

РОССИЙСКИЙ СЕГМЕНТ ГЛОБАЛЬНОЙ ИНФРАСТРУКТУРЫ LCG (LHC Computing GRID)

Ильин В.А., НИИЯФ МГУ, ilyin@sinp.msu.ru

Кореньков В.В., ОИЯИ, korenkov@cv.jinr.ru

Солдатов А.А., РНЦ «Курчатовский институт», saa@kiae.su

В статье описывается проект создания мировой инфраструктуры типа GRID, предназначенной для обработки и анализа экспериментальных данных с Большого адронного коллайдера (LHC) – крупнейшего ускорителя в области физики элементарных частиц. Многие ведущие центры и университеты мира (в том числе в России) активно участвуют в этом проекте, уникальном как по масштабам обрабатываемой информации, так и с компьютерно-технологической точки зрения. Использование GRID - технологий здесь не дань моде, а жизненная необходимость. Уникальность проекта состоит ещё и в том, что мы не можем ждать, когда GRID - технологии будут доведены до уровня применения в прикладных областях. Уже осенью 2003 года первый вариант глобальной инфраструктуры GRID должен выполнять полномасштабное моделирование экспериментов на LHC

Введение

Осенью 2001 года в ЦЕРН принят проект создания глобальной информационно-вычислительной инфраструктуры, основной задачей которой будет обработка, хранение и анализ экспериментальных данных с Большого адронного коллайдера (далее мы будем использовать аббревиатуру БАК или LHC).

ЦЕРН – это Европейский центр ядерных исследований, расположенный недалеко от г. Женевы на границе между Швейцарией и Францией. На протяжении почти 50-ти лет ЦЕРН является одной из крупнейших мировых лабораторий в области физики высоких энергий – науке, изучающей структуру и взаимодействие элементарных частиц. Основным методом все эти годы было ускорение элементарных частиц до максимально больших энергий и затем их столкновение. В результате такого столкновения происходит что-то вроде микровзрыва, после которого во все стороны разлетаются «осколки» - множество различных элементарных частиц. Вот по «фотографиям» этих разлетающихся «осколков» и определяются характеристики взаимодействия элементарных частиц. В этой статье мы не собираемся ни описывать эти эксперименты, ни перечислять выдающиеся открытия, сделанные в ЦЕРН или в других центрах по физике высоких энергий (эту информацию можно найти на сайте <http://www.cern.ch>). Для лучшего понимания Но к некоторым словам выше по тексту стоит дать

комментарии, которые будут иметь полезное отношение к дальнейшему изложению.

Общая характеристика вычислительных задач БАК

Во-первых, заметим, что слова «*максимально большие энергии*» означают *максимально достижимые при применении всех возможных технологий, имеющихся в мире на данный момент*. Причем часто при этом создаются новейшие технологии, которые в дальнейшем находят применение в других областях науки и в промышленности. Одним из современных примеров могут служить источники синхротронного излучения, которые широко применяются, например, в физике твердого тела и материаловедении.

Как следствие повышения энергии сталкивающихся частиц (а за 40 лет она возросла в миллион раз!) увеличиваются размеры ускорителя и детекторов - приборов, «фотографирующих» акт столкновения. Если первые ускорители могли размещаться на рабочем столе физиков, то в настоящее время это уже гигантские установки. Например, ускоритель ЛЭП, которые работал в ЦЕРНе до 2000 года, располагался в подземном туннеле радиуса 27 км на глубине нескольких десятков метров. А новый ускоритель БАК (Большой адронный коллайдер), который сооружается в том же туннеле и начнет работу в 2007 году, будет иметь самую большую в мире систему сверхпроводящих магнитов. Соответственно и детекторы представляют собой огромные установки. Например, CMS - один из четырех детекторов БАК, будет иметь вес 12.5 тысяч тонн. Причем, весь объем детектора будет заполнен разнообразной электроникой, регистрирующей разлетающиеся частицы. Слово «*фотографирование*», использованное выше, конечно же, сейчас не означает фотографирование в обычном смысле. Процесс регистрации пролетающих частиц происходит с помощью различных физических приборов, информация с которых снимается электроникой. Эта информация проходит несколько этапов обработки в реальном времени, и затем записывается в базу данных. Соответственно увеличению и усложнению регистрирующих приборов, рос и объем информации, снимаемой электроникой для каждого акта столкновения (далее мы будем использовать термин *событие* для совокупной информации, полученной для данного акта столкновения и записанной в базу данных после обработки в реальном времени). Например, для ускорителя БАК объем события будет составлять 1-2 Мбайта для детекторов общего назначения (ATLAS и CMS), и до Гбайта для специализированного детектора ALICE.

Объем в Мбайты конечно не может удивить современника – такого же масштаба файлы можно получать с помощью цифровых фотоаппаратов. Уникальным является частота поступления событий – для БАК это будет скорость порядка 100-200 событий в секунду для каждого детектора. Дело в том, что по одной «фотографии» невозможно сделать никаких выводов о взаимодействии элементарных частиц. Это квантовое взаимодействие, а значит, получаемая информация имеет статистический характер. Другими

словами, необходимо обработать много событий, чтобы можно было сделать сколько-нибудь достоверный вывод. В зависимости от природы изучаемого явления и условий его наблюдения в экспериментах на БАК будет требоваться набор статистики (набор событий) от нескольких месяцев и до нескольких лет. Рабочий год ускорителя составляет около 100 дней. Поэтому для проектирования и моделирования работы ускорителя в расчетах исходят из продолжительности рабочего года, равной 10^7 сек. Таким образом, объем годовой базы данных по событиям для ускорителя БАК будет больше 10^9 Мбайт для каждого из детекторов общего назначения, ATLAS и CMS. Примерно такого же объема будут базы данных для специализированных детекторов ALICE и LHCb. 10^9 Мбайт означает миллион Гбайт или тысяча Тбайт. Следующей единицей является *Пбайт (Петабайт)* – тысяча Тбайт. В итоге, совокупный объем годовых данных, которые будут получаться, начиная с 2007 года, в экспериментах на ускорителе БАК, будет составлять десятки Пбайт.

