Архитектура, протоколы, сервисы GRID

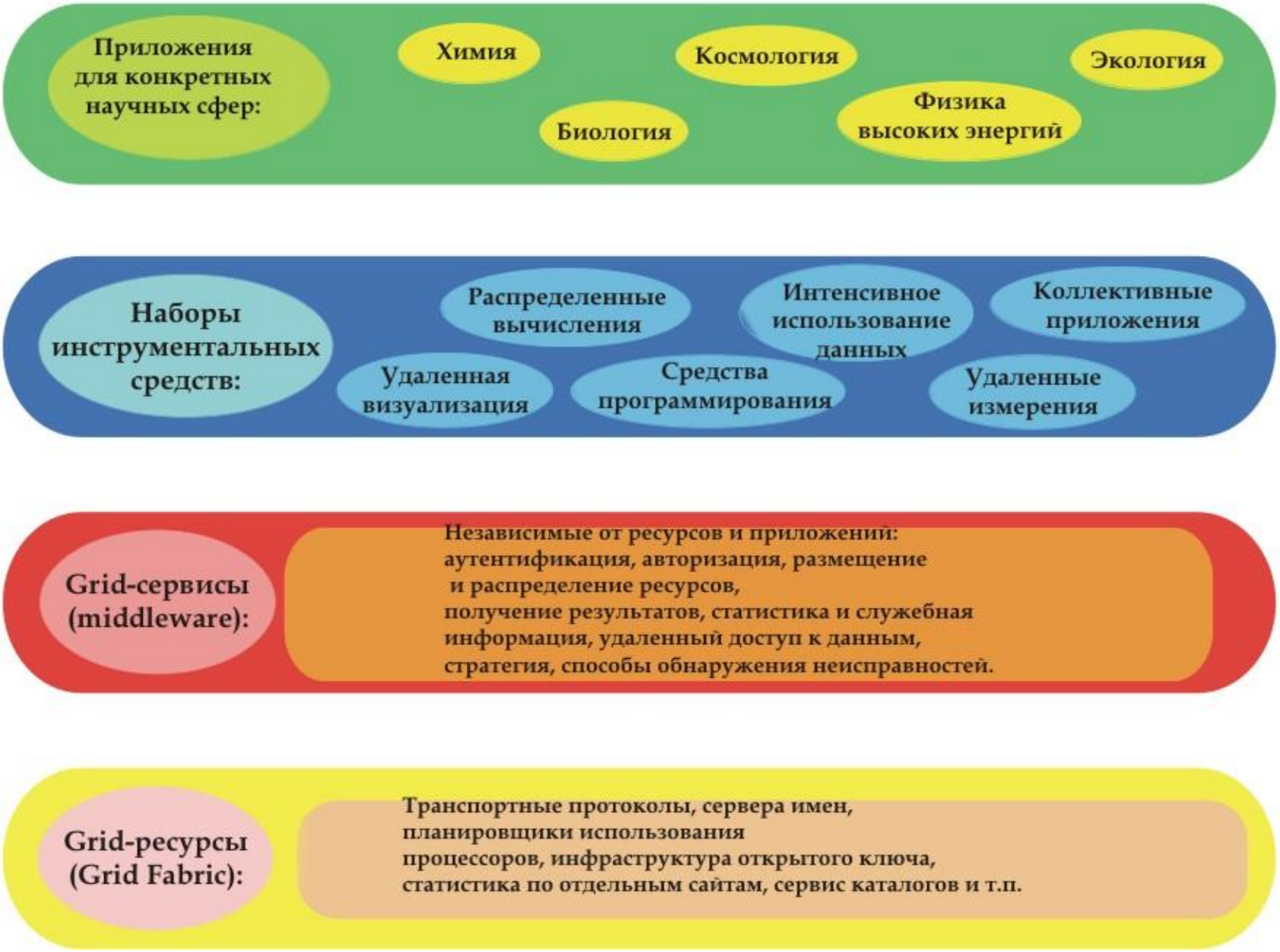
Кореньков В.В.

Некоторые Требования

- Идентификация
- Авторизация&правила
- Поиск ресурсов
- Описание ресурсов
- Резервирование ресурсов
- Распределённые алгоритмы
- Доступ к удалённым данным
- Высоко-скоростная пересылка данных
- Гарантирование производительности

- Обнаружение несанкционированного доступа
- Распределение ресурсов
- Счета и оплата
- Обнаружение неполадок
- Эволюция систем
- Мониторинг
- И т.д.
- И т.д.
- …

Grid-архитектура с точки зрения программного обеспечения



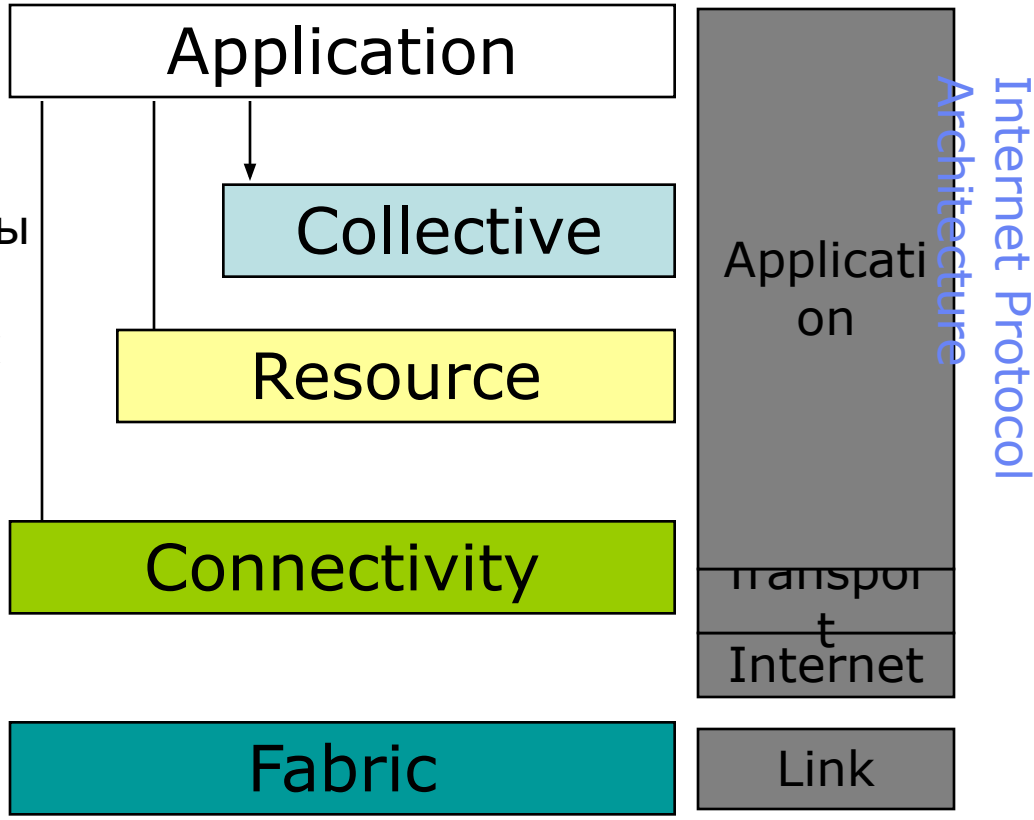
Многоуровневая Архитектура Grid (По Аналогии с Архитектутой Интернета)

