ИНТЕРНЕТ-ТЕХНОЛОГИИ

Лекция № 10

Распределенные вычислительные системы

Юрий Иванович Молородов

yumo@ict.sbras.ru

Содержание курса

1. Гетерогенные Grid-системы на базе BOINC
2. Метакомпьютинг
3. GRID системы

Распределенные вычисления

Методы распределенных вычислений (РВ) становятся всё более популярными и востребованными. Особое место среди них занимает платформа BOINC (Berkeley Open Infrastructure for Network Computing) и вычислительные системы на основе *Grid* –системам.

BOINC предоставляет возможность гибкой настройки клиентской части, регулируя максимальный размер загружаемых файлов, время выполнения рабочих заданий, загрузку CPU или GPU, выделяемый объем оперативной памяти и дискового пространства.

Гетерогенная Grid-система на базе BOINC

Платформа организации распределенных вычислений *BOINC* - активно развивающаяся СППО с открытым исходным кодом. Она стала основой большого числа независимых проектов добровольных вычислений (наиболее популярные из них - Climateprediction.net , SETI@home и Einstein@home). Платформа *BOINC* отличается простотой в установке, настройке и администрировании. Она обладает хорошими возможностями по масштабируемости, простоте подключения новых вычислительных узлов, использованию дополнительного ПО, интеграции с другими *Grid* – системами.

Гетерогенная Grid-система на базе BOINC

Наиболее эффективно использование *Grid* при выполнении следующих задач:

- анализ и обработка независимых наборов данных;
- решение задач, обладающих хорошей степенью параллелизации по данным.

Существует большое количество систем промежуточного программного обеспечения (СППО), предназначенных для организации, сопровождения и управления *Grid* .

По назначению СППО можно условно разделить на две группы.

Гетерогенная Grid-система на базе BOINC

Первая группа содержит системы, предназначенные для объединения относительно небольшого числа высокопроизводительных вычислителей (кластеров). К таким системам относятся, например, Condor , Globus Toolkit, Unicore.

Вторая группа СППО содержит системы, цель которых заключается в объединении в *Grid* большого числа (до сотен тысяч) вычислителей, каждый из которых обладает относительно невысокой производительностью.

Гетерогенная Grid-система на базе BOINC

Такие *Grid* - сети также называют системами распределенных вычислений (*distributed computing*) и системами добровольных вычислений (volunteer computing). Их специфика заключается, в высокой вероятности недоступности отдельных вычислительных узлов.
Наиболее популярной СППО этой группы является платформа *BOINC* (Berkeley Open Infrastructure for Network Computing)
Другой пример - разработка НИВЦ МГУ им. М.В. Ломоносова система *X-Com*.

Гетерогенная Grid-система на базе BOINC

Платформа BOINC имеет архитектуру "*клиент-сервер*. *Клиентская часть* может работать на компьютерах с различными аппаратными и программными характеристиками. Ключевым объектом системы является проект - автономная сущность, в рамках которой производятся распределенные вычисления.

BOINC-сервер поддерживает одновременную работу большого числа независимых проектов. Каждый вычислительный узел может одновременно производить вычисления для нескольких BOINC- проектов.

Проект однозначно идентифицируется своим URL-адресом (*URL - Universal Resource Locator*).

(URI - Uniform Resource Identifier)

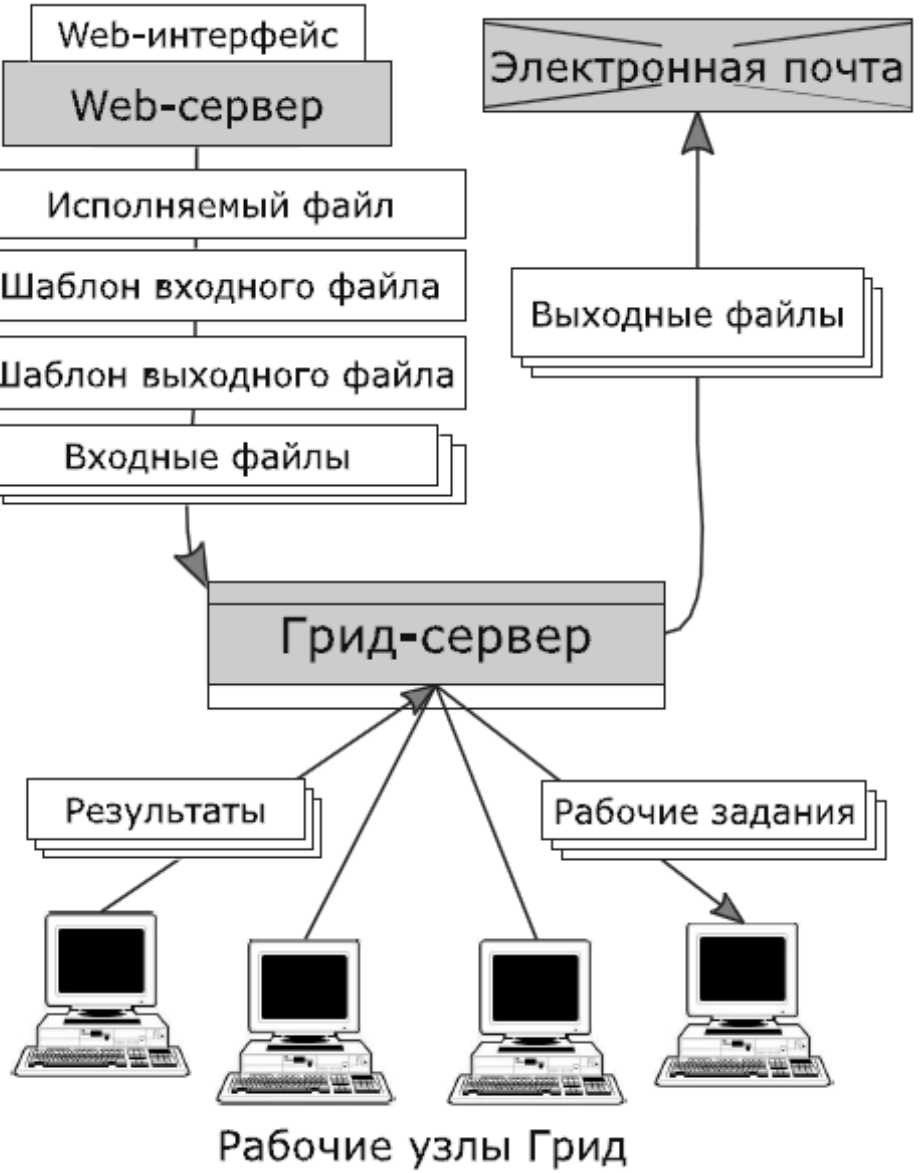
Архитектура системы на базе BOINC

Рабочий процесс в *Grid* