Использование грид-технологий для обработки и распределенного хранения сверхбольших объемов данных (сотни петабайт)

Кореньков В.В.
Объединенный институт ядерных исследований
Г. Дубна
korenkov@cv.jinr.ru

Аннотация

Дается краткий обзор развития глобальной грид-инфраструктуры проекта WLCG (World-wide LHC Computing Grid или Всемирный грид для Большого адронного коллайдера). Особое внимание уделяется архитектуре распределенного хранения сверхбольших объемов данных. Дается краткая информация о проектах развития средств мониторинга и управления распределенными файловыми системами, выполненных в ОИЯИ: система мониторинга сервиса передачи файлов FTS развитие системы управления распределенными данными эксперимента ATLAS, **с**истема мониторинга центров уровня Tier3 для анализа данных, глобальная система мониторинга передачи данных в инфраструктуре проекта WLCG. Обозначены новые решения и перспективы в обработке «Больших Данных»

1. Грид-инфраструктура для обработки и хранения данных Большого адронного коллайдера (WLCG)

Развитие исследований в физике высоких энергий, астрофизике, биологии, науках о Земле и других научных отраслях требует совместной работы многих организаций по обработке большого объема данных в относительно короткие сроки. Для этого необходимы географически распределенные вычис-лительные системы способные передавать и прини-мать данные порядка сотен терабайт в сутки, одно-временно обрабатывать сотни тысяч задач и долго-временно хранить сотни петабайт данных.

Современные грид-инфраструктуры обес-печивают интеграцию аппаратурных и программных ресурсов, находящихся в разных организациях в масштабах стран, регионов, континентов в единую вычислительную среду, позволяющую решать задачи по обработке сверхбольших объемов данных, чего в настоящее время невозможно достичь в локальных вычислительных центрах. Наиболее впечатляющие результаты по организации глобальной инфраструктуры распределенных вычислений получены в проекте WLCG (Worldwide LHC Computing Grid или Всемирный грид для Большого адронного коллайдера) в ЦЕРНе при обработке данных с экспериментов на LHC (Large Hadron Collider) или БАК (Большой адронный коллайдер). На семинаре 4 июля 20012 года, посвященном наблюдению бозона Хигса, директор ЦЕРНа Р.Хойер дал высокую оценку грид-техноло-гиям и их значимости для мировой науки. Без организации грид-инфраструктуры на LHC было бы невозможно обрабатывать и хранить колоссальный объем данных, поступающих с коллайдера, а значит и совершать научные открытия. Сегодня уже ни один крупный научный проект не осуществим без использования распределенной инфраструктуры для обработки данных. Задача организации компью-тинга для экспериментов на БАК была совершенно беспрецедентной, поскольку требовалось:

- обеспечить быстрый доступ к массивам данных колоссального объема;
- обеспечить прозрачный доступ к географически распределенным ресурсам;
- создать протяженную надежную сетевую инфраструктуру в гетерогенной среде.

Была разработана базовая модель компьютинга для экспериментов БАК как иерархическая централизованная структура региональных центров, включающая в себя центры нескольких уровней. Суть распределенной модели архитектуры компьютерной системы состоит в том, что весь объем информации с детекторов БАК после обработки в реальном времени и первичной реконструкции (восстановления треков частиц, их импульсов и других характеристик из хаотического набора сигналов от различных регистрирующих систем) должен направляться для дальнейшей обработки и анализа в региональные центры разных уровней или ярусов (Tier's): Tier0 (CERN) $\Rightarrow$ Tier1 $\Rightarrow$ Tier2 $\Rightarrow$ Tier3 $\Rightarrow$ компьютеры пользователей.

Уровни различаются по масштабу ресурсов (сетевые, вычислительные, дисковые, архивные) и по выполняемым функциям:

Tier0 (ЦЕРН) - первичная реконструкция событий, калибровка, хранение копий полных баз данных;

Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.

Tier1 - полная реконструкция событий, хранение актуальных баз данных по событиям, создание и хранение наборов анализируемых событий, моделирование, анализ;

Tier2 - репликация и хранение наборов анализируемых событий, моделирование, анализ.

В рамках этого проекта были проработаны требования к ресурсам и функции региональных центров уровней Tier0, Tier1, Tier2. Разработанная модель была реализована и успешно функционирует с момента запуска Большого адронного коллайдера в 2009 году. Ежегодно собираются и обрабатываются данные объемом в десятки и даже сотни петабайт.

В настоящее время проект WLCG объединяет более 150 грид-сайтов, более 300 000 ЦПУ, более 250 Пбайт систем хранения данных на дисках и ленточных роботах. С начала 2012 года до начала сентября на грид-инфраструктуре WLCG было выполнено более 430 миллионов задач обработки и анализа данных с экспериментов на LHC, которые использовали 10.5 миллиардов часов процессорного времени в единицах HEPspec06.

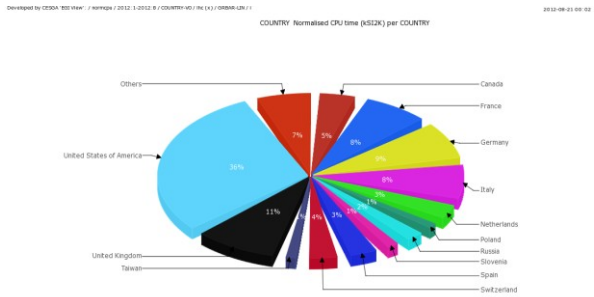


Рис. 1. Распределение процессорного времени проекта WLCG по странам за 2012 год до сентября.

На рис.1 показано распределение процессорного времени по странам WLCG, в котором: США - 36%, Великобритания – 11%, Германия – 9%, Франция и Италия – по 8%, Канада- 5%, Швейцария – 4%, Испания и Нидерланды по 3%, Россия - 2 %.

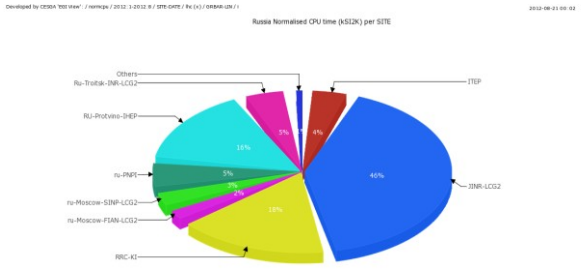


Рис. 2. Распределение процессорного времени проекта WLCG по российским грид-сайтам в течение 2012 года: ОИЯИ (Дубна) – 46%, НИЦ «Курчатовский институт» - 19%, ИФВЭ (Протвино) -16%, ПИЯФ (Гатчина) и ИЯИ (Троицк) – по 5%, ИТЭФ 4%, НИИЯФ МГУ – 3%, ФИАН-2%

На российских грид-сайтах, участвующих в обработке и анализе данных экспериментов на LHC [9,10,12] за 2012 год до начала сентября выполнено 14.5 миллионов задач, которые использовали более 220 миллионов часов процессорного времени в единицах HEPspec06 (из них ОИЯИ более 100 миллионов часов). На рис. 2 показана статистика по российским центрам.

2 Архитектура подсистемы хранения данных

Подсистема управления данными включает три сервиса, поддерживающие доступ к файлам:

- ресурс хранения данных (Storage Element, SE),
- сервис каталогов (Catalog Services, CS),
- планировщик передачи данных (Data Scheduler, DS).

Общая структура каталогов подсистемы управления данными представлена на Рис 3.

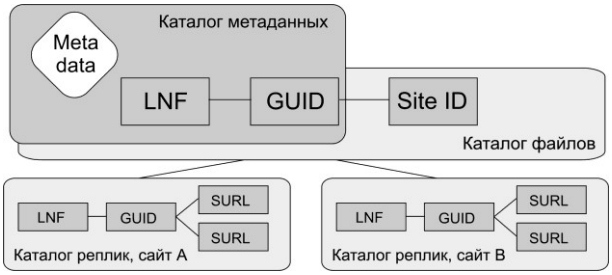
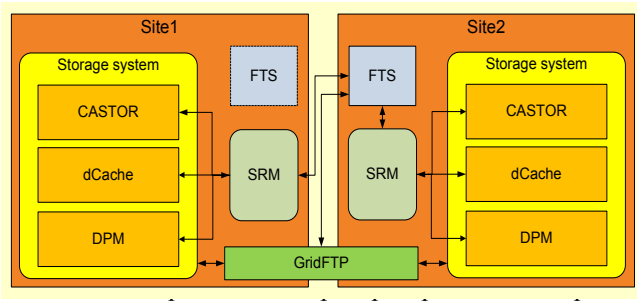


Рис.3 Каталоги данных

Таблицы соответствия между первоначальными названиями файлов (Logical File Name, LFN), уникальными идентификационными номерами (Glbally Unique Identifier, GUID) и названиями файлов на конкретных сайтах (Site URL, SURL) хранятся в каталоге LCG File Catalog (www.gridpp.ac.uk/wiki/LCG_File_Catalog), объединяющем в себе функции файлового каталога и каталога реплик. Сервис каталога файлов и реплик прослеживает наличие файлов с данными на разных ресурсах хранения данных и обеспечивает сопоставление логических имен файлов (LNF) , под которыми к ним обращается пользователь, их уникальным в гриде универсальным именам (GUID), под которыми файлы (возможно в виде нескольких копий – реплик, SURL) хранятся в ресурсных центрах грид-инфраструктуры.

Для организации хранения данных в среде грид используются различные системы и сервисы. Наиболее распространены следующие из них: Castor [5], dCache [6], DPM [7], для взаимодействия с которыми разработан специальный сервис SRM (Storage Resource Manager) [7]. За перемещение данных на физическом уровне отвечает GridFTP— протокол, разработанный в проекте Globus [4]. Система dCache [6] ориентирована на хранение больших объёмов экспериментальной информации. Для доступа к файлам dCache используется собственные протоколы (например, DCAP), gridFTP или любой протокол доступа к файлам.

Для обеспечения необходимой надежности, производительности и организации взаимодействия между остальными сервисами управления данными был создан сервис передачи данных — FTS (File Transfer Service) [11], основные обязанности которого: обеспечение надежных и удобных механизмов передачи файлов типа «точка-точка», контроль и мониторинг передач, распределение ресурсов сайта между различными организациями, управление запросами пользователей. Ежедневно с помощью FTS между различными сайтами передаются сотни тысяч файлов, а объемы составляют сотни терабайт в день. FTS предназначен для надежной пересылки файлов между крупными хранилищами данных, в первую очередь, между центрами уровня Tier0 и Tier1. На рис. 4 представлена схема функционирования сервиса FTS



среде получила распределенная файловая система Xrootd, направленная на высокопроизводительный, масштабируемый и отказоустойчивый доступ к хранилищам данных многих видов.

Доступ к данным основан на масштабируемой архитектуре, протоколе связи, а также наборе плагинов и инструментов. Xrootd обеспечивает удобный для пользователя и быстрый доступ к данным любого вида. Данные должны быть организованны в виде иерархической файловой системы, как пространства имен, основанные на концепции каталога.

Xrootd включает такие сложные функции, как аутентификация/авторизация, интеграция с другими системами, создание новых иерархий и федераций глобально распределенных данных.

3. Развитие средств мониторинга и управления распределенными системами хранения в ОИЯИ

ОИЯИ активно участвует развитии европей-ской и российской грид-инфраструктуры, в первую очередь в проекте WLCG [13]. Ядром этой инфраструктуры является Центральный информационно - вычислительный комплекса (ЦИВК) ОИЯИ, базирующемся на распределенной модели хранения и обработки данных. ЦИВК ОИЯИ организован как единый информационно-вычислительный ресурс, предназначенный для обеспечения всех направлений деятельности Института. В настоящее время вычислительный комплекс ЦИВК состоит из 2582 64-х битных процессоров и системы хранения данных общей емкостью 1800 Тбайт. Счетные ресурсы и ресурсы для хранения данных используются как локальными пользователями ОИЯИ, так и пользователями международных проектов распределенных вычислений, в первую очередь экспериментов на Большом адронном коллайдере (ATLAS, CMS, ALICE). Доступ к данным обеспечивается программным обеспечением dCache и Xrootd, которые обеспечивают доступ к данным как локально, так и глобально для пользователей виртуальных организаций WLCG. Созданы средства мониторинга [17], которые помогают решать задачу эффективного использования системы хранения и балансировки нагрузки на дисковые пулы. Все системы хранения построены с использованием аппаратного механизма RAID6.

