ВЫЧИСЛИТЕЛЬНАЯ СРЕДА ДЛЯ Большого Адронного Коллайдера

А.Е. Шевель *
Петербургский Институт Ядерной Физики
188350, Гатчина, Ленинградской обл.,
Россия

10 февраля 2000

Аннотация

В статье обсуждаются ряд проблем связанных с организацией компьютерной обработки данных, которые начнут поступать с крупнейшего в мире ускорителя LHC в международном исследовательском центре CERN (`http://www.cern.ch/`). Освещаются ряд вопросов связанных с построением такого типа вычислительной среды.

ВВЕДЕНИЕ

В CERN строится новый ускоритель заряженных частиц Large Hadron Collider – LHC (Большой Адронный Коллайдер – БАК; [`http://wwwlhc01.cern.ch/`]), который начнет работать в 2005 году.

Важными проблемами в обработке результатов сложнейших измерений на ускорителе являются как сложность самих измерений, так и значительные объемы данных (при-

*e-mail:*Andrei.Chevel@pnpi.spb.ru*

мерно 10**15 байтов вгод, Petabyte), поступающих с измерительных установок. Это обусловлено тем, что изучаются весьма редкие события в микромире. В проведении измерений участвуют тысячи учёных и специалистов из различных областей знаний. Сам процесс проведения измерений является сложнейшим технологическим процессом. Однако в данной статье, мы обсудим лишь основные фрагменты строящейся компьютерной организации процесса моделирования и обработки результатов измерений, которые могут рассматриваться как отдельная задача при проведении экспериментов на ускорителях.

Уже сейчас идет подготовка экспериментов на БАК. Для этого образованы несколько международных коллабораций (временных коллективов): 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/`). Каждая из коллабораций разработала более или менее реалистичные сценарии обработки данных.

Имеется несколько крупных исследовательских программ, которые координирует CERN в отношении компьютинга: Persistent Object Manager for HEP (`http://wwwinfo.cern.ch/asd/rd45/`) и пакет GEANT4 (`http://wwwinfo.cern.ch/asd/geant4/geant4.html`).
Пакет GEANT4 представляет собой набор инструментальных программ моделирующих прохождение элементарных частиц сквозь материю. Проект Модели Сетевого Анализа в Региональных Вычислительных Центрах – Models of Networked Analysis at Regional Centres for LHC

Experiments MONARC (`http://www.cern.ch/MONARC/`). Создание библиотек прикладных программ для экспериментов на LHC, как дополнение CERNlib, (`http://wwwinfo.cern.ch/asd/lhc++/index.html`) и другие.

СЛОЖНОСТЬ ПРОБЛЕМЫ

Все проводимые эксперименты в CERN довольно нетривиальны по набору детекторов разнообразных излучений, которые собраны в громадные измерительные компьютеризированные комплексы. Количество детекторов может измеряться сотнями тысяч. В разработке и изготовлении, как правило уникального оборудования, принимают участие десятки тысяч специалистов из сотен университетов и институтов по всему миру.

Предполагается, что только физическим анализом полученных измерений будут заниматься около 5 тысяч физиков на планете. Очевидно, что нет возможности всех собрать вместе. Большая часть физиков будут заниматься анализом данных в своих институтах на своих рабочих местах. Естественно, что они должны иметь относительно простые способы доступа к данным, полученным в CERN. Это первая сторона вопроса.

Как уже было отмечено, в измерениях придется снимать показания с сотен тысяч детекторов. Общий ожидаемый поток данных с измерительных установок составит примерно 5 PB в год (PB означает PetaByte = 10**9 Mbytes). Планируется, что продолжительность жизни и работы экспериментальных установок составит около 20 лет.

Таким образом масштабы объемов данных и потребной вычислительной мощности столь значительны, что никакая отдельно взятая страна или организация не будет в состоянии обеспечить все необходимое, т.е. потребуется аккумулировать ресурсы всего мирового сообщества: технические ресурсы, финансовые возможности, интеллектуальный потенциал.