"Координация многочисленных ресурсов": специфические сервисы

"Совместное использование одних ресурсов": доступ по договору, использование под контролем

"Коммуникация": коммуникация (Internet протоколы) & защищённость

"локальный контроль над ресурсами": Доступ и контроль ресурсов



Уровень связи: Протоколы & Сервис

- Коммуникация
  - Internet протоколы: IP, DNS, routing, etc.
- Защищённость: Grid Security Infrastructure (GSI)
  - Единая идентификация, авторизация и защищённая передача сообщений
  - Однократный логин, делегирование, идентификация
  - Public key technology, SSL, X.509, GSS-API
  - Инфраструктура поддержки: централизованная выдача сертификатов, управление сертификатами и ключами, …

Уровень ресурсов: Протоколы & Сервис

- Grid Resource Allocation Management (GRAM)
  - Удалённые ресурсы : выделение, резервирование, мониторинг и управление компьютерными ресурсами
  - GridFTP протокол (FTP расширения)
  - Высокоскоростной доступ к данным и пересылка
- Grid Resource Information Service (GRIS)
  - Доступ к информации
- В проекте: доступ к каталогам, доступ к библиотеке програм, Catalog access, code repository access, и т.д.
- Всё пострено на уровне: GSI & IP

Общий Уровень: Протоколы & Сервис

- Распределение ресурсов (e.g., Condor Matchmaker)
  - Поиск и выявление ресурсов
- Каталог реплик
- Сервис копирования
- Сервис по одновременному резервированию и выделению
- И т.д.

Пример:
Data Grid Архитектура

App	Приложение, специфичное для какой-то области
Collective (App)	Выбор реплики, управление заданием, виртуальный каталог данных, …
Collective (Generic)	Каталог реплик, управление репликами, выделение ресурсов, выдача сертификатов, каталоги метаданных
Resource	Доступ к данным, доступ к компьютерам, доступ к информации о сети,..
Connect	Коммуникации, поиск сервиса (DNS), идентификация, авторизация, делегация
Fabric	Системы хранения данных, кластеры, сети, ...

Основные протоколы

- Глобус (The Globus Toolkit™) основан на четырёх основных протоколах
  - Уровень связи:
    - *защищённость*: Grid Security Infrastructure (GSI)
  - Уровень ресурсов:
    - *Управление ресурсами*: Grid Resource Allocation Management (GRAM)
    - *Информационный сервис*: Grid Resource Information Protocol (GRIP)
    - *Пересылка данных*: Grid File Transfer Protocol (GridFTP)
- Также основные протоколы ‘общего’ уровня
  - Информационный сервис, управление репликами, и т.д.

Grid Security Infrastructure (GSI)

- Глобус использует протоколы и APIs GSI для создания защищённости
- GSI протоколы расширяют стандартные протоколы public key
  - Стандарты: X.509 & SSL/TLS
  - Расширения: X.509 Proxy Certificates & Delegation
- GSI расширяет стандартное GSS-API

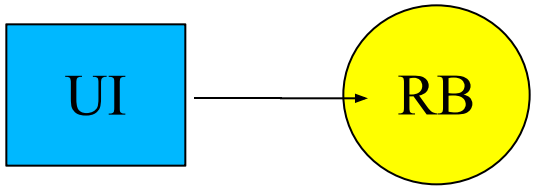
Управление ресурсами

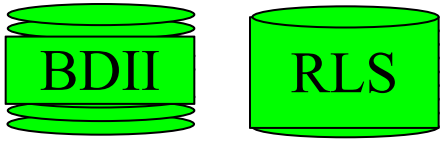
- The Grid Resource Allocation Management (GRAM) протокол и API позволяет запуск программ на удалённых компьютерах, управление этими программами – несмотря на локальные особенности и неоднородность
- Resource Specification Language (RSL) используется для передачи информации/требований на удалённый ресурс
- Многоуровневая архитектура позволяет конкретным приложениям специфицировать требования выделения ресурсов в терминах GRAM
  - Используется в Кондоре, PBS, MPICH-G2, …

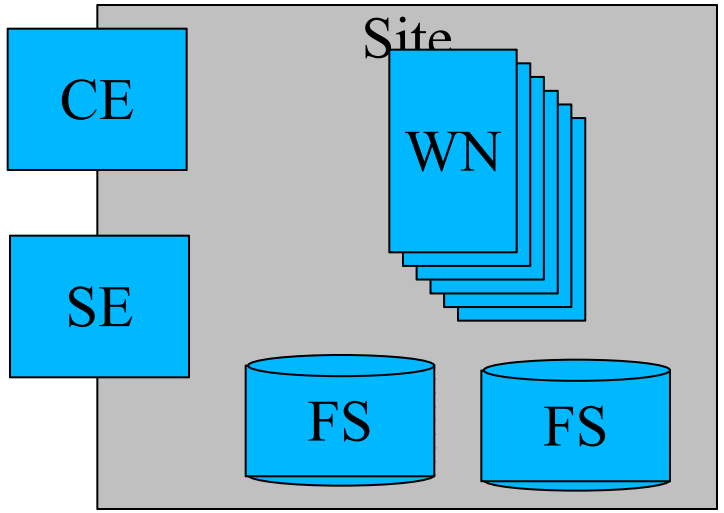
Доступ к данным и их пересылка

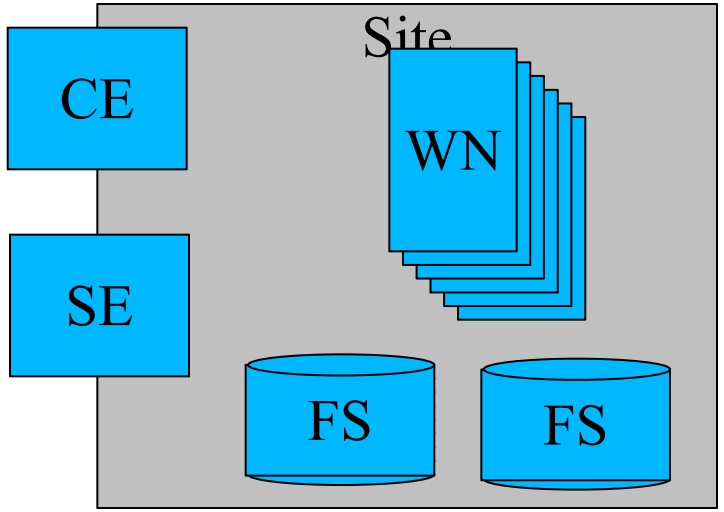
- GridFTP: расширенная версия популярного FTP протокола для доступа к данным на Grid
- Надёжный, эффективный, гибкий, параллельный, одновременный, и т.д.:
  - Пересылка данных третьими лицами, пересылка неполных файлов
  - Параллельность, striping (e.g., на параллельных файловых системах PVFS)
  - Надёжная, возобновляемая пересылка данных
- Соответствующее воплощение
  - Существующие клиенты и серверы: wuftpd, ncftp
  - Гибкие, расширяемые библиотеки в Глобусе (Globus Toolkit)

Запуск заданий в грид



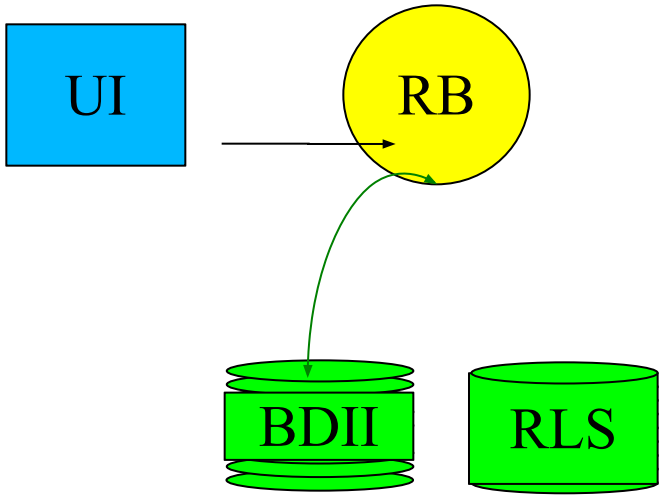


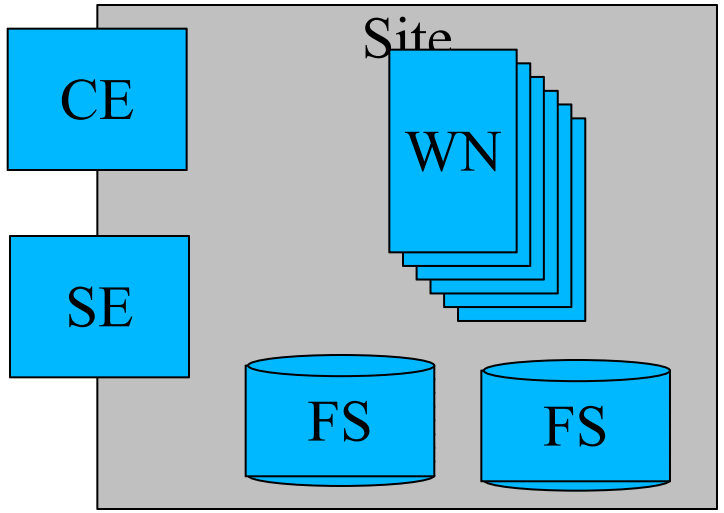


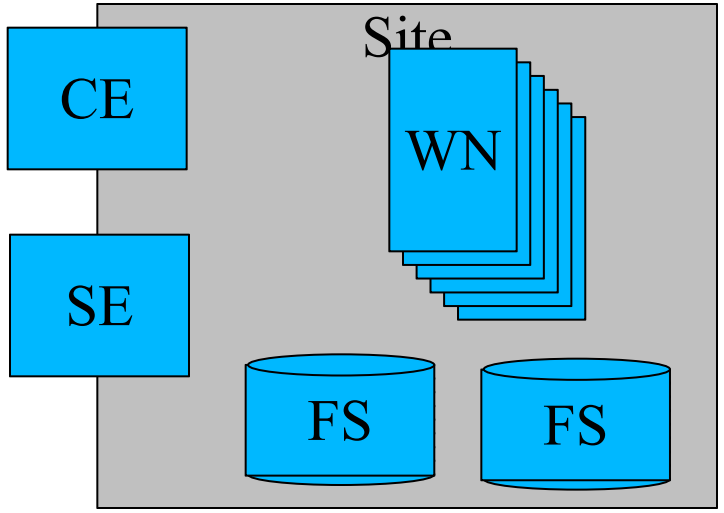


UI – Интерфейс пользователя
RB – Брокер ресурсов
BDII – Информационная база данных по ресурсам
RLS – Сервер реплик файлов
CE – Компьютерный элемент
SE – Элемент хранения данных
WN – рабочая нода
FS – файловый сервер
MyProxy – сервер продление действия сертификата пользователя

Запуск заданий в грид

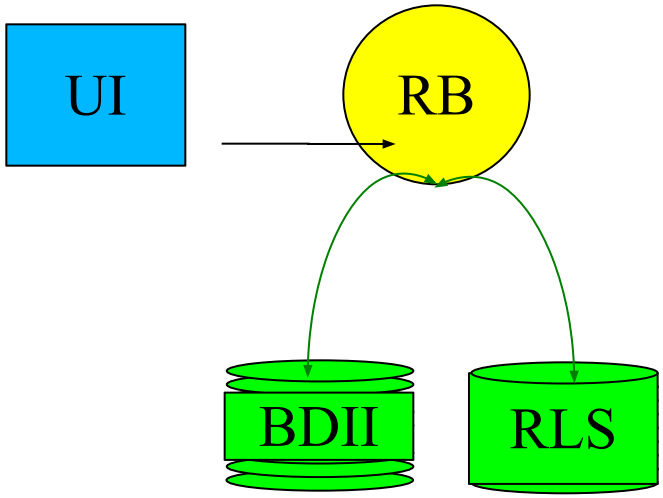


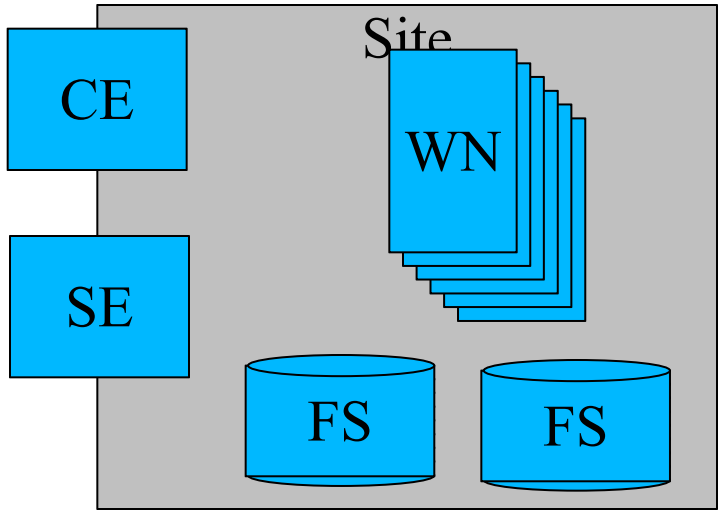


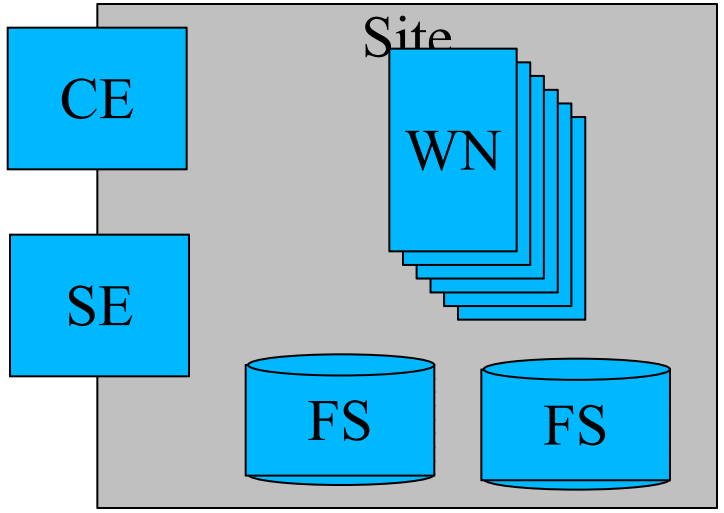


UI – Интерфейс пользователя
RB – Брокер ресурсов
BDII – Информационная база данных по ресурсам
RLS – Сервер реплик файлов
CE – Компьютерный элемент
SE – Элемент хранения данных
WN – рабочая нода
FS – файловый сервер