Такой объем баз данных также может не удивить экспертов. В настоящее время объем корпоративных баз данных крупнейших фирм и организаций достигает десятков Тбайт. Понятно, что к 2007 году базы данных масштаба Пбайта будут, если не рядовым явлением, то технологически не уникальным. Проблема физики высоких энергий в целом и проекта БАК в частности состоит в том, что в настоящее время ни в какой другой области науки и промышленности не стоит вопрос о разработке технологий создания баз данных масштаба десятков Пбайт, и операций с ними. А в физике высоких энергий уже сейчас объем баз данных действующих экспериментов (например, ВаВаг в США и BELLE в Японии) составляет величину порядка Пбайта. А для экспериментов на строящемся в ЦЕРНе ускорителе БАК уже сейчас требуется создание баз данных по моделирующим событиям порядка сотен Тбайт, а в 2004 году порядка Пбайта. Наконец уже сейчас необходимо провести проектирование компьютерной системы, с помощью которой в 2005-2007 годах физики проведут полномасштабное моделирование будущих экспериментов, а, начиная с 2007 года, будет идти процесс создания баз экспериментальных данных и обеспечиваться их физический анализ.

Наконец, выше неоднократно использовался термин «*международный*» в применении к проекту БАК. Уже более тридцати лет основной организационной формой экспериментов в физике высоких энергий является *международное сотрудничество*. Это означает, что данный эксперимент проводится на всех этапах, включая проектирование и создание соответствующих установок (в нашем случае ускорителей и детекторов), сообществом лабораторий и институтов со всего мира, согласившихся работать в едином коллективе для достижения определенных научных целей. И если в 70-х годах международные коллективы включали несколько лабораторий и институтов, то в 90-е годы, например, в каждом из четырех экспериментов на ускорителе ЛЭП участвовали сотни ученых из десятков лабораторий и институтов мира. А в каждом из международных коллективов (далее будем использовать термин *коллаборация*) по проектированию и

созданию детекторов на ускорителе БАК уже сейчас участвуют несколько тысяч физиков из сотен лабораторий и институтов мира.

Это, само по себе уникальное явление в научном мире, для проектирования компьютерной системы БАК является, возможно, наиболее сложным обстоятельством. Действительно, надо будет обеспечить доступ к базам данных сотен активных пользователей. *Активный* – означает, что такой пользователь будет практически ежедневно работать с полной базой данных или с существенной ее частью (не вдаваясь в детали, напомним, что достоверную физическую информацию можно получать только в результате анализа статистически значимого количества событий). Таким образом, возникает дилемма: либо 1) базы данных создаются и хранятся в одном месте (значит в ЦЕРНе), а пользователи приезжают для работы в ЦЕРН или работают в своих институтах за удаленным терминалом, подключенным через глобальные линии связи к ЦЕРНу; либо 2) базы данных как-то передаются в институты, где работают физики. До настоящего времени всегда реализовывался первый вариант, хотя периодически обсуждалась целесообразность и второго подхода. Например, так было в конце 80-х годов при проектировании компьютерных систем для экспериментов на ускорителях ЛЭП в ЦЕРНе и Теватрон в Фермилабе (США). Однако выяснилось, что передавать полные базы данных даже в несколько институтов мира и организовывать их хранение там очень дорого, а для создания распределенных баз данных и организации работы с ними не было технологической основы. В общем, как с финансовой стороны, так и технологической, в предыдущие годы более обоснованным оказывался первый подход к построению компьютерной системы – по принципу «пользователь к данным».

Здесь необходимо сделать важное замечание. В конце 80-х годов в физике высоких энергий интенсивно обсуждалась также проблема эффективной организации обмена текстовой информацией в рамках широких международных коллабораций. Тогда это было связано с началом работы ускорителей ЛЭП в ЦЕРНе, Тэватрон в Фермилабе и SLC в СЛАКе (США), HERA в ДЭЗИ (Германия) и TRISTAN в КЕК (Япония). Вот для такой задачи, гораздо более узкой и простой, нежели обработка и анализ данных, удалось разработать технологию глобального поиска и обмена информацией, которая получила название WWW (World Wide Web – всемирная паутина). В контексте нашего обсуждения можно сказать, что удалось решить задачу глобально распределенного создания общих документов, без собирания авторов в одном месте. Важно отметить ключевой момент революции WWW – стандартизация протокола обмена/транспортировки текстовой информации (HTTP) и разработка метаязыка для создания текстовых файлов (HTML). В результате были созданы *браузеры* – пользовательские программы по использованию этой технологии, которые могут работать на компьютерах разных производителей, на разных операционных системах и понимать друг друга! Отметим, что *стандартизация*, в данном случае, означала согласие мирового сообщества использовать единый протокол и метаязык.