Наконец, ответим зачем понадобилась эта всепланетная организация, которая стоит немалых денег. Исчерпывающий и краткий ответ на такой вопрос вряд ли возможен, поэтому автор хотел бы отметить главные по его 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 разрабатываются и применяются в таких научных исследованиях. Наиболее популярный пример побочной разработки – это всемирная паутина WWW, появившаяся в CERN 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 пойдёт в бизнес, правительственные организации, народное хозяйство.

Возвращаясь к компьютерным проблемам заметим, что для обработки должна быть использована распределенная вычислительные ресурсы (компьютеры, сети передачи

данных, программное обеспечение для доступа к распределённым данным), которые составят специализированную информационную и вычислительную среду. Доступ к этой специализированной среде предположительно будет относительно свободным для всех научных работников, студентов, аспирантов, которые будут связаны с обработкой и анализом данных с экспериментов БАК.

Рассматривается несколько сценариев распределенной обработки данных и моделирования экспериментов. Общие черты таких сценариев следующие.

Главный вычислительный кластер для обработки данных реализуется в CERN. В этом центре можно будет решать любые задачи, но нельзя будет решать их все целиком. Предполагается, что будет реализовано несколько центров-сателлитов (может быть 4-7), которые будут иметь реальную мощность на уровне 7%-20% от главного. Часть из этих центров начали складываться в Японии (KEK; `http://www.kek.jp/`), Франции (IN2P3; `http://www.in2p3.fr.:80/`), в Италии (INFN: `http://www.infn.it/`).

ОСНОВНЫЕ ПОДХОДЫ К РЕШЕНИЮ ПРОБЛЕМЫ

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

Естественно, что пропагандируется широкое использова-

ние технологии ОО и C++, как основного языка программирования. Трудно представить, что все до единой программы будут написаны на C++, но не исключено, что большая часть будет написана с использованием этого языка.

Решение должно быть достаточно масштабируемым, т.е. работать как в CERN, так и в Региональном Центре. Большинство пакетов должны работать на любых вычислительных установках: от 1 до 1000 пользователей, от notebook до крупных серверов, с объемами дисковой памяти от нескольких GB до PetaByte.

Objectivity/DB + HPSS тестируется для проекта RD-45 (`http://wwwinfo.cern.ch/asd/rd45/index.html`). Среди используемых в настоящее время находятся такие продукты как библиотека LHC++, NAG C Library, OpenInventor, IRIS Explorer, STL, OpenGL. Окончательный выбор Объектной Системы Управления Базой Данных для БАК будет произведен в 2001 году. В конце того же года будет производиться выбор поставщика этой системы.

Планируемая иерархия данных в тестируемом Objectivity/DB следующая:

- Федерация баз данных.
- База данных.
- Контейнер.
- Страница.
- Объект.

Одна из проблем выбора базы данных состоит в том, что данные должны согласованно поддерживаться (вместе с соответствующими связями) в течение нескольких десятков

лет. Естественно, что необходимо организовать лицензирование коммерческого программного обеспечения таким образом, чтобы его можно было использовать примерно в равных условиях примерно в ста университетах и институтах.

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

МОДЕЛЬ ПРЕДСТАВЛЕНИЯ ДАННЫХ

Здесь рассматривается одна из вероятных моделей данных, которые будут получены в результате измерений или в результате моделирования экспериментов.

Такую модель необходимо иметь уже сейчас, поскольку во всех моделирующих программах необходимо участь особенности использования тех или иных баз данных. Здесь имеется в виду федеративная база данных в смысле Objectivity/DB, которая допускает распределённые наборы объектов как в смысле географическом, так и в смысле разнообразия устройств для хранения данных (ленты, диски, прочее), а также серверов данных. В то же время поддерживается согласованное и логически универсальное представление полной распределённой базы данных. Различные объекты могут содержать указатели друг на друга (ассоциации), которые позволяют производить поиск по всей базе данных. Эта модель включает 4 функционально различных групп объектов:

- **Исходные (сырые) данные** (raw data); около 1MB/событие; скорее всего запомненных на магнитной ленте только в CERN;