Запуск заданий в грид

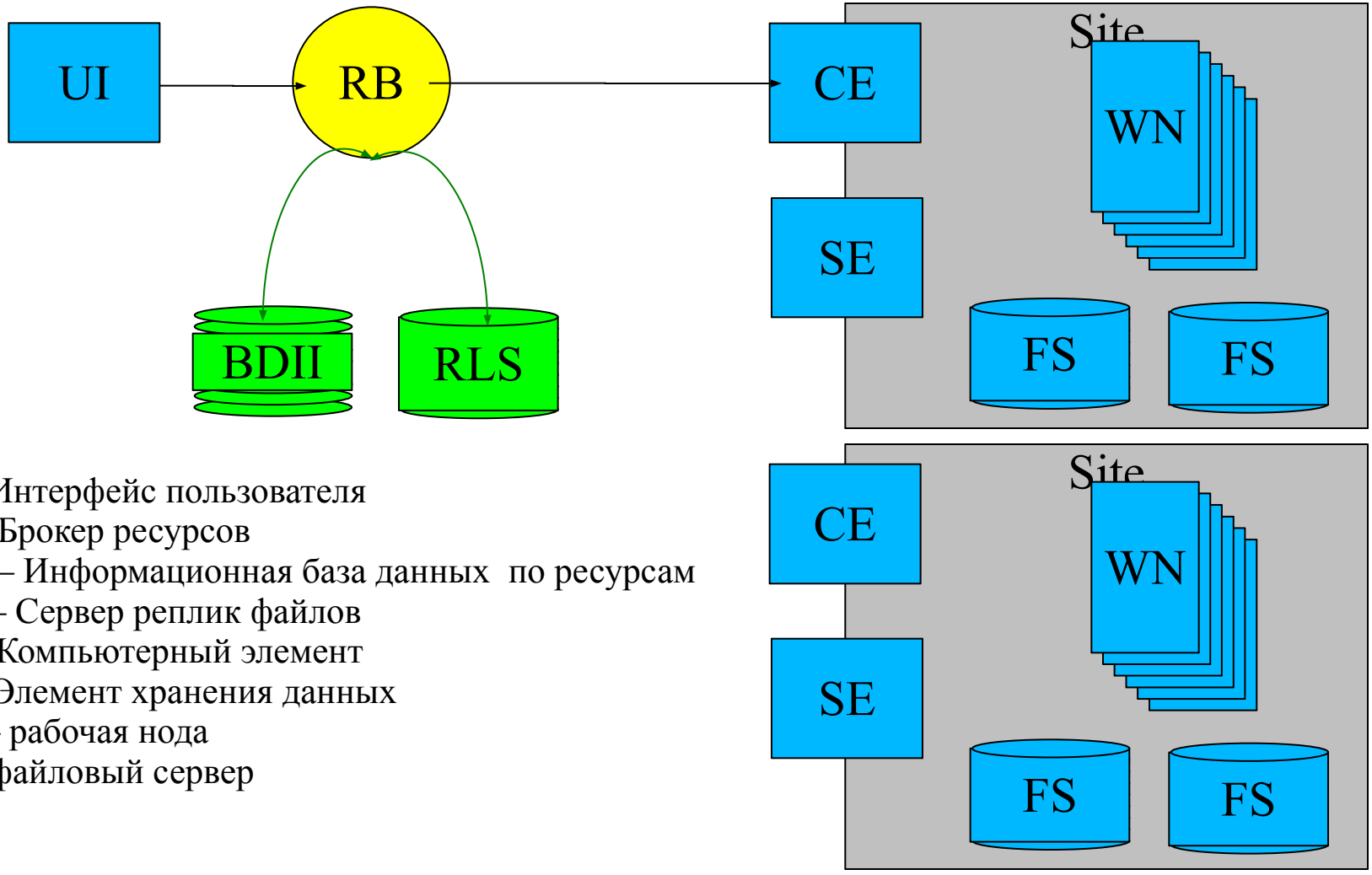






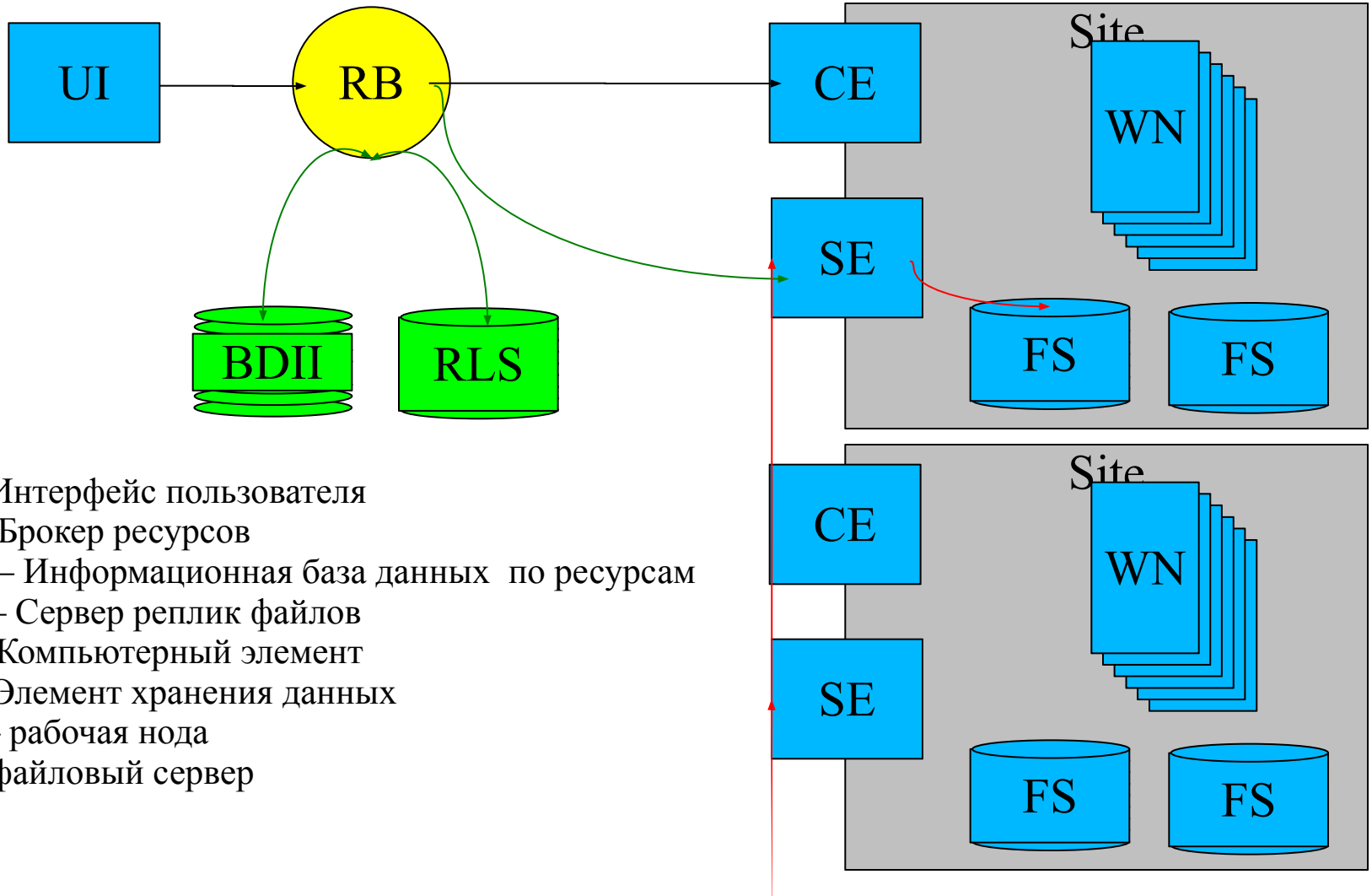
UI – Интерфейс пользователя
RB – Брокер ресурсов
BDII – Информационная база данных по ресурсам
RLS – Сервер реплик файлов
CE – Компьютерный элемент
SE – Элемент хранения данных
WN – рабочая нода
FS – файловый сервер

