СОЗДАНИЕ РОССИЙСКОГО СЕГМЕНТА ЕВРОПЕЙСКОЙ ИНФРАСТРУКТУРЫ EU DATAGRID

В.А. Ильин, научно-исследовательский институт ядерной физики (НИИЯФ МГУ), Москва, ilyin@theory.sinp.msu.ru
В.В. Кореньков, Объединенный институт ядерных исследований (ОИЯИ), Россия, 141980, г. Дубна Московской области, ул.Жолио-Кюри, 6, korenkov@cv.jinr.ru

CREATING OF RUSSIAN SEGMENT OF EUROPEAN INFRASTRUCTURE EU DATAGRID

V. Ilyin, Skobeltsyn Institute of Nuclear Physics, Moscow State University, Moscow, Russia, ilyin@theory.sinp.msu.ru
V. Korenkov, Joint institute for nuclear research, Joliot-Curie 6, 141980 Dubna, Moscow region, Russia, korenkov@cv.jinr.ru

The intensive development of the network, computer and information technologies has created a basis for a global integration of calculations and high-speed communication links. The concept GRID was formulated that guesses the creation of a computer infrastructure of a new generation oriented to a qualitatively new level of access to informational and computing resources at a global level.

We have a major interest in the development of the GRID – technologies and the creation DataGRID infrastructure in Russia.

В настоящее время в мире интенсивно развивается концепция *GRID* - компьютерной инфраструктуры нового типа, обеспечивающей *глобальную интеграцию информационных и вычислительных ресурсов* на основе создания и развития управляющего и оптимизирующего программного обеспечения (middleware) нового поколения. Основа технологии *GRID* состоит в создании набора стандартизированных служб для обеспечения надежного, совместимого, дешевого и безопасного доступа к географически распределенным высокотехнологичным информационным и вычислительным ресурсам - отдельным компьютерам, кластерам и суперкомпьютерным центрам, хранилищам информации, сетям, научному инструментарию и т.д. ([1],[2])

Существуют два основных направления развития GRID технологий - вычислительный (computational) GRID, и DataGrid для интенсивных операций с базами данных (data intensive GRID). В вычислительном GRID создаваемая инфраструктура нацелена на достижение максимальной скорости расчетов за счет глобального распараллеливания вычислений.

В других задачах важным аспектом является работа с массивами данными - такие проекты попадают под рубрику DataGRID. Существует большое количество прикладных областей, в которых обеспечивается хранение, обработка и анализ огромных массивов информации от сотен Терабайт до Петабайт с одновременным доступом нескольких тысяч пользователей со всего мира.

Для решения таких задач создаются DataGrid-инфраструктуры различного уровня, и развивается соответствующее промежуточное ПО, задачей которого будет обеспечивать эффективные, стандартные и прозрачные методы доступа к данным для осуществления кэширования данных, тиражирования и миграции файлов в гетерогенной среде. Для этого необходимо обеспечить управление универсальным пространством имен, эффективный перенос данных между вычислительными узлами, синхронизацию удаленных копий, доступ и кэширование данных на глобальном уровне, а также интерфейс к системам управления массовой памятью.

В настоящее время в мире реализуются масштабные проекты, основная цель которых - апробация разрабатываемых *GRID* технологий и их развитие. Сейчас эти проекты в основном создаются в различных научных приложениях. Например, в США реализуется проект создания крупнейшего распределенного суперкомпьютера *TeraGRID (http://www.teragrid.org),* проекты *iVDGL (http://www.ivdgl.org)*, *PPDG (http://www.ppdg.org)* и *GriPhyn (http://www.griphyn.org)*, а в Европе проекты *EU-DataGRID (http://www.eu-datagrid.org), EuroGRID (http://www.eurogrid.org), CrossGRID (http://www.cyfronet.krakow.pl/crossgrid), DataTag*. *(http://www.dattag.org)*.

Во многих европейских странах созданы и реализуются национальные *GRID* проекты. С 2003 года начинает реализовываться шестая рамочная программа Европейской Комиссии, одной из основных задач которой является создание и развитие *GRID* инфраструктур в Европе, направленная на широкое внедрение *GRID* технологий, разрабатываемых и развиваемых в научных исследованиях, в промышленности и других областях жизни современного общества.