- **Отфильтрованные данные** (ESD data – Event Summary Data) – объекты данных после первичной фильтрации и геометрической реконструкции; объём около 0.1 MB/событие.
- **Объект физического анализа** (AOD data – Analysis Object Data) – подмножество ESD; связано с ESD посредством ассоциации AOD->ESD; объём около 0.01MB/событие.
- **TAG** – небольшой набор существенной информации, описывающей 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 проводятся соответствующие испытания. Естественно, следует учитывать опыт получаемый сейчас в действующих коллаборациях типа BaBar (`http://www.slac.stanford.edu/BFROOT/`), в других проектах, например, COMPASS (`http://wwwcompass.cern.ch/`).

АРХИТЕКТУРА

Одно из важных проблем как распределить вычислительную нагрузку между вычислительными ресурсами CERN 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", т.е. сколько часов в год система будет реально доступной для пользователей;
  - гибкость реагирования на изменение загрузки в течение дняили года.

Общими (всеми разделяемыми представлениями) на сегодняшний день являются:

- Обработка данных такого объёма и сложности может быть реализована только объединёнными усилиями всех специалистов наций.
- Пока имеется довольно сильная неопределённость в том, какую пропускную способность каналов связи можно будет реально себе позволить.
- Главные вычислительные ресурсы должны располагаться в CERN.
- Дополнительные Региональные Центры – вычислительные центры (с условным названием типа Tier1) на уровне 10-20% от вычислительной мощности в CERN.

- Вспомогательные Региональные Центры – вычислительные центры (с условным названием типа Tier2), которые могут быть меньше Tier1, но выполнять 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 по представлениям что и кто должен делать.

Чтобы оценить как важны региональные центры, следует заметить, что по существующим представлениям CERN обеспечит не более половины потребной вычислительной мощности и ёмкости 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 PB на дисках падает вдвое примерно за 17 месяцев. Наконец, удельная цена одного PB на лентах падает вдвое примерно за 25 месяцев.

Основной особенностью оборудования и операционной среды является гетерогенность: видимо будут использоваться все типы операционных систем. Одной из распространенных операционных сред предполагается Linux.

Предполагается, что региональный центр должен обеспечивать примерно следующее.

- Вся необходимые виды сервисов для выполнения физического анализа.
- Все виды физических объектов, а также теги и результаты калибровки.
- Заметную часть сырых данных.
- Хранить копию всех калибровочных констант.
- Отличные возможности по связи с CERN и с региональными пользователями.
- Наличие людей, которые могли бы принимать брать на себя часть работы по разработке, тестированию и поддержанию общего программного обеспечения.
- Отличная поддержка сервиса по обучению, документированию и пр. для всех пользователей Регионального Центра.

Регион С.Петербурга

Как было отмечено выше, РЕГИОНАЛЬНЫЕ КОМПЬЮТЕРНЫЕ ЦЕНТРЫ для обработки данных с БАК начинают разворачиваться во многих странах. Аналогичные процессы происходят в России. Информацию о Российской части проекта можно почерпнуть на сайте `http://www.pnpi.spb.ru/RRCF/RRCF.html`. Здесь уместно

добавить, что в свете общероссийской направленности данного проекта полезно обрисовать региональные особенности С.Петербурга.

В С.Петербурге имеется несколько 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 (`http://www.pnpi.spb.ru`). Возможно, в физическом анализе получаемых данных начнут участвовать отдельные факультеты или кафедры, которые будут вести соответствующие курсы лекций и привлекать студентов и аспирантов.

Какие это ресурсы?

- В первую очередь высокоскоростной канал на CERN для получения данных.
- Роботизированную внешнюю память: как минимум робот на пару сотен TB с ленточными картриджами по 40/100 GB на картридж.
- Вычислительные ресурсы для вычислений как таковых: достаточное количество процессоров, чтобы обеспечить мощность около 90K единиц SPECint95 (о SPECint можно посмотреть `http://www.specbench.org/spec/`).

