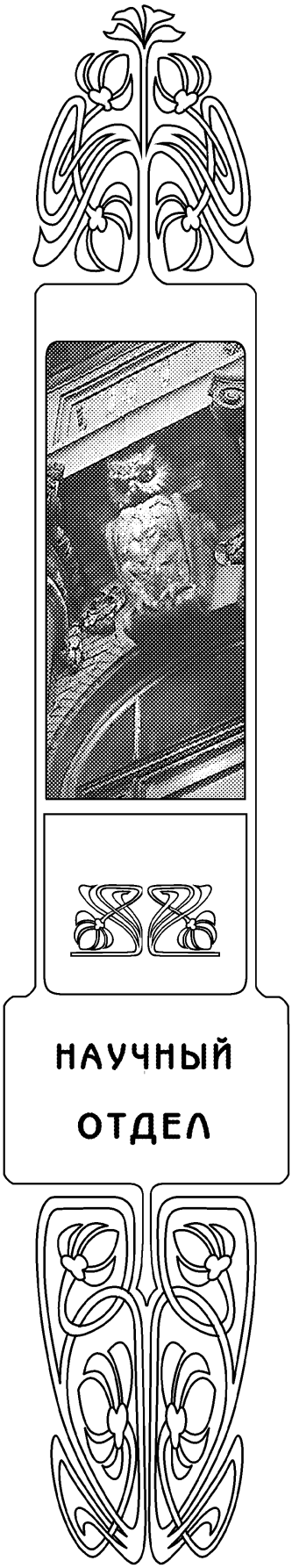
ИНФОРМАТИКА



УДК 519.689

ПРИМЕНЕНИЕ ПРОГРАММНОГО КОМПЛЕКСА gLite ДЛЯ ОРГАНИЗАЦИИ РАСПРЕДЕЛЕННОГО ХРАНИЛИЩА ДАННЫХ

В.М. Соловьев, М.Г. Щербаков*

Саратовский государственный университет,
ПРЦНИТ,
*кафедра математической кибернетики и компьютерных наук
E-mail: svm@sgu.ru, *mihgen@gmail.com

В статье дается введение в использование инфраструктуры grid для организации распределенного хранилища данных. Существует большое число различных технологий построения grid-систем для задач географически распределенного хранения данных и распределенных вычислений. Одна из таких технологий, описываемая в данной работе, программное обеспечение промежуточного слоя gLite. Описаны архитектура, принцип работы и основные сервисы gLite.

Ключевые слова: хранилище, распределенный, gLite, грид, srm, dpm.

The Using of gLite Middleware for Organization of Distributed Data Storage

V.M. Solovyev, M.G. Scherbakov

This paper gives an introduction to using the grid technology for distributed data storage. There are many different technologies of grid systems deployment for geographically distributed data storages and distributed computation purposes. One of these technologies, gLite middleware is described in this paper. The architecture, the functioning, and the main gLite services are presented here.

Key words: storage, distributed, gLite, grid, srm, dpm.

ВВЕДЕНИЕ

В наноиндустрии при проведении уникальных и дорогостоящих экспериментов часто требуется долговременно хранить очень большие объемы «сырых» необработанных экспериментальных данных (десятки и сотни Тбайт). Такие данные обычно сопровождаются описанием параметров экспериментов, данными имитационного моделирования, инструкциями по эксплуатации оборудования, методиками проведения экспериментов, научными отчетами и комментариями, а также другими цифровыми материалами. Для дальнейшей обработки экспериментальных данных к ним должен быть организован удобный доступ. Кроме того, нужны поиск в среде необработанных экспериментальных данных и сопровождающих их цифровых материалов, управление жизненным циклом данных, включая создание цифровых материалов, передачу, хранение и организацию доступа к ним. Таким образом, пользователю необходимо масштабируемое прозрачное хранилище гетерогенных данных с гарантированным качеством сервиса: требуемым уровнем защиты, сохранности, удобства, скорости доступа к данным и т.д. Кроме того, нужен унифицированный механизм обмена данными разного типа (файлами, каталогами файлов, массивами, таблицами), совместимый с распределенными вычислительными сервисами (grid-сервисами). Такому распределенно-

му хранилищу данных соответствует многоуровневая архитектура, содержащая уровень интерфейсов пользователя для управления данными, уровень программного интерфейса для grid-сервисов сбора, обработки и управления потоками данных и ресурсный уровень.