Запуск заданий в грид



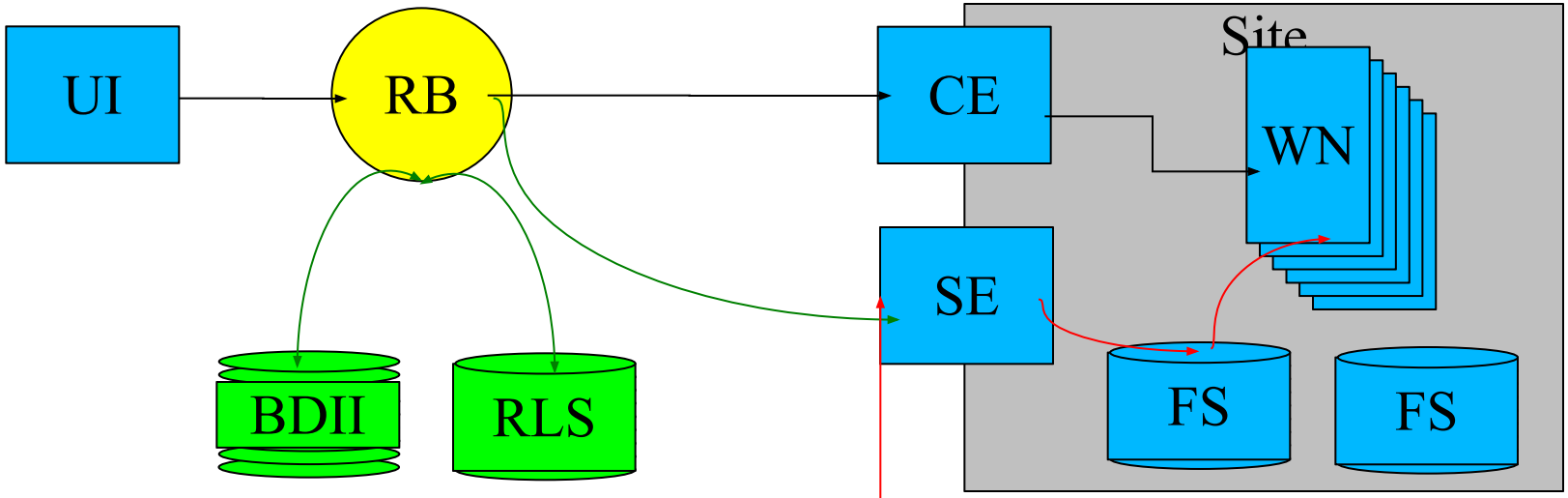
UI – Интерфейс пользователя
RB – Брокер ресурсов
BDII – Информационная база данных по ресурсам
RLS – Сервер реплик файлов
CE – Компьютерный элемент
SE – Элемент хранения данных
WN – рабочая нода
FS – файловый сервер

Запуск заданий в грид



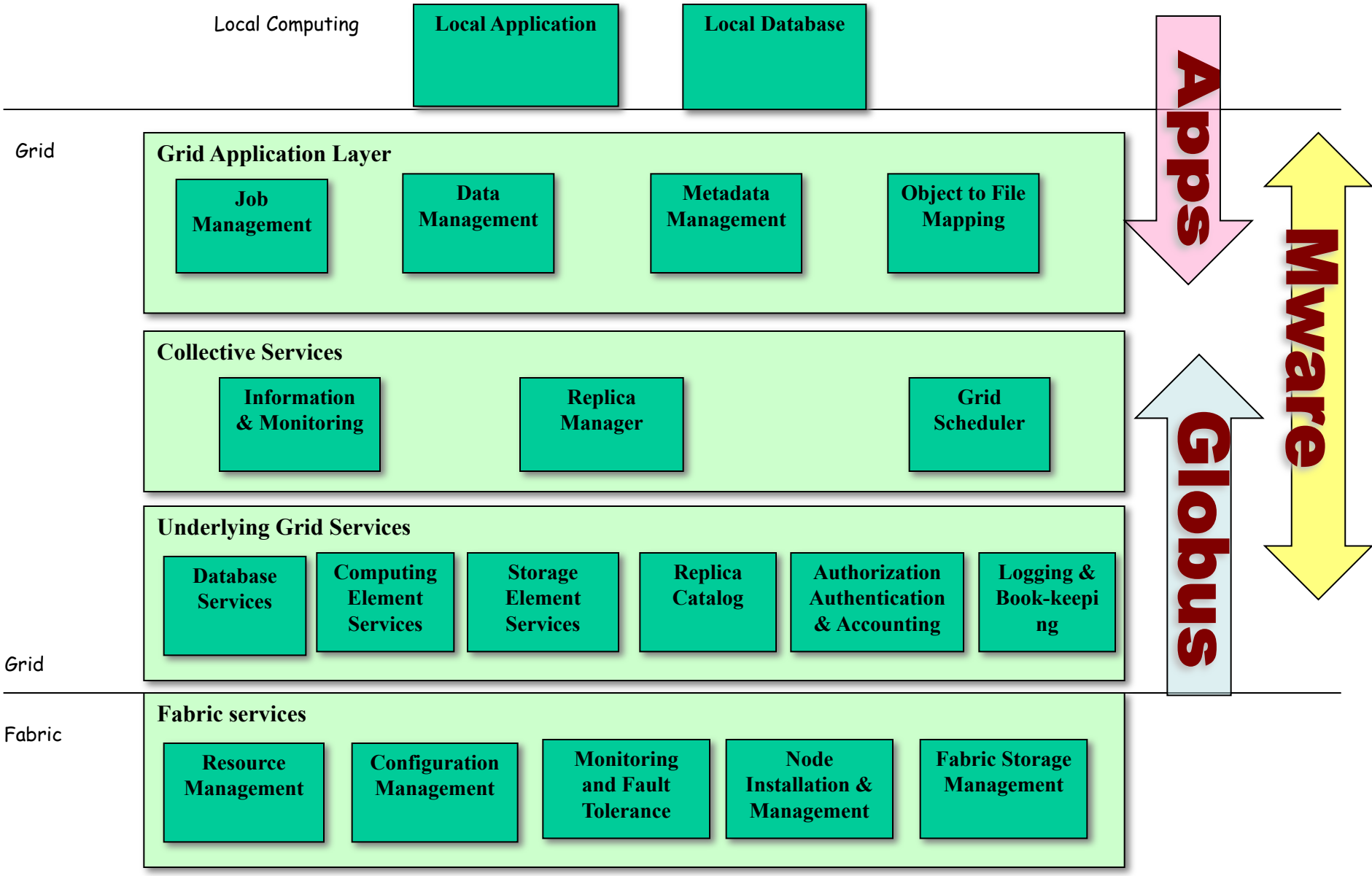
UI – Интерфейс пользователя
RB – Брокер ресурсов
BDII – Информационная база данных по ресурсам
RLS – Сервер реплик файлов
CE – Компьютерный элемент
SE – Элемент хранения данных
WN – рабочая нода
FS – файловый сервер

Запуск заданий в грид



UI – Интерфейс пользователя
RB – Брокер ресурсов
BDII – Информационная база данных по ресурсам
RLS – Сервер реплик файлов
CE – Компьютерный элемент
SE – Элемент хранения данных
WN – рабочая нода
FS – файловый сервер

DataGrid Architecture



LHC Computing Grid Project (LCG)

Основной задачей проекта LCG является создание глобальной инфраструктуры региональных центров для обработки, хранения и анализа данных физических экспериментов LHC.
Новейшие технологии GRID являются основой построения этой инфраструктуры.