Рассмотрим их подробнее.

СВЯЗЬ

Для справки можно указать таблицу объемов данных, которые возможно прокачать через линию связи определённой пропускной способности.

Канал с пропускной способностью Mbit/sec	Максимальный объём передачи за 1 секунду в MB	Максимально возможный объем передачи за один час в MB	Максимально возможный объем передачи за 24 часа в TB	Время передачи файла объёмом 100 GB в часах
32	3.2	11520	0.3	8.7
64	6.4	23040	0.6	4.3
155	15.5	55800	1.3	1.8
622	62.2	223920	5.4	0.5

Таблица 1: Таблица объёмов данных и времени их передачи при передаче через различные каналы связи

Когда автор говорит о канале связи, то имеется в виду любого вида канал любой физической природы и технической организации. В том числе полезно рассмотреть временны́е окна для каналов типа точка–точка, например, CERN–ИВВБД или CERN–ПИЯФ.

Из таблицы видно, что имея недельное окно в месяц даже на небольшой скорости (32 Mbit/sec) можно прокачать около 2 TB.

Сравнивая с традиционными способами транспортировки данных, которые превалировали в прошлом, т.е. перевозку картриджей с магнитными лентами, следует обратить внимание на следующее. Обычный картридж, например для устройства DLT 8000, имеет ёмкость 40 GB. Средняя скорость записи на ленту составляет около 5 MB/sec. 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 объём физически значимой информации составляет 100 GB, 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 потребуется около 12 часов (не считая времени перевозки) на копирование. Задержка в передаче данных традиционным способом составит в лучшем случае 3-4 дня.

Если сравнить эти времена с временем передачи такого же объёма данных по каналу 32 Mbit/sec (около 9 часов), то становится очевидным, что эти величины сравнимы.

Естественно, что организации временны́х каналов вполне может оказаться нетривиальной во всех отношениях.

С другой стороны, если дальняя связь (в CERN) будет реализована в виде временны́х окон, то региональная связь должна быть весьма быстрой постоянно, 10-155 Mbit/sec.

РОБОТИЗИРОВАННАЯ ВНЕШНЯЯ ПАМЯТЬ

Если иметь в виду, что данные будут передаваться из CERN, то полезно все передачи кешировать, т.е. записывать все переданные массивы на магнитную ленту (картридж), которая является по сей день носителем с наименьшей стоимостью хранения одного GB.

Поскольку данные будут, в основном, передаваться в один институт, то и ленточный робот следует устанавливать там. Среди прочих материалов на эту тему полезно посмотреть доклад авторов I.Augustin, J.P.Baud, R.Tuebbicke, P.Vande Vyvre озаглавленным **"PASTA - The LHC Technology**

Tracking Team for Processors, Memory, Architectures, Storage and Tapes - Run II". Working Group (d): Storage Management Systems. Status report - 05 July 1999 - Version 1.4 (http://pcbunn.cithep.caltech.edu/pubs/PASTA.html)

ВЫЧИСЛЕНИЯ

Вычислительная мощность будет строиться по очевидным соображениям минимальной стоимости на основе PC кластеров. Согласно имеющимся оценкам (`http://nicewww.cern.ch/~les/monarc/base_config.html`) в региональном вычислительном центре потребуется суммарная производительность около 90K единиц SPECint95. Принимая во внимание, что один процессор Pentium-III/733 имеет примерно 35 единиц, то 90K единиц SPECint95 смогут дать в сумме 2571 таких процессоров. Если иметь в виду 4-х процессорные конфигурации, то потребуется примерно 642 машины.

По всей видимости эта вычислительная мощность будет распределена по нескольким институтам и университетам. Иными словами, будет несколько вычислителей меньшей производительности. Если предположить, что вычислительная мощность будет рассеяна по четырём вычислительным кластерам, то получим по 160 машин в кластере. Учитывая, что к 2005 году производительность микропроцессоров вырастет примерно в 3-4 раза, то получим, что к тому времени будет достаточно около 40 машин на кластер. Иными словами, число машин вполне представимое для размещения в комнате площадью 30-40 квадратных метров с кондиционером.

