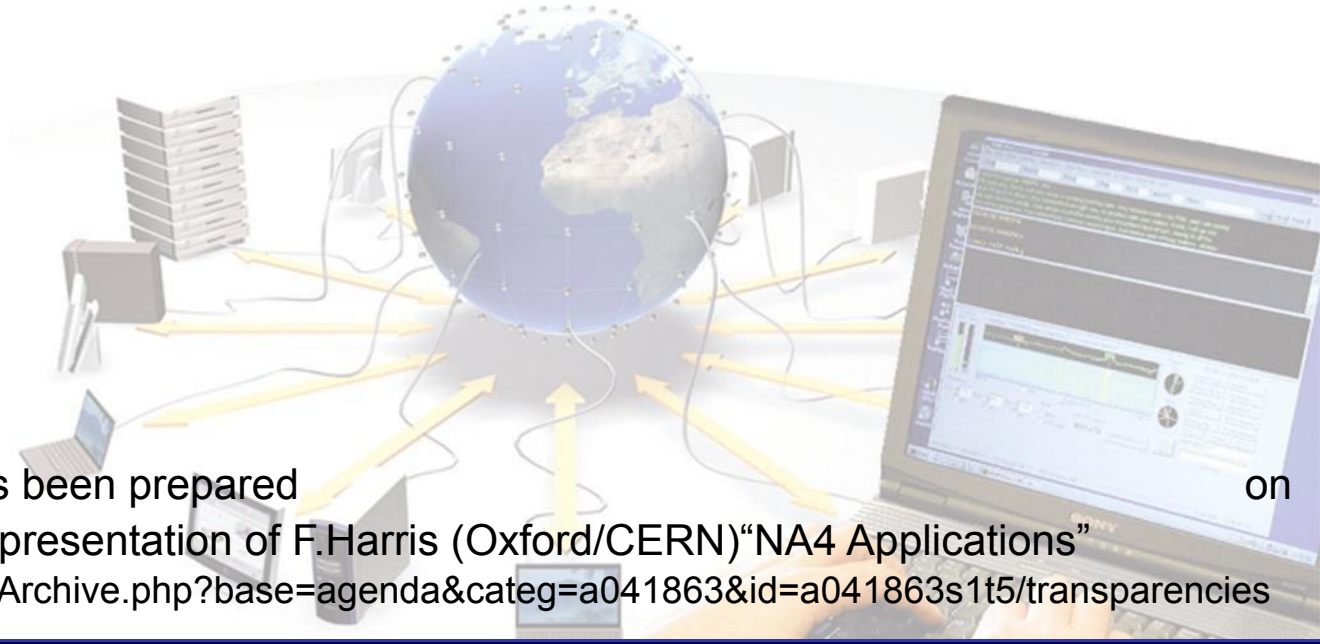
Enabling Grids for E-science in Europe

www.eu-egee.org

NA3 Induction Courses in the Dubna Conference
June 28, 2004

EGEE Applications

E.Tikhonenko (JINR, NA4 Manager for Russia),
N.Zaikin (JINR, NA3 Manager)



A presentation has been prepared on basis of the presentation of F.Harris (Oxford/CERN)"NA4 Applications"
http://agenda.cern.ch/askArchive.php?base=agenda&categ=a041863&id=a041863s1t5/transparencies

EGEE is a project funded by the European Union under contract IST-2003-508833

Talk Outline

Enabling Grids for E-science in Europe

- NA4 basic goals and the directions of activities
- Organizational structure
- Participants
- NA4 sub-tasks:
  - *biomed*
  - *HEP*
  - *'generic' приложения*
  - *testing*
  - *Industry Forum*
- Milestones and deliverables
- RDIG-EGEE participation in NA4
- Conclusions
- Glossary & Useful links

План лекции

- Основные цели и составляющие работы NA4
- Организационная структура
- Участники
- Направления работ подгрупп NA4:
  - *биомедицинские приложения*
  - *приложения физики высоких энергий*
  - *‘базовые’ приложения*
  - *тестирование*
  - *промышленный форум*
- Этапы работы и ожидаемые результаты
- Взаимодействие с другими рабочими группами проекта EGEE
- Участие RDIG-EGEE в NA4
- Заключение

Группа NA4: Идентификация и поддержка приложений в среде проекта EGEE

Enabling Grids for E-science in Europe

Цели работ по идентификации и поддержке приложений:

- определение набора существующих пользовательских приложений из широкого спектра прикладных областей – научной, промышленной и коммерческой;
- создание для каждой новой отрасли хорошо подготовленных групп для поддержки и развертывания (размещения) приложений, что, в свою очередь, создаст прочную основу для расширения сообщества EGEE;
- сосредоточение работы на начальном периоде действия проекта в хорошо сформулированных прикладных областях – физике частиц и науках о жизни (в частности, биомедицине). Эти два научных сообщества уже приобщены к грид-технологиям и с самого начала проекта готовы к развертыванию реальных сложных приложений

