

Программный пакет



*Дмитрий Заборов
(ИТЭФ)*

Дубна, 2006

- Вычислительный GRID – форма распределенной вычислительной сети, предполагающая совместное координированное использование различными географически распределенными организациями вычислительных мощностей, программ, данных и систем их хранения, а также сетевых ресурсов.

- ... перед простым набором независимых вычислительных кластеров
 - ✓ Автоматическое распределение нагрузки между кластерами
 - ✓ Доступ к системам хранения данных не зависит от географического положения за исключением потребления сетевых ресурсов
 - ✓ Унифицированная среда приложений и др.

- ... перед одним большим "супер-кластером"
 - ✓ Большая гибкость в плане организации и финансирования
 - ✓ Отсутствие необходимости выбора места строительства

- “Lightweight Middleware for Grid Computing”
- Программный пакет для организации географически распределенных вычислительных сетей
- Позволяет объединять в единую систему:
 - вычислительные кластеры
 - системы хранения данных
- www.glite.org

- gLite – программное обеспечение GRID, разрабатываемое в рамках проекта EGEE (www.eu-egee.org), финансируемого Евросоюзом
- В пакете gLite использованы наработки, полученные в проектах
 - LCG (LHC Computing GRID)
 - EDG (European DataGrid)
 - Globus
 - и др.

- Condor (входит в gLite WMS)
- Системы хранения данных CASTOR, dCache, и др.
- Базы данных Oracle, MySQL
- Языки XML, WSDL
- Axis, Tomcat (для Web-сервисов)
- и прочие

- Многие сетевые протоколы gLite определены с помощью языка WSDL (Web Service Definition Language)
 - WSDL – XML формат для описания сетевых сервисов как набора конечных точек (endpoints), оперирующих над сообщениями
 - Операции и сообщения описываются абстрактно и могут быть физически реализованы с использованием различных протоколов (SOAP, HTTP, и др.)
 - Существует взаимно однозначное соответствие между WSDL описанием сервиса и java-описанием его интерфейса
 - <http://www.w3.org/TR/wsdl>

- gLite 3.0 в большей степени основан на LCG 2.7, чем на gLite 1.5 (предыдущая версия gLite)
- Некоторые разработанные в рамках EGEE компоненты, в том числе gLite FireMan Catalog, не вошли в релиз gLite 3.0.
- Некоторые другие разработанные в EGEE компоненты, в том числе информационная система R-GMA, напротив, уже ранее были включены в пакет LCG

- gLite 3 в настоящее время является стандартом для WLCG/EGEE GRID (старое название - LHC Computing GRID)
- Переход с пакета LCG на gLite официально состоялся в мае 2006

Job management Services

- Workload Management
- Computing Element
- Logging and Bookkeeping

Data management Services

- File and Replica catalog
- File Transfer and Placement Services
- gLite I/O

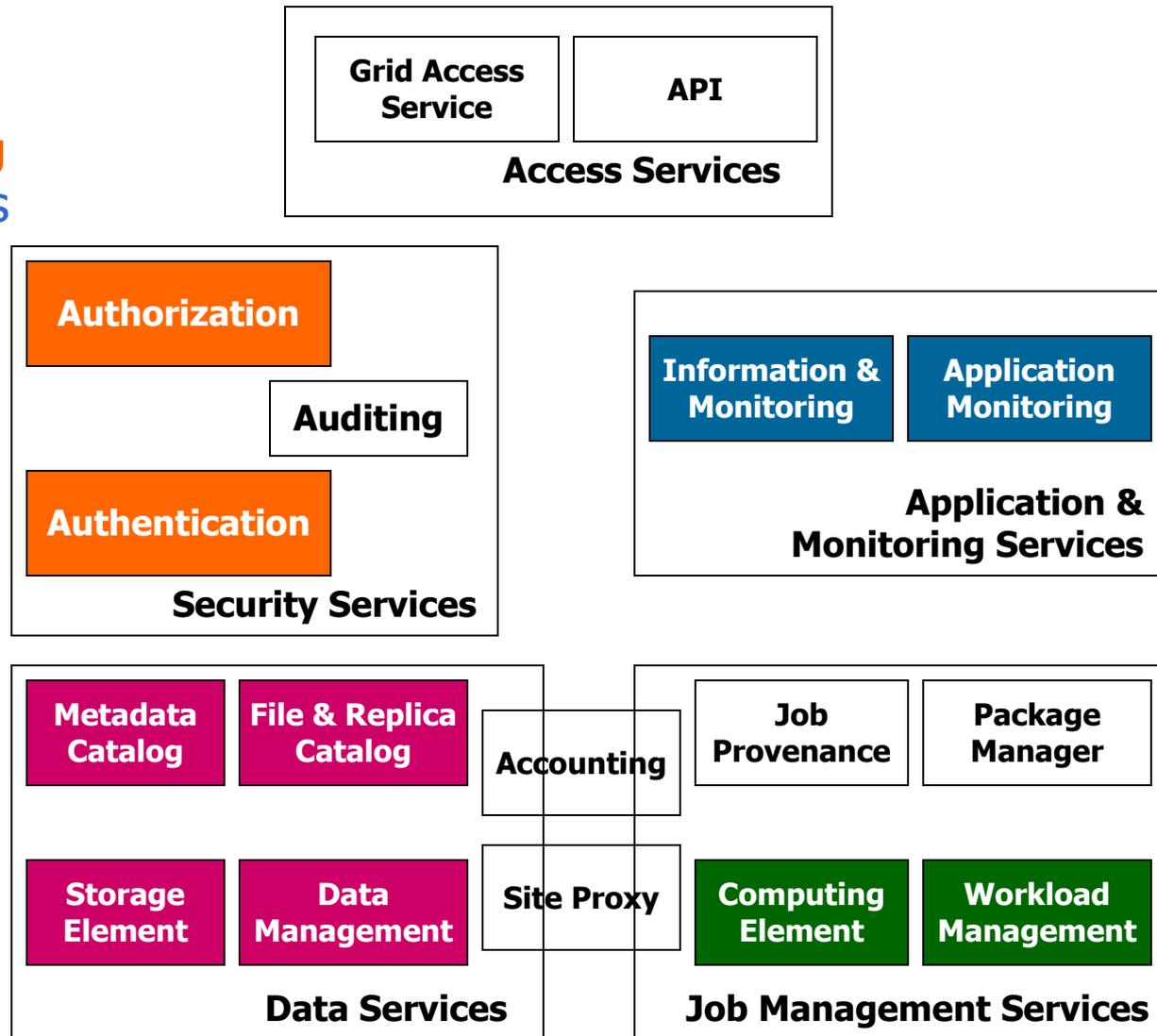
Information Services

- R-GMA
- Service Discovery

Security

Deployment Modules

- Distribution available as RPM's, Binary Tarballs, Source Tarballs, APT cache



Job management Services

- LCG WMS
- LCG Computing Element
- LCG L&B Service
- gLite WMS
- gLite Computing Element
- gLite L&B Service

Data management Services

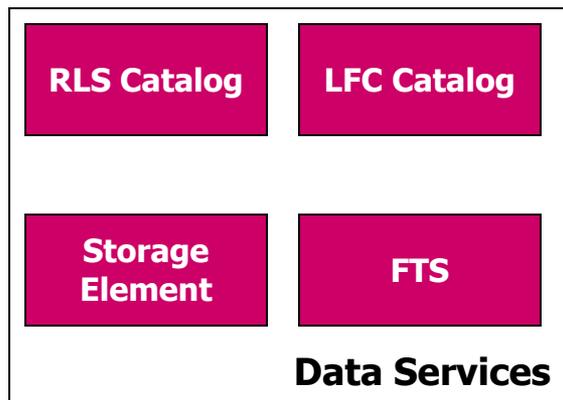
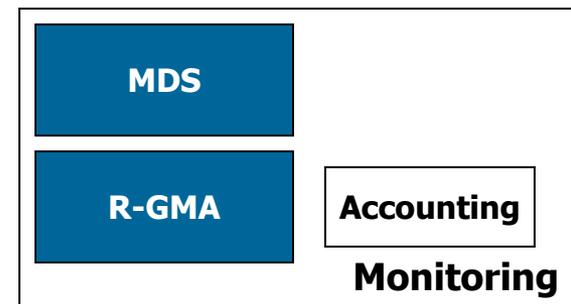
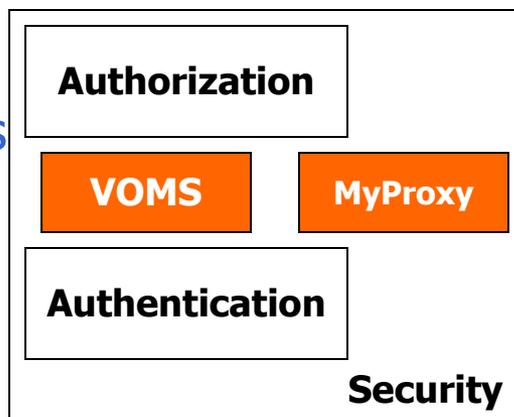
- Storage Element
- "old" RLS catalog
- "new" LFC catalog
- File Transfer Service

2 Information Systems

- Globus MDS
- gLite R-GMA

Security Services

- VOMS Server
- MyProxy Server



- Инфраструктура безопасности gLite 3.0 в целом соответствует принятой во многих GRID-проектах т.н. GRID Security Infrastructure (GSI)
- Новизна связана с использованием VOMS-сервиса (Virtual Organisation Membership Service), хранящего информацию о принадлежности пользователей к определенным группам и виртуальным организациям и о роли в них

- Аутентификация = опознавание пользователя
- Аутентификация в GRID основана на использовании асимметричной криптографии и цифровых сертификатов стандарта X.509 v3
- Обычно аутентификация осуществляется при установке соединения (с сервером)
- В качестве имени пользователя используется текстовая строка определенного формата (Distinguished Name, DN)
- Максимальная длина ключей в gLite 3.0 ограничена 2048 битами (ограничение Java 1.4.x)

- Криптография с публичным ключом, или асимметричное шифрование – форма криптографии, позволяющая обмениваться зашифрованными сообщениями не имея доступа к секретному ключу корреспондента
- Реализуется асимметричная криптография с помощью пары математически связанных друг с другом ключей: приватного и публичного
- Для расшифровки данных, зашифрованных публичным ключом, необходим соответствующий приватный ключ
- Асимметричное шифрование может использоваться для создания цифровых подписей и в процедурах аутентификации

- Сертификат является «цифровым паспортом» пользователя (или ресурса) GRID
- Сертификаты выдаются специальными органами (“Certification Authority”)
- При создании запроса на получение сертификата пользователь генерирует пару ключей
- Физически сертификат представляет собой файл, содержащий информацию о пользователе, его публичный ключ, период действия сертификата и цифровую подпись выдавшей его Certification Authority
 - Цифровую подпись невозможно подделать простым копированием, т.к. процедура «подписывания» включает в себя шифрование тела сертификата с использованием приватного ключа СА.
 - С другой стороны, для проверки подписи достаточно иметь копию сертификата СА (которая содержит ее публичный ключ)

Certificate:

Data:

Version: 3 (0x2)

Serial Number: 1365 (0x555)

Signature Algorithm: sha1WithRSAEncryption

Issuer: C=CH, O=CERN, OU=GRID, CN=CERN CA

Validity

Not Before: Jul 13 15:16:30 2005 GMT

Not After : Jul 13 15:16:30 2006 GMT

Subject: C=CH, O=CERN, OU=GRID, CN=Dmitry Zaborov 9003

Subject Public Key Info:

Public Key Algorithm: rsaEncryption

RSA Public Key: (1024 bit)

Modulus (1024 bit):

00:a9:60:44:9b:ae:91:13:1b:6b:40:bf:ad:11:66:
7b:65:09:f4:e4:a1:c1:56:05:8c:80:2e:44:7c:a9:
81:df:33:89:ac:a5:08:43:fc:88:91:71:07:f2:13:
8e:49:6f:56:5c:75:56:91:4f:c0:f4:f8:f9:34:0c:
20:cd:3a:14:f9:05:a8:e6:f7:d9:91:94:40:11:5b:
4c:d4:b9:10:f3:07:d3:8c:4e:5f:eb:6a:64:21:4c:
4f:69:85:86:21:7a:ea:0b:f8:8c:81:73:9f:84:b3:
db:32:4e:dc:32:c4:4e:37:4a:d7:24:4c:28:8c:a3:
ad:97:37:37:17:a3:b0:af:d7

Exponent: 65537 (0x10001)

X509v3 extensions:

...