Ресурсный уровень хранилища отвечает за физическое хранение данных. В качестве репозитория хранения цифровых материалов могут выступать файловые системы Unix и Windows, архивные системы хранения HPSS, бинарные большие объекты в DBMS (DB2, Oracle, MS SQL Server), доступные через SQL-запросы объекты баз данных DB2 или Oracle, библиотеки на магнитных лентах. Для получения требуемого уровня масштабирования хранилище может быть географически распределенным, эффективно использующим grid-технологии. Для создания таких географически распределенных систем в рамках проекта GRID EGEE [1] был создан программный комплекс gLite [2], позволяющий объединять вычислительные кластеры (grid computing) и кластерные системы хранения данных (data grid). Проект EGEE (Enabling Grids for E-SciencE) создавался европейским сообществом и направлен на создание grid-технологий для научных исследований, работающих в режиме 24/7. Этот проект ставит своей целью обеспечить академические и промышленные исследования доступом к основным вычислительным ресурсам, независимо от того, где они находятся. Во главе проекта EGEE стоит CERN, Европейская организация по ядерным исследованиям. Проект включает свыше 70 институтов-партнеров в Европе, Азии и Соединенных Штатах.

Программный комплекс gLite создавался с учетом опыта предшествующих исследовательских проектов LCG [3], DataGrid [4], DataTag [5], Globus [6], GriPhyN [7], iVDGL [8] и вобрал в себя ряд разработанных в этих проектах компонент. Он устанавливается на операционную систему Scientific Linux – релиз Linux, который создан совместными усилиями Fermilab и CERN при поддержке различных лабораторий и университетов. Базовый дистрибутив Scientific Linux «собран» на основе Enterprise Linux, скомпилированного из открытых исходных текстов.

1. АРХИТЕКТУРА ХРАНИЛИЩА ДАННЫХ НА ОСНОВЕ gLite

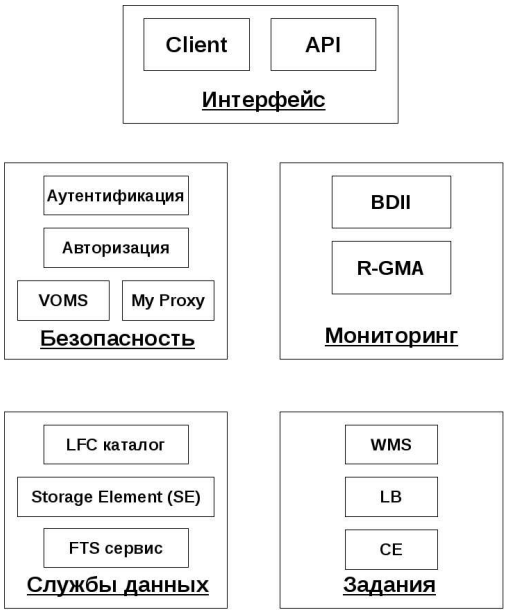


Рис. 1. Архитектура хранилища данных

Типовая архитектура хранилища данных на основе gLite может включать пять основных компонент [10], реализующих следующие функции:

управление заданиями (Job Management),

учет использования ресурсов (Monitoring),

хранение данных (Data Services),

обеспечение безопасности и сетевого мониторинга (Security),

организация интерфейса (Access).

Основу работы программного комплекса gLite составляет поддержка виртуальных научных организаций VOMS (Virtual Organisation Membership Service) на основе цифровых сертификатов. Основной единицей хранения в такой grid-системе является файл, который может иметь реплики на отдельных серверах grid-системы. Архитектура хранилища данных на основе gLite приведена на рис. 1.

