Вычислительная среда для Большого адронного коллайдера

*В статье обсуждаются ряд проблем, связанных с организацией компьютерной обработки данных, которые начнут поступать с крупнейшего в мире ускорителя в Международном исследовательском центре ЦЕРН (**www.cern.ch**). Освещается ряд вопросов, связанных с построением такого типа вычислительной среды.*

***А. Е. Шевель**. Петербургский институт ядерной физики*

Введение

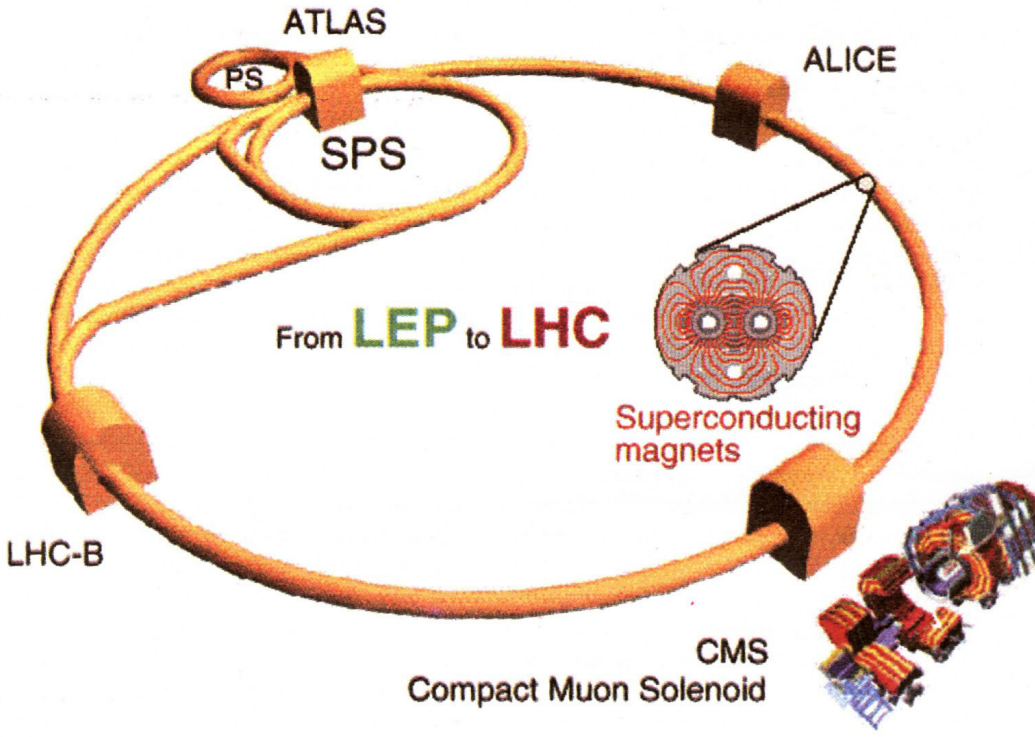
В ЦЕРНе в рамках проекта LHC ведутся работы по созданию БАК – Большого адронного коллайдера (Large Hadron Collider, **www.cern.ch/lhc**), направленные на решение фундаментальных проблем физики частиц высоких энергий. Одна из важнейших задач проекта – регистрация и определение массы так называемого Хиггсовского бозона. Достижение положительного результата может стать краеугольным камнем на пути к созданию единой теории, объединяющей электрослабые и сильные взаимодействия, и, соответственно, к значительному продвижению в познании природы. Это один из самых грандиозных международных проектов, объединяющий весь мировой опыт создания и эксплуатации больших экспериментальных установок.

Начало работы коллайдера намечено на 2005 год. Но уже сейчас идет подготовка будущих экспериментов. Для этого образовано несколько международных коллабораций (временных коллективов). В их числе ATLAS (**http://atlasinfo.cern.ch/Atlas/Welcome.html**), CMS (**http://cmsinfo.cern.ch/cmsinfo/Welcome.html**), Alice (**http://www.cern.ch/ALICE**), LHCb (**http://lhcb.cern.ch**).

Важнейшая проблема, общая для всех коллабораций, – подготовка вычислительной среды для обработки данных, поступающих с измерительных установок. Сложность экспериментальных установок и редкость событий, изучение которых необходимо для уточнения строения материи, делают процесс обработки данных очень сложным и трудоемким.

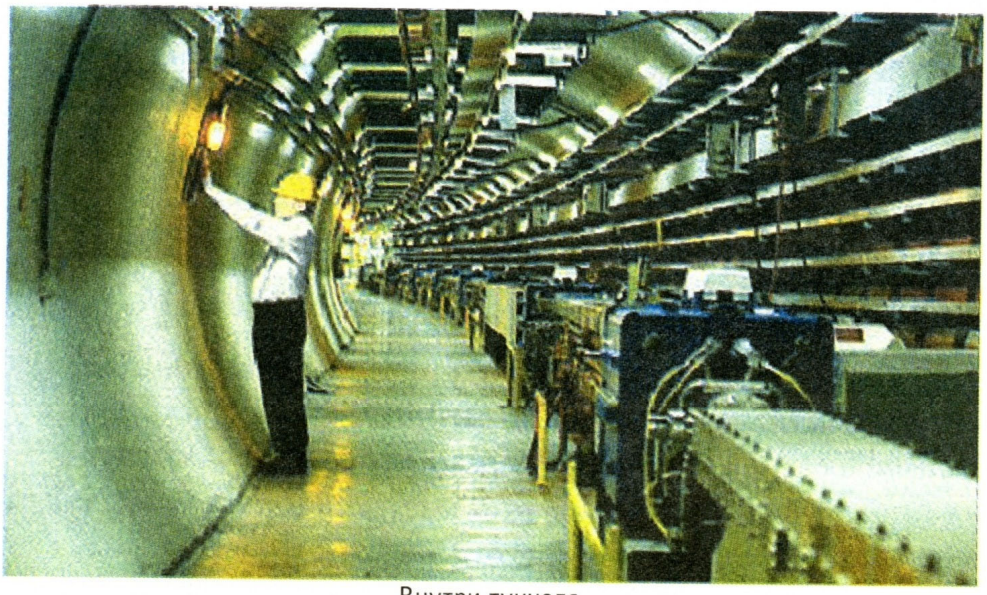
Множество компьютеров выполняют огромную вычислительную работу еще во время проведения измерений на самих экспериментальных установках. Однако мы сейчас сосредоточим внимание на компьютерной организации процессов моделирования экспериментов и на обработке результатов измерений. Каждая из коллабораций разработала более или менее реалистичные сценарии обработки данных, но всем им нужна единая среда для доступа к данным.

Общая схема Большого адронного коллайдера



ЦЕРН координирует несколько крупных исследовательских программ в области компьютинга. В их числе:

- объектно-ориентированная база данных Persistent Object Manager for HEP (**wwwinfo.cern.ch/asd/rd45**);
- пакет GEANT4 (**wwwinfo.cern.ch/asd/geant4/geant4.html**), представляющий собой набор инструментальных программ, моделирующих прохождение элементарных частиц сквозь материю;
- проект MONARC (Models of Networked Analysis at Regional Centres for LHC Experiments (**www.cern.ch/MONARC**);



Внутри туннеля

• набор библиотек инструментальных программ LHC++ для обработки экспериментов на LHC, который должен заменить старую библиотеку CERNlib (**wwwinfo.cern.ch/asd/lhc++/index.html**).

Сложность проблемы

Все проводимые в ЦЕРНе эксперименты весьма нетривиальны по набору детекторов разнообразных излучений, собранных в громадные компьютеризированные измерительные комплексы. Количество детекторов может измеряться сотнями тысяч. В разработке и изготовлении уникального, как правило, оборудования принимают участие десятки тысяч специалистов из сотен университетов и институтов всего мира.

Предполагается, что только физическим анализом полученных измерений будут заниматься около 5 тысяч физиков на планете. Очевидно, что нет возможности собрать всех их вместе. Большая часть физиков будет заниматься анализом данных в своих институтах на своих рабочих местах. Естественно, что у них должны быть относительно простые способы доступа к данным, получаемым в ЦЕРНе. Это первая сторона вопроса.

В ходе экспериментов придется снимать показания с сотен тысяч детекторов. Общий ожидаемый поток данных с измерительных установок составит примерно 5 Пбайт в год при планируемом сроке эксплуатации экспериментальных установок около 20 лет. Единица измерения петабайт (Пбайт), означает 1015 байт и пока еще нечасто встречается на практике. Она примерно соответствует количеству информации, которое за год передает в эфир круглосуточный канал телевидения высокой четкости.