- Основной сертификат пользователя (точнее его приватный ключ) для повышения безопасности защищен паролем
- При работе в GRID пользователь создает Proxy-сертификат
 - Proxy-сертификат не защищен паролем, что позволяет использовать его для делегирования полномочий пользователя запускаемым от его имени процессам
 - Proxy-сертификат имеет сравнительно короткий период действия (по умолчанию 12 часов)
 - При создании proxy-сертификата (команды типа `grid-proxy-init` или `voms-proxy-init`) необходимы основной сертификат пользователя и его приватный ключ (для подписи)
 - Обладание proxy-сертификатом дает права его обладателя (на период его действия)

- Иногда периода действия проху-сертификата может оказаться недостаточно (например при запуске очень «долгих» 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
- MyProxy используется при копировании файлов с помощью FTS-сервера

- Авторизация = проверка наличия прав на совершение запрошенной операции
- В gLite 3.0 существуют два механизма авторизации:
 - `grid-mapfile` механизм
 - Использование VOMS сервиса (Virtual Organisation Membership Service)
- Права пользователя могут устанавливаться в зависимости от его принадлежности к той или иной виртуальной организации, группе и роли (с помощью VOMS)

- При выполнении задач на рабочих нодах производится “отображение” DN пользователя на локальный User ID, а его группы на локальный Group ID. Эти id используются также при доступе к файлам по незащищенной версии протокола rfiо (если таковая поддерживается близлежащим SE).
- В подсистемах каталогов и управления данными вместо DN могут фактически использоваться так называемые виртуальные id

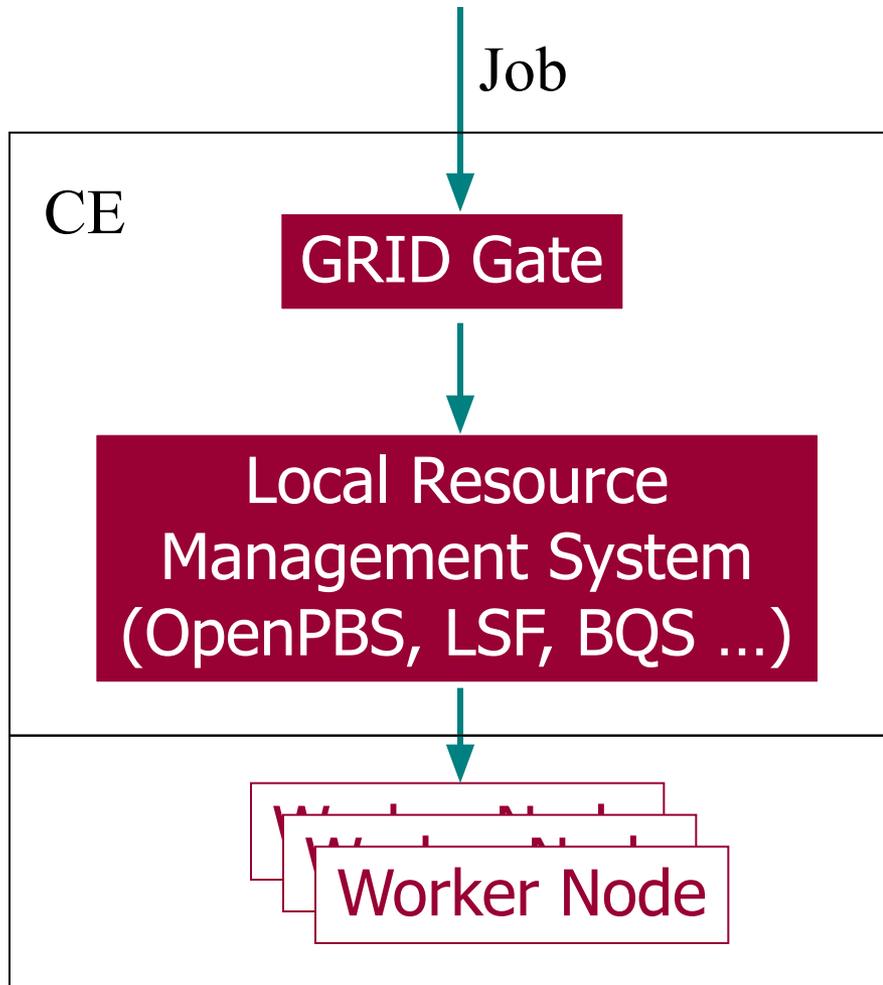
- User Interface (UI) – компьютер, служащий в качестве точки доступа к GRID
- Чтобы сделать из компьютера UI достаточно установить и настроить соответствующее программное обеспечение
 - Предусмотрен вариант установки в пользовательскую директорию, не требующий прав администратора
- UI позволяет:
 - Запускать и отменять «задачи»
 - Получать информацию о ресурсах GRID, статусе запущенных задач, их «истории» и т.п.
 - Копировать, реплицировать и удалять файлы из GRID

- Working Node (WN, Рабочий Нод) – компьютер, служащий для выполнения вычислений (задач) в GRID
- Рабочие ноды объединяются в кластеры с помощью одной из batch-систем (таких как LSF или PBS), образуя единый вычислительный элемент (Computing Element, CE)
- По составу установленного клиентского программного обеспечения GRID WN весьма близок к UI

- В распоряжении пользователя имеются все необходимые для работы в GRID команды, которые можно запускать как с UI из командной строки (Command Line Interface, CLI), так и использовать в скриптах
- Также существуют различные инструменты, обладающие графическим интерфейсом или Web-интерфейсом
- Кроме того, взаимодействие с GRID может осуществляться посредством различных API (программных интерфейсов приложений), т.е. «изнутри» программы пользователя