1.1. Подсистема безопасности и сетевого мониторинга (Security)

Подсистема безопасности и сетевого мониторинга обеспечивает безопасный доступ к ресурсам в незащищенных сетях общего использования (Internet). В gLite работа этой подсистемы основана на инфраструктуре grid-безопасности (Grid Security Infrastructure, GSI), которая разработана Globus Alliance [6]. Подсистема предоставляет сервисы аутентификации, конфиденциальности передачи информации и делегирование прав. Под делегированием прав подразумевается следующее. Пользователю нужно лишь один раз пройти процедуру аутентификации, а далее система сама обеспечивает

аутентификацию его на всех ресурсах, которыми он собирается воспользоваться. Инфраструктура GSI, в свою очередь, основана на широко используемой технологии открытых криптографических ключей (Public Key Infrastructure, PKI). В качестве идентификаторов пользователей и ресурсов в GSI используются цифровые сертификаты стандарта X.509 [9] (стандарт международной организации International Telecommunication Union, ITU [10]), подписываемые центром выдачи сертификатов (Certification Authority, CA). Пользовательский сертификат, приватный ключ которого защищается паролем, используется для генерации так называемого прокси-сертификата (Proxy certificate). Прокси-сертификат не защищен паролем, что позволяет использовать его для делегирования полномочий пользователя запускаемым от его имени процессам, и имеет сравнительно короткий период действия (по умолчанию 12 часов).

Авторизация пользователей в grid-системе может осуществляться двумя способами. Первый способ основан на механизме grid-mapfile. Файл grid-mapfile содержит список Subject Names (имя пользователя, атрибуты из сертификата) и локальные аккаунты, в которых тот или иной пользователь должен отображаться. Таким образом, при авторизации список Subject Name из сертификата пользователя проверяется на соответствие локальным аккаунтам в системе, и, если соответствие найдено, пользователь может работать в системе с правами соответствующего аккаунта. Второй способ основывается на механизме VOMS. Авторизация здесь определяет возможности пользователя на доступ к ресурсам и сервисам. Для работы создается сертификат. Пользователи grid-инфраструктуры обычно разделяются по принадлежности к виртуальным организациям (Virtual Organisation, VO), объединяющим группы людей с похожими задачами. Члены виртуальной организации могут делиться на группы и подгруппы в иерархическом порядке.

При использовании gLite может возникнуть ситуация, когда период действия proxy-сертификата может оказаться недостаточным, например, при копировании файлов с помощью FTS-сервера или при запуске долговременных задач. Для таких случаев пользователю предоставляется возможность поместить долгоживущий сертификат в защищенное хранилище MyProxy. Использование сертификата, хранящегося на MyProxy-сервере, требует знания пароля, заданного при помещении сертификата на сервер.

1.2. Подсистема пользовательского интерфейса (Access)

Точкой доступа в grid-систему служит пользовательский интерфейс (User interface, UI). Пользовательский интерфейс может быть установлен на любом компьютере с операционной системой Linux или Windows и имеющим пользовательский сертификат. С помощью UI пользователь может быть аутентифицирован и авторизован в хранилище и получить доступ к ресурсам хранилища. В частности, пользователь может получать информацию о ресурсах, запускать вычислительные задачи, копировать файлы со своего компьютера через UI в хранилище, и наоборот, удалять файлы, делать реплики цифровых материалов между серверами и т.д.

1.3. Подсистема управления заданиями (Job Management)

Задачей подсистемы управления заданиями (Workload Management System, WMS) является принятие запросов на запуск заданий, поиск подходящих ресурсов и контроль их выполнения. Благодаря работе WMS, сложность управления приложениями и ресурсами скрыта от пользователей. Их взаимодействие с WMS ограничено описанием характеристик и требований запроса через ориентированный на пользователя язык высокого уровня — язык описания заданий (Job Description Language, JDL) и к направлению такого запроса через предоставленные интерфейсы.

1.4. Подсистема хранения данных (Data Services)

Подсистема хранения данных состоит из двух частей: устройств хранения данных и сервисов управления данными. Устройства хранения (Storage Elements, SE) могут управлять как простыми дисковыми серверами, так и большими дисковыми массивами или серверами хранения данных на магнитных лентах. Элементы SE поддерживают различные протоколы доступа к файлам, основными из которых являются GSIFTP (GSI-secure FTP, защищенный FTP-протокол) и RFIO (протокол

прямого доступа к файлам). Для элементов SE, управляющих простыми дисковыми серверами, используется сервис Disk Pool Manager (DPM). Большинство устройств хранения данных управляются сервисом Storage Resource Manager (SRM), предоставляющим единый интерфейс доступа к данным на дисках и магнитных лентах. Обычно DPM сервер устанавливается на том же узле, что и SRM сервер.