Таким образом, масштаб объемов данных и потребной вычислительной мощности столь значителен, что никакая отдельно взятая организация или даже страна не будут в состоянии обеспечить все необходимое, то есть потребуется аккумулировать возможности всего мирового сообщества: технические ресурсы, финансовые возможности, интеллектуальный потенциал.

Наконец, ответим, зачем понадобилась эта всепланетная организация, которая стоит немалых денег. Полный ответ на такой вопрос может занять на значительный срок немало умнейших голов. Поэтому автор хотел бы отметить лишь главные, по его личному мнению, моменты.

Во-первых, в результате таких крупномасштабных исследований фундаментальных свойств материи значительно продвинутся все представления о возможностях и ограничениях окружающего нас мира. Кроме того, следует учитывать ценнейший шлейф сопутствующих побочных разработок и новых технологий, которые первоначально создаются и применяются в таких научных исследованиях. Наиболее популярный пример побочной разработки – это всемирная паутина WWW, появившаяся в ЦЕРНе в 1989 году.

Во-вторых, такое всемирное исследовательское объединение является отличным полигоном для подготовки молодых специалистов высшей по мировым меркам квалификации в самых разных областях. Ведь подавляющее большинство студентов и аспирантов после завершения подготовки дипломов и диссертаций пойдут в бизнес, правительственные организации, народное хозяйство.

Возвращаясь к компьютерным проблемам, заметим, что для обработки данных должны будут использоваться распределенные вычислительные ресурсы (компьютеры, сети передачи данных, программное обеспечение для доступа к распределенным данным), которые составят специализированную информационную и вычислительную среду. Доступ к этой среде предположительно будет относительно свободным для всех научных работников, студентов, аспирантов, которые связаны с обработкой и анализом данных с экспериментов БАК.

Сейчас рассматривается несколько сценариев распределенной обработки данных и моделирования экспериментов, но во всех просматриваются общие черты. Главный вычислительный кластер для обработки данных реализуется в ЦЕРНе. В этом центре можно будет решать любые задачи, но нельзя будет решать их все. Поэтому предполагается создать несколько центров-сателлитов (может быть, 4–7), которые будут иметь реальную мощность на уровне 7–20% от мощности главного. Часть из этих центров начали складываться в Японии (KEK, **www.kek.jp**), во Франции (IN2P3, **http://www.in2p3.fr:80**), в Италии (INFN, **www.infn.it**).

Основные подходы к решению проблемы

Одной из основных предпосылок эффективной реализации распределенного проекта является использование общепринятых мировых стандартов и широко распространенного программного обеспечения, поставляемого коммерческими компаниями.

Естественно, что пропагандируется широкое использование объектно-ориентированной технологии ОО и C++ в качестве основного языка программирования. Трудно предположить, что все до единой программы будут написаны на C++, но не исключено, что большая их часть будет создана с использованием этого языка.

Решение должно быть достаточно масштабируемым, то есть работать как в ЦЕРНе, так и в региональных центрах. Большинство пакетов должно работать на любых вычислительных установках: от 1 до 1000 пользователей, от notebook до крупных серверов, от нескольких гигабайт до петабайт дисковой памяти.

В рамках проекта RD45 сейчас тестируется ООБД Objectivity/DB (**www.objectivity.com**) в связке с HEP Mass Storage (HEPMSS, **http://home.cern.ch/h/hepmss/www**). В числе используемых в настоящее время находятся такие продукты, как библиотека LHC++, NAG C Library, OpenInventor, IRIS Explorer, STL, OpenGL. Окончательный выбор объектно-ориентированной СУБД для БАК

и ее поставщика произойдет в 2001 году.

При использовании Objectivity/DB планируемая иерархия данных выглядит так:

- федерация баз данных;
- база данных;
- контейнер;
- страница;
- объект.

Одна из проблем выбора базы данных состоит в том, что данные вместе с соответствующими связями должны согласованно поддерживаться в базе в течение нескольких десятков лет. Естественно, что необходимо организовать лицензирование коммерческого программного обеспечения таким образом, чтобы его можно было использовать примерно на равных условиях примерно в ста университетах и институтах.

Важным моментом является также то, какими порциями будут хранится данные и какова будет глобальная стратегия доступа к ним.