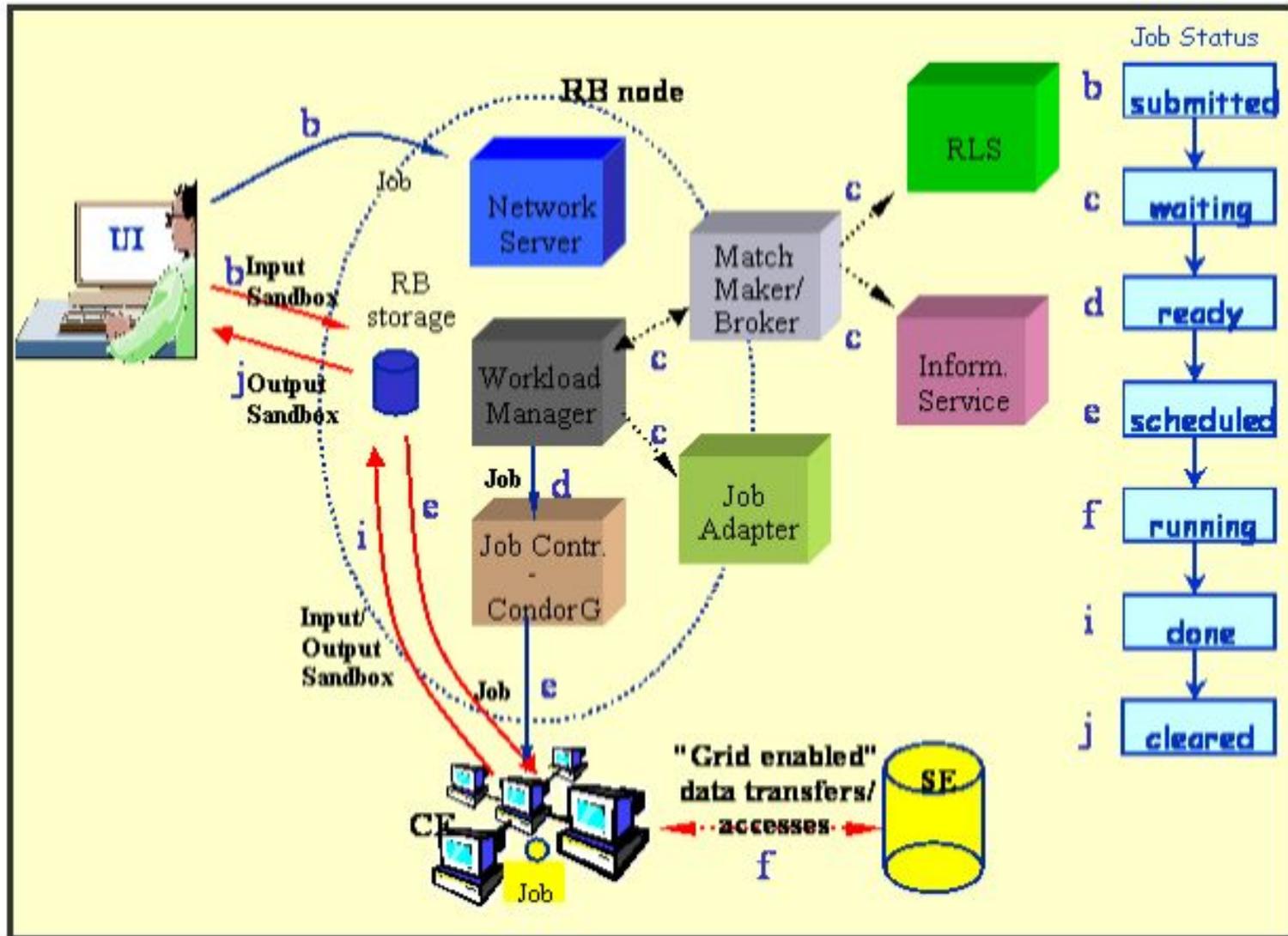
- Задача (англ. Job) – совокупность действий по обработке и передаче данных, рассматриваемая в комплексе в качестве единицы вычислительной работы
- С точки зрения пользователя запуск задачи представляет собой запуск программы с помощью соответствующих средств GRID
- В настоящее время официально поддерживается лишь единственный тип задач – простая пакетная задача (batch job)
- Предусмотрены средства для запуска интерактивных, “checkpointable-” и MPI-задач

- Задача должна быть описана на языке описания задач JDL
- JDL файл служит аргументом команды запуска задачи
- Язык JDL, используемый в LCG/gLite GRID, основан на языке ClassAd (Classified Advertisement), разработанном проектом Condor
- Синтаксис языка JDL представляют собой выражения типа
`attribute = value;`
- Язык чувствителен к количеству пробелов и символам табуляции



- CE – совокупность вычислительных ресурсов, локализованных в одном месте (“сайте”)
- В gLite 3.0 существуют два типа CE: LCG CE и gLite CE
 - В LCG CE роль GRID Gate выполняет Globus Gatekeeper
 - gLite GRID Gate происходит из CondorC
- В информационной системе gLite каждый CE фактически соответствует одной очереди LRMS

- WMS - Система Управления Рабочей Нагрузкой - отвечает за:
- Принятие запросов на постановку задач
- Выбор СЕ, наилучшим образом подходящего для выполнения задачи
- Передачу файлов, составляющих т.н. Input и Output Sandbox (небольшие файлы параметров и т.п., не файлы данных!)
- Отслеживание состояния задач
- Ведение логов

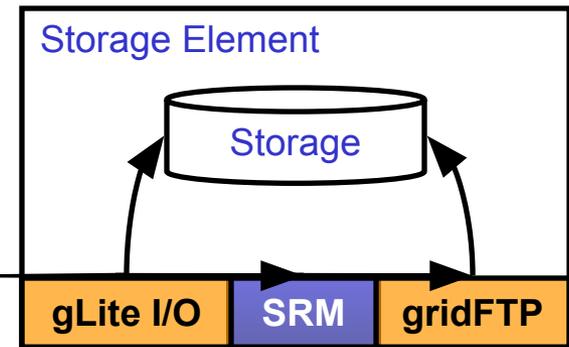


- **WMPoxy/Network Server** – интерфейс к функциональности WMS
- **Workload Manager** – главный компонент системы
- **Resource Broker** – вспомогательный компонент
- **Information SuperMarket (ISM)** – хранит информацию необходимую для сопоставления ресурсов задачам (или наоборот)
- **Job Controller & Job Adapter** – ответственны за CondorC submission file и wrapper script, передачу input/output sandbox'ов
- **CondorC** – постановка, удаление задач
- **DAGMan** – DAG Manager
- **Logging and Bookkeeping** – мониторинг задач
- **Log Monitor** – просмотр лог-файлов CondorC для активных задач и т.п.

- В gLite 3.0 две системы WMS: LCG WMS и gLite WMS
- LCG WMS может направлять задачи лишь на LCG CE, в то время как gLite WMS способна работать с обоими типами CE
- gLite WMS является более новой и более совершенной разработкой по сравнению с LCG WMS
- Преимущества gLite WMS в эффективности должны быть особенно велики при запуске массивных параллельных вычислений

Виды SE в gLite 3.0

- Classic SE
 - состоит из GridFTP сервиса и “небезопасного” RFIО сервиса, дающих доступ к единственному диску или дисковому массиву
 - Не имеет SRM (Storage Resource Manager) интерфейса
 - не рекомендуется к установке
- Mass Storage System (такие как CASTOR)
 - Доступ по GridFTP и/или rfiо
- dCache Disk pool manager
 - Помимо dCache сервера может включать несколько машин с дисковыми массивами
 - Доступ по GridFTP и/или dcap/rfiо
- LCG Disk pool manager (DPM)
 - Более легкая альтернатива dCache, разработан для LCG
 - Рекомендуется в качестве замены Classic SE



- GUID (Grid Unique ID)
 - Уникальный идентификатор файла
- LFN (Logical File Name)
 - “Human-readable” псевдоним для GUID
 - Файл может иметь несколько имен
- SURL (Storage URL)
 - URL физической копии файла (реплики)
 - Включает в себя адрес SE
 - Файл может иметь несколько реплик
- TURL (Transport URL)
 - URL, готовый для обращения по одному из транспортных протоколов (например GridFTP)
 - Является производным от SURL

- GSIFTP/GridFTP
 - Позволяет помещать файлы на SE и получать/скачивать их с SE
 - Поддерживается всеми типами SE
 - Аутентификация по сертификату
 - Из командной строки или через API
- Прямой доступ
 - secure rfiio (с аутентификацией по сертификату)
 - insecure rfiio (аутентификация в пределах локальной сети по uid)
 - dCache Access Protocol (gsidcap)
 - Различные типы SE могут поддерживать один или несколько протоколов
 - только через API (GFAL, rfiio API, ...)