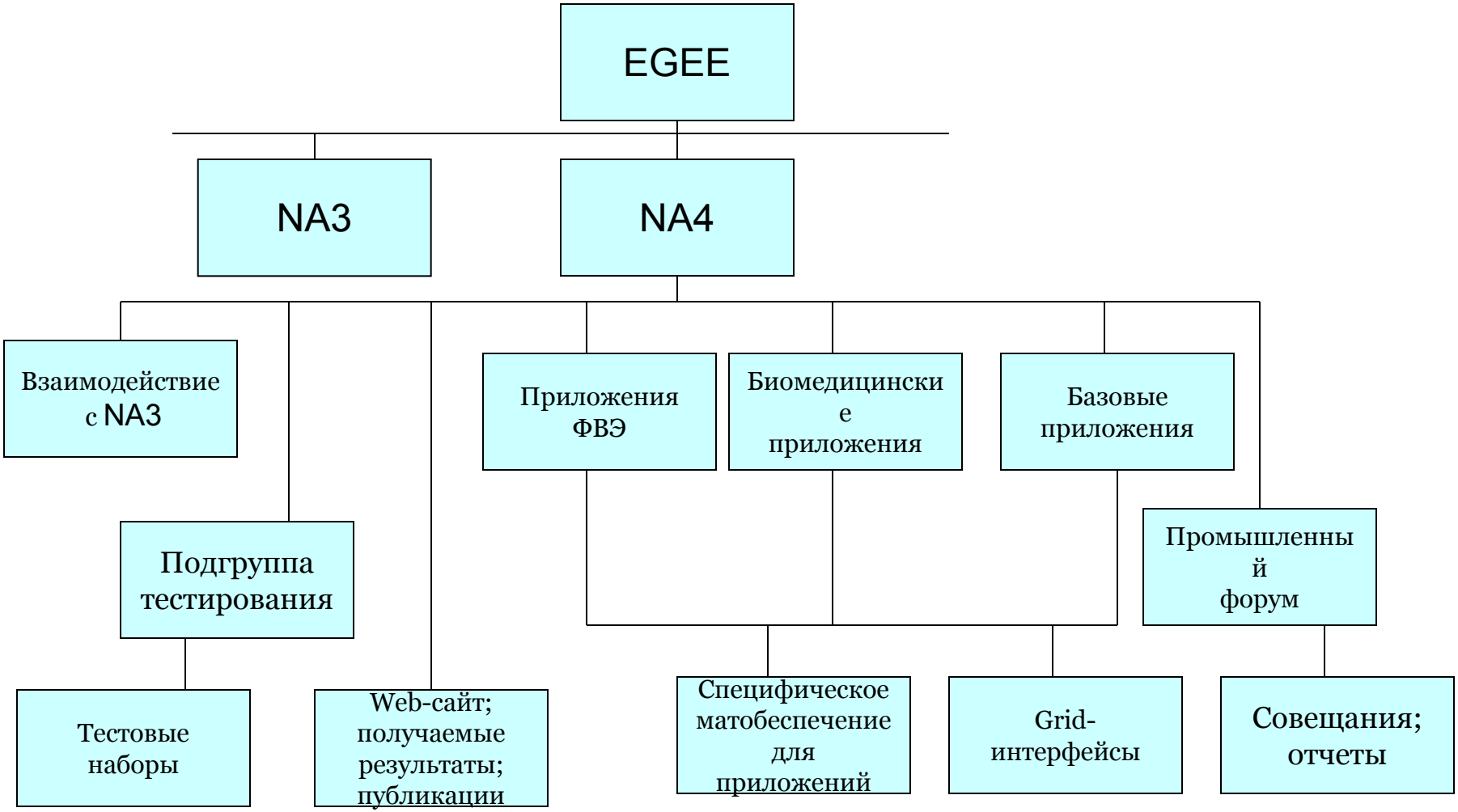
Группа NA4: Идентификация и поддержка приложений в среде проекта EGEE

Enabling Grids for E-science in Europe

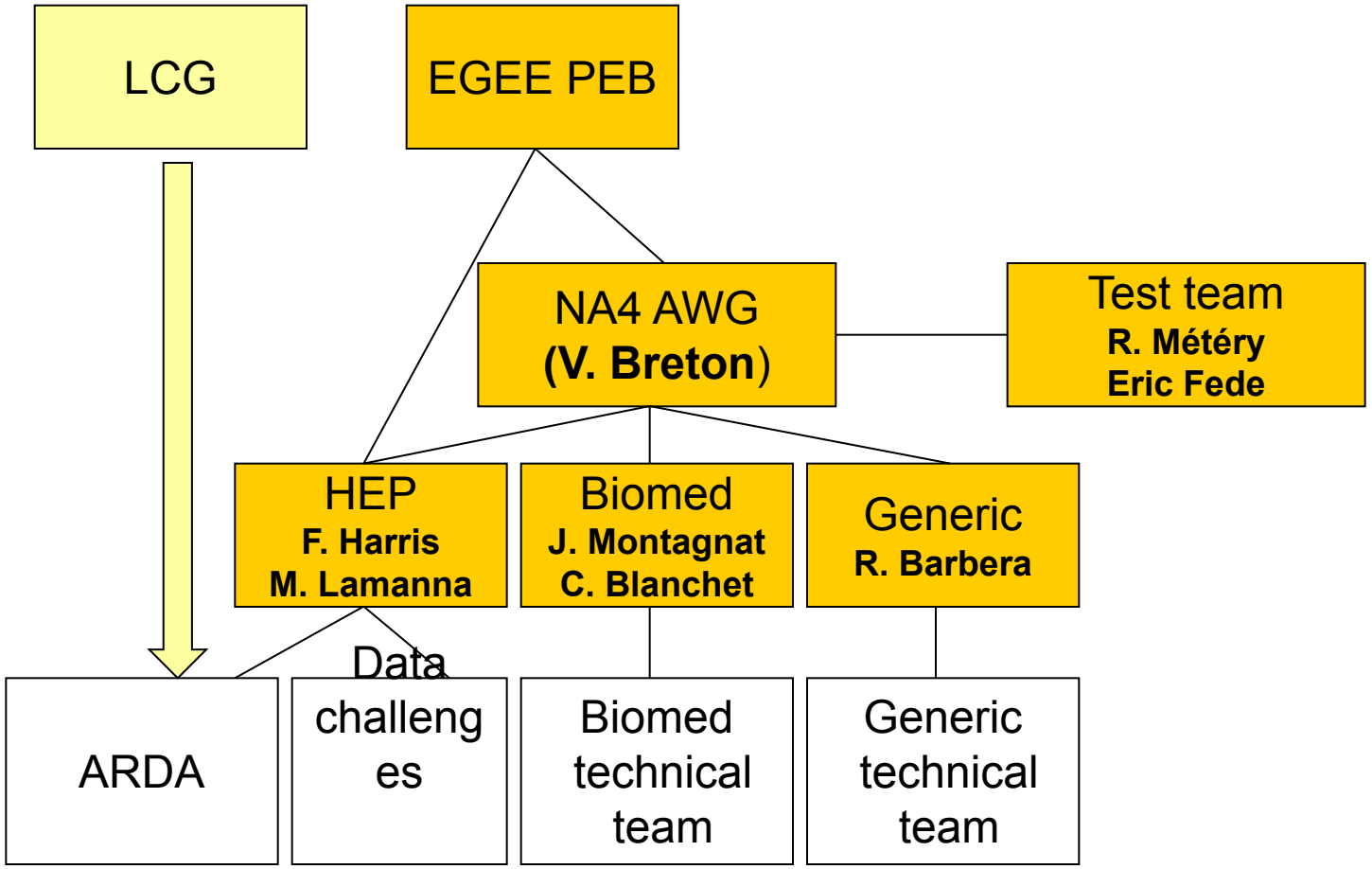
Основные составляющие работы:

- Результатом работы группы NA4 будут являться программные приложения – прикладные пакеты, развернутые в инфраструктуре EGEE и доступные для работы в grid-среде соответствующим сообществам пользователей
- Для развертывания этих приложений может понадобиться специальное ПО для обеспечения интерфейса к grid. Необходимо собрать существующую документацию из проекта EDG и других проектов (LCG, ARDA, GridLab, Healthgrid, …) для выработки общего решения
- Процесс развертывания приложений в инфрастуктуру EGEE будет происходить в рамках виртуальных организаций, объединяющих соответствующих пользователей
- Инфраструктура EGEE будет расширяться; с появлением новых пользователей им будет оказываться поддержка и будет организовываться обучение; также будут создаваться новые виртуальные организации

Организационная структура NA4



NA4: руководство и взаимодействие



NA4: роли партнеров в проекте и финансирование

Enabling Grids for E-science in Europe

Федерация	Роль	FTE Funded	FTE Unfunded
CERN	Приложения ФВЭ (координация)	4	4 (9)
UK+Ireland	Взаимодействие с NA3	0,5	0,5
Italy	Базовые приложения (координация)	2	2
France	Общая координация; биомедицинские приложения; подгруппа тестирования; контакты с промышленными партнерами	7	7
Northern Europe	Базовые приложения	1	1
Germany + Switzerland	Базовые приложения	1	1
Central Europe	Базовые приложения	1	1
South West Europe	Биомедицинские приложения	2	2
Russia	Приложения ФВЭ; биомедицинские приложения; приложения ядерной физики	5.7	0.3
		24,2	**18.8**

специфика биомедицинских приложений

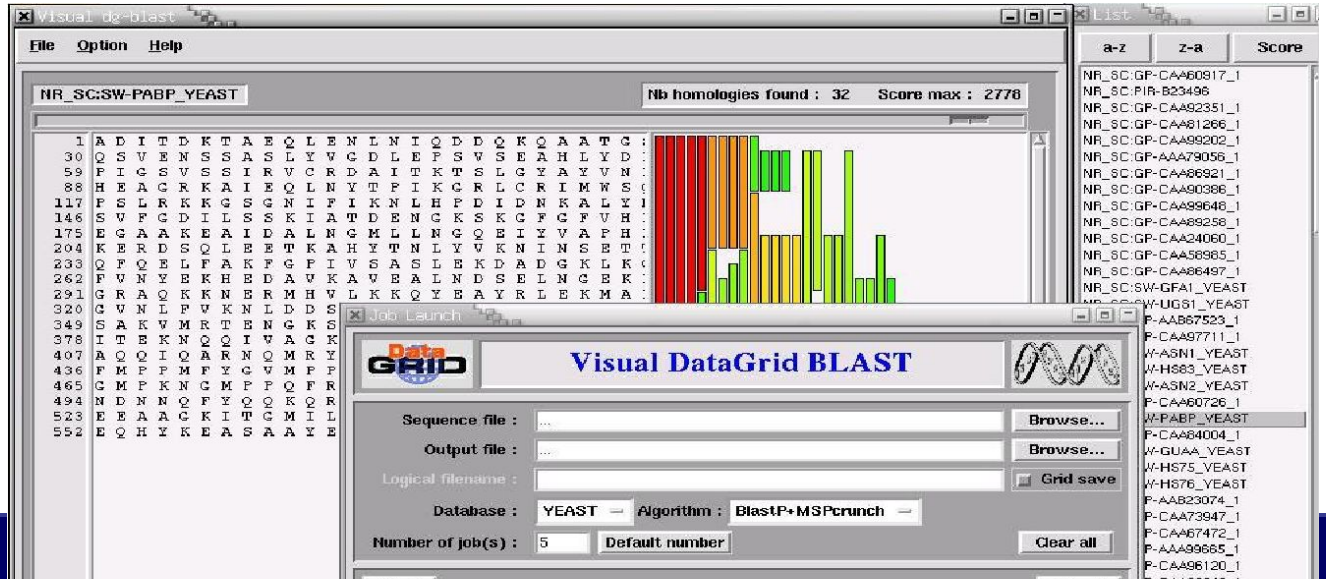
Enabling Grids for E-science in Europe

- Сложные требования по данным
  - Гетерогенные форматы данных
  - Частая обновляемость данных
  - Сложные наборы данных (медицинские записи)
  - Ограничения на безопасность и конфиденциальность
  - Необходимость длительного хранения данных
- Сложные требования по обработке данных
  - Биоинформатика (геномика, протеомика, …): распределенные базы данных
  - Медицинские(просмотр снимков, эпидемиология...): распределенные базы графических данных
  - Использование параллельных алгоритмов для обработки медицинских графических данных и для моделирования
  - Интерактивные приложения
  - Ограничения на безопасность и конфиденциальность