Проект LCG осуществляется в две фазы.
1 фаза (2001-2005 гг.) - *создание прототипа и разработка проекта системы (LCG TDR).*
2 фаза (2005-2007 гг.) - *создание инфраструктуры LCG, готовой к обработке, хранению и анализу данных на момент начала работы ускорителя в 2007 году.*

Состав программного обеспечения (GRID middleware) для проекта LCG

Пакет GLOBUS Toolkit.

Пакет VDT (Virtual Data Toolkit), разработанный в американских GRID проектах: PPDG, GriPhyN и iVDGL.

Этот пакет представляет собой набор надстроек над библиотекой инструментальных средств GLOBUS, позволяющих реализовывать распределенную вычислительную систему, но без GRID сервисов.

Он включает в себя пакет Condor/Condor-G, который используется в качестве распределенной системы запуска заданий в пакетном режиме.

Набор компонент, разработанных в проекте

EU DATAGRID : *ресурс-брокер* (обеспечивающий сервис по распределению заданий), информационная служба, replica catalog.

LHC Computing Grid Project
http://lcg.web.cern.ch/LCG/

Applications | Fabric | Grid Deployment | Grid Technology | ARDA

CERN Home > The LHC Computing Grid Project (LCG)

Find
- All CERN
- IT Department
- LCG

LCG Home
Operations Centre
Project Structure
Calendar
Job Opportunities
Contact Us...
LCG Logos

Operating Committees

Implementation (PEB)
GRID Deployment Board
Architects Forum

High Level Committees

Overview (POB)
Software & Computing Committee (SC2)
Computing Resources Review Board

LHC Computing Grid Project

The world's largest and most powerful particle accelerator, the Large Hadron Collider (LHC), is being constructed at CERN, the European Organization for Nuclear Research, near Geneva on the border between France and Switzerland.

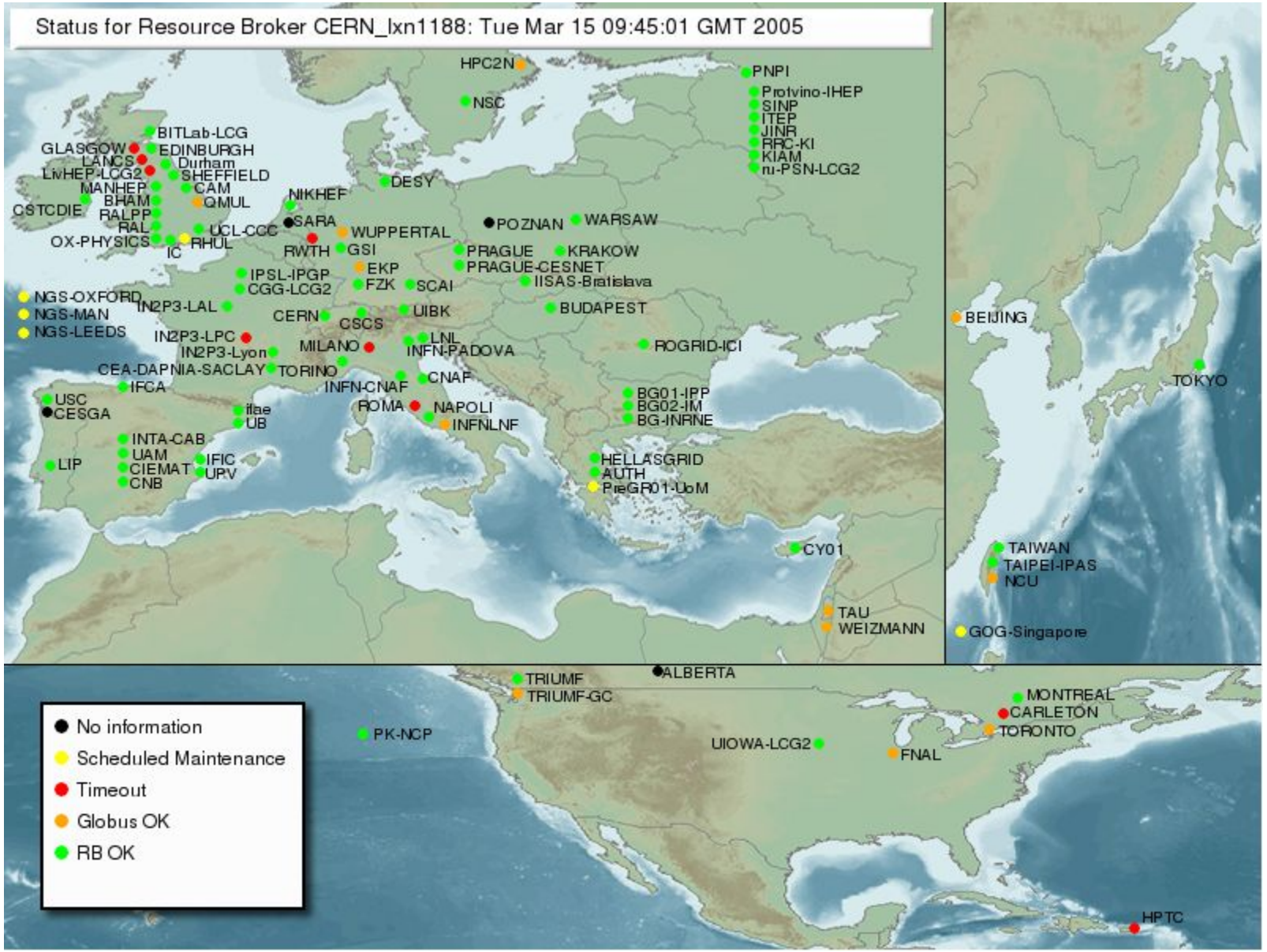
The accelerator will start operation in 2007 and will be used to answer the most fundamental questions of science by some 6,000 people from universities and laboratories all around the world. The computational requirements of the experiments that will use the LHC are enormous: 12-14 PetaBytes of data will be generated each year, the equivalent of more than 20 million CDs. Analysing this will require the equivalent of 70,000 of today's fastest PC processors.

The goal of the LCG project is to meet these unprecedented computing needs by deploying a worldwide computational grid service, integrating the capacity of scientific computing centres spread across Europe, America and Asia into a virtual computing organisation.

