Объединение Грид–сегментов с различной инфраструктурой♣

© А. Демичев, В. Ильин, А. Крюков, Л. Шамардин.

НИИ ядерной физики МГУ, 119992, Москва
demichev@theory.sinp.msu.ru, ilyin@theory.sinp.msu.ru, kryukov@theory.sinp.msu.ru, shamardin@theory.sinp.msu.ru

Аннотация

Предложена схема объединения сегментов Грид–систем, предназначенных для распределенной обработки (сверх)больших массивов данных. Эти сегменты могут быть основаны на различных спецификациях (инфраструктуре) соответствующих Грид–служб. Конкретным применением предложенной схемы может стать Грид–система для обработки экспериментальных данных с крупнейшей установки для изучения свойств элементарных частиц — Большого Адронного Коллайдера (БАК).

1 Введение

Главной целью данной работы является выработка схемы объединения различных сегментов Грид–систем [1], предназначенных для распределенной обработки (сверх)больших массивов данных. Необходимость в объединении таких сегментов может возникнуть по ряду причин, в том числе:

- расширение общей Грид–системы за счет сегментов, построенных на основе спецификаций, протоколов и стандартов, отличающихся от используемых в основной части Грида;
- подключение к глобальной Грид–системе ("интер–Гриду") распределенной системы масштаба корпорации или университета ("экстра–Грид");
- постепенная ("посегментная") модернизация программного обеспечения промежуточного слоя (middleware) глобальной Грид–системы.

Примером последней ситуации (на котором мы, в основном, и сконцентрируемся в дальнейшем) является планируемая модернизация ПО промежуточного слоя распределенной вычислительной системы для обработки результатов с БАК. Большой Адронный Коллайдер (Large Hadron Collider — LHC), запуск которого планируется в 2007 году в Европейском центре ядерных исследований (ЦЕРН, Женева), станет крупнейшей в мире установкой для изучения свойств элементарных частиц. Данные экспериментов БАК позволят существенно развить представления о строении микромира, они сыграют важную роль также в получении новых знаний о строении и эволюции Вселенной. Уникальной особенностью проекта БАК является огромный объем регистрируемых экспериментальных данных, порядка 12–14 петабайт в год (что в 1000 раз превышает поток данных с предшествующей подобной установки — ускорителя LEP). Проблема обработки в режиме реального времени такого потока информации до настоящего времени не встречалась. Кроме того, обрабатывать и анализировать эти экспериментальные данные будут исследователи из многих стран, общим числом порядка десяти тысяч из сотен институтов и университетов.

Для решения указанных задач обработки данных будут использоваться Грид–технологии. Распределенная программно–аппаратная компьютерная среда Грид позволяет объединять вычислительные мощности различных организаций или подразделений внутри одной крупной организации [1]. Такие структуры должны обеспечивать общедоступный, надежный и сравнительно простой доступ к вычислительным Грид–ресурсам, а также к устройствам хранения и обработки данных. Следует отметить, что на данном этапе развития Грид–технологий более или менее успешно пройден этап испытаний систем, характеризующихся большой степенью однородности, как по ресурсам, так и по характеру прикладных задач. Интенсивная работа по развертыванию такой системы для задач БАК ведется в настоящее время в рамках международных проектов LCG (LHC Computing Grid) [2] и EGEE (Enabling Grids for E–science in Europe) [3].

Однако, существующий подход к построению Грид–систем, базирующийся на программном обеспечении промежуточного слоя Globus Toolkit 2 (GT2) [4], [5], имеет определенные недостатки. Это связано с недостаточной гибкостью и надежностью отдельных компонентов системы, а также плохой совместимостью с мощными и развитыми

Труды 6ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL2004, Пущино, Россия, 2004.

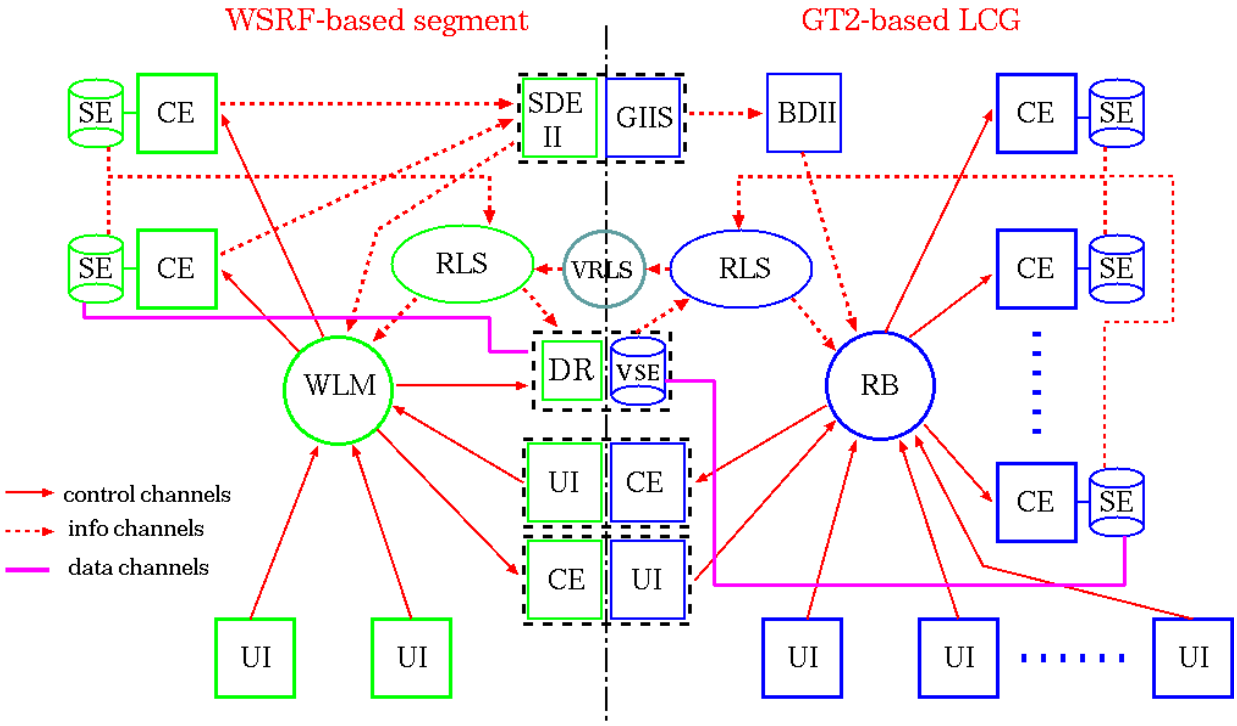

Рис. 1. Общая схема Грида для БАК с WS-сегментом.

средствами современных Веб–технологий. Последнее обстоятельство является особенно важным в связи с тем, что в соответствии с общеевропейским проектом EGEE [3], Грид для БАК должен стать основой системы распределенных вычислений и обработки данных общенаучного назначения. Разработанная Globus Alliance Открытая архитектура Грид–сервисов (OGSA) [6] должна обеспечить повышение надежности, гибкости, масштабируемости и защищенности распределенных систем. Это программное обеспечение — в соответствии с требованиями OGSA — поддерживает платформо–независимую интеграцию распределенных ресурсов на базе Java и XML технологий и протокола SOAP.

Наиболее полной реализацией принципов OGSA в настоящее время является Globus Toolkit 3 (GT3). Оказалось, однако, что спецификации инфраструктуры GT3 — OGSI (Open Grid Services Infrastructure) — тоже оказались несвободными от недостатков и в настоящее время ожидается переход на технологию WSRF (Web Services Resource Framework) [7], который будет реализован в Globus Toolkit 4 (GT4; ожидаемое время выпуска бета–версии — июнь 2004 г.).

В дальнейшем, для определенности, мы будем иметь ввиду переход от GT2 Грид ПО к технологии, основанной на использовании стандартов Веб–служб (WS–стандартов, например, WSRF).

2 Возможные способы эволюционного перехода к новым технологиям в крупномасштабных Грид–системах

В случае небольших Грид–систем, наиболее простым и надежным способом перехода к новым технологиям для ПО промежуточного слоя является его одновременная замена на всех узлах распределенной системы. Однако, в случае очень крупных систем (в частности, в случае Грид–системы для БАК) такое решение может оказаться неприемлемым по следующим причинам:

- время перехода к новой работающей версии всей системы может оказаться слишком большим с точки зрения тех практических задач, для решения которых создавалась данная Грид–система (хотя в рабочем режиме вся установка БАК должна заработать в 2007 году, Грид система необходима уже сейчас для тестирования отдельных компонент установки, прикладного программного обеспечения и т.д.);
- необходимы большие усилия и затраты на одновременное переобучение многочисленного обслуживающего персонала и пользователей системы.

Поэтому эволюционный, постепенный путь перехода к новой технологии может оказаться более подходящим. Несмотря на то, что такой путь требует дополнительных затрат для обеспечения совместимости частей системы с различным ПО

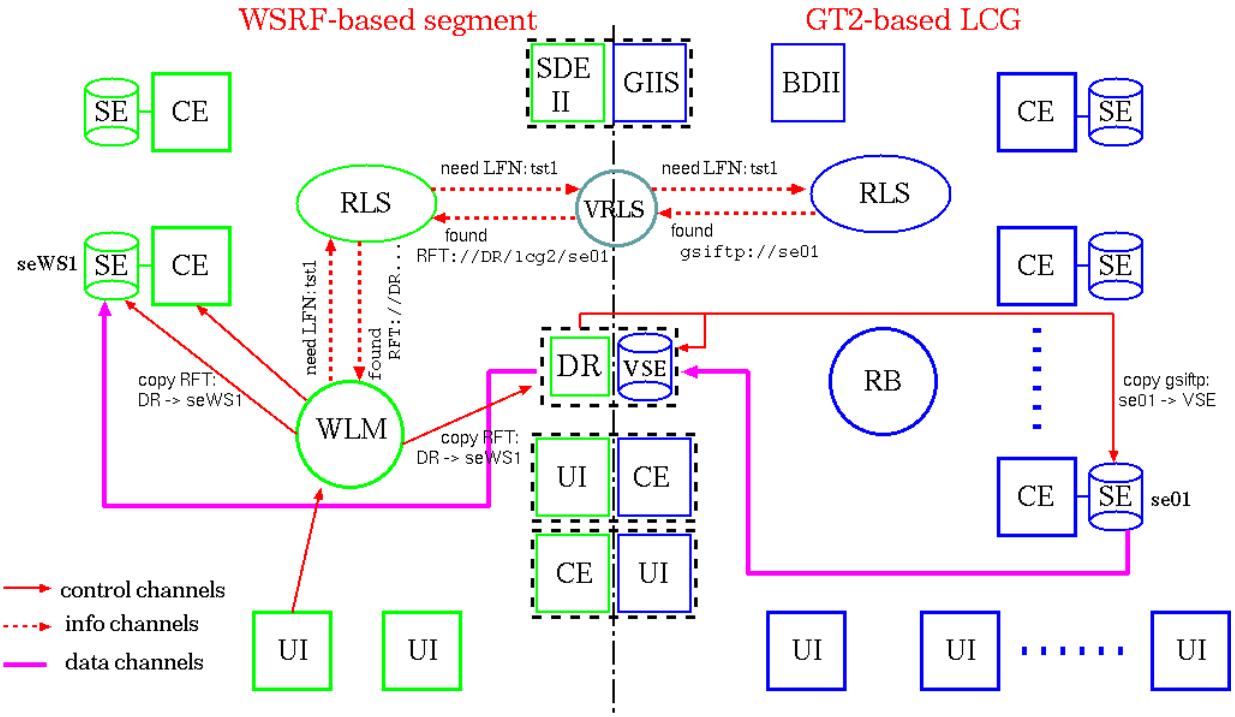

Рис. 2. Запуск задания на ресурсе WS-сегмента в случае, когда данные для его выполнения находятся на накопителе в GT2-Грид системе.

промежуточного слоя, он позволяет осуществить переход постепенно, с сохранением полной функциональности большей части Грида практически на каждой стадии перехода.

Эволюционный путь перехода может быть реализован следующими способами.