Россия имеет уникальную возможность полномасштабно включиться в этот революционный процесс создания новейшей компьютерной технологии *XXI* века ([3]). Прогресс, достигнутый в области метакомпьютинга и распределенных вычислений и уже имеющийся опыт участия ряда российских научных организаций в международных *GRID* проектах, в особенности в области физике высоких энергий – проект *EU DataGRID,* позволит успешно развивать это важнейшее направление.

Особенно важным является участие России в крупнейшем международном научном проекте создания Большого адронного коллайдера (БАК) в ЦЕРН (Женева, Швейцария) на основе межправительственного соглашения. Для обработки данных экспериментов на этом ускорителе

создается уникальная мировая компьютерная система на основе применения *GRID* технологий – проект *LHC Computing GRID*.

Россия активно участвует в этом крупнейшем *GRID* проекте, создавая национальный сегмент (проект РИВК-БАК). Масштаб ресурсов создаваемой инфраструктуры должен составить порядка 50-100 Терафлопс вычислительной мощности, дисковых массивов для хранения информации на уровне Петабайта и линий связи пропускной способности более Гигабит/сек.

Проект EU DataGrid

Одним из наиболее масштабных в настоящее время является проект *EU DataGRID*, в котором участвуют крупные научные центры 14 европейских стран (в том числе и России) с целью создания глобальной инфраструктуры нового поколения для обработки огромных массивов информации в области физики высоких энергий, биологии и систем наблюдений за Землей.

Общим во всех этих исследованиях является разделение данных по различным базам, распределенным по всем континентам. Основная их цель — улучшение эффективности и скорости анализа данных посредством интеграции глобально распределенных процессорных мощностей и систем хранения данных, доступ к которым будет характеризоваться динамическим распределением по grid-инфрастуктуре, что предполагает управление репликацией и кэшированием.

Можно выделить две основные категории в работе с данными физических экспериментов: «производство» данных и их анализ конечным пользователем. Производство данных включает получение экспериментальных данных, распределенное моделирование физических событий, реконструкцию событий и частичную переработку. Анализ данных конечным пользователем включает интерактивный и удаленный анализ. Наиболее часто используемые данные потребуется хранить в памяти с наиболее быстрым доступом. В процессе анализа будут создаваться новые сложные объекты событий, которые будут сохраняться для дальнейшего анализа. Значительное количество времени будет затрачиваться на чтение объектов (их поиск и чтение из дискового кэша или с ленты). В силу независимости событий, их обработка предполагает крупномодульный параллелизм, основанный на высокой степени свободы в управлении вводом/выводом, что позволит обрабатывать события параллельно на различных вычислительных узлах. Задачи управления данными будут состоять в организации стандартного и быстрого переноса файлов из одной системы хранения в другую. Важными задачами также являются управление распределенным иерархическим кэшем, обеспечение проблем безопасности и прав доступа для пользователей.

В проект EU DataGrid вовлечено множество организаций, специалистов по программному обеспечению и ученых. Архитектура создавае-

мой grid-инфраструктуры должна быть достаточно простой, гибкой, масштабируемой, отвечающей требованиям распределенной обработки.

Проект состоит из нескольких рабочих пакетов (Work Packages, WP):