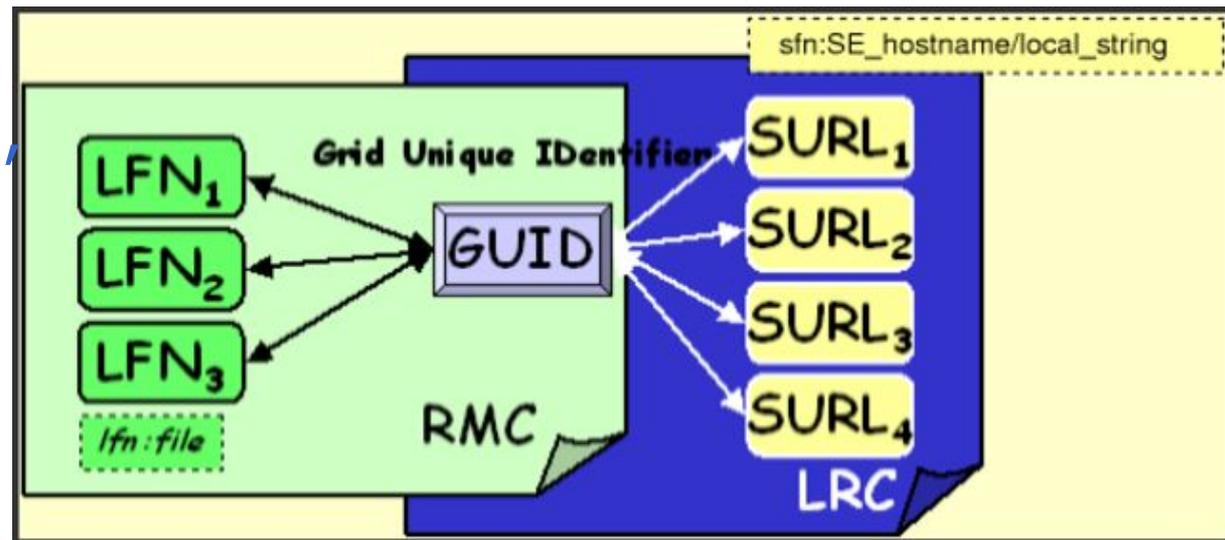
- Права доступа к файлам в WLCG/EGEE GRID определяются с помощью списков управления доступом (Access Control List, ACL)
- ACL состоит из базовой и расширенной частей
- Базовая часть полностью аналогична стандартным правам UNIX (пользователь, группа, другие)
- Расширенные ACL позволяют задать права для дополнительных пользователей и групп
- У директорий помимо собственного ACL есть также т.н. default ACL. Новые файлы и директории автоматически получают default ACL родительской директории

- Различные сайты (операционные центры) GRID могут поддерживать или не поддерживать те или иные VO
- В частности, могут существовать CE и SE, зарезервированные для конкретной виртуальной организации

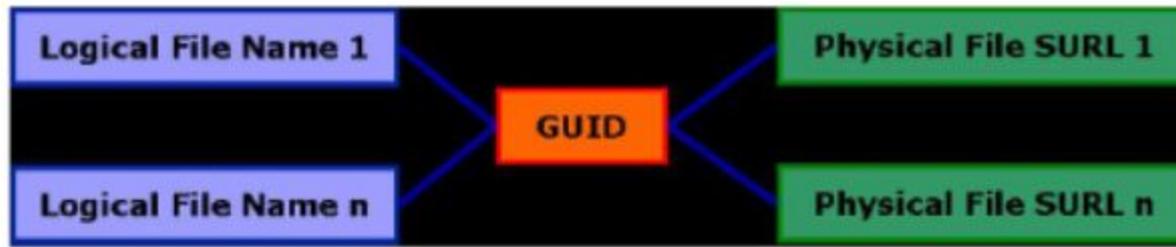
- Файлы хранящиеся на различных SE объединяются в единую “файловую систему” путем ведения их каталога
- В gLite 3.0 две службы каталогов:
- старый RLS Каталог (Replica Location Server)
- и новый LFC каталог (LCG File Catalog)
- Эти два каталога **не** синхронизованы
- Пользователь вынужден выбирать с каким из этих двух каталогов он будет работать
- Рекомендуется выбирать LFC

- RLS каталог состоит из двух сервисов:
- Local Replica Catalog (LRC)
 - Хранит список реплик GRID файла, т.е. Связывает GUID и SURL
 - Также хранит системные метаданные (размер файла и т.п.)
- Replica Metadata Catalog (RMC)
 - Хранит список имен файла, т.е. связывает GUID с LFN
 - Также позволяет хранить пользовательские метаданные

Недостаток
RLS каталога в том,
что он допускает
неавторизованный
доступ

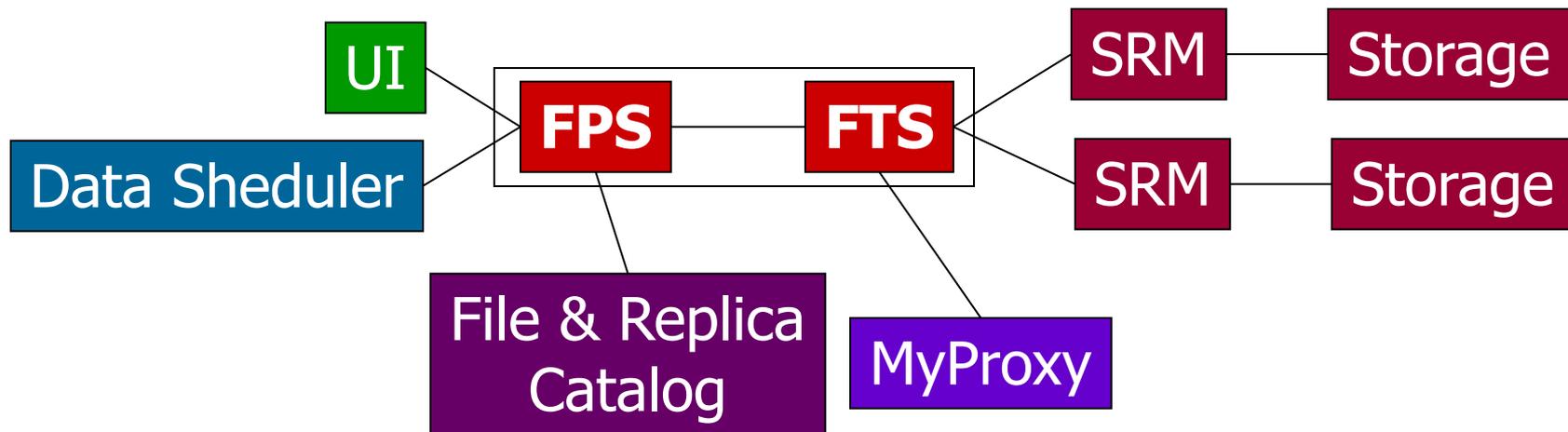


- LCG File Catalog (LFC) объединяет в себе функции файлового каталога и каталога реплик, т.е. обеспечивает связь между GUID, LFN и SURL файлов
- В LFC любой файл имеет основное имя (main LFN), а также может иметь псевдонимы (аналогично символическим ссылкам в UNIX)