Модель представления данных

Здесь мы рассмотрим одну из вероятных моделей для данных, которые будут получены в результате измерений или в результате моделирования экспериментов.

Такую модель необходимо иметь уже сейчас, поскольку во всех моделирующих программах необходимо учесть особенности использования тех или иных баз данных. Федеративная база данных по терминологии Objectivity/DB допускает распределенные наборы объектов как в географическом отношении, так и в смысле разнообразия серверов и устройств для хранения данных (ленты, диски и пр.). При этом должно поддерживаться согласованное и логически универсальное представление полной распределенной базы данных. Различные объекты могут содержать указатели друг на друга (ассоциации), которые позволяют производить поиск по всей базе данных. Эта модель включает четыре функционально различные группы объектов:

- Первичные данные (raw data) – около 1 Мбайт/событие; хранятся, скорее всего, на магнитной ленте только в ЦЕРН.
- Отфильтрованные данные (ESD data – Event Summary Data) – объекты данных после первичной фильтрации и геометрической реконструкции; объем около 0,1 Мбайт/событие.
- Объект физического анализа (AOD data – Analysis Object Data) – подмножество ESD; связано с ESD посредством ассоциации AOD->ESD; объем около 0,01 Мбайт/событие.
- TAG – небольшой набор существенной информации, описывающий искомое физическое событие, которое позволяет производить первоначальную селекцию данных AOD для последующей обработки.

Данные этих различных типов организованы в уникальных контейнерах (файлах). Соответствующее программное обеспечение позволяет однозначно определить, какие контейнеры нужны для выполнения конкретного задания.

Многие проблемы в организации данных уже поняты, и уже ведутся соответствующие испытания. Естественно, следует учитывать опыт, получаемый сейчас в действующих коллаборациях типа BaBar (**www.slac.stanford.edu/BFROOT**), в других проектах, например COMPASS (**http://wwwcompass.cern.ch**).

Архитектура

Одна из важных проблем – как распределить вычислительную нагрузку между ресурсами ЦЕРН и региональными вычислительными центрами. Здесь принимаются во внимание следующие важные параметры:

- общая удельная стоимость оборудования (вычислители и системы хранения данных);
- количество специалистов, необходимых для поддержания всей системы в работоспособном состоянии;
- наконец, эффективность, выраженная в следующих терминах:
- среднее время выполнения одного типичного задания;
- доступность, то есть сколько часов в год система будет реально доступной для пользователей;
- гибкость реагирования на изменение загрузки в течение дня или года.

Общими (всеми разделяемыми) представлениями на сегодняшний день являются следующие:

- Обработка данных такого объема и сложности может быть реализована только объединенными усилиями всего международного сообщества специалистов.
- Пока имеется довольно сильная неопределенность в том, какую пропускную способность каналов связи можно будет реально себе позволить.
- Главные вычислительные ресурсы должны располагаться в ЦЕРНе.
- Дополнительные региональные центры (с условным названием Tier1) должны располагать вычислительной мощностью на уровне 10–20% от имеющейся в ЦЕРНе.
- Вспомогательные региональные центры (с условным названием Tier2) могут быть меньше Tier1, но выполняют конкретные важные задачи.

Вычислительные центры класса Tier1 и Tier2 могут объединяться для выполнения специфических задач, например обслуживания конкретной коллаборации, или для выполнения специального вида работ для всех экспериментов на БАК.

Весьма важно, чтобы к началу выполнения экспериментов на БАК была бы достигнута высокая степень координации между различными региональными центрами. Здесь имеется в виду как совместимость по оборудованию и программному обеспечению, так и в части согласования, что и кто должен делать. Чтобы оценить, как важны региональные центры, следует заметить, что, по существующим представлениям, ЦЕРН обеспечит не более половины необходимой вычислительной мощности и емкости запоминающих устройств.

Обратим внимание на изменение удельной стоимости разных компонентов. Удельная цена на процессоры падает в два раза примерно за 18 месяцев. Удельная цена одного петабайта на дисках падает вдвое примерно за 17 месяцев. Наконец, удельная цена одного петабайта на лентах уменьшается настолько же примерно за 25 месяцев.

Основной особенностью оборудования и всей операционной среды является гетерогенность: в ней, видимо, будут использоваться все типы операционных систем. В качестве одной из распространенных операционных сред, вероятно, будет использоваться Linux.