BLAST: биоинформатика в испытательной модели EDG

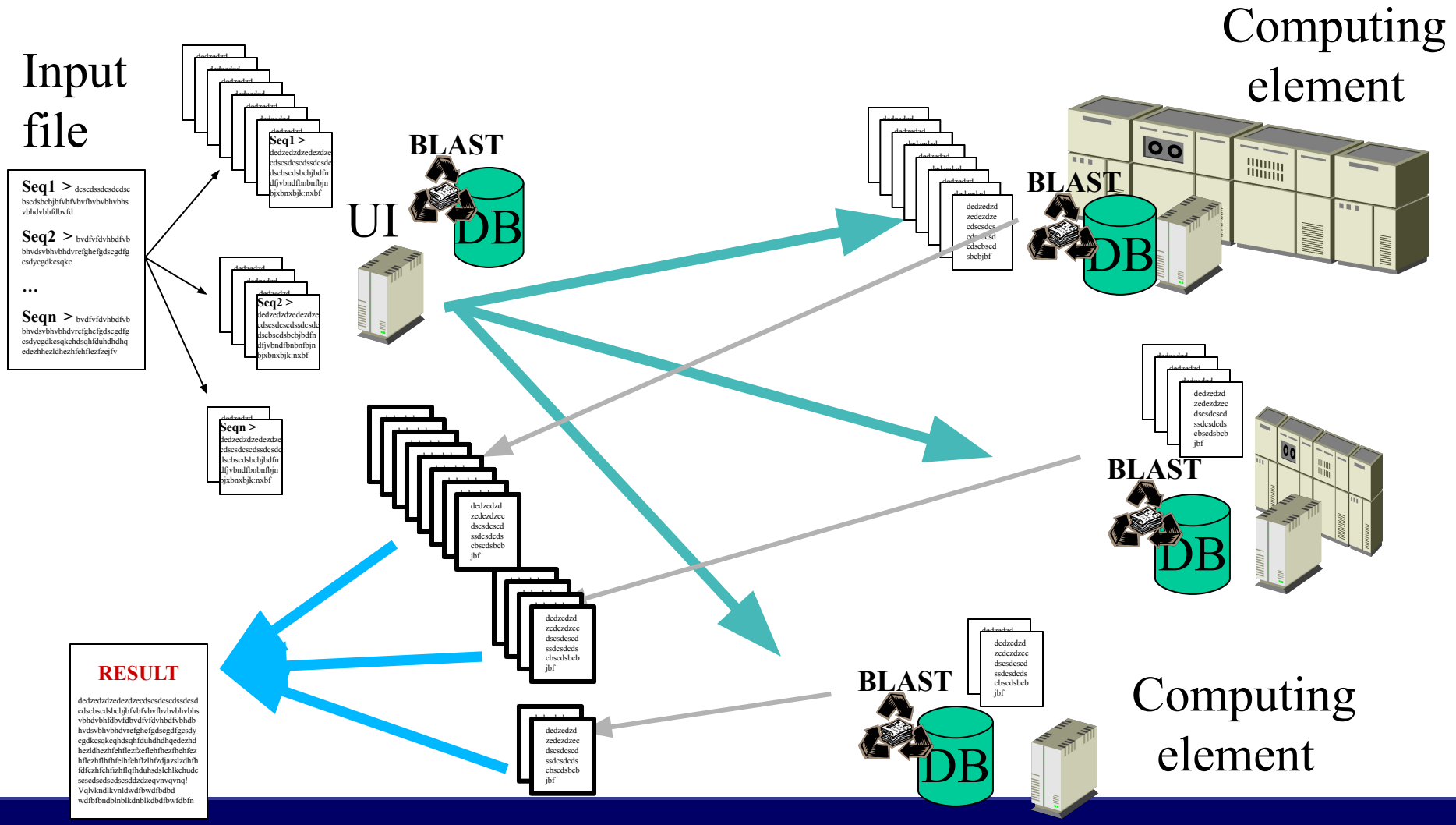
Enabling Grids for E-science in Europe

- Приложение BLAST - первый шаг в анализе новых последовательностей при сравнении ДНК- или белковых последовательностей с последовательностями, хранящимися в частных и публичных базах данных; может рассматриваться как идеальное grid-приложение:
  - Требует ресурсы для хранения баз данных и запуска задачи
  - Позволяет производить сравнение одной или нескольких последовательностей вместо параллельной работы с несколькими базами данныхl
  - Большое сообщество пользователей