Сотрудники ОИЯИ приняли активное участие в развитии глобальных систем мониторинга и управления распределенными хранилищами данных. Представлены наиболее значимые проекты в этом направлении.

3.1. Система мониторинга сервиса передачи файлов FTS

Была разработана системы мониторинга сервиса передачи файлов FTS [11]. Интерфейс системы состоит из нескольких модулей. У пользователей есть возможность начать свою работу с системой непосредственно из интересующего его модуля, либо с главной страницы, на которой пред-ставлены общие отчеты, позволяющие определить состояние сервиса и возможные источники проблем. Система предоставляет возможности получения широкого спектра отчетов, рейтингов, статисти-ческих выкладок и определения коэффициента корреляции для пары ошибок. Практически все отчеты системы мониторинга сервиса передачи данных снабжены перекрестными ссылками, что очень удобно для детализации результатов. В системе реализован механизм оповещения при сбоях, позволяющий администратору сервиса соз-дать свои собственные наборы правил (триггеры), при срабатывании которых будут выполнены определенные действия (отправлены сообщения посредством web-интерфейса, электронной почты, коротких сообщений sms и т.д.). Триггеры можно создавать для каналов передачи, грид-сайтов, хостов и виртуальных организаций. Реализованы три типа триггеров: (1) при превышении числа ошибок определенного уровня, (2) при изменении уровня ошибок более, чем на заданную величину и (3) при превышении процента неудачных передач определенного уровня. Если пользователь работает с триггерами типа 1 и 2, то он может указать идентификационный номер определенной ошибки, чтобы отслеживать только ее развитие. При работе с каналами, грид-сайтами и хостами, пользователь может указать виртуальную организацию для получения необходимых параметров.

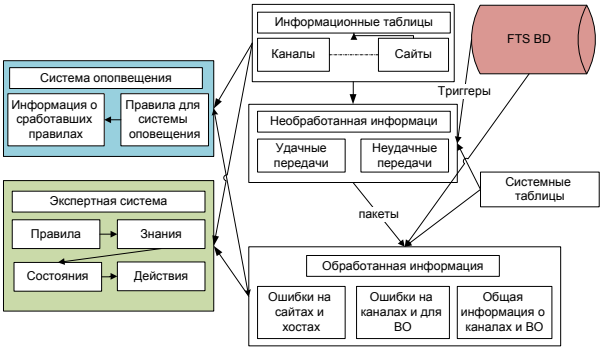


Рис. 5 Модель данных системы мониторинга FTS

Информация о срабатывании триггера может быть получена из специальной таблицы, что существенно упрощает использование механизма оповещений. Благодаря механизму оповещений значительно упрощается работа администраторов сервиса. На рис. 5 представлена модель данных системы мониторинга сервиса FTS

Предоставляется следующая информация о каналах передачи данных сервиса FTS с детализацией по грид-сайтам и виртуальным организациям (выдаваемая информация относится к выбранному пользователем промежутку времени):

- количество передач файлов;
- абсолютное и относительное число успешных и неуспешных передач;
- выявленные причины возникающих ошибок (несколько первых в цепочке) и их количественное соотношение в общем числе ошибок;
- средний размер переданных файлов;
- среднее время передачи;
- средняя скорость передачи данных в канале;
- объем переданных и полученных данных.

3.2. Развитие системы управления распределенными данными эксперимента ATLAS

Система управления распределенными данными DQ2 коллаборации ATLAS отвечает за репликацию, доступ и учет данных на распределенных грид-сайтах, обеспечивающих обработку данных коллаборации. Она также реализует политику управления данными, определенную в вычислительной модели ATLAS.