Аналогично и технологии GRID (которые упрощенно можно характеризовать как глобально распределенные вычисления) также должны основываться на принятии мировым сообществом соответствующих стандартов. В этом смысле иногда говорят, что GRID это следующий этап WWW.

Распределенная иерархическая модель

Прежде чем мы начнем содержательно говорить о применении GRID технологий в построении компьютерной системы БАК, необходимо рассказать о результатах моделирования этой системы, проведенной в конце 90-х годов в рамках международного проекта MONARC (<http://monarc.web.cern.ch/MONARC/docs/phase2report/Phase2report.pdf>).

Основным выводом была *распределенная модель* архитектуры системы - 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 для анализа в *региональные центры*.

Таким образом, был обоснован иерархический принцип организации информационно-вычислительной системы БАК, предполагающей создание центров разных ярусов (Tier's):

$Tier0(CERN) \Rightarrow Tier1 \Rightarrow Tier2 \Rightarrow Tier3 \Rightarrow$ компьютеры пользователей
Ярусы должны различаться как по масштабу вычислительных и архивных ресурсов, так и по выполняемым функциям:

<i>Tier0 (ЦЕРН)</i>	<i>первичная реконструкция событий, калибровка, хранение копий полных баз данных</i>
<i>Tier1</i>	<i>полная реконструкция событий, хранение актуальных баз данных по событиям, создание и хранение наборов анализируемых событий, моделирование, анализ</i>
<i>Tier2</i>	<i>репликация и хранение наборов анализируемых событий, моделирование, анализ</i>
<i>Tier3</i>	<i>кластеры отдельных исследовательских групп</i>

Планируется создание нескольких (4-6) Tier1 центров по каждому из 4-х экспериментов БАК. Количество Tier2 центров планируется в количестве порядка 25-ти для каждого из экспериментов. Было предложено примерно равное распределение всех компьютерных ресурсов по ярусам:
 $Tier0 = \sum Tier1 = \sum Tier2$.

В ЦЕРНе будет создан комплекс Tier0+Tier1 в виде единой вычислительной системы, совместно используемой (разделяемой) всеми четырьмя коллаборациями БАК. По коллаборациям масштаб ресурсов информационно-вычислительной системы БАК на конец 2006 года

оценивается следующим образом (первая цифра по эксперименту в целом, вторая по соответствующей части в регионах вне ЦЕРНа):

	ALICE	ATLAS	CMS	LHCb
CPU (KSI-95)	1758/934	1944/1254	2180/1565	925/700
Disk (Пбайт)	1.6/1.1	2.6/2.2	4.0/3.2	1.1/0.77
Tape (Пбайт)	4.7/1.5	20.0/11.0	10.5/6.5	2.8/1.6

Здесь и далее используются следующие сокращения для характеристики компьютерных ресурсов:

- *CPU* - процессорные мощности вычислительных ферм (кластеров), в качестве единицы измерения здесь использовано KSI-95 (1000 SPECint95/s, <http://www.spec.org/>). Для сравнения – скорость соответствующих вычислений на персональном компьютере на процессоре Pentium III и частотой 1 GHz оценивается примерно в 40-50 SI-95,
- *Disk* – пространство для активного хранения данных на жестких дисках в составе файловых серверов,
- *Tape* - пространство для архивного хранения данных на ленточных носителях, с использованием автоматизированных библиотечных систем (роботов).

Приведенная выше оценка была сделана в 2000 году на основе рекомендаций проекта MONARC. Конечно, в результате развития проекта создания информационно-вычислительной системы БАК происходит уточнение его параметров и архитектуры. Можно отметить, что, если оценивать по 2002 году, то для каждого эксперимента БАК вычислительные мощности должны включать порядка 10 тысяч двухпроцессорных компьютеров Pentium-IV/2GHz. Если к концу 2006 году мощность процессоров вырастет на порядок (что прогнозируется), то для каждого из экспериментов БАК потребуется вычислительная мощность, соответствующая кластеру из примерно тысячи двухпроцессорных компьютеров. Такой уровень представляется вполне реалистичным для 2006-2007 гг., тем более что планируется 2/3 этой мощности размещать в региональных центрах. Основная проблема, как мы уже отмечали выше – организация когерентной работы системы региональных центров с распределенной по этим же центрам базой данных (событий).

Необходимо пояснить, почему мы говорим об обычных персональных компьютерах. Дело в том, что все операции с отдельным событием (реконструкция, преобразование в различные форматы, наконец – анализ) выполняются на персональных компьютерах уже сегодня достаточно быстро – за минуты. Важным обстоятельством является то, что операции с данным событием выполняются независимо от других событий. Таким образом, нет необходимости выполнять эти операции одновременно на нескольких

процессорах. Соответственно, вычислительные кластеры можно строить на простом коммуникационном оборудовании и с примитивной архитектурой - один компьютер распределяет задачи по вычислительным узлам, которые не связаны между собой. В физике высоких энергий принято называть такие вычислительные кластеры *фермами*. В результате, фермы можно строить полностью на основе оборудования массового производства, что сильно снижает финансовые затраты (в 2-3 раза минимум).