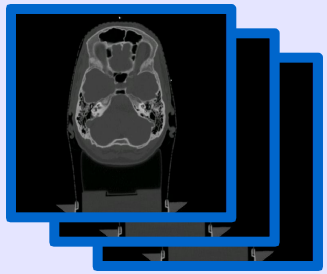
Гридификация приложения BLAST

Enabling Grids for E-science in Europe



Моделирование Монте-Карло в рентгенотерапии

Enabling Grids for E-science in Europe



Scanner slices: DICOM format

Concatenation

Image: text file

Binary file: Image.raw Size 19M

Retrieving of root output files from CEs

the CE

GATE

Computing Element

Storage Element

CCIN2P3

RAL

NIKHEF

MARSEILLE

User interface

Database

Anonymisation

Эксперименты на LHC

Enabling Grids for E-science in Europe

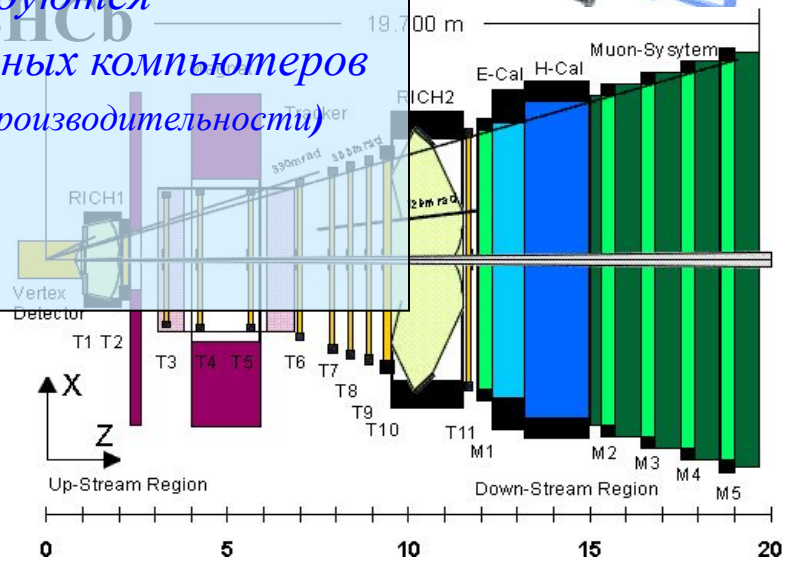
ATLAS

CMS

ALICE

При ожидаемой скорости записи сырых данных потребуются ресурсы для хранения данных порядка ***десятков и сотен ПБ****.*

Для обработки данных потребуются ***сотни тысяч*** *персональных компьютеров (максимальной на текущий момент производительности)*



Обработка данных и вычисления в физике высоких энергий

Enabling Grids for E-science in Europe

детектор

Триггер 1-го уровня

Отбор событий и первичная реконструкция

суммарные данные по событиям

обработанные данные

«сырые» данные

Пакетная обработка данных

Реконструкция событий

моделирование физических событий

объекты для физического анализа (выделенные по физическим каналам)

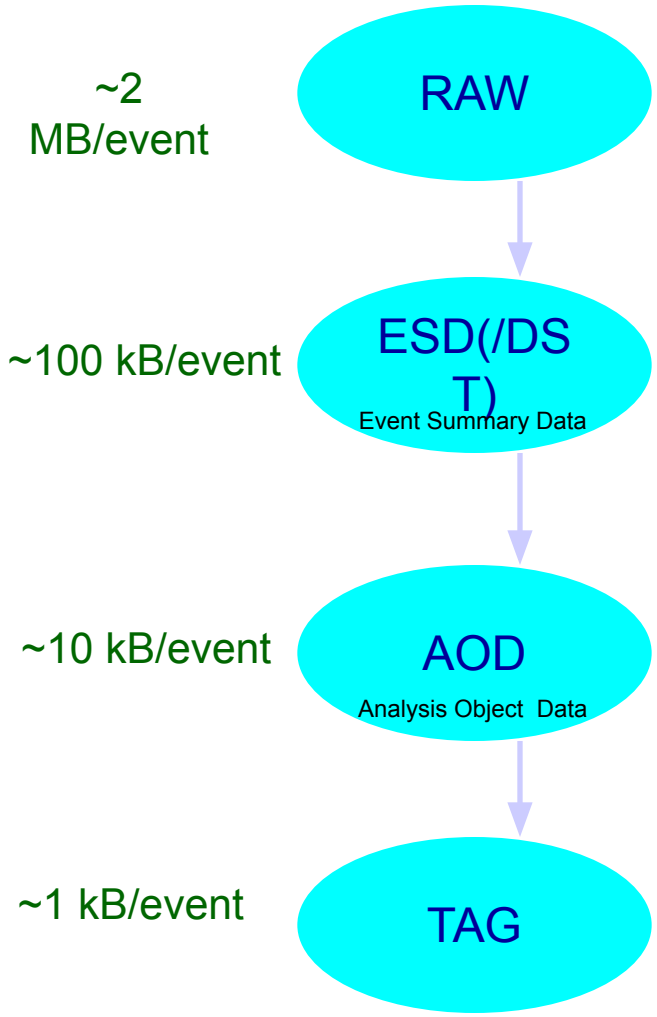
интерактивный физический анализ

Иерархия данных

Enabling Grids for E-science in Europe

На 2 порядка уменьшается объем данных по сравнению с исходным потоком

"RAW, ESD, AOD, TAG"



События, удовлетворяющие условиям триггера; записываются системами сбора данных (DAQ)

Съем цифровых отсчетов с детекторов

Реконструированная информация

L2+L3

Первичная стадия реконструкции на уровне создания базовых кластеров и трек-сегментов

Геометрическая реконструкция событий: траектории частиц, импульсы и энергии

Информация для анализа

Физическая реконструкция: установление соответствия частиц и треков, уточнение характеристик струй, поиск распадных вершин

Указатель события (классификационная информация для быстрого выбора нужного события)