- Файлы организуют древовидную структуру (по их основным LFN)
- Поддерживаются системные метаданные и краткое описание файла (одна строка)
- Безопасность обеспечивается в полном объеме

- File Transfer Service – низкоуровневый сервис, используется для копирования файлов между системами хранения данных
- File Placement Service также ответственен за регистрацию скопированных файлов в каталоге



* FPS не вошел в релиз gLite 3.0

User Tools

Data Management (Replication, Indexing, Querying)

lcg_utils: CLI + C API

edg-rm: CLI + API

Cataloging

GFAL C API

Storage

GFAL C API

File I/O

GFAL C API

Data transfer

(GFAL C API)

EDG

edg-rmc
edg-lrc
CLI + API

LFC

LFC
CLI + API

SRM

SRM
APIClassic
SE

RFIO

rfio
API

DCAP

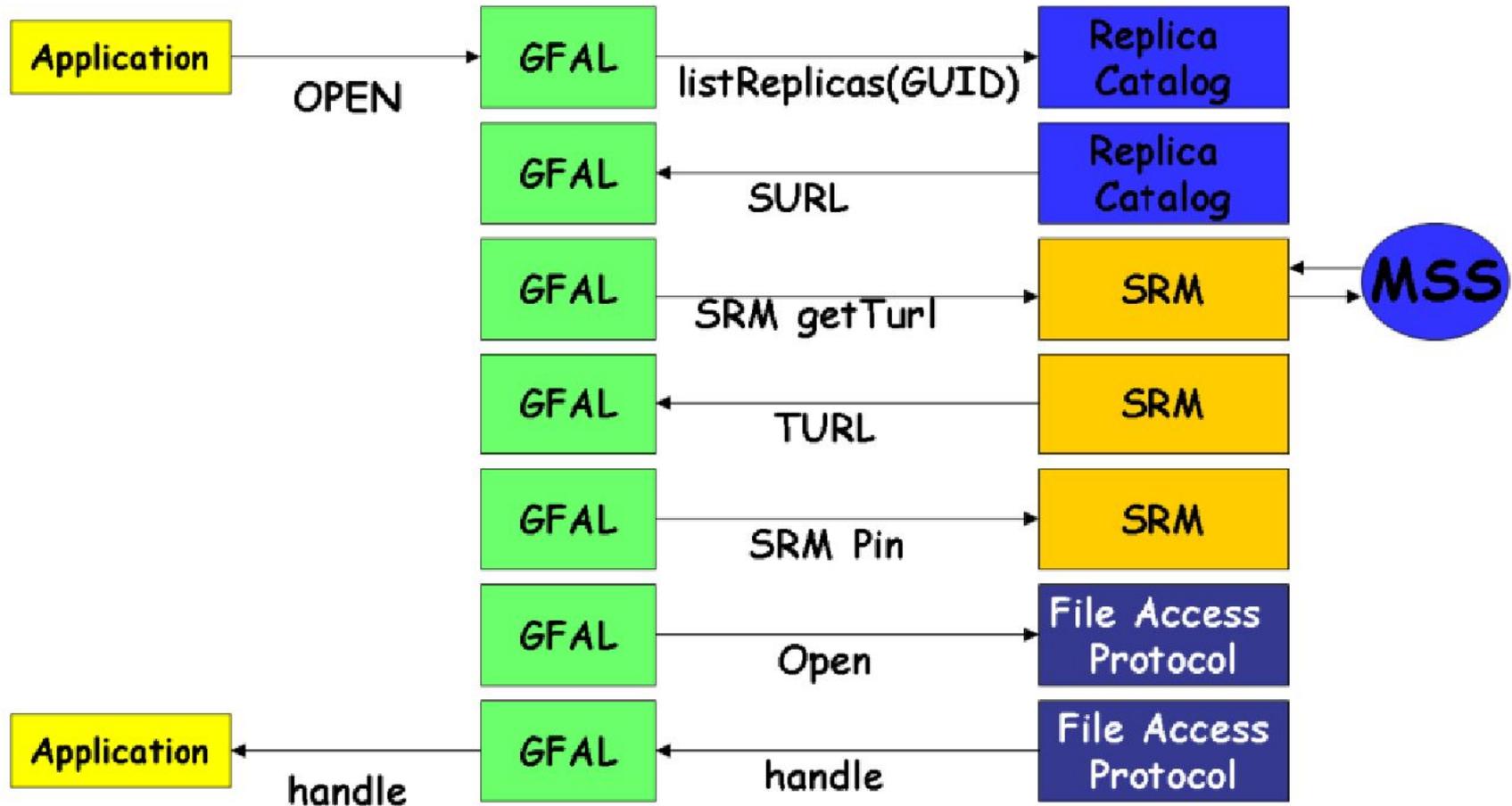
gsidcap
API

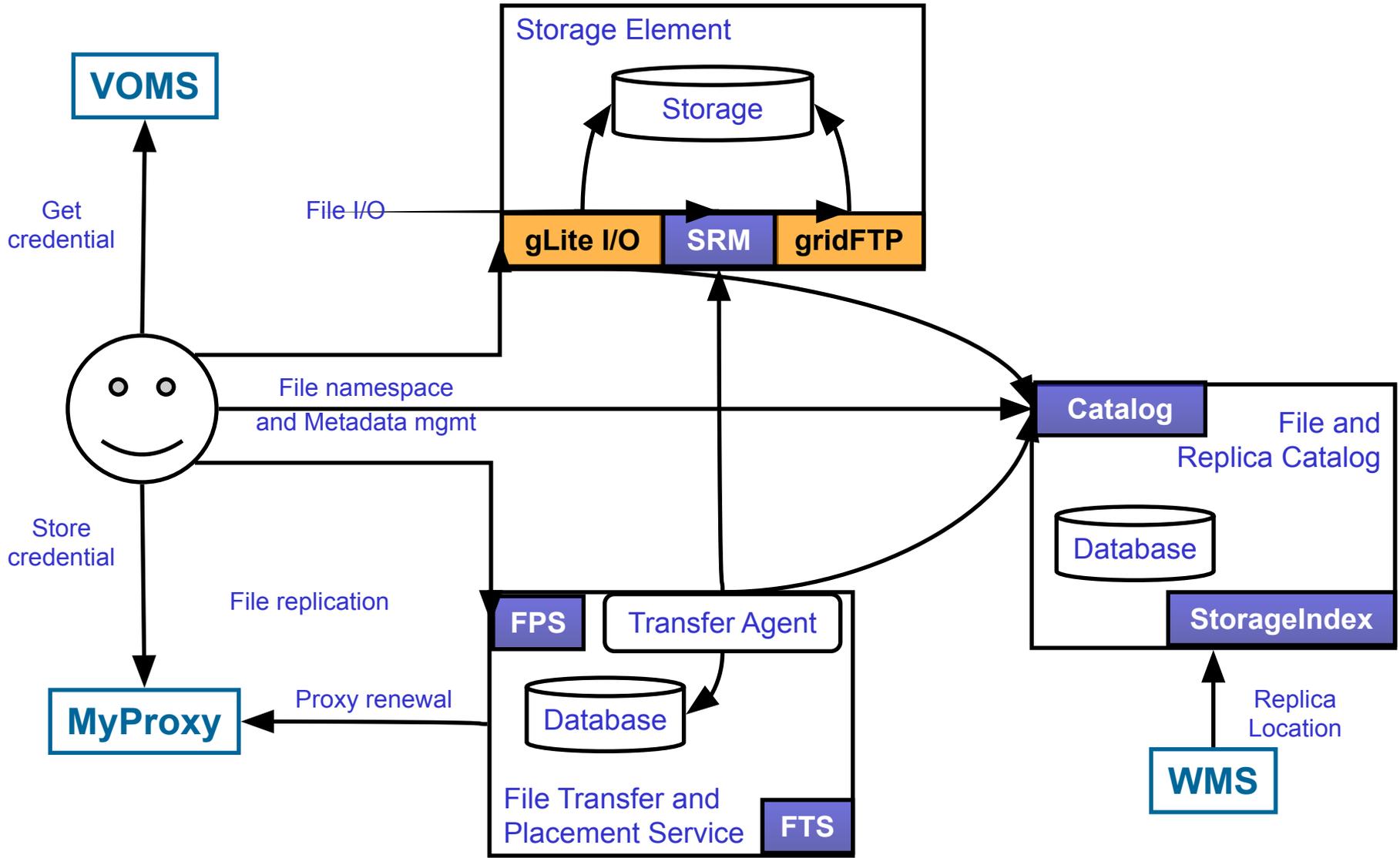
GridFTP

edg-gridtp
Globus
API

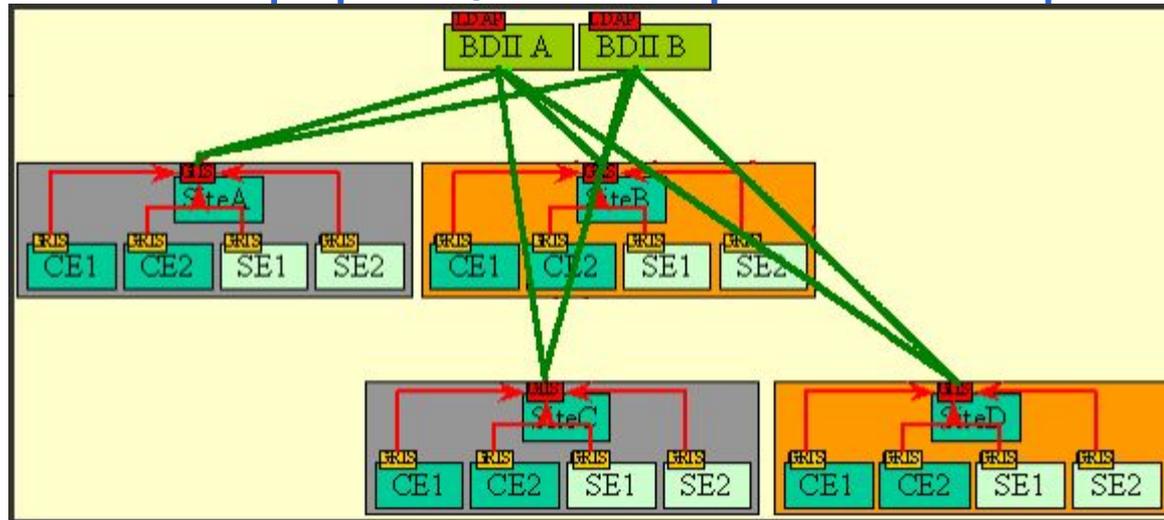
Other

Other
API



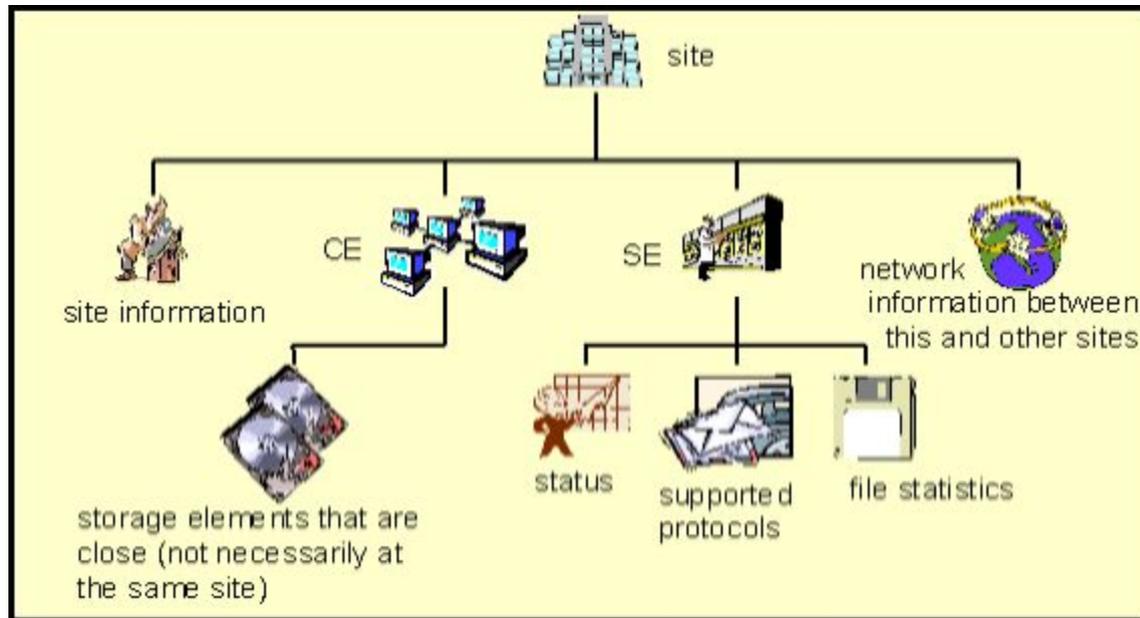


- MDS представляет собой централизованную систему, в которой вся информация собирается от краев к центру