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

ТЕХНОЛОГИЧЕСКИЙ ПРОЦЕСС ОБРАБОТКИ

В свете вышеописанного полезно представить в общих чертах возможный технологический процесс обработки данных с БАК, например, в регионе С.Петербурга. автор исходит из того, что передача данных из CERN будет происходить в основном с использованием компьютерного канала связи.

Предположим, что уже имеется группа для физического анализа конкретной проблемы, в одной из коллабораций (Alice, Atlas, etc.). Естественно, что для выполнения анализа потребуется переписывать некоторые виды данных из CERN в С.Петербург в централизованное хранилище. По всей видимости совсем не будут передаваться сырые измерительные данные. Видимо будет передаваться небольшая часть данных **ESD**, а чаще будут передаваться данные типа **AOD**. Относительно небольшая порция данных должна будет поддерживаться полностью соответствующей той, что в CERN. Это базы данных с геометрией экспериментальных установок и данных калибровки (своя для каждой коллаборациии). Естественно, следует учесть, что там же может находиться программное обеспечение, которое будет храниться, по всей видимости, в стандарте **CVS**.

Процесс передачи данных можно было бы представить в виде следующей последовательности шагов. Исходным отправным пунктом для передачи данных является, конечно, CERN, В каждой коллаборации будет каталог (оглавление) имеющихся в CERN файлов с данными типа **ESD**. Эти каталоги будут доступны в каждом институте, участвующем в физическом анализе. Человек, занимающийся анализом, будет описывать запросы на группы файлов (или части файлов) в соответствии с правилами языка запросов. Готовая спецификация будет переправляться на специальный сервер CERN, который перешлёт запрошенные данные, куда вы попросили. Очевидно, что если запрошенного файла нет или возникли другие проблемы при выполнении запроса, то соответствующее сообщение должно поступить тому, кто запросил файл или группу файлов.

Наличие централизованного роботизированного хранилища избавит от дополнительных пересылок, в тех случаях, если те же данные потребуются или могут потребоваться другой группе исследователей или другому отдельному физику. Поскольку заранее неизвестно, какие экспериментальные данные потребуются, а какие нет, то необходимо будет запоминать всё, что передано из CERN. Иными словами, централизованное хранилище будет играть роль мощного прокси сервера для региона.

Следующим шагом должен быть отбор полезных событий и использование готовой программы анализа физических результатов. Отбор полезных событий можно производить на компьютерных мощностях (вычислительном кластере), которые могут находиться в непосредственной близости к хранилищу. Сейчас мы оставим в стороне вопрос откуда по-

явятся вычислительные ресурсы в централизованном хранилище. Отобранные данные могут иметь заметно меньший объём, что позволит передать их по региональной сети (по С.Петербургу и области) с меньшими затратами времени.

Содержательный физический анализ данных может производится на том вычислительном кластере, где находится физик, т.е. на своём рабочем месте в университете или институте.

Организационные соображения

Из предыдущего видно, что хорошо бы обозначить место центрального хранилища данных в С.Петербурге, которое было бы уже в значительной степени готовым к выполнению описанных задач: наличие персонала, опыта использования соответствующего оборудования и программного обеспечения, а также опыта обслуживания большого количества потребителей на недискриминационной основе.

Заключение

Нетрудно видеть, что описываемые в данной статье проблемы весьма важны, нетривиальны и перспективны. Методы, разработанные и опробованные при решении такой задачи, наверняка будут востребованы во всех областях применения компьютерных технологий.

Масштаб задач таков, что участие любого института, лаборатории или продвинутой компании не будет лишним. Требуется масса компонентов: компьютеры и роботизированные установки внешней памяти, линии и устройства

связи, техническая и учебная литература, а также многое другое.

Естественно, что появление такой платформы есть лишь необходимая предпосылка для полноценного участия в физическом анализе данных с крупнейшего в мире ускорителя заряженных частиц, т.е. участия в новейших исследованиях фундаментальных свойств материи.