Предполагается, что региональный центр должен обеспечивать примерно следующее:

- Поддержку всех служб, необходимых для выполнения физического анализа.
- Хранение данных по всем видам физических объектов, а также тегов и результатов калибровки.
- Хранение заметной части первичных данных.
- Хранение копий всех калибровочных констант.
- Отличные возможности по связи с ЦЕРНом и с региональными пользователями.
- Наличие персонала, который мог бы взять на себя часть работы по созданию, тестированию и поддержанию общего программного обеспечения.
- Поддержку обучения, документирования и другого сервиса для всех пользователей регионального центра.

Регион Санкт-Петербурга

Как было отмечено выше, региональные компьютерные центры для обработки данных БАК начинают разворачиваться во многих странах, и Россия тоже участвует в этом процессе. Информацию о российской части проекта можно почерпнуть на сайте **www.pnpi.spb.ru/RRCF/RRCF.html**.

В Санкт-Петербурге имеется несколько

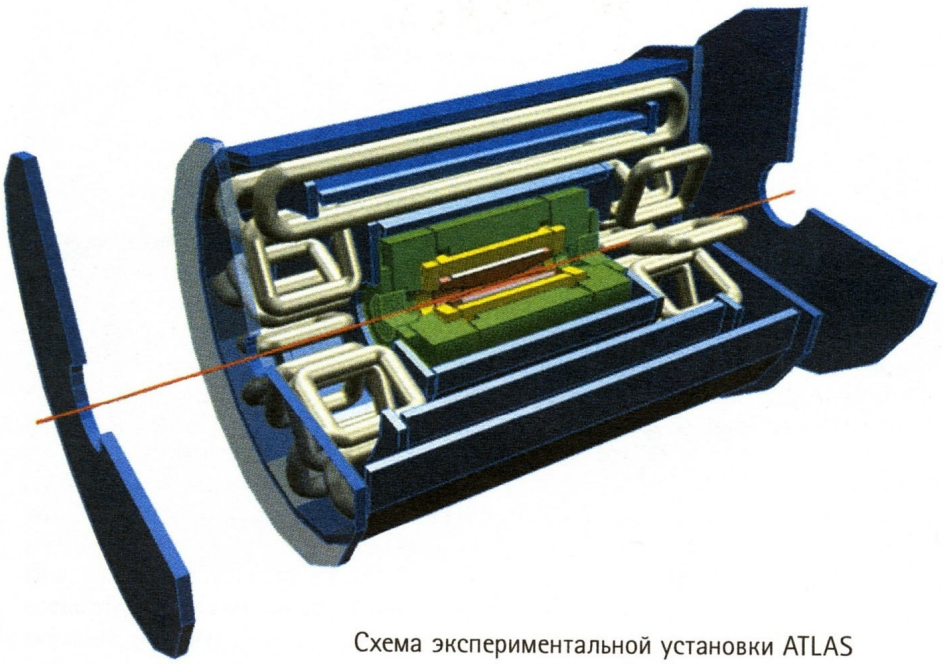


Схема экспериментальной установки ATLAS

институтов и университетов, которые, по всей видимости, примут участие в обработке данных и моделировании будущих экспериментов на БАК. В первую очередь это Петербургский институт ядерной физики (**www.pnpi.spb.ru**). Возможно, в физическом анализе получаемых данных станут участвовать отдельные факультеты или кафедры, которые будут вести соответствующие курсы лекций и привлекать студентов и аспирантов.

С высокой степенью вероятности Институт высокопроизводительных вычислений и баз данных (**www.csa.runnet.ru/ihpcdb**) будет играть роль технологической вычислительной базы в регионе Санкт-Петербурга для задач обработки данных БАК. Туда же должны стекаться те фрагменты данных, которые будут требоваться для обработки. Естественно, что этот институт должен иметь соответствующие ресурсы для выполнения этой задачи. Какие же это ресурсы?

- В первую очередь, высокоскоростной канал в ЦЕРН для получения данных.
- Роботизированная внешняя память: как минимум робот на пару сотен терабайт с ленточными картриджами по 40–100 Гбайт.
- Компьютерные ресурсы для вычислений как таковых: достаточное количество процессоров, чтобы обеспечить мощность около 90 K единиц SPECint95 (о SPECint можно посмотреть **www.specbench.org/spec**).

Обсудим эти ресурсы подробнее.