В проекте MONARC были даны оценки также и на пропускную способность линий связи в рамках жесткой иерархической распределенной модели. Согласно этим оценкам, линии связи Tier1-Tier0 должны быть порядка 1.5 Гбит/сек для каждого из экспериментов ALICE, ATLAS и CMS, и 0.3 Гбит/сек для эксперимента LHCb. Соответственно, линии связи Tier2-Tier1 должны быть порядка 0.622 Гбит/сек. Конечно, эти оценки также подвергаются изменениям со временем. В частности, применение GRID технологий может существенно изменить спектр требований. И обратно – успех GRID инициативы во многом зависит от достижения определенно высоких параметров используемых каналов передачи данных. Поэтому в настоящее время указывается порядок мощности требуемых линий связи – 1-2 Гбит/сек для каждого эксперимента для линий связи ЦЕРНа с региональными центрами, а также между основными региональными центрами.

Наконец, для завершения описания вычислительной задачи БАК необходимо отдельно обсудить задачу моделирования событий. Эта задача включает в себя генерацию акта столкновения элементарных частиц, симулирование отклика регистрирующей аппаратуры детекторов, симулирование обработки информации в режиме реального времени, реконструкцию события и его запись в базу данных. Моделирование необходимо по ряду причин. Одна из них состоит в тестировании работы создаваемых алгоритмов обработки данных, и другого программного обеспечения. Необходимо также проверить насколько проектируемый эксперимент способен дать статистически значимые результаты для тех или иных физических явлений с учетом систематических и статистических ошибок, возникающих от всех подсистем детектора и процедур обработки данных. И в том и другом случае требуется количество моделируемых событий по порядку величины то же, что и годовая статистика при работе ускорителя. Более точно, уже в 2004-2005 годах требуется создание баз данных моделирующих событий на уровне не меньше 20% от годовой статистики 2007 года. А в 2005-2006 годах на 50% уровне. Если учесть, что функция создания моделирующих событий в основном отдается в региональные центры, то уже в 2003 году требуется проектирование распределенной информационно-вычислительной системы сложности, сравнимой с той системой, которая будет создаваться к началу работы ускорителя в 2007 году.

Роль GRID технологий

В проекте MONARC была предложена достаточно обоснованная распределенная модель жесткой иерархии региональных центров. Трудно сказать сейчас, насколько эта модель была бы эффективной при функционировании информационно-вычислительной системы БАК, если бы она так и была бы создана. Но в 1999 году внимание разработчиков этой системы привлекла концепция GRID, первоначально предложенная в академической среде в США. Мы не будем здесь останавливаться на изложении этой концепции. Читатель может найти соответствующий материал в других статьях этого выпуска журнала «Открытые системы». Мы также в конце статьи даем ссылку на материалы Международной конференции ACAT'2002, где можно найти дополнительную информацию. В этой же главе мы обсудим роль GRID технологий для информационно-вычислительной системы БАК.

Внимательный читатель мог заметить, что иерархическая модель проекта MONARC имеет структуру дерева. В частности, актуальная база данных по событиям (то есть та, с которой работают пользователи) полностью хранится в каждом из Tier1 центров. А Tier2 центры, в сущности, являются промежуточными станциями в процессе обращения пользователей к актуальным базам данных. Не предусматривается никаких связей между Tier1 центрами, тем более между Tier2 центрами. Да, в общем-то, в такой модели горизонтальные связи и не нужны. В результате появляется в большой степени дублирование баз данных. А проблема столкновения многих пользовательских задач в одном центре переводится на региональный уровень.

Применение GRID технологий позволит разрешить эти две проблемы. Действительно, актуальные базы данных можно хранить распределено – по всем Tier1 центрам, а возможно и по части Tier2 центров. В этом случае пользователь запускает свою задачу в сеть (grid) этих региональных центров, которая обходит их, обрабатывая необходимый набор событий, а результаты (готовый материал для конечного анализа) будет отсылаться обратно пользователю, или в какой-то близко расположенный или хорошо доступный центр.

Точно такой же эффект даст применение GRID технологий и в задаче создания баз данных. В этом случае роль пользователя выполняет сама коллаборация, которая запускает задачи, например, реконструкции событий на выполнение в сеть (grid) региональных центров, обладающих значимыми вычислительными ресурсами (Tier1 и Tier2 центры).

Наконец, задача создания и хранения копий баз данных также может решаться с помощью GRID технологий на той же основе.

Подчеркнем другие очевидные дивиденды применения GRID технологий. Во-первых, мы получаем возможность более эффективного использования компьютерных ресурсов, задействованных для экспериментов на БАК. Действительно, одни и те же ресурсы можно использовать в разное время под разные задачи, в том числе и предоставлять пользователям из других

регионов, когда в данном регионе нагрузка падает. Последнее может происходить, например, в связи с неравномерностью загрузки в дневное и ночное время (что актуально для задач анализа) и несовпадением этих периодов в регионах, расположенных в разных часовых поясах. Приведем еще один пример в этом направлении. В эксперименте ALICE основные задачи будут связаны с анализом событий при столкновении ядер тяжелых элементов (например, свинца), в то время как для остальных экспериментов основные задачи связаны с событиями при столкновении протонов. Конечно, сеансы работы БАК с разными пучками сталкивающихся частиц (протонов или тяжелых ядер) будут разнесены по времени. Поэтому, использование GRID технологий может обеспечить мобилизацию компьютерных ресурсов гораздо большего числа региональных центров в периоды работы БАК с пучками тяжелых ядер, чем в случае жесткой привязки региональных центров к данному эксперименту.

Далее, применение GRID технологий может повысить надежность хранения данных, так как в этом случае базы данных распределены между региональными центрами. Можно предусмотреть хранение 2-3 копий каждого события в такой распределенной базе, причем каждое событие будет актуально, то есть может использоваться в пользовательских задачах. В таком случае можно будет совместить задачу повышения надежности хранения данных и оптимизацию доступа к ним пользователям.

