наступления этих событий низка по сравнению со временем передачи информации по коммуникационным каналам.

Основная идея обеспечения безусловного прохождения пользовательским заданием всех этапов обработки состоит в многократном резервировании различных, независимо функционирующих компонентов системы.

В предложенной системе управления грид брокер каждого задания обеспечивает поддержку его актуального состояния задания в грид, блокируя отказы и сбои аппаратно-программных компонентов и обеспечивая автоматическое восстановление функционирования, за исключением случая отказа управляющей машины, где запущен брокер, при котором по требованию обеспечения отсутствия несанкционированного доступа после восстановления управляющей машины необходима новая аутентификация пользователя и повторный запуск задания на выполнение.

Ациклическая структура менеджеров характеризуется свойством самовосстановления и не снижает уровень безопасности, обеспечиваемый средствами пакета GT.

Литература

1. ***Корнеев В.В., Семенов Д.В.*** Распределенный метапланировщик грид // Вычислительные методы и программирование. – 2010. – Т. 11. – С. 69–76.

2. ***Семенов Д.В., Корнеев В.В.*** Оптимальная структура распределенного метапланировщика грид // Научный журнал «Вычислительные методы и программирование». – 2012. – Т. 13. – С. 146–152.

3. Создание распределенной инфрастуктуры для суперкомпьютерных приложений / ***Г.И. Савин, В.В. Корнеев, Б.М. Шабанов и др.*** // Программные продукты и системы. – 2008. – № 2. – С. 2–7.

4. Руководство программиста грид. – URL: http://www.jscc.ru/informat/grid1.zip (дата обращения: 04.05.2014).

5. ***Foster I., Kesselman C.*** Globus: a metacomputing infrastrusture toolkit. – URL: http://www.globus.org (дата обращения: 04.05.2014).

6. Руководство программиста. Суперкомпьютер МВС 15000ВМ. – URL: www.jscc.ru (дата обращения: 04.05.2014).

Статья поступила
6 мая 2014 г.

Корнеев Виктор Владимирович – доктор технических наук, профессор, заместитель директора ФГУП «НИИ «Квант» (г. Москва). *Область научных интересов*: вычислительные системы, микропроцессорные системы.

Семенов Дмитрий Викторович – старший научный сотрудник ФГУП «НИИ «Квант» (г. Москва). *Область научных интересов*: вычислительные системы, комплексы и сети.

Телегин Павел Николаевич – кандидат технических наук, заведующий отделом Межведомственного суперкомпьютерного центра РАН (г. Москва). *Область научных интересов*: параллельное программирование, распределенные вычисления, инструментальные средства программирования, автоматическое распараллеливание программ, эффективность выполнения программ, архитектуры высокопроизводительных вычислительных систем. **E-mail: telegin@jscc.ru**

Шабанов Борис Михайлович – кандидат технических наук, заместитель директора Межведомственного суперкомпьютерного центра РАН (г. Москва). *Область научных интересов*: высокопроизводительные вычислительные системы, архитектура ЭВМ, использование суперкомпьютеров в научных расчетах, грид-технологии.