В 2010 году возникла необходимость разработать новую архитектуру сервиса удаления данных для обеспечения целостности распределенного хранения информации эксперимента ATLAS. Сервис удаления данных один из основных сервисов DQ2. Этот распределенный сервис взаимодействует с различным промежуточным программным обеспечением грид и DQ2 каталогами для обслуживания запросов на удаление [14]. Кроме того, сервис организует балансировку нагрузки, обеспечивая масштабируемость и отказоустойчивость системы DQ2, корректную обработку исключений, возникающих в процессе работы, стратегию повтора операций в случае возникновения отказов. Разработка включала построение нового интерфейса между компонентами сервиса удаления (основанного на технологии веб-сервисов), создание новой схемы базы данных, перестройку ядра сервиса, разработку интерфейсов с системами массового хранения, и развитие системы мониторинга работы сервиса. Сервис разработан, внедрен и поддерживается сотрудниками ОИЯИ. Данные эксперимента ATLAS распределены более, чем на 100 грид-сайтах с общим объемом дискового пространства более 150 петабайт, в котором хранятся сотни миллионов файлов. Недельный объем удаляемых данных составляет 2 Пб (20 000 000 файлов). Созданный сервис обеспечивает целостность хранения информации в географически распределенной среде.

3.3. Система мониторинга центров уровня Tier3 для анализа данных.

Для анализа данных экспериментов БАК стали использоваться разнообразные вычислительные ресурсы (серверы, кластеры, суперкомпьютеры) центров уровня Tier3, которые находятся вне централизованного управления и планирования и на которые не распространяются какие-либо единые требования, касающиеся технических решений. Для этих центров характерно большое разнообразие систем хранения данных и систем пакетной обработки задач. Было выполнено исследование центров уровня Tier3 для систематизации и обеспечения средств интеграции с центрами уровня Tier2. В результате этого исследования было выявлено около 40 различных вариантов конфигураций программно-аппаратных комплексов Tier3. Необходимо было реализовать все варианты комплексов Tier3 для создания дистрибутивов и внедрения системы локального мониторинга для сбора информации о функционировании каждого Tier3 центра [15,16]. Для этой цели в ОИЯИ было разработана архитектура тестовой инфраструктуры на базе виртуальных кластеров, что позволило промоделировать все возможные на данный момент конфигурации Tier3 центров и выработать рекомендации по системе сбора информации для глобального мониторинга Tier3-центров.

Проект глобального мониторинга Tier3 центров (T3mon) направлен на разработку программного комплекса для мониторинга Tier3 сайтов, как с точки зрения локального администратора сайта, так и с точки зрения администратора виртуальной организации ATLAS.

Реализация этого проекта имеет огромное значение для координации работ в рамках виртуальной организации, так как обеспечивается глобальный взгляд на вклад Tier3 сайтов в вычислительный процесс. Схема функционирования основных вариантов Tier3 центров и их взаимодействие с системой глобального мониторинга представлена на рис. 6.

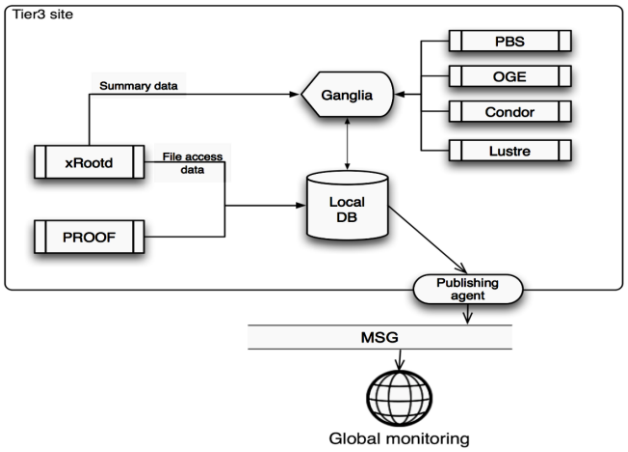


Рис. 6. Схема функционирования основных вариантов Tier3 центров и их взаимодействие с системой глобального мониторинга.

3.4. Глобальная система мониторинга передачи данных в инфраструктуре проекта WLCG.