Этими примерами мы ограничимся, а в заключение этой главы подчеркнем два момента. Во-первых, описанные выше дивиденды очевидно привлекательны, но на настоящий момент GRID технологии еще не разработаны настолько, чтобы с уверенностью можно было на них рассчитывать. Правильнее говорить, что наши примеры являются описаниями того, что данная прикладная область требует от разработчиков GRID технологий.

Во-вторых, применение GRID технологий не перечеркивает результаты проекта MONARC. Вернее, соответствующие выводы о распределенной иерархической модели находят свое дальнейшее развитие. Действительно, иерархическая структура региональных центров остается, в основном, в отношении к выполняемым базовым функциям. Если говорить несколько упрощенно, то теперь Tier1 это тот региональный центр, в котором проходят последние этапы создания актуальных баз данных и их хранение (хотя бы и частичное). А Tier2 это такой региональный центр, в котором производятся моделирующие события, и в котором сосредоточены компьютерные ресурсы значимые для всей системы в целом.

Проект LCG

Как мы уже говорили в начале статьи, осенью 2001 года принят пятый проект БАК, основная задача которого состоит в проектировании и создании

информационно-вычислительной системы. Этот проект получил название “*LHC Computing GRID*” (*LCG*), тем самым подчеркивается особая роль GRID технологий. В проекте LCG можно выделить две компоненты – ресурсы и программное обеспечение. К первой компоненте собственно и относятся вопросы архитектуры системы региональных центров, по которой принята распределенная иерархическая модель.

Что касается программного обеспечения, то оно, в свою очередь, подразделяется на две части. В первую входит прикладное программное обеспечение, специфичное для каждого из четырех экспериментов БАК. Например, программы симуляции отклика регистрирующей аппаратуры детекторов, программы реконструкции треков частиц и др. Во вторую часть входят программы и пакеты общего для всех четырех экспериментов назначения. Например, это программы автоматической инсталляции прикладного ПО в региональных центрах, иерархические файловые системы для организации хранения данных в роботизированных библиотеках с автоматической подкачкой затребованных файлов на дисковые массивы и др. Такое ПО называют *общими решениями (common solutions)*. Кстати, можно сказать, что система региональных центров также является общим решением, так как каждый из четырех экспериментов БАК будет использовать ресурсы распределенной иерархической системы региональных центров, одни и те же или разные (в некоторых региональных центрах), но организованные по одной схеме и с использованием общих технологических решений.

Проект LCG состоит из двух фаз. Первая фаза должна завершиться к 2005 году созданием полномасштабного прототипа и разработкой проекта рабочей системы (LCG TDR - Technical Design Report). Вторая фаза – это создание собственно рабочей информационно-вычислительной системы БАК, готовой к обработке и анализу экспериментальных данных на момент начала их поступления в 2007 году.

На 2003 год выделен подпроект LCG-1. Поставлена задача создания к осени 2003 года первой инфраструктуры типа GRID (прототипа), на которой в начале 2004 года будут проведены первые массовые вычислительные работы по созданию баз моделирующих событий.

Таким образом, необходимо уже сейчас (в начале 2003 года) определить тот состав GRID программного обеспечения (GRID middleware), который может быть использован в поставленной производственной задаче LCG-1. К сожалению, ни один из проектов по разработке такого программного обеспечения не вышел уровень, приемлемый для данного приложения, прежде всего в отношении надежности и устойчивости длительной работы.

В настоящее время в первой рабочей группе проекта LCG вырабатываются соответствующие рекомендации. Скорее всего, на этапе LCG-1 будет использован пакет VDT с добавленными GRID сервисами высокого уровня, созданными в европейском проекте EU DataGRID (EDG, <http://eu-datagrid.web.cern.ch>). Пакет VDT (Virtual Data Toolkit) разработан в американских GRID проектах: PPDG – *The Particle Physics Data Grid* (<http://www.ppdg.net/>), GriPhyN – *Grid Physics Network*

(<http://www.griphyn.org/>), и iVDGL – the *International Virtual Data Grid Laboratory* (<http://www.ivdgl.org/>). Этот пакет представляет собой набор надстроек над библиотекой инструментальных средств GLOBUS, позволяющих реализовывать распределенную вычислительную систему, но практически без каких либо GRID сервисов. Он также включает в себя пакет Condor/Condor-G, который используется в качестве распределенной системы запуска заданий в пакетном режиме. Из проекта EDG планируется взять *ресурс-брокер* (обеспечивающий сервис по распределению заданий), информационная служба, replica catalog и некоторые другие разработки.

В качестве основы промежуточного программного обеспечения для этих проектов выбран набор инструментальных средств Globus (<http://www.globus.org>), который *de facto* стал стандартом.

Каждая компонента из этого списка достаточно хорошо разработана и отлажена. Однако в единый пакет они пока еще не объединены, и это представляет собой одну из срочных задач для проекта LCG.