- WP1: Work Load Management System - управление рабочей загрузкой (распределенное планирование и управление ресурсами);
- WP2: Data Management - управление данными (создание интегрированного инструментария и инфраструктуры промежуточного слоя для согласованного управления и разделения петабайтных объемов данных с эффективным использованием ресурсов);
- WP3: Grid Monitoring / Grid information Systems - мониторинг (доступ к информации о состоянии и об ошибках в grid-инфраструктуре);
- WP4: Fabric Management – управление ресурсами вычислительных комплексов (серверов, дисковых массивов, систем массовой памяти, кластеров, состоящих из десятков тысяч вычислительных узлов);
- WP5: Storage Element - управление массовой памятью (создание глобального grid-интерфейса к существующим системам управления массовой памятью);
- WP6: Testbed and demonstrators – тестирование и отладка взаимодействия всех компонентов и сервисов географически распределенных сегментов grid-инфрастуктуры;
- WP7: Network Monitoring - создание виртуальной частной сети, объединяющей вычислительные ресурсы и ресурсы данных, участвующие в отладке grid-инфрастуктуры;
- WP8-WP10: High Energy Physics Applications, Earth Observation, Biology - создание для всех рассматриваемых отраслей (физики высоких энергий, биологии и наблюдения Земли) приложений, осуществляющих прозрачный доступ к распределенным данным и высокопроизводительным вычислительным ресурсам;

В качестве основы промежуточного программного обеспечения для проекта EU Data Grid выбран набор инструментальных средств Globus. Toolkit (*http://www.globus.org*).

В настоящий момент разработана многоуровневая архитектура DataGRID, основные уровни и компоненты которой изображены на рис. 1.

Каждая компонента в этой архитектуре состоит из пакетов, протоколов, сервисов и интерфейсов с другими компонентами.

Например, Grid Scheduler, который разрабатывается в рамках WP1 (Workload Management System), состоит из следующих частей:

User Interface (UI) – обеспечивает доступ пользователей к инфраструктуре GRID (с помощью языка описания заданий JDL);

Resource Broker (RB) – диспетчер GRID ресурсов;