Связь

Когда автор говорит о канале связи, то имеется в виду канал любой физической природы и технической организации. В том числе полезно рассмотреть временные окна для каналов типа «точка-точка», например, ЦЕРН-ИВВБД или ЦЕРН-ПИЯФ.

Для справки мы приводим таблицу объемов данных, которые возможно прокачать через линию связи определенной пропускной способности. Из нее видно, что, имея недельное окно в месяц, даже на небольшой скорости (32 Мбит/с) можно прокачать около 2 Тбайт.

Таблица объемов данных и времени их передачи по различным каналы связи

Канал с пропускной способностью, Мбит/с	Максимальный объем передачи за один час, Гбайт	Максимальный объем передачи за 24 часа, Тбайт	Время передачи файла объемом 100 Гбайт, ч
32	11,5	0,3	8,7
64	23,0	0,6	4,3
155	55,8	1,3	1,8
622	223,9	5,4	0,5

Для сравнения рассмотрим традиционный способ транспортировки данных, который превалировал в прошлом, – перевозку картриджей с магнитными лентами. Обычный картридж, например, для устройства DLT 8000, имеет емкость 40 Гбайт. Средняя скорость записи на ленту составляет около 5 Мбайт/с. Таким образом, время записи одного картриджа составит примерно 2 часа. Как нетрудно догадаться, это и есть минимальное время копирования информации из ленточного робота на ваш картридж. Если предполагать, что минимальный объем физически значимой информации составляет 100 Гбайт, то потребуется три картриджа, или 5–6 часов для копирования. После перевозки магнитной ленты потребуется столько же времени для перезаписи данных на диск. Таким образом, в сумме на копирование потребуется около 12 часов (не считая времени перевозки). Задержка в передаче данных традиционным способом составит в лучшем случае 3–4 дня.

Если сравнить эти времена с временем передачи такого же объема данных по каналу 32 Мбит/с (около 9 часов), то становится очевидным, что эти величины сравнимы. Естественно, что организация временных каналов вполне может оказаться во всех отношениях нетривиальной задачей. С другой стороны, если дальняя связь (в ЦЕРН) будет реализована в виде временных окон, то региональная связь должна быть весьма быстрой (10–155 Мбит/с) постоянно.

Роботизированная внешняя память

Если иметь в виду, что данные будут передаваться из ЦЕРН, то полезно все передачи кешировать, то есть записывать все переданные массивы на магнитную ленту (картридж), которая является по сей день носителем с наименьшей стоимостью хранения одного гигабайта. Поскольку данные будут, в основном, передаваться в один институт, то и ленточный робот следует устанавливать там.

Вычисления

Вычислительная мощность по очевидным соображениям минимизации стоимости будет строиться на основе PC-кластеров. Согласно имеющимся оценкам (**nicewww.cern.ch/~les/monarc/base_config.html**), в региональном вычислительном центре потребуется суммарная производительность около 90 K единиц SPECint95. Принимая во внимание, что один процессор Pentium-III/733 имеет примерно 35 единиц, то 90 K единиц

SPECint95 смогут дать в сумме 2571 таких процессоров. Если использовать четырехпроцессорные конфигурации, то потребуется примерно 650 машин.

По всей видимости, эта вычислительная мощность будет распределена по нескольким институтам и университетам. Иными словами, будет несколько вычислителей меньшей производительности. В случае четырех вычислительных кластеров мы получим по 160 машин в кластере. С учетом того, что к 2005 году производительность микропроцессоров вырастет примерно в 3–4 раза, к тому времени будет достаточно около 40 машин на кластере. Иными словами, число машин вполне представимое для размещения в комнате площадью 30–40 м^2 с кондиционером.

Структура каждого вычислителя будет включать кластер на базе PC, выполняющий основной объем вычислений, и традиционный сервер (Sun, HP или другой), на котором идут такие рутинные приложения, как всевозможное администрирование, СУБД, поддержание X-сессий и прочие ординарные задачи.

Технологический процесс обработки

В свете описанного выше полезно представить в общих чертах технологический процесс обработки данных с БАК, например, в регионе Санкт-Петербурга.

Предположим, что уже имеется группа для физического анализа конкретной проблемы, которая переписывает необходимые данные из ЦЕРНа в централизованное хранилище в Санкт-Петербурге. Наличие такого централизованного роботизированного хранилища избавит от дополнительных пересылок, если те же данные потребуются другой группе исследователей или другому отдельному физику. Вероятнее всего, передача данных из ЦЕРНа произойдет с использованием компьютерного канала связи.