Планируется, что в начале июня пакет программного обеспечения инфраструктуры LCG-1 будет заморожен с тем, чтобы работы по созданию инфраструктуры завершались в стабильных и неизменяемых условиях. В июле 2003 года должен заработать прототип LCG-1. Контрольный тест – совместная непрерывная работа в течение недели 4-5 региональных центров. Конечно, в этом тесте будут задействованы достаточно малые ресурсы центров, по 3-5 вычислительных узла. Следующим этапом, осенью 2003 года, будет непрерывная работа прототипа в течение месяца, включая 2-3 центра яруса Tier2 и 5-6 Tier1 центров. В начале 2004 года начнется массовый перевод ресурсов участвующих региональных центров в инфраструктуру LCG-1. Соответственно этому процессу в экспериментах БАК начнутся сеансы массового производства баз данных моделирующих событий с использованием построенной инфраструктуры, то есть с интенсивным использованием GRID технологий.

Схема объединения ресурсов тех региональных центров, на основе которых будет создаваться инфраструктура LCG-1, составляет основную задачу второй рабочей группы. Составляется график, по которому ежемесячно, начиная с февраля 2003 года, будут подключаться Tier1 центры в Италии, Франции, Великобритании, Германии, США и др. С мая-июня начнут подсоединяться Tier2 центры, в частности и Россия.

В третьей рабочей группе обсуждаются вопросы обеспечения безопасности работы в создаваемой инфраструктуре, в частности сертификация и авторизация пользователей. В четвертой рабочей группе определяются стандарты конфигурации ресурсов в региональных центрах, в частности операционная система (скорее всего Linux 7.3), организация системного администрирования, инсталляции программного обеспечения и другие вопросы организации функционирования региональных центров в инфраструктуре LCG-1. Наконец в пятой рабочей группе обсуждается пользовательский интерфейс. Дальнейшие детали можно найти на сайте проекта LCG – <http://www.cern.ch/lcg>.

В заключение этой главы отметим важнейшую черту проекта LCG. Это *производственный* проект, в нем не предполагается осуществлять собственные научные или технологические разработки. Специальная группа должна будет постоянно оценивать разработки других проектов, в том числе и в части GRID программного обеспечения, и далее возможность и целесообразность их применения в LCG. В положительном случае соответствующие разработки передаются в группу тестирования и далее в группу внедрения. Если же выяснится, что требуется создание какого-то нового ПО, возможно создание группы для разработки технического задания, которое необходимо передавать на реализацию в какой-то научно-технологический проект. В качестве стартового пакета, как мы уже отмечали, выбрана комбинация базового пакета VDT и набора сервисов высокого уровня из проекта EDG. А для пакетирования этого набора ПО создается специальная группа разработчиков.

Участие России в Европейском проекте EU DataGRID

С начала работы проекта EDG (EU DataGRID), то есть с января 2000 года, российские институты по физике высоких энергий приняли участие в этих работах. Здесь мы остановимся в основном на участии в шестом рабочем пакете, основными задачами которого являются глобальные испытания разрабатываемого GRID ПО и демонстрация его работоспособности в полигонных условиях (WP6 Testbed and Demonstration). В этих работах принимали участие в основном следующие институты: ИТЭФ (Москва), ИФВЭ (Протвино), НИИЯФ МГУ (Москва) и ОИЯИ (Дубна). В отдельных работах принимали участие сотрудники ПИЯФ РАН (Гатчина), РИЦ «Курчатовский институт» и ИПМ им. Келдыша РАН. Кроме того, в десятом рабочем пакете (WP10 Biology Applications) принимали участие российские биологические институты и АНО «Наука и Общество». В некоторых разработках по другим (собственно научно-технологическим) рабочим пакетам принимали участие отдельные сотрудники из ИТЭФ, ИФВЭ, НИИЯФ МГУ и ОИЯИ.

Основным результатом нашего участия в проекте EDG стало получение опыта работы с новейшим программным обеспечением типа GRID. Полученный опыт найдет непосредственное применение в практической работе по созданию российского сегмента LCG-1 уже в этом году. Именно этот опыт позволяет говорить, что наши коллективы готовы к участию в уникальном инфраструктурном проекте мирового масштаба, каковым является проект LCG.

Итак, впервые в России созданы виртуальные организации типа GRID (VO – Virtual Organizations) для решения конкретных прикладных задач. Виртуальные организации являются основной формой объединения ресурсов, уже имеющихся в GRID. Они позволяют подключать данный ресурс к решению разных прикладных задач разными группами

пользователей с обеспечением безопасности и независимости их одновременной работы.

Освоена технология создания информационных серверов GIIIS, собирающих информацию о локальных вычислительных ресурсах и ресурсах по хранению данных (создаваемых GLOBUS службой GRIS на каждом узле распределенной системы) и передающих эту информацию в динамическом режиме в вышестоящий сервер GIIIS. Таким образом, освоена и протестирована иерархическая структура построения информационной службы GRIS-GIIIS. Организован общий информационный сервер GIIIS (ldap://lhc-fs.sinp.msu.ru:2137), который передает информацию о локальных ресурсах российских институтов на информационный сервер GIIIS (ldap://testbed1.cern.ch:2137) проекта EU DataGRID.

В НИИЯФ МГУ создан Сертификационный центр (Certification authority, CA). Его сертификаты принимаются всеми участниками проекта EU DataGRID. Разработана схема подтверждения запросов на сертификаты с помощью расположенных в других организациях Регистрационных центров (Registration authority, RC), заверяющих запросы пользователей электронной подписью с помощью сертификата GRID. Разработаны программы постановки и проверки электронной подписи, а также пакет программ для автоматизации работы Сертификационного центра.