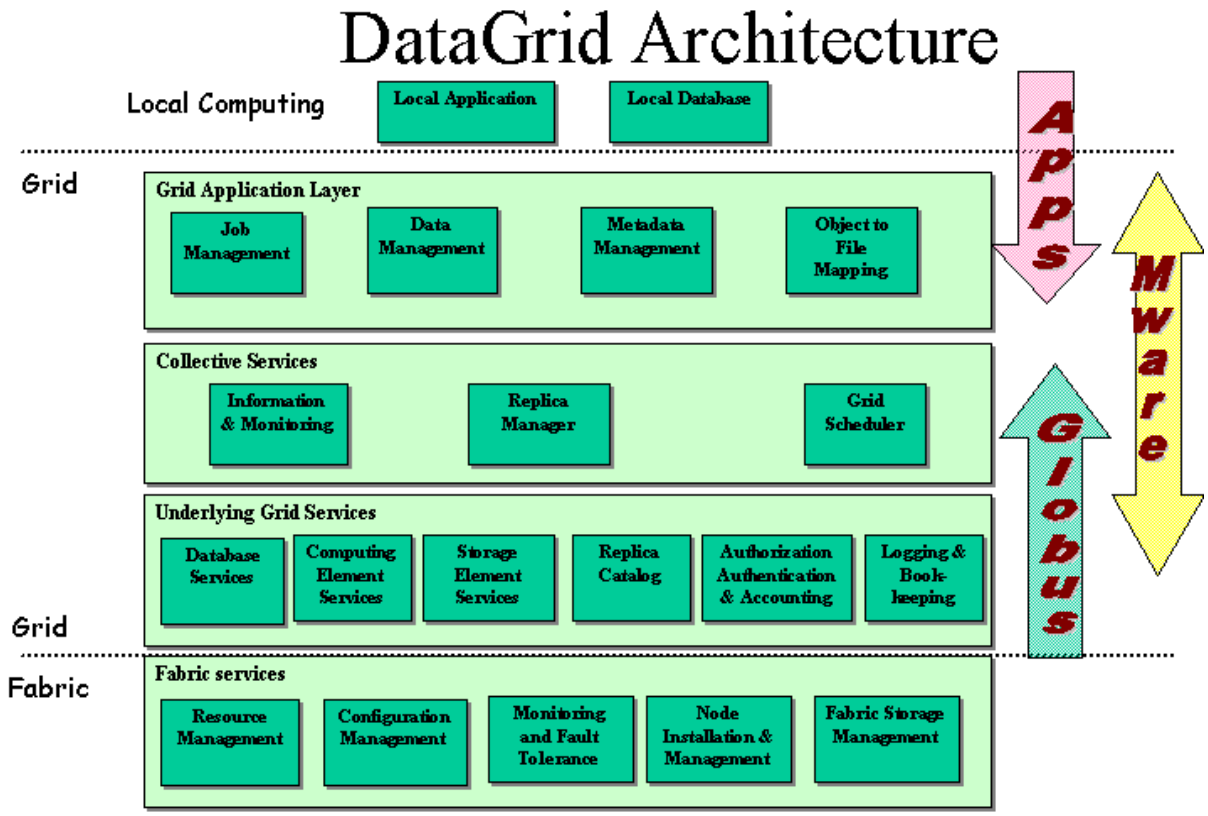

Рис. 1

Job Submission System (JSS) – интерфейс к системе пакетной обработки заданий;

Information Index (II) – LDAP – сервер, который используется в качестве фильтра для выбора ресурсов;

Logging and Bookkeeping (LB) - база данных для хранения информации о заданиях.

Эта компонента имеет интерфейсы ко многим другим пакетам и сервисам (Globus Gatekeeper, Replica Catalog, Information Systems, Network monitoring).

Большое внимание в проекте уделяется созданию средств управления данными, компоненты которых затрагивают несколько слоев архитектуры GRID (file and replica management, metadata access, data security).

Пакет WP2 (Data Management) состоит из нескольких компонент:

Replica Manager – управляет процессом тиражирования данных в среде GRID, включая оптимизацию запросов , интерфейс к сервису Replica Catalog;

Replica Catalog – обеспечивает преобразование логических имен файлов в соответствующие физические имена;

GDMP (GRID Data Mirroring Package) – используется для создания реплик некоторых типов файлов, синхронизации этого процесса с поддержкой и актуализацией Replica Catalog;

Spitfire – обеспечивает сервис доступа к реляционным базам данных посредством промежуточного ПО в инфраструктуре GRID.

Средствами управления тиражирования копии файлов или метаданных помещаются в распределенный иерархический кэш. Для выполнения этой задачи необходимо обращение к блоку пересылки данных в промежуточном слое, который, в свою очередь, будет использовать средства доступа к данным или указатели к метаданным, хранящимся под управлением тех или иных систем управления хранением данных или метаданных. Перечисленные компоненты должны обеспечивать надлежащие механизмы безопасности.

Стержневой проблемой управления данными в инфраструктуре data grid является гетерогенность репозиториев данных. Задача должна решаться для различных систем хранения данных: системы управления типа HPSS, Castor, UniTree (*http://www.unitree.com*) или Enstore (*http://www-isd.fnal.gov/enstore/ design.html*); дисковые системы типа DPSS (*http://www-itg.lbl.gov/DPSS*); распределенные файловые системы; базы данных. При такой гетерогенной организации хранения данных очень сложным является решение проблемы наименования и доступа. При иерархической организации управления памятью (Hierarchical Storage Management — HSM) обеспечивается автоматический и прозрачный доступ к хранилищу данных, состоящему из лент, промежуточного дискового хранилища данных и дисков быстрого доступа. В подобной иерархической системе данные переносятся сначала с лент на локальный дисковый кэш до начала grid-переноса данных. При этом запросы должны группироваться таким образом, чтобы достичь оптимального монтирования лент, что требует организации внутренних каталогов и механизмов переноса данных с ленты на диск.

Тиражирование данных может рассматриваться как процесс управления копиями. Это также есть стратегия кэширования, при которой идентичные файлы доступны в нескольких местах grid-инфраструктуры. Главная цель тиражирования — достижение более быстрого доступа к данным за счет их местонахождения в локальном кэше или в ближайшей копии. Иначе говоря, осуществлять перенос файла по всей глобальной сети для каждого единичного запроса не приходится. Каждая реплика должна синхронизироваться с другими репликами. Качество реплики зависит от протоколов обновления и сетевых параметров grid-инфраструктуры. Должна быть также выработана стратегия обновления и создания реплик. Создание реплик особенно актуально при объемах данных порядка нескольких петабайт.