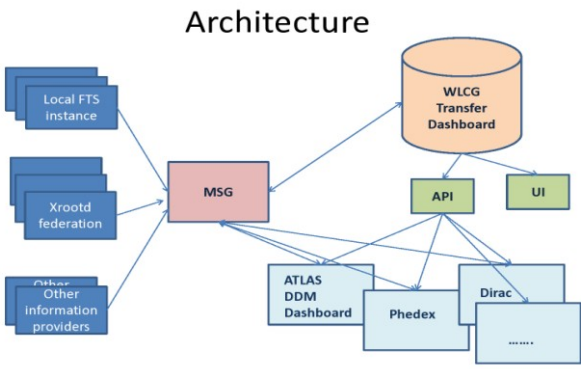


Рис. 7 Архитектура универсальной системы мониторинга передачи файлов

В рамках совместного проекта РФФИ-ЦЕРН «Глобальная система мониторинга передачи данных в инфраструктуре проекта WLCG», разработан прототип универсальной системы мониторинга [18] передачи файлов, способной собирать подробную информацию:

- о каждой передаче файлов (около 1 Петабайта в день),
- независимо от метода осуществления передачи (несколько протоколов и сервисов передачи файлов, FTS, Xrootd),
- уровень ресурсного центра (Tier0, Tier1, Tier2, Tier3)
- принадлежности данных определенной виртуальной организации;
- передавать с высокой степенью надежности собранную информацию в центральное хранилище;
- обрабатывать собранные данные для предоставления различным потребителям;
- предоставлять пользовательские и программные интерфейсы для получения данных.

На рис. 7. представлена архитектура универсальной системы мониторинга передачи файлов в грид-среде проекта WLCG. Система позволяет полностью удовлетворить потребности в информации различных типов пользователей и администраторов инфраструктуры WLCG.

4 Эволюция модели компьютинга и хранения данных БАК

Для обработки данных БАК требуется распределенное управление данными и поддержка очень высоких скоростей передачи огромных массивов данных. В этом направлении постоянно совершенствуются сервисы и программные продукты.

Происходит эволюция инфраструктуры и модели компьютинга экспериментов на БАК:

- переход от иерархической структуры к сетевой, а в идеале к полно-связной, где возможны связи между центрами всех уровней;
- развитие средств распределенного управления данными, поддержка очень высоких скоростей передачи огромных массивов данных;
- создание мощных и разнообразных центров уровеня Tier3 для индивидуального анализа данных;
- развитие и применение средств виртуализации и облачных вычислений (проект «Helix Nebula – научное облако»)

Изменяется и модель размещения данных – осуществлен переход к концепции динамического размещения данных и созданию дополнительных их копий и удалению не используемых копий. Эволюция распределенной инфраструктуры и модели компьютинга постоянно развивается в направлении конвергенции технологий.

4.1 Новые решения и перспективы в обработке «Больших Данных»

Постоянно растущие объёмы научных данных ставят новые задачи перед технологиями распределённых вычислений и Грид [20]. Набирающая размах революция Больших Данных ведёт к открытиям в самых различных областях науки, включая нанотехнологии, астрофизику, физику высоких энергий, биологию и медицину. Недавнее открытие бозона Хиггса на БАК является наилучшим примером этого прогресса. Новые проекты и разработки преображают исследования, основанные на данных, расширяя границы применения Больших Данных и требуя массивной обработки данных новыми методами. При обработке массивов данных объемами порядка петабайтов, современные учёные имеют дело не с отдельными файлами, а с наборами данных. В результате, единицей обработки петабайтного массива в системах Грид становится вычислительная задача, состоящая из многих подзадач. Разбиение обработки на отдельные задачи ведёт к созданию контрольных точек с высокой степенью детальности, что подобно разбиению большого файла на небольшие пакеты TCP/IP при передаче по сетям. Передача больших файлов

маленькими пакетами обеспечивает отказоустойчивость путём повторной пересылки утерянных пакетов. Подобным же образом обеспечивается отказоустойчивость обработки данных в системах Грид: прерванные в результате сбоев задания могут быть автоматически перезапущены, чем достигается качество обработки. С другой стороны, необходимость повторного исполнения отдельных заданий ведёт к непредсказуемому снижению эффективности использования вычислительного ресурса. Опыт компьютерной обработки данных на БАК позволяет изучить связь между отказоустойчивыми решениями и эффективностью, и разработать новые масштабируемые подходы к применению технологий Грид для обработки объемов данных, превышающих сотни петабайт.