1. Создание Веб–сервисных (WS) надстроек над GT2–службами; примером может служить GRAM (Grid Resource Allocation Manager) в GT3 (OGSI–надстройка над GT2–службой) или служба RFT (Reliable File Transfer) — с надстройкой над GridFTP. На заключительном этапе и "внутренние" части служб должны быть реализованы на основе WS–технологий.
2. Создание "WS–слоев", то есть функционально связанных наборов служб на основе Веб–служб, взаимодействующих с остальной Грид–системой; например, это может быть слой мониторинга Грид–системы на основе спецификаций WSMF (Web Service Management Framework) [8].
3. Создание сегментов на основе WS–технологий и спецификаций и объединение их с GT2–Грид–системой с помощью специальных граничных интерфейсных служб.

Достоинством последнего из перечисленных способов является то, что в процессе разработки, реализации, развертывания и тестирования служб на основе новой технологии работающий основной GT2–сегмент полностью сохраняет свою функциональность и может использоваться для других текущих задач. При этом новый сегмент является сравнительно небольшим, но реально работающим “полигоном” для тестирования и отладки новой инфраструктуры распределенных вычислений. После этапа отладки может быть создано несколько таких сегментов, и после переобучения с их помощью персонала и пользователей можно достаточно быстро и безболезненно перейти к новой инфраструктуре всей Грид–системы.

3 Общая схема подключения сегмента с отличающейся инфраструктурой Грид–служб

Общая схема подключения сегмента с другой инфраструктурой служб приведена на рис. 1. При этом для обозначения компонент системы (общее введение в Грид–технологии и архитектуру см., например, в [1], [5]; существующую в настоящее время архитектуру LCG можно найти в документах на сайтах [2]) использованы следующие аббревиатуры: CE — computing element; SE — storage element; RB — Resource Broker; WLM — Workload Manager; DR — Data Router; VSE — Virtual Storage Element; UI — User Interface; WN — Working Nodes; SDE — Service data Elements; SDE II — Information Index based on SDE; GIIS — Grid Index Information Service; BDII — Berkeley DB Information Index; RLS — Replica Location Service; VRLS — Virtual Replica Location Service. (Замечание: для диспетчеров запуска заданий в разных сегментах используются различные наименования и, соответственно, аббревиатуры,

чтобы подчеркнуть, что они могут использовать совершенно разные алгоритмы распределения заданий по ресурсам.)

Граница между сегментами с различной инфраструктурой обозначена на рисунке штрих–пунктирной линией. Компонентами, которые позволяют объединить WS–сегмент с основным GT2–Гридом являются специальные граничные службы, обозначенные штриховыми прямоугольниками и виртуальная служба репликации данных (VRLS). Задачей этих граничных служб является трансформация протоколов, форматов и т.д., на которых основана инфраструктура одного сегмента в соответствующий набор форматов и протоколов другого сегмента.

Например, при запуске задания из WS–сегмента (то есть, с какого–либо UI в левой половине рисунка) диспетчер WLM сначала проверяет возможность запуска заданий в том же сегменте. Если подходящего ресурса в этом сегменте не обнаружено, то задание направляется на граничный CE. Задачами этого CE (то есть, соответствующей веб–службы) являются:

- трансляция языка описания заданий и спецификации необходимых ресурсов, принятого в WS–сегменте (например, в случае GT3 это так называемый Resource Specification Language (RSL),основанный на XML) в язык (JDL), понятный диспетчеру и CE в GT2–сегменте;
- запуск задания с соответствующего граничного интерфейса пользователя UI (направления запроса диспетчеру RB), то есть преобразование команд WS–клиента в команды GT2–клиента;
- согласование инфраструктур, обеспечивающих безопасность вычислений в обоих сегментах (с сохранением свойства делегирования прав во всей Грид–системе).

При запуске заданий из GT2–сегмента, они могут быть направлены для выполнения в WS–сегмент с помощью соответствующей пары граничных служб CE–UI. При этом RB использует информацию о ресурсах WS–сегмента, полученную от граничных служб SDE II — GIIS, которые преобразуют информацию о ресурсах в формате WS–сегмента (например, в случае GT3 это может быть XML–формат SDE) в формат, основанный на протоколе LDAP и “понятный” для BDII и RB.