Тиражирование метаданных требует использования механизма связи на каждом grid-узле. Инструментальный набор средств Globus предоставляет две возможности: сокеты и коммуникационную библиотеку Nexus более высокого уровня. В подсистеме коммуникации должны быть реализованы различные протоколы тиражирования (синхронные и асинхронные методы обновления). Replica Manager обеспечивает службы доступа высокого уровня и оптимизирует глобальную пропускную способность с использованием grid-кэшей. Анализ запроса пользователя приводит к оптимальному выполнению этого запроса, а в соответствии с анализом множества запросов принимается решение о создании или уничтожении реплики. Replica Manager осуществляет глобальное кэширование, в то время как за создание локальных кэшей отвечают системы массовой памяти.

Связующим элементом в grid-системе является служба управления метаданными — каталогами с именем и указателем на расположение единичных или реплицированных файлов, информацией по мониторингу (статус, пропускная способность и т.п.), информацией по конфигурации grid (описание сетей, коммутаторов, кластеров, узлов и ПО), стратегиями гибкого динамического управления. Именно эта служба обеспечивает интеграцию разнообразных, децентрализованных и гетерогенных составляющих grid.

Многие аспекты обеспечения безопасности в grid-инфраструктуре тесно связаны с управлением данными, в особенности, организация grid-кэшей и стратегия синхронного тиражирования. В распределенной системе хранения данных, включающей реплики, запрос оптимизируется за счет существования нескольких копий файлов. Оптимальная схема выполнения запроса зависит от ряда динамических и статических факторов: размер файла, к которому требуется доступ; уровень загрузки данных для обслуживания запрашиваемого файла; метод/протокол, по которым будет осуществлен доступ и перенос файла; пропускная способность сети, расстояние и трафик внутри grid; стратегия управления удаленным доступом.

Российский сегмент DataGRID

С 2000 года в российских институтах по физике высоких энергий (ОИЯИ, НИИЯФ МГУ, ИТЭФ, ИФВЭ и др.) осуществляется проект создания российского сегмента европейской инфраструктуры DataGRID, которая должна стать основой для построения глобальной распределенной компьютерной системы для обработки, хранение и анализа данных с ускорителя БАК (LHC Computing GRID). Российский сегмент DataGRID является функциональной подсистемой этой европейской инфраструктуры. В России впервые созданы виртуальные организации типа GRID для решения конкретных прикладных задач.

Среди основных результатов можно отметить следующие.

Создана сетевая инфраструктура, объединяющая ядерно-физические центры Московского региона с пропускной способностью от 30 до 1000 Мв/сек.

В крупных центрах созданы компьютерные инфраструктуры, состоящие из вычислительных кластеров (суммарно более 200 процессоров), дисковых массивов емкостью около 10 ТВ, ленточных библиотек, а также средств визуализации.

Освоена технология создания информационных серверов GIIS, собирающих информацию о локальных вычислительных ресурсах и ресурсов по хранению данных (создаваемых GLOBUS службой GRIS на каждом узле распределенной системы) и передающих эту информацию в динамическом режиме в вышестоящий сервер GIIS. Таким образом, освоена и протестирована иерархическая структура построения информационной службы GRIS-GIIS. Организован общий информационный сервер GIIS (ldap://lhc-fs.sinp.msu.ru:2137), который передает информацию о локальных ресурсах российских институтов на информационный сервер GIIS (ldap://testbed001.cern.ch:2137) европейского проекта EU DataGRID (рис. 2).

Рис. 2

В НИИЯФ МГУ создан Сертификационный центр (Certification authority, CA) для российского сегмента. Сертификаты этого центра прини-

маются всеми участниками европейского проекта EU DataGRID. Разработана схема подтверждения запросов на сертификаты с помощью расположенных в других организациях Регистрационных центров (Registration authority, RC), заверяющих запросы пользователей электронной подписью с помощью сертификата GRID. Разработаны программы постановки и проверки электронной подписи, а также пакет программ для автоматизации работы Сертификационного центра.