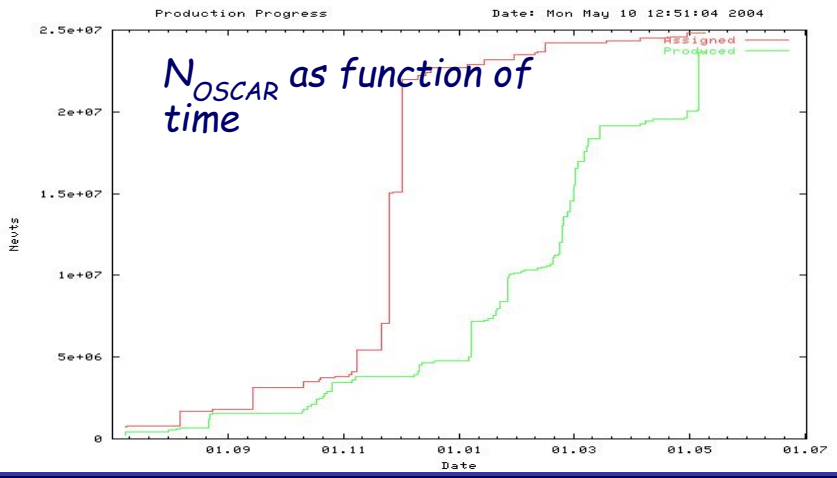
Специфика приложений ФВЭ

Enabling Grids for E-science in Europe

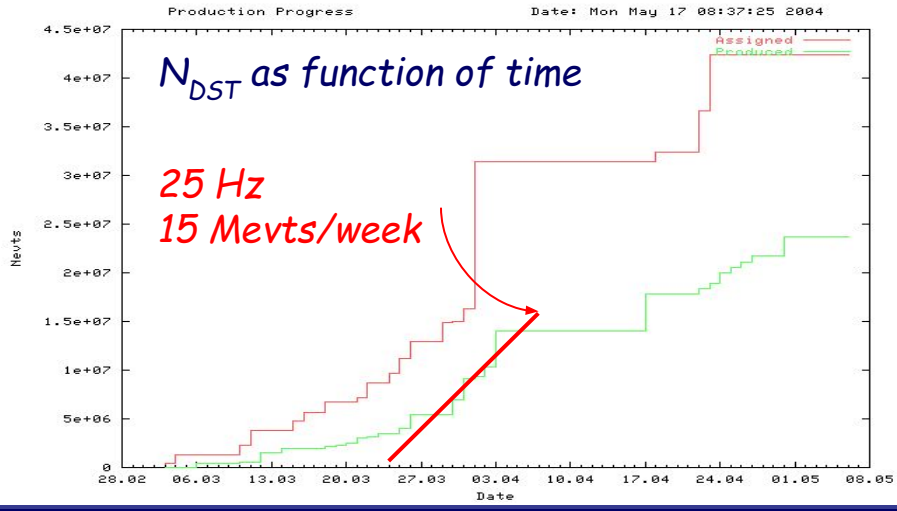
- Требования по данным
  - **Колоссальные объемы данных (десятки и сотни Петабайт)**
  - **Данные типа WORM (писать единожды, читать многократно)**
  - **Структуризация данных с последующим извлечением информации из данных (data mining)**
  - **Продолжительное время хранения данных, а также необходимость создания копий данных в разных странах мира**
- Требования к обработке данных
  - **Обработка данных подразделяется на 2 типа – регулярное производство данных и «нерегулярный» анализ данных**
    - Производство (моделирование) данных происходит систематически; при этом производятся наборы данных порядка ~ 10**9 физических событий.
    - Анализ физических данных (на наборах данных порядка 10**7 событий) проводится произвольным образом и в индивидуальном порядке многими сотнями отдельных пользователей
  - **Высокий уровень параллелизма обработки на уровне событий, который может быть описан ориентированным графом с указанием последовательности обработки**
  - **Поскольку интерактивная работа очень важна при анализе данных, необходимо предусмотреть возможность спасения сессий с сохранением информации об источнике данных («проверяемость», provenance)**
  - **Необходимость глобального доступа к базам данных экспериментов для получения значений констант, условий работы и т.д.**

Характеристики CMS Data Challenge DC04

- Pre-Challenge Production
  - Использование набора инструментальных средств OCTOPUS, объединяющего средства производства данных CMS (CMS production tools) с grid-средствами.
  - В результате 8-ми месяцев непрерывного производства данных:
    - просчитано 750 000 заданий
    - при затратах производительности порядка 3500 KSI2000 - месяцев
    - получено 700 000 файлов
    - объем полученных данных - 80 TB
  - Производство данных с использованием пакета OSCAR (на основе Geant 4)
    - За 6 месяцев произведено 16 миллионов событий