В случае, когда для выполнения заданий необходимы большие массивы данных (как в случае Грида для БАК), основной проблемой становится управление передачей данных между сегментами с различной инфраструктурой. Возможная схема передачи данных представлена на рис. 2.

Для простоты и наглядности на рисунке изображен только случай, когда данные с SE в GT2–сегменте должны быть скопированы на SE в WS–сегменте, при этом в данном конкретном примере запуск задания и его выполнение осуществляются в WS–сегменте. Диспетчер WLM запрашивает RLS о местоположении необходимых данных (LFN — Logical File Name). Если данные уже находятся в WS–сегменте, то они просто копируются на нужный SE — это простой вариант. Если данных в WS–сегменте нет, то запрос переправляется в RLS GT2–сегмента (преобразование форматов метаданных о копиях файлов с данными осуществляет VRLS). Получив информацию о местоположении необходимых файлов с данными, WLM выдает команду (так называемая, передача данных, контролируемая "третьей стороной" (third party transfer)) граничным службам DR — VSE на копирование данных с соответствующего SE в GT2–сегменте на VSE, а затем на SE того CE, на котором будет выполняться задание.

Аналогичным образом осуществляется передача данных в случае, когда задание выполняется в GT2–сегменте, а данные находятся в WS–сегменте.

4 Заключение

Предполагается, что основная Грид–система для обработки экспериментальных результатов БАК постепенно будет модернизирована так, чтобы она удовлетворяла требованиям и спецификациям OGSA/WSRF. Предлагаемая нами схема может обеспечить постепенный переход к новой инфраструктуре с сохранением функциональных свойств большей части Грид–системы почти на всех этапах модернизации. В заключение, отметим, что многие из проблем обсуждавшихся в данной работе актуальны не только при модернизации Грид–систем, но и в случае объединения экстра– и интер–Грида (экстра–Грид — это Грид в масштабах корпорации, крупной компании или университета, а интер–Грид — это глобальная система распределенных вычислений; более подробно об этих понятиях см., например, в [5]).

Литература

[1] Ian Foster and Carl Kesselman, “The Grid: Blueprint for a New Computing Infrastructure”, Morgan Kaufmann, 1999;
Ian Foster and Carl Kesselman, “The Grid 2: Blueprint for a New Computing Infrastructure”, 2nd Edition, Morgan Kaufmann, 2003.

[2] http://www.cern.ch/LCG; http://lcg.jinr.ru;
В. Ильин, В. Кореньков, А. Солдатов "Российский сегмент глобальной инфраструктуры LCG", *Открытые системы*, No.1 (2003), http://www.osp.ru/os/2003/01/056.htm

[3] EGEE Project, 2003. http://egee-intranet.web.cern.ch/egee-intranet/gateway.html

[4] The Globus Alliance, 1999. http://www.globus.org

[5] "Introduction to Grid Computing with Globus" IBM Red book, 2003. http://www.redbooks.ibm.com/
[6] The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Ian Foster, Carl Kesselman, Jeffrey M. Nick, Steven Tuecke, 2002. http://www.globus.org/research/papers/ogsa.pdf; Globus Project: Open Grid Services Architecture, 2002. http://www.globus.org/ogsa
[7] The Globus Alliance: WS–Resource Framework, 2004. http://www.globus.org/wsrf
[8] Web services management framework, version 2.0, 2003. http://devresource.hp.com/drc/specifications/wsmf/WSMF-Overview.jsp

Integration of Grid segments with diverse infrastructure

A. Demichev, V. Ilyin, A. Kryukov, L. Shamardin

The scheme for integration of segments of Grid systems dedicated to distributed processing of the huge amounts of data is purposed. These segments may be based on different infrastructure specifications for the corresponding Grid services. The specific application for this scheme may be the Grid system for processing the data from the world's largest and most powerful particle accelerator, the Large Hadron Collider (LHC).

♣ Работа выполнена при поддержке РФФИ (грант 04-07-90342) и фонда INTAS (грант INTAS-CERN-03-52-4297).