LCG Links
- Project Overview
- Project Planning
- Documents/Presentations
- Meetings
- LCG User's Overview
- Grid Application Group (GAG)
- Technology Tracking
- Regional Centre Resources
- Requirements (RTAGs)

External Links
- LHC Experiments
- Industrial Collaboration
- European Grid Projects
- Other Grid Projects
- In the Press

Участие российских институтов в проектах LCG/EGEE



LCG Grid Operations Centre

LCG infrastructure (node types)

- RB Resource Broker node
- MON Node that runs GOUT
- LCFG LCFGng server
- MDS Regional GIIS node
- BDII Information Index replacement
- RLS Relplica location service node (RMC + LRC)
- CE Computing Element, gateway to computing resource
- SE Storage Element
- WN Worker node, Farm node that provides the computing cycles for the CE
- PROX Proxy renewal service node
- IC Information Catalogue, used by RGMA
- NM Network monitoring node
- VOMS Virtual organization management service server
- VO Virtual organization server
- UI User interface node

Модель служб

- Система включает в себя несколько постоянных служб и потенциально много временных служб
- Все службы обеспечивают определенные интерфейсы Грид-служб
  - надежность, управление временем исполнения, доступностью, авторизацией, оповещением, обновлением, управляемостью
- Интерфейсы для управления требованиями Грид-служб
  - исполнение, регистрация, открытие, время выполнения ...

Надежное и безопасное управление распределенными ресурсами

Базовые службы Грид

- поиска сервисов Discovery Services
- регистрации сервисов Registry Services
- управления именами Name Space Management Services
- аутентификации Authentication Services
- авторизации Authorization Servcies
- ресурсов Resource Services
- резервирования Reservation Services
- брокера запросов Brokering Services
- планирования заданий Scheduling services
- балансировки загрузки Load Balancing services
- отказоустойчивости Fault Tolerance Services
- событий и оповещений Event and Notification Services
- протоколирования Loggin Services
- мониторинга Instrumental and Monitoring
- биллинга Accounting Services
- кеширования и репликаций Data Caches and Data Replication Services
- поиска метаданных Metadata Search Services
- транзакций Transaction Services
- администрирования Administration Services

Open GRID Service Architecture - OGSA

- *Общепринятая точка зрения*: следующее поколение глобально-распределенных систем будет основано на Грид-сервисах с открытой архитектурой (OGSA)
- **OGSA:**
  - основной объект - **Грид-служба** (≠ Web-service)
  - обширный набор служб, которые VO могут комбинировать различными способами для создания Грид-систем с заданными свойствами;
  - определяет стандарты методов создания, наименования, поиска экземпляров служб и т.п.;
  - предполагает платформо-независимую интеграцию распределение ресурсов на основе технологий ***Java*** и ***XML,*** а также протокола ***SOAP***.

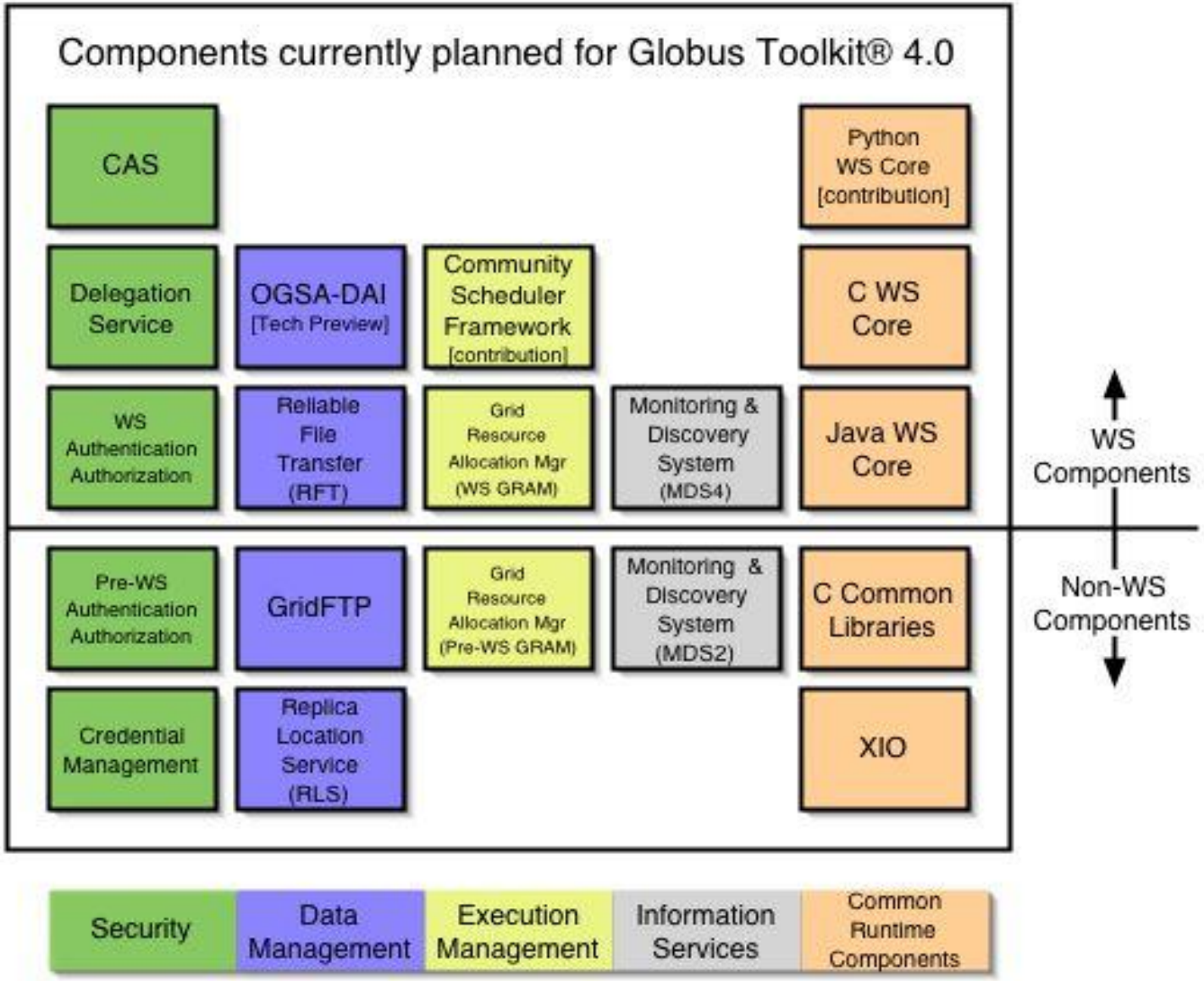
WS – GT3 – WSRF – GT4 - ...

Открытая архитектура GRID-служб (OGSA)

- Эволюция технологии Grid привела к возникновению **Open Grid Services Architecture**, которая определяет стандартные механизмы для создания, именования и обнаружения экземпляров Grid-служб.
- OGSA поддерживает создание и применение служб для виртуальных организаций (VO), предлагая общее представление вычислительных ресурсов, сетей, баз данных, программ, трактуя их как службы, предлагающие свои возможности посредством обмена сообщениями.
- OGSA представляет эволюцию технологий Grid и Web-служб.
- Поддерживая временные, сохраняющие состояния экземпляры служб, OGSA значительно расширяет возможности Web-служб, при минимальной доработке имеющихся технологий. OGSA обеспечивает реализацию концепций Grid, позволяя при этом использовать инструментарий Web-служб.
- Службы и абстракции OGSA предлагают строительные блоки, которые применяются для реализации Grid-служб более высокого уровня.