Подсистема хранения данных реализована на основе файловых каталогов. Как и в обычных компьютерах, в grid-системе основной единицей хранения данных является файл. Grid-система может состоять из нескольких географически удаленных групп серверов (сайтов). В этой среде цифровые материалы могут иметь реплики на различных сайтах. Поскольку все реплики должны быть идентичными, в хранилище нельзя произвольно изменять файлы, а разрешается только читать файлы, создавать новые и удалять их. Если пользователь использует логические имена файлов, то ему не требуется знать, на каком узле физически расположен его файл. Файлы в grid-системе именуются через Grid Unique IDentifier (GUID), Logical File Name (LFN), Site URL (SURL) и Transport URL (TURL). Логическое имя файла (LFN) и GUID не содержат информацию о физическом местоположении файла. Логическое имя LFN обычно представляет строку вида /grid/<vo-name>/<my-directory>/<file-name>, где <vo-name> – название виртуальной организации, <my-directory> – каталог или группа каталогов пользователя, <file-name> – имя файла. Соответствие между LFN, GUID и SURL хранится в файловом каталоге LFC (LCG File Catalogue).

1.5. Подсистема учета использования ресурсов (Monitoring)

Подсистема учета использования ресурсов — это информационная система. На каждом узле запускаются так называемые информационные провайдеры (Grid Resource Information Server, GRIS), которые собирают информацию о запущенных grid-сервисах на узле, их состоянии, а также некоторую динамическую информацию от сервисов. Например, объем неиспользованного дискового пространства SE. На каждом сайте стоит коллектор информации – сервис Berkeley Database Information Index (BDII). На самом верхнем уровне grid-системы находится top-BDII, который собирает информацию со всех сайтов. Таким образом, информацию обо всех ресурсах grid-системы можно получить в одном месте. Информационные провайдеры GRIS и BDII-сервисы работают на основе LDAP протокола. Недостаток такой схемы — сложность поиска информации. Решением проблемы является использование службы R-GMA (Relation Grid Monitoring Architecture), в которой информация представляется в виде пространственно-распределенной реляционной базы данных. Служба R-GMA входит в комплекс gLite и может использоваться одновременно с BDII. Она использует реляционную модель данных: данные представляются в виде таблиц, каждая запись представлена строкой (tuple) в таблице, обращаться к базе данных можно с помощью языка запросов SQL (Structured Query Language).

2. СЕРВИСЫ gLite

Схематически взаимодействие сервисов gLite показано на рис. 2.

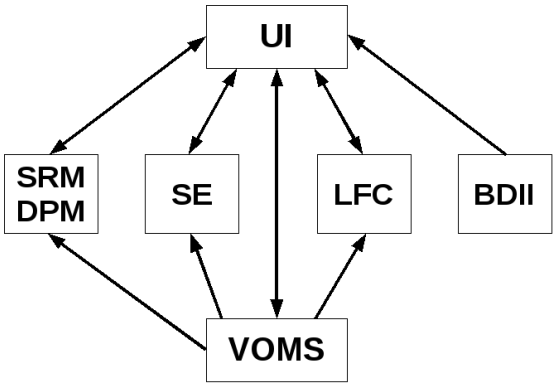


Рис. 2. Сервисы gLite

2.1. Безопасность

Как уже было отмечено, в качестве идентификаторов пользователей и ресурсов в grid-системе используются цифровые сертификаты стандарта X.509. В работе с сертификатами и в процедуре их выдачи-получения задействованы три компоненты.

Центр сертификации (Certification Authority, CA) – специальная организация, обладающая полномочиями выдавать (подписывать электронным способом) цифровые сертификаты.

Владелец сертификата – пользователь, или grid-ресурс, который пользуется сертификационными услугами CA. Компонента CA включает в сертификат данные, предоставляемые владельцем (имя, организация и т.д.), и заверяет его своей цифровой подписью. Владельцу сертификата присваивается уникальное имя (Distinguished Name, DN).

Пользователи или ресурсы, проводящие аутентификацию других grid-объектов. Они полагаются на информацию из сертификатов, при получении его от идентифицируемых объектов. Они могут принимать или отвергать сертификаты, подписанные CA.