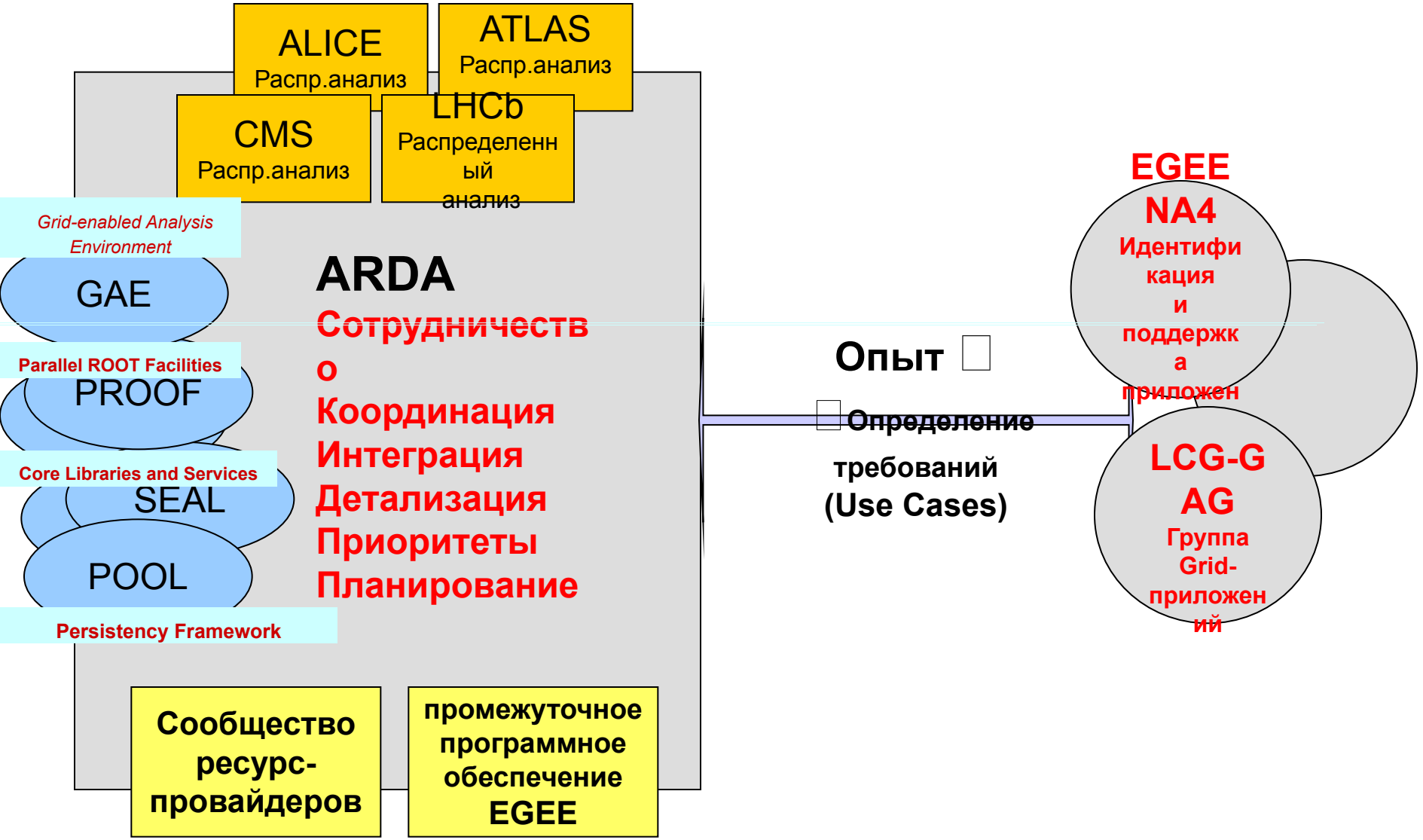


- Data Challenge
  - Поставленная задача: воспроизвести полную последовательность действий по реконструкции и распределению (размещению) данных на частоте 25 Гц
  - В результате удалось выполнить эту задачу в течение ограниченного периода времени; при этом:
    - В Tier-0 на 500 ЦПУ выполнялось 2200 заданий в день и производились данные со скоростью 4 MB/c;
    - затем данные передавались в соотв.Tier-1
    - регистрация данных (с POOL-метаданными) в RLS (Replica Location Service) происходила со скоростью 0.4 файла в секунду



ARDA :A Realisation of Distributed Analysis for LHC

Enabling Grids for E-science in Europe



NA4 «базовые» приложения

Enabling Grids for E-science in Europe

- Основная задача - привлечение новых научных и промышленных сообществ, заинтересованных в использовании инфраструктуры, которая будет создана в ходе проекта EGEE.
- Хорошо организованный портал GENIUS может служить прекрасным инструментом для внедрения в среду промежуточного ПО EGEE новых приложений – в значительной степени потому, что на портале создан очень простой и доступный пользовательский интерфейс, что особенно важно при привлечении новых пользователей, не имеющих опыта работы в grid-среде.
- GILDA – это полный набор элементов grid (испытательная модель, сертификация, виртуальная организация, система мониторинга, веб-портал) и приложений, который целиком посвящен задаче распространения знаний о grid-технологиях. Поэтому он успешно используется на обучающих курсах в рамках проекта EGEE. **Его можно также считать идеальной испытательной моделью для портирования новых базовых приложений.**

Портал GILDA (http://gilda.ct.infn.it)

Вопросник по базовыми приложениям

Enabling Grids for E-science in Europe

- Чтобы получить информацию и узнать о первых требованиях от новых сообществ, заинтересованных в использовании инфраструктуры EGEE, был разработан вопросник, который доступен по адресу (http://alipc1.ct.infn.it/grid/egee/na4/questionnaire/na4-genapp-questionnaire.doc)
- С уже поступившими сведениями можно ознакомиться по адресу (http://alipc1.ct.infn.it/grid/egee/na4/questionnaire):
  - **Астрофизика (изучение эволюции галактики с помощью искусственного спутника Планк)**
  - **Система наблюдения Земли (озоновые карты, сейсмология, климат)**
  - **Электронные библиотеки (проект DILIGENT)**
  - **Grid – поисковые серверы (поисковый сервер проекта GRACE (Gravity Recovery and Climate Experiment))**
  - **Промышленные приложения (проект SIMDAT – grid-приложения в автомобильной, фармацевтической, авиа-космической промышленности и метеорологии)**
- Также был проявлен интерес из нескольких других сфер: вычислительной химии (Италия и Чехия), гражданского проектирования (Испания), и геофизики (Швейцария и Франция)

Задачи Промышленного форума в проекте EGEE

Enabling Grids for E-science in Europe

- Основная роль Промышленного форума – вовлечение в проект партнеров из различных сфер промышленности.
- Членами Промышленного форума EGEE могут быть компании любого уровня, основной или частичный бизнес которых развернут в Европе.
- Промышленным форумом будет руководить исполнительный комитет, состоящий из участников проекта EGEE и представителей промышленности.
- http://public.eu-egee.org/industry-forum/information

NA4 группа тестирования

Enabling Grids for E-science in Europe

Будут разработаны 3 типа тестов, основанных на требованиях пользователей и опыте работы LHC DCs и ARDA :

- **Тесты по работоспособности сервисов:** набор тестов по проверке работоспособности EGEE-сервисов. При этом должны проверяться все виды grid-сервисов: запуск и управление заданием, управление файлами, информационный сервис, ….
- **Тесты по оценке функциональности:** для проверки, все ли необходимые функциональные возможности доступны: например, создание, перенос или удаление файлов, восстановление при ошибках и т.п.
- **Тесты для оценки рабочих характеристик:** для возможности оценить испытательную модель с точки зрения конечный пользователь-приложение. Часть таких тестов будет посвящена временным оценкам (время запуска задачи, время репликации какого-то количества файлов, …), другие – оценкам масштабируемости (например, какое количество заданий может быть принято таким-то сервисом, ...), некоторые – менее конкретным оценкам (возможность использования информации, доступ к сообщениям об ошибкам,…).
- Эти работы будут проводиться в тесном взаимодействии с ARDA , JRA1 и SA1