Предложенная схема CA+RC и пакет программ приняты в ЦЕРНе и других участниках европейского проекта EU DataGRID. Инсталлирована и протестирована программа репликации файлов и баз данных GDMP (**GRID Data Mirroring Package),** которая создана для выполнения удаленных операций с распределенными базами данных. Она использует сертификаты GRID и работает по схеме клиент-сервер, т.е. репликация изменений в базе данных происходит в динамическом режиме. Сервер периодически оповещает клиентов об изменениях в базе, а клиенты пересылают обновленные файлы с помощью команды GSI-ftp. Текущая версия GDMP работает с объектно-ориентированнной базой данных Objectivity DB, а также создается версия с динамической репликацией обычных файлов. Программа GDMP активно используется для репликации в ЦЕРН распределенной базы смоделированных данных, создаваемой в ОИЯИ (Дубна), НИИЯФ МГУ и других институтах по физике высоких энергий для эксперимента LHC-CMS. Программа GDMP рассматривается в качестве GRID стандарта для репликации изменений в распределенных базах данных.

Проводятся работы по исследованию сетевого трафика процесса тиражирования данных между серверами ОИЯИ и ИФВЭ (Протвино).

В ОИЯИ выполнен комплекс работ по мониторингу сетевых ресурсов, узлов, сервисов и приложений.

Сотрудники ОИЯИ принимают участие в развитии средств мониторинга для вычислительных кластеров с очень большим количеством узлов (10.000 и более), используемых в создаваемой инфраструктуре EU DataGRID.

В рамках задачи Monitoring and Fault Tolerance (Мониторинг и устойчивость при сбоях) они участвуют в создании системы корреляции событий (Correlation Engine). Задача этой системы - своевременное обнаружение аномальных состояний на узлах кластера и принятие мер по предупреждению сбоев.

С помощью созданного прототипа Системы корреляции событий (Correlation Engine) ведется сбор статистики аномальных состояний узлов на базе вычислительных кластеров ЦЕРН. Производится анализ полученных данных для выявления причин сбоев узлов. Этот этап позволит получить первый опыт в предсказании сбоев. На втором этапе предусмотрено расширение прототипа Correlation Engine с учетом полученных результа-

тов и испытание системы автоматизированного предупреждения сбоев на практике.

Этот прототип установлен на вычислительных кластерах в ЦЕРН и ОИЯИ, где производится сбор статистики аномальных состояний узлов.

Эти разработки включены в создаваемую архитектуру системы глобального мониторинга (GMA – Grid Monitoring Architecture).

Специалистами НИИЯФ МГУ и ОИЯИ совместно с сотрудниками INFN (Италия) разработана и апробирована новая схема интеграции инструментальных пакетов IMPALA/BOSS и GRID-технологий для автоматизации процесса массовой генерации событий эксперимента LHC-CMS.

В сотрудничестве с Институтом прикладной математики имени М.И. Келдыша программа *Metadispetcher* установлена в российском сегменте инфраструктуры EU DataGRID.

Программа *Metadispetcher* предназначена для планирования запуска заданий в среде распределенных компьютерных ресурсов типа GRID.

Было проведено ее тестирование, по результатам которого программа была доработана для обеспечения эффективной передачи данных средствами GLOBUS. Ведутся работы по сравнительному анализу компонент *Metadispetcher и* Resource Broker.

В сферу работ по развитию российского сегмента инфраструктуры EU DataGRID включаются постепенно все большее количество российских институтов и университетов.

Создание национальной инфраструктуры GRID является стратегически важной задачей для многих направлений современной науки и высокотехнологических прикладных областей.

Литература

[1] J. Foster, K. Kesselman. GRID: a Blueprint to the New Computing Infrastructure. Morgan Kaufman Publishers, 1999

[2] Виктор Коваленко, Дмитрий Корягин. Вычислительная инфраструктура будущего. «Открытые системы», 1999, № 11-12

[3] А.В. Жучков, В.А. Ильин, В.В. Кореньков, «Некоторые аспекты создания глобальной системы распределенных вычислений в России», труды Всероссийской научной конференции «Высокопроизводительные вычисления и их приложения», сс. 227-231, Черноголовка, 2000