Базовую безопасность в распределенном хранилище данных обеспечивают следующие сервисы: аутентификация (определяет «Кто я в хранилище?»); авторизация (определяет «Есть ли у меня доступ к ресурсам и сервисам?»); защита (обеспечивает целостность и конфиденциальность данных). Как уже отмечалось, для работы с grid-системой создается временный самоподписываемый сертификат Proxy certificate, со временем действия, равным 12 часам. Пользователи grid-инфраструктуры обычно разделяются по принадлежности к виртуальным организациям (ВО). Система VOMS используется для управления информацией о роли и привилегиях пользователей внутри виртуальной организации. Эта информация предоставляется grid-сервисам через расширение прокси-сертификата. Расширение представляет собой мини-сертификат, выдаваемый сервером VOMS по запросу пользователя и содержащий информацию о его членстве в ВО и его права. Авторизующие сервисы затем могут расшифровать расширение, и пользователю будут присвоены соответствующие права, например, позволяющие ему выполнять только определенные действия. Процедура авторизации в grid-системе схематично показана на рис. 3.

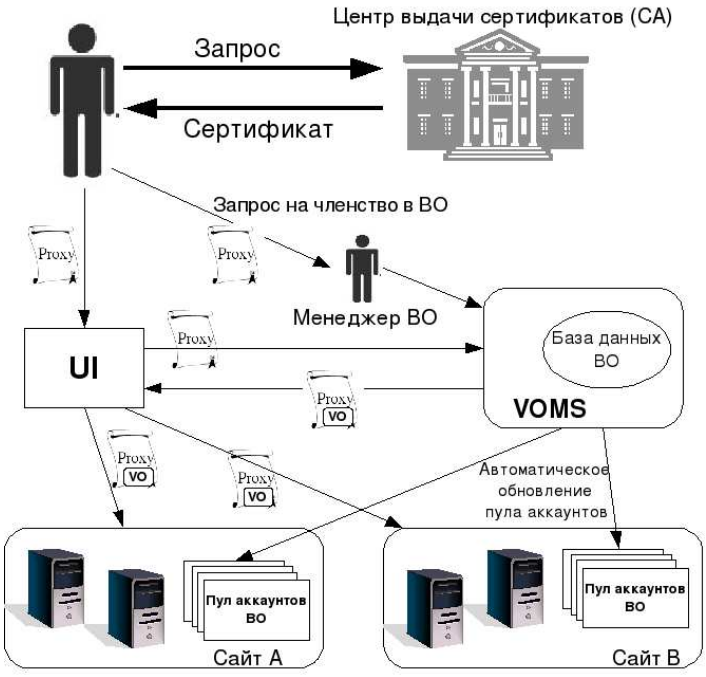


Рис. 3. Авторизация на основе прокси-сертификатов

2.2. Сервисы хранения данных

Сервисы хранения данных позволяют пользователю или приложению сохранять данные с целью последующего их использования. Они обеспечивают: гетерогенность – данные хранятся на различных устройствах (диски, ленты), использующих различные методы доступа; распределенность – данные хранятся на различных сайтах, где отсутствует общая разделяемая файловая система (данные могут перемещаться между сайтами); администрирование различных доменов – данные хранятся там, куда обычному пользователю доступ запрещен. Для реализации распределенного хранения цифрового материала необходимы следующие службы: сервис организации общего интерфейса к устройствам SRM; сервис определения местоположения файлов (LFC); сервис управления и надежной передачи файлов (File Transfer Service, FTS).

На имена файлов и каталоги пользователя можно создавать символические ссылки, подобно то-

му, как это делается в UNIX. Таким образом, одному физическому файлу могут соответствовать несколько логических имен. Идентификатор GUID [11] представляет собой число вида 93bd772a-b282-4332-a0c5-c79e99fc2e9c и служит для однозначной идентификации файла в рамках всего хранилища. Идентификатор хранения SURL имеет вид: srm://<srm-host>/<path>, где <srm-host> – имя сервера SRM; <path> – путь к файлу. Поскольку обычно на сайте используется один SRM-сервер, то SURL отражает сайт, где находится файл или одна из его реплик. Идентификатор транспорта TURL имеет вид: <protocol>://<host>/<path>, где <protocol> – протокол доступа к файлу, например gsiftp; <host> – имя узла, на котором физически расположен файл; <path> – путь к файлу. Все данные о соответствии между GUID, LFN и SURL располагаются в файловом каталоге LFC. Кроме того, каждому файлу можно поставить в соответствие мета-информацию (краткое описание), и эти данные также хранятся в файловом каталоге.

В целом подсистема хранения работает следующим образом. Пусть пользователь хочет прочитать содержимое файла, зная его логическое имя Logical File Name. Условная схема этой процедуры изображена на рис.4. Во-первых, пользователь обращается к серверу LFC (файловому каталогу), передавая LFN. Сервер LFC находит соответствующее значение уникального grid-идентификатора этого файла (GUID) и возвращает пользователю список адресов элементов хранения файла (SURL-адреса), соответствующих этому GUID. Во-вторых, пользователь выбирает из списка SURL-адресов ближайший к нему SRM-сервер и обращается к нему, передавая в качестве параметра полученный SURL. Сервер SRM, взаимодействуя с DPM-сервером (в случае дискового ресурса хранения данных), возвращает пользователю адрес протокола доступа к файлу (TURL-адрес), где содержится протокол доступа и имя узла. В-третьих, пользователь, используя указанный протокол доступа, получает доступ к файлу, расположенному на устройстве хранения (SE). Такая сложная схема реализуется автоматически с помощью команд UI (высокоуровневый метод доступа), но возможно и выполнение каждого шага отдельно (низкоуровневый метод доступа). В пользовательских интерфейсах с графической оболочкой операции по работе с файлами выглядят не намного сложнее работы с файлами в пределах одного компьютера. Файловый каталог LFC позволяет закреплять за файлом дополнительные права доступа с помощью списков управления доступом (Access Control List, ACL). Списки ACL состоят из базовой и расширенной частей. Базовая часть полностью аналогична стандартным правам UNIX (пользователь, группа, другие и т.д.), расширенная часть ACL позволяет задать права для дополнительных пользователей и групп. У каталога, помимо собственного списка ACL, есть также так называемый ACL по умолчанию (default). Новые файлы и каталоги автоматически получают default ACL родительского каталога.

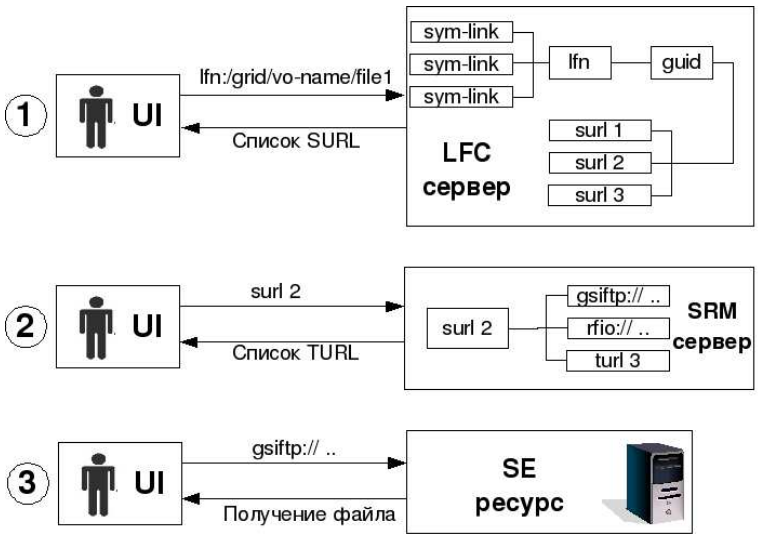


Рис. 4. Процедура получения файла из grid-системы (с целью упрощения, DPM сервер на рисунке не показан)

Другим сервисом, относящимся к хранению данных, является grid-сервис для передачи данных (File Transfer Service, FTS). С помощью этого сервиса пользователь может запускать задания по расписанию на репликацию файлов между сайтами или другие задачи. При этом служба FTS автоматически выполняет взаимодействие с SRM-серверами сайтов по протоколу gridFTP.

2.3. Интерфейс прикладного программирования (API)

Работа с grid-системой, управляемой программным обеспечением gLite, возможна не только с помощью стандартных пользовательских интерфейсов. Программный комплекс gLite поддерживает

работу с интерфейсом прикладного программирования (Application Programming Interfaces, API) для доступа к ресурсам grid-системы как для работы с распределенными вычислениями, так и с распределенным хранилищем. Это позволяет программистам создавать программы, непосредственно работающие с grid-ресурсами как изнутри grid-системы, так и извне. В качестве базовых API-функций используется библиотека Grid File Access Library (GFAL). Библиотека позволяет реализовать удалённую работу с файлами и ее компоненты могут быть включены в программы на языке C/C++ и Java. Состав и структура API приведена на рис.5.

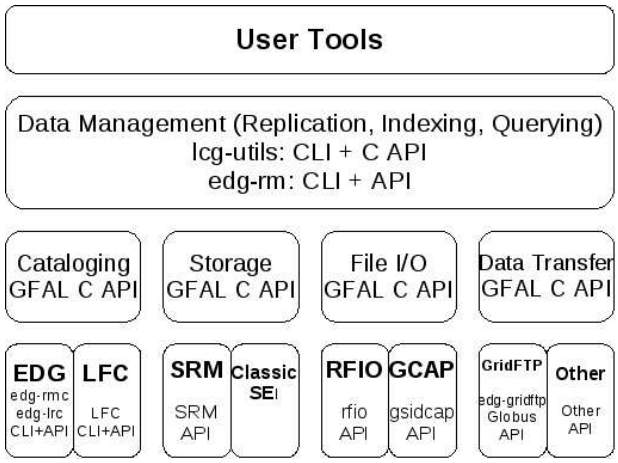


Рис. 5. Структура API

ЗАКЛЮЧЕНИЕ

Современные достижения в области grid-технологий опираются на развитие промежуточного программного обеспечения (ППО). Проект EGEE, реализуя двухфазный подход, использует ППО своего предшественника – проекта EDG (European Data Grid). В EGEE выполнена модернизация большей части исходных компонент и был создан новый продукт – gLite, который может устанавливаться в grid-инфраструктуру, обеспечивая предпроизводственный сервис. Пакет gLite является законченным решением для grid-систем и включает как базовые низкоуровневые программы, так и службы высокого уровня. Программный комплекс gLite обеспечивает защищенный и прозрачный доступ к распределенному хранилищу данных на основе распределенной файловой системы. Он соответствует требованиям SOA (Service Oriented Architecture). Поэтому при необходимости его можно связать с другими grid-службами и обеспечить реализацию стандартов GRID (Web Service Resource Framework – WSRF, OASIS, Open Grid Service Architecture – OGSA). Он является одним из лучших базово-инструментальных средств, совместимых с планировщиками PBS, Condor и LSF. Программный комплекс gLite разработан с учетом свойств интероперабельности и содержит базовые службы, облегчающие построение grid-приложений для любых прикладных областей.

Разработка и установка gLite поддерживается программой EGEE по распределенной т-инфраструктуре (тренировочной инфраструктуре). Эта программа предоставляет по Internet онлайновую документацию, учебные фильмы, организует дистанционные семинары. Обучение можно пройти и на специальном тестовом стенде GILDA, который имеет собственный сертификационный центр (CA). На стенде пользователи и системные администраторы могут проверить все аспекты развертывания и эксплуатации gLite.

Библиографический список

1. EGEE — Enabling Grids for E-sciencE. http://eu-egee.org/

2. gLite — Lightweight Middleware for Grid Computing. http://cern.ch/glite/

3. Worldwide LHC Computing Grid. http://cern.ch/LCG/

4. The DataGrid Project. http://www.edg.org/

5. DataTAG — Research & Technological Development for a Data TransAtlantic Grid. http://cern.ch/datatag/

6. The Globus Alliance. http://www.globus.org/

7. GriPhyN — Grid Physics Network. http://www.griphyn.org/

8. iVDgL — International Virtual Data Grid Laboratory. http://www.ivdgl.org/

9. Стандарт X.509: http://www.itu.int/rec/T-REC-X.509-200508-I.

10. gLite 3.1 User's guide. https://edms.cern.ch/file/722398/gLite-3-UserGuide.pdf

11. Стандарт UUID. http://www.ietf.org/rfc/rfc4122.txt