Инсталлирована и протестирована программа репликации файлов и баз данных GDMP (GRID Data Mirroring Package), которая создана для выполнения удаленных операций с распределенными базами данных. Она использует сертификаты GRID и работает по схеме клиент-сервер, т.е. репликация изменений в базе данных происходит в динамическом режиме. Сервер периодически оповещает клиентов об изменениях в базе, а клиенты пересылают обновленные файлы с помощью команды GSI-ftp. Программа GDMP активно используется для репликации в ЦЕРН распределенной базы моделирующих событий, создаваемой в ОИЯИ, НИИЯФ МГУ и других институтах по физике высоких энергий для эксперимента CMS. Программа GDMP рассматривается в качестве GRID стандарта для репликации изменений в распределенных базах данных.

Сотрудники ОИЯИ принимали участие в развитии средств мониторинга для вычислительных кластеров с очень большим количеством узлов (10000 и более). В рамках задачи Monitoring and Fault Tolerance (Мониторинг и устойчивость при сбоях) они участвуют в создании системы корреляции событий (Correlation Engine). С помощью созданного прототипа Системы корреляции событий (Correlation Engine) ведется сбор статистики аномальных состояний узлов на базе вычислительных кластеров ЦЕРН. Производится анализ полученных данных для выявления причин сбоев узлов. Этот этап позволит получить первый опыт в предсказании сбоев. На втором этапе предусмотрено расширение прототипа Correlation Engine с учетом полученных результатов и испытание системы автоматизированного предупреждения сбоев на практике. Эти разработки включены в создаваемую

архитектуру системы глобального мониторинга (GMA – Grid Monitoring Architecture).

Специалистами НИИЯФ МГУ и ОИЯИ совместно с сотрудниками INFN (Италия) разработана и апробирована новая схема интеграции прикладных инструментальных пакетов IMPALA/BOSS и GRID-технологий для автоматизации процесса массовой генерации событий эксперимента CMS.

В участвующих российских институтах налажен процесс инсталляции актуальных версий пакета EDG. Были проведены тесты совместной работы российских центров с ЦЕРНом и центром в г.Падуя (Италия) в среде EDG.

В архитектуре EDG в настоящий момент не предусмотрена одновременная установка в разных центрах компоненты по распределению заданий по удаленным ресурсам, так называемый Resource Broker (разработка первого рабочего пакета WP1 EDG). Во всех тестах используется Resource Broker, установленный в ЦЕРНе. Однако ясно, что для оптимизации распределения задач в условиях достаточно большого количества центров и ресурсов потребуются когерентная работа нескольких программ Resource Broker, установленных в разных центрах. В НИИЯФ МГУ инсталлирован Resource Broker и проведены его тестирование.

В сотрудничестве с Институтом прикладной математики им. Келдыша РАН проведена инсталляция программы *Metadispatcher* в российском сегменте инфраструктуры EU DataGRID. Эта программа предназначена для планирования запуска заданий в среде распределенных компьютерных ресурсов типа GRID. Было проведено ее тестирование, по результатам которого программа была доработана для обеспечения эффективной передачи данных средствами GLOBUS. Ведутся работы по сравнительному анализу компонент *Metadispatcher* и Resource Broker.

В декабре 2003 года научно-технологический проект EU DataGRID закончится. Поэтому сейчас идет подготовка нового Европейского проекта, EGEE (Enabling Grids for E-science and industry in Europe, <http://www.cern.ch/egee-ej>), который в мае 2003 года будет подан на финансирование в Шестую рамочную программу Европейской комиссии (6th Framework Programme, <http://www.cordis.lu/fp6>). Но этот проект будет уже не научно-технологический, а инфраструктурный. Предполагается, что в результате работ по проекту EGEE будет создан прототип инфраструктуры типа GRID, который станет ядром будущего Европейского GRID. Этот проект будет тесно связан с проектом LCG. В сущности, они будут проектами-партнерами. Основные причины этому – одинаковые цели (создание глобальных GRID инфраструктур) и одна технологическая база (во многом базирование будет на разработках проекта EDG). Российские институты примут активное участие в этом проекте.

Региональный центр БАК в России

В 1999 году в России начаты работы по созданию регионального центра БАК, получившем название *Российского информационно-вычислительного*

комплекса по обработке и анализу данных экспериментов на Большом адронном коллайдере (РИВК-БАК). Для решения организационных вопросов в Миннауки РФ (в рамках Подкомитета РФ-ЦЕРН) в 1999 году был создан Научно-координационный совет (НКС РИВК-БАК).

РИВК-БАК должен стать составной частью инфраструктуры LCG, создаваемой в рамках единой концепции для всех 4-х экспериментов - ALICE, ATLAS, CMS и LHCb. Основу РИВК-БАК составят вычислительные центры российских институтов.

Основными функциями РИВК-БАК будет обеспечение условий для физического анализа данных, доступ к актуальным базам данных в глобальной инфраструктуре региональных центров LCG и создание баз моделирующих событий.

В соответствии с этими базовыми функциями РИВК-БАК является кластером институтских центров уровня Tier2. НКС РИВК-БАК принял концепцию российского регионального центра, согласно которой суммарный уровень ресурсов по участвующим институтам будет порядка 70% от уровня ресурсов канонического центра Tier1 проекта MONARC. Предполагается участие РИВК-БАК в распределенном хранении актуальных баз данных на уровне 5%.