OGSA (Open Grid Services Architecture)

Components for Globus Toolkit 4



gLite http://www.glite.org

gLite Services for Release 1
Software stack and origin (simplified)

- Computing Element
  - Gatekeeper (Globus)
  - Condor-C (Condor)
  - CE Monitor (EGEE)
  - Local batch system (PBS, LSF, Condor)
- Workload Management
  - WMS (EDG)
  - Logging and bookkeeping (EDG)
  - Condor-C (Condor)
- Storage Element
  - File Transfer/Placement (EGEE)
  - glite-I/O (AliEn)
  - GridFTP (Globus)
  - SRM: Castor (CERN), dCache (FNAL, DESY), other SRMs
- Catalog
  - File and Replica Catalog (EGEE)
  - Metadata Catalog (EGEE)
- Information and Monitoring
  - R-GMA (EDG)
- Security
  - VOMS (DataTAG, EDG)
  - GSI (Globus)
  - Authentication for C and Java based (web) services (EDG)

Main Differences to LCG-2

- **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)**
- **Distributed and re-factored file & replica catalogs**
- **Secure catalogs (based on user DN; VOMS certificates being integrated)**
- **Scheduled data transfers**
- **SRM based storage**
- **Information Services:**
  R-GMA with improved API and registry replication
- **Prototypes of additional services**
  - **Grid Access Service (GAS)**
  - **Package manager**
  - **DGAS based accounting system**
  - **Job provenance service**
- **Move towards Web Services**

Standards

- **Web Services Fast moving area**
  - **Follow WSRF and related standards but are not early adopters**
  - **WS-I compatibility is a target**
    - **Challenging to write WSDL which is WS-I compatible AND can be processed by all the tools**
  - **Industry strength tooling not always available**
  - **Trying to keep back from the bleeding edge**

- **Work on standards bodies**
  - **Active contributions to**
    - **GGF OGSA-WG**
      - **GMA in OGSA**
      - **Data Design team**
    - **GGF INFOD-WG**
    - **OASIS WS-N**
    - **GGF GSM-WG (SRM)**
      - **Co-chairing WG**
    - **Replica Registration Service**
  - **And following many, many others**
  - **Adopting mature standards is a goal**

Основные подсистемы gLite

Вычислительный элемент (Computing Element – CE) – это служба, представляющая ресурсный узел грид и выполняющая на нем функции управления заданиями (запуск, удаление и т.д.). Обращения к CE могут исходить либо от интерфейса пользователя, либо от Менеджера загрузки (Workload Manager –WM), который распределяет задания по множеству CE.

В gLite функциональность CE расширена по сравнению с аналогичной службой LCG-2. Если в LCG-2 CE может работать только в соответствие с Push моделью (WM самостоятельно принимает решение о посылке задания на CE), то в gLite возможен режим работы CE также и в Pull модели, когда CE запрашивает задание у WM.

Помимо функций управления заданиями CE также вырабатывает информацию о состоянии ресурсов. В Push модели ее публикует информационная служба, и она используется WM для выбора CE, на котором будет запускаться задание. В Pull модели информация встраивается в посылаемое WM сообщение "CE доступен".

Основные подсистемы gLite

- **Подсистема управления данными (Data Management Subsystem - DM)** включает три службы, поддерживающие доступ к файлам: элемент памяти (Storage Element – SE), службы каталога (Catalog Services – CS) и диспетчер данных (Data Scheduling –DS). Все службы работают с данными на файловом уровне, в противоположность, например, системам баз данных, которые оперируют такими элементами как записи и поля.

  В распределенной среде грид пользовательские файлы могут храниться во множестве экземпляров – реплик, размещенных в разных местах, и задача CS и DS состоит в том, чтобы сделать процесс управления репликами прозрачным для пользователя, так чтобы приложения получали доступ к файлам по их именам или дескрипторам метаданных.

  Доступ к данным файлов реально происходит через SE, но DM поддерживает также концепцию виртуальных наборов данных. Это открывает новые интересные возможности, основанные на абстракции глобальной файловой системы: при навигации по файлам клиентское приложение может быть устроено как командная оболочка Unix, используя команды смены директорий, просмотра файлов и т.п. Защита файлов обеспечивается в DM списками контроля доступа ACL (Access Control Lists).

Основные подсистемы gLite

- **Подсистема учета (Accounting Subsystem - DGAS)** аккумулирует информацию об использовании ресурсов грид отдельными пользователями, группами пользователей и виртуальными организациями. Собранная информация позволяет построить общую картину деятельности в грид, на основе которой может формироваться политика распределения ресурсов и взиматься плата за их использование.

  Подсистема протоколирования (Logging and Bookkeeping - LB) отслеживает выполняющиеся в разных точных грид шаги обработки задания, фиксируя происходящие с ним события (запуск, распределение на подходящий CE, начало выполнения и т.д.) и запоминая их. Информация о событиях (протокол) поставляется компонентами WM и CE, для чего в эти компоненты встраиваются обращения к LB.

  Протоколы собираются в два приема. Вначале события передаются в локальную службу (*Locallogger*) и записываются в файл на диске. *Locallogger* отвечает за передачу протокола одному из серверов хранения (Bookkeeper), который "укрупняет" события, давая общую картину изменения состояний задания (*Submitted, Running, Done…*). Помимо того, Bookkeeper сохраняет различные атрибуты задания: его описание (JDL); CE, на котором оно выполнялось; коды завершения и т.д. Как протокол состояний, так и протокол событий можно получить либо с помощью специального интерфейса WM, либо через уведомления при определенных изменениях состояния, например, при окончании задания.

Основные подсистемы gLite

- **Подсистема информационного обслуживания и мониторинга грид (Relational Grid Monitoring Architecture - R-GMA**) решает задачу сбора и управления данными о состоянии грид, получая информацию от множества распределенных источников – поставщиков. В ее основе лежит разработанная одной из групп Global Grid Forum (GGF-PERF) схема "Потребитель-Поставщик", описывающая способ взаимодействия этих компонентов. Поскольку схема достаточно общая, она применима как для хранения данных о грид (какие ресурсы и службы доступны, каковы их характеристики), так и для мониторинга приложений.

  R-GMA представляет собой реляционную реализацию этого общего подхода. При наличии множества распределенных поставщиков с точки зрения информационных запросов R-GMA действует как одна большая реляционная база данных.

Основные подсистемы gLite

- **Подсистема управления загрузкой (Workload Management System - WMS)** состоит из ряда компонентов, ответственных за распределение заданий между ресурсами грид, а также обеспечивающих управление заданиями. Центральной компонентой является Менеджер загрузки (WM), который получает от своих клиентов запросы по управлению заданиями. В частности, обрабатывая запрос типа "запуск" WM определяет подходящий для выполнения CE, принимая во внимание требования и предпочтения, указанные в описании задания.

  Система безопасности рассматривается как средство защиты Web-служб и будет реализовываться в виде дополнительных модулей, размещаемых в контейнерах (Apache Axis, Tomcat). Разработаны предложения по архитектуре безопасности: https://edms.cern.ch/document/487004/ и сформулированы основные цели: модульность, расширяемость, соответствие развивающимся стандартам Web-служб (WS-Security).