Этапы работы и ожидаемые результаты

Enabling Grids for E-science in Europe

В течение первых 6-ти месяцев должна быть завершена миграция первых приложений в структуру EGEE:

- HEP DCs для 4 экспериментов LHC и эксперимента D0
- биомедицина – моделирование с помощью приложения GATE в ядерной медицине + иные приложения
- плюс первые ‘базовые’ приложения

В течение первых нескольких месяцев работы проекта будет выработано определение общего интерфейса для приложений (что особенно важно для новых приложений; здесь будет очень существенным использование портала GENIUS)

В течение первого полугодия будет создан документ по целевой стратегии (который необходим в контексте использования новых приложений в инфраструктуре EGEE)

К концу 3-го квартала работы проекта будет подготовлен отчет о процессе миграции приложений

- По всем приложениям будут даны оценки по действующим и опытным сервисам LCG (текущему и «новому» промежуточному ПО)

Взаимодействие NA4 с другими группами EGEE и иными партнерами (1)

Enabling Grids for E-science in Europe

- SA1 - функционирование grid
  - Как ввести новые виртуальные организации в LCG из других доменов?
  - Как организовать процесс интеграции в LCG новых ресурсов (сайтов) из новых прикладных областей?
  - Рациональность тестовых процедур
  - Сотрудничество с национальными проектами (например, использование мониторинга приложений, разработанного в Великобритании в проекте GridPP)
- NA3 - обучение
  - Оценка требований к курсам
  - Подготовка и проведение курсов
- JRA1 - промежуточное программное обеспечение
  - Обобщение всех исходных требований приложений и мониторинг (с обратной связью к промежуточному ПО) степени удовлетворения этих требований (этот процесс тщательно прорабатывается в подгруппе PTF-Project Technical Forum в рамках группы JRA1)
- JRA2 - обеспечение качества
  - NA4 имеет своего представителя в этой группе для определения процесса мониторирования качества сервисов EGEE

Взаимодействие NA4 с другими группами EGEE и иными партнерами (2)

Enabling Grids for E-science in Europe

- JRA3 - безопасность
  - Безопасность данных для медицинских (и других) приложений
  - Безопасность сайтов
- SA2,JRA4 – организация сети
  - Обеспечение глобальных требований приложений ФВЭ в LCG
  - Биомедицинские и другие приложения могут иметь иные глобальные требования
  - NA4 будет предоставлять информацию по определению требований для отдельных приложений, особенно в проблемных ситуациях
- LCG
  - NA4/HEP представлены в группе grid-приложений (GAG) проекта LCG
    - Это требования от экспериментов ФВЭ и формирование обратной связи в промежуточное программное обеспечение. Некоторые члены группы GAG входят в состав PTF (Project Technical Forum) группы JRA2.

Участие RDIG-EGEE в NA4

Enabling Grids for E-science in Europe

- Приложения ФВЭ:
  - Институт теоретической и экспериментальной физики (Москва) (отв.по LHCb)
  - Институт физики высоких энергий (Протвино) (отв.по ATLAS)
  - Курчатовский институт (Москва)
  - Научно-исследовательский институт ядерной физики (Москва) (отв. по CMS)
  - С.-Петербургский институт ядерной физики (Гатчина)
  - Объединенный институт ядерных исследований (Дубна) (отв. по ALICE и CMS)
- Биологические приложения
  - Институт математических проблем биологии (Пущино)
- Приложения ядерной физики (FusionGrid)
  - Курчатовский институт (Москва)

Основная задача (NA4.4.2) – миграция приложений в инфраструктуру EGEE

Заключение

Enabling Grids for E-science in Europe

- Деятельность группы NA4 на данном этапе базируется на следующих моментах:
  - **Эксперименты ФВЭ предполагают использовать окружение LCG-2 для своих Data Challenges**
  - **ARDA успешно разворачивает свою работу и ждет появления первого прототипа нового промежуточного математического обеспечения**
  - **Биомедицинские приложения готовы для развертывания в среде LCG-2 и опытных сервисов**
  - **Подгруппа «базовых» приложений очень активно взаимодействует с GILDA и NA3**
  - **Подгруппа тестирования ведет свою работу совместно с JRA1 и ARDA**
  - **Промышленный форум налаживает контакты с различными компаниями (см. доклады на конференции EGEE в Корке)**
- 14-16 июля в Катанье намечено проведение открытого совещания NA4, **на котором планируется обсудить проблемы промежуточного п/о, функционирования, безопасности и сетевые вопросы.**
- NA4 Web-сайт http://egee-na4.ct.infn.it

Некоторые термины

Enabling Grids for E-science in Europe

- Data Challenge – крупномасштабные сеансы массового моделирования и обработки физических событий в распределенной среде с использованием grid-технологий; проводятся в ряде экспериментов ФВЭ с целью подготовки и оптимизации рабочей стадии экспериментов
- deployment – развертывание; внедрение, размещение (например, системы, ПО на системе или платформе)
- disseminate - распространять знания
- errors recovery – восстановление при ошибках
- gridification - «гридификация» - развертывание (приложения) в grid-среде
- GRID-services: Job submission and management - запуск и управление заданием; files management - управление файлами; Information service – информационный сервис
- testbed - испытательная модель
- virtual organization (VO) – объединение пользователей, организаций и ресурсов в новый административный домен в рамках grid-инфраструктуры

. . . *should be continued* . . .

Полезные ссылки

- http://lcgapp.cern.ch/project/ – **LCG Project - Applications Area (POOL, GEANT4, SEAL, …)**
- http://www.gridpp.ac.uk/ – **The Grid for UK Particle Physics**

. . . should be continued . . .