Литература

[1] Ian Foster and Carl Kesselman, “The Grid: Blueprint for a New Computing Infrastructure,” Morgan Kaufmann, 1999, http://www.mkp.com/grids

[2] L. Robertson, J. Knobloch. LHC Computing Grid Technical Design Report.CERN-LHCC-2005-023.[Электронный ресурс]. 2005. : http://cdsweb.cern.ch/record/840543/files/lhcc-2005-024.pdf

[3] Worldwide LHC Computing Grid http://lcg.web.cern.ch

[4] Проект Globus: http://www.globus.org

[5] Сайт проекта Castor: http:// www.castor.org

[6] Сайт проекта dCache : http:// www. dcache.org

[7] Сайт проекта DPM: www.gridpp.ac.uk/wiki/Disk_Pool_Manager

[8] Сайт проекта SRM (Storage Resource Manager) (www.gridpp.ac.uk/wiki/SRM).

[9] В. Кореньков, Е. Тихоненко. Концепция GRID и компьютерные технологии в эру LHC // Физика элементарных частиц и атомного ядра, т. 32, вып.6, 2001, с.1458-1493.

[10] В. Ильин, В. Кореньков, А.Солдатов. Российский сегмент глобальной инфраструктуры LCG, Открытые системы // №1, 2003, с. 56-60.

[11] В. Кореньков, А. Ужинский. Система мониторинга сервиса передачи данных (FTS) проекта EGEE/WLCGВычислительные методы и программирование: Новые вычислительные технологии, том 10, 2009, с.96-100.

[12] В. Ильин, В. Кореньков. Компьютерная грид-инфраструктура коллаборации RDMS CMS// в сборнике «В глубь материи: физика XXI века глазами создателей экспериментального комплекса на Большом адронном коллайдере в Женеве» – М: Этерна, 2009, с. 361-372.

[13] V. Korenkov. GRID ACTIVITIES AT THE JOINT INSTITUTE FOR NUCLEAR RESEARCH // in Proc. of the 4th Intern. Conf. «Distributed Computing and Grid-Technologies in Science and Education, GRID-2010», ISBN 978-5-9530-0269-1, Dubna, 2010, p. 142-147.

[14] D. Oleynik, A. Petrosyan, V. Garonne, S. Campana, ATLAS DQ2 Deletion Service, труды конференции CHEP’2012, Нью-Йорк, США, 21-25 мая 2012

[15] J. Andreeva, D. Benjamin, S. Campana, A. Klimentov, V. Korenkov, D. Oleynik, S. Panitkin, A. Petrosyan Tier-3 Monitoring Software Suite (T3MON) proposal //ATL-SOFT-PUB-2011-001, CERN, 2011

[16] S. Belov, I. Kadochnikov, V. Korenkov, M. Kutouski1, D. Oleynik, A. Petrosyan on behalf of the ATLAS Collaboration. VM-based infrastructure for simulating different cluster and storage solutions used on ATLAS Tier-3 sites // ATL-SOFT-PROC-2012-057, 2012.

[17] В. В. Кореньков, В. В. Мицын, П. В. Дмитриенко Архитектура системы мониторинга центрального информационно-вычислительного комплекса ОИЯИ // Информационные технологии и вычислительные системы, 2012, №3, стр. 3-14

[18] Портал по развитию грид-технологий в ОИЯИ http://grid.jinr.ru/

[19] Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC Author(s) CMS Collaboration 31 Jul 2012. - 59 p. Note Comments: Submitted to Phys. Lett. B In Press

[20] Сайт международной конференции GRID-2012 Доклад Ваняшина А. В. Advancements in big data processing http://grid2012.jinr.ru/docs/VaniachineBigData.pdf

Using of grid-infrastructure for distributed processing and storage of enormous amount of data (hundred’s of PB)

Vladimir Korenkov
Joint Institute for Nuclear Research

A short review of the development of the global Grid-infrastructure within the WLCG project (Worldwide LHC Computing Grid or Worldwide Grid for the Large Hadron Collider) is presented. Particular attention is paid to the architecture of the distributed storage of enormous data volumes. A brief information is given on the projects devoted to the development of tools of monitoring and management of the distributed file systems performed at JINR: a system of monitoring the FTS file transfer service, development of the control system for distributed data of the ATLAS experiment, a system of monitoring the centers of Tier3 level for data analysis, a global system of monitoring data transfer in the infrastructure of the project WLCG. Advanced solutions and perspectives in the processing of "Big Data” are indicated.