Следующими шагами могут быть отбор полезных событий и использование готовой программы анализа физических результатов. Отбор полезных событий можно производить на компьютерных мощностях (вычислительном кластере), которые находятся в непосредственной близости к хранилищу. Отобранные данные могут иметь заметно меньший объем, что позволит передать их по региональной сети с меньшими затратами времени. Поэтому физический анализ данных может производиться уже на том вычислительном кластере, где находится физик. ■

А. Е. Шевель, **Andrei.Chevel@pnpi.spb.ru**,
Петербургский институт ядерной физики, **www.pnpi.spb.ru/comp_home.html**

news

Оплата почты через Интернет

Компания E-Stamp, занимающаяся оплатой почтовых пересылок посредством Интернета, и eBay, один из крупнейших устроителей аукционов во Всемирной Сети, собираются заключить трехгодичный договор с целью продвижения использования компьютеризированных почтовых платежей. На данный момент всего две компании – E-Stamp и Stamps – получили лицензию правительства Соединенных Штатов на осуществление платежей такого типа. eBay же уступает по популярности лишь Amazon. E-Stamp собирается выплачивать eBay порядка $10 миллионов в год для саморекламы среди участников аукционов. В целом система заметно облегчает оплату почтовых услуг путем замены традиционной системы оплаты посредством марок на печать индикатора оплаты непосредственно на корреспонденцию, что в первую очередь должно облегчить жизнь небольшим и средним компаниям.

Работа на колесах

Одной из новинок этого года, привлекшей к себе большое внимание, оказались компьютерные системы, предназначенные для использования в автомобилях. IBM и Motorola собираются объединиться для создания беспроводной интернет-технологии, предназначенной специально для использования ее в автомобилях. Помимо таких стандартных функций, как просмотр почты и курса акций, приема и отправки сообщений и прослушивания музыки, предполагается оснащение автомобилей такими устройствами, как глобальная позиционная система (GPS), позволяющая определять свое месторасположение, или специальными устройствами, которые будут предназначены для повышения безопасности движения, куда будут включены противоугонные системы. Помимо IBM и Motorola, подобной деятельностью также занимаются другие гиганты компьютерного мира, например Sun и Hewlett Packard. Естественно, не обошлось без компании Microsoft, которая сотрудничает с Clarion, Daewoo, Intel и другими компаниями и собирается предоставить для автомобильных устройств свою операционную систему, основанную на Windows CE. Интересно, кто доверит свой автомобиль Windows...

IBM устанавливает рекорд по количеству полученных патентов

За предыдущий год компания IBM получила рекордное число патентов – 2756 штук, что побивает рекорд прошлого года, равный 2658. Патенты составляют важную часть новой стратегии Big Blue, которая заключается в продаже технологий и компонентов другим компаниям, в список которых входят производитель персональных компьютеров Dell Computers, и в производстве сетевых устройств Cisco Systems. «Треть всех инвестиций, сделанных в 1999, году уже находятся на рынке и представлены в виде продуктов», – заявил Nick Donofrio, вице-президент технологий и производства компании IBM. Наиболее важными из этих патентов представляются технология, позволяющая компьютерам потреблять меньше питания, и лицензия на ПО, осуществляющее поиск в Интернете.

AOL может выйграть спор за Net TV

Сможет ли AOL то, что не удалось Microsoft, – создать он-лайн-телесистему? Цель объединения AOL и медиа-гиганта Time Warner вполне прозрачна – создание системы интерактивного телевидения. Основная задача AOL – это создать популярную он-лайн-службу, предназначенную исключительно для телезрителей, но предыдущая попытка не увенчалась успехом. С начала 90-х годов Microsoft весьма агрессивно представляла себя в качестве лидирующего провайдера служб, так или иначе связанных с телевидением. У AOL же уже есть около 20 миллионов клиентов, на которых можно повлиять. Вдобавок, многочисленные возможности Time Warner могут использовать интерактивное телевидение для продвижения и раскрутки, например, музыкальных альбомов. Планы у AOL поистине наполеонские: уже заключены договоры с Hughes Network Systems, Sun, Gateway, Intel и другими некомпьютерными компаниями, что говорит об огромной аудитории, на которую рассчитан новый проект.