Как уже отмечалось выше, летом 2003 года РИВК-БАК будет подключен к пилотному варианту инфраструктуры LCG-1. С этой целью в каждом из участвующих институтов будет выделены определенные (небольшие) ресурсы и организованы необходимые GRID сервисы. Дальнейшее подключение ресурсов РИВК-БАК в инфраструктуру LCG-1 будет происходить в соответствии с планами экспериментов БАК.

Следует отметить, что, в отличие от других стран, в РИВК-БАК будет реализовываться модель архитектуры, принятая в LCG в целом. Другими словами в РИВК-БАК будет реализовываться GRID инфраструктура институтских центров. В других странах, где будут несколько центров, будет реализовываться иерархическая модель проекта MONARC. Типичным примером может служить Германия, где в Карлсруэ будет создан центр Tier1, а остальные центры (Tier2), будут связаны только с центром в Карлсруэ.

В настоящий момент РИВК-БАК использует следующие ресурсы.

В ИТЭФ, ИФВЭ, НИИЯФ МГУ и ОИЯИ созданы компьютерные инфраструктуры, состоящие из вычислительных кластеров (суммарно более 200 процессоров), дисковых массивов емкостью около 10 ТВ, ленточных библиотек, а также средств визуализации. Эти ресурсы должны быть удвоены в 2003 году, в соответствии с планами участия РИВК-БАК в создании баз моделирующих событий в экспериментах БАК. Отметим, что в этих работах в 2002 году принимали участие и другие институты – ФИАН и НИВЦ МГУ.

В 2003 году планируется подключение к работам по проекту РИВК-БАК, а значит и к работам по проекту LCG, и других институтов по физике высоких энергий. Среди них ИЯИ РАН, МИФИ, РИЦ «Курчатовский институт» и ПИЯФ РАН.

Одним из наиболее важных ресурсов являются линии связи. Очевидно, что без линий передачи данных достаточной пропускной способности невозможно выполнять запланированные работы по проекту РИВК-БАК. В настоящее время минимальным уровнем пропускной способности линии связи для участвующего института является уровень в несколько Мбит/сек. Это определяется тем, что объем часто передаваемых данных составляет сотни Гбайт. Оптимальным в 2002 году был уровень в 20-30 Мбит/сек для подключаемых институтов. В 2003 году требования к линиям связи увеличатся примерно в два раза. Это относится как к линиям связи между российскими институтами, так и к линии связи с ЦЕРНом. В целом требуемый уровень линий связи успешно обеспечивался в рамках Межведомственной государственной программы «Компьютерные сети нового поколения». Пока еще имеются существенные проблемы у ИФВЭ и ПИЯФ РАН. Однако в обоих случаях выполняются проекты по кардинальному изменению ситуации к лету 2003 года, когда линии связи для этих институтов достигнут мощности 100 Мбит/сек. В целом имеются хорошие перспективы роста существующей сетевой инфраструктуры до требуемого в 2007 году уровня в 1-2 Гбит/сек, конечно при выделении соответствующего финансирования, прежде всего по программе «Компьютерные сети нового поколения».

В 2002 году существенно улучшилась ситуация и с международным каналом связи с ЦЕРНом и другими центрами мира, участвующими в проекте БАК. Использовались все существующие линии: FASTNet (Москва – StarTAP в Чикаго) для связи с американскими центрами, а также с ЦЕРНом; линия RUNNet на североевропейскую сеть NORDUNet; а также иногда использовалась линия общего доступа в Интернет, созданная в рамках программы «Компьютерные сети нового поколения». Все эти линии имели к концу 2002 года мощность 155 Мбит/сек. В 2003 году в соответствии с планами участия в проекте LCG необходима будет полоса пропускания порядка 70 Мбит/сек для международного канала.

Заключение

В настоящее время в мире интенсивно развивается концепция *GRID* - компьютерной инфраструктуры нового типа, обеспечивающей *глобальную интеграцию информационных и вычислительных ресурсов* на основе создания и развития управляющего и оптимизирующего программного обеспечения (middleware) нового поколения.

Россия имеет уникальную возможность полномасштабно включиться в этот революционный процесс создания новейшей компьютерной технологии XXI века. Прогресс, достигнутый в области метакомпьютинга и распределенных вычислений и уже имеющийся опыт участия ряда российских научных организаций в международных *GRID* проектах, в особенности в области физике высоких энергий позволит успешно развивать это важнейшее направление.

Особенно важным представляется участие России в крупнейшем международном научном проекте создания Большого адронного коллайдера, для обработки данных экспериментов на котором создается уникальная мировая компьютерная система на основе применения GRID технологий – проект LHC Computing GRID.

В качестве основной ссылки на литературу по всем вопросам, обсуждаемых в этой статье, мы отсылаем читателя к материалам Международной конференции ACAT'2002, которая состоялась в Москве (в МГУ и ОИЯИ) в июне 2002 года:

<http://acat02.sinp.msu.ru>

Первая секция этой конференции, «Very Large Scale Computing and GRID», была полностью посвящена теме GRID и применению этих новейших технологий в физике высоких энергий и других областях науки. Пленарные доклады были сделаны лидерами основных GRID проектов мира, включая доклады С. Kesselman (GLOBUS), L. Robertson (LCG) и P. Kunszt (EDG). Были представлены пленарные доклады по организации вычислительного процесса в основных международных экспериментах по физике высоких энергий, а также в астрофизике. В секционных докладах были представлены многие новые разработки по GRID и применению этих технологий. Наконец, читатель найдет материалы учебных курсов, прочитанных на этой конференции, по GRID/GLOBUS/CONDOR/EDG.