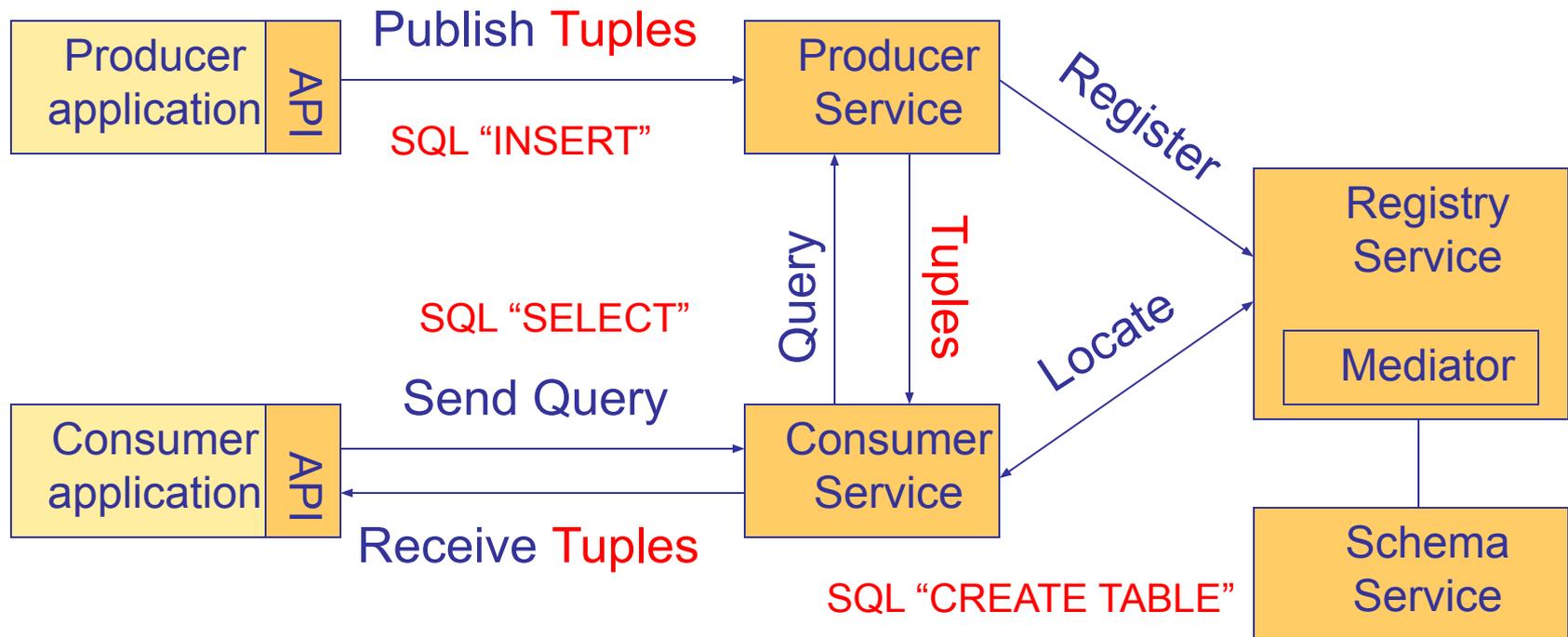
- На каждом CE и SE запущен GRIS (GRID Resource Information Server), в котором публикуется информация о ресурсе
- Информация со многих GRIS собирается другими сервисами, GIIS (GRID Index Information Server)
- Роль центральных GIIS обычно выполняет Berkley Database Information Index (BDII)

- Информация, публикуемая в MDS, организована в соответствии с определенными правилами (GLUE Schema) и образует древовидную структуру (Directory Information Tree)



- MDS основана на использовании OpenLDAP, "open source" реализации протокола LDAP – Lightweight Directory Access Protocol

- R-GMA (Relational Grid Monitoring Architecture) – реализация предложенной GGF (Global GRID Forum) архитектуры мониторинга GRID
- Основана на разделении субъектов на производители данных, потребители данных и *единый* реестр



- Хотя R-GMA и является централизованной системой, данные в ней не собираются на одном сервере, а образуют распределенную базу данных, объединяемую в одно целое Реестром
- Предусмотрена возможность дублирования Реестра (для повышения надежности)

- R-GMA обладает большей гибкостью чем MDS
- Главным недостатком MDS является то, что безопасность доступа к информации в ней не обеспечивается
- В настоящее время в WLCG/EGEE GRID для обнаружения ресурсов и публикации их статуса используется MDS, а для мониторинга, подсчета использованных ресурсов (Accounting) и публикации пользовательской информации – R-GMA. Также R-GMA дублирует информацию MDS
- В будущем должен произойти полный переход на R-GMA

- Computing Element
 - Grid Gate (Globus/Condor)
 - Condor-C (Condor)
 - Local batch system (PBS, LSF, Condor, ...)
- Workload Management
 - LCG WMS (LCG, EDG)
 - gLite WMS (EGEE)
 - Logging and bookkeeping (EDG, EGEE)
 - Condor-C (Condor)
- Storage Element
 - GridFTP (Globus)
 - File Transfer/Placement (EGEE)
 - SRM: Castor (CERN), dCache (FNAL, DESY), DPM (LCG)
- Catalog
 - RLC Catalog (EDG, LCG, Globus)
 - LFC Catalog (LCG)
- Information and Monitoring
 - MDS (Globus)
 - R-GMA (EDG, EGEE)
- Security
 - GSI (Globus)
 - VOMS (DataTAG, LCG, EDG, EGEE)
 - GSI Authentication for C and Java based (web) services (EDG)

- Новые компоненты:
- gLite WMS/LB
- gLite CE
- gLite/LCG WN
- gLite/LCG UI
- FTS
- FTA

- **Workload Management System** works in push and pull mode
- **Computing Element** moving towards a VO based scheduler guarding the jobs of the VO (reduces load on GRAM)
- Re-factored **file & replica catalogs**
- **Secure catalogs** (based on user DN; VOMS certificates being integrated)
- Scheduled **data transfers (FTS)**
- **SRM** based storage
- Information Services: **R-GMA** with improved API, **Service Discovery** and **registry replication**
- Move towards **Web Services**

- gLite 3 является логическим продолжением и развитием технологий LCG-2 и gLite 1.5
- Отличия gLite 3.0 от LCG 2.7
 - Новые подсистемы gLite WMS, LB и CE
 - Смешанные LCG/gLite UI и WN
 - Новые сервисы FTS и FTA
- gLite 3 должен стать важным шагом к полному переходу LHC Computing GRID от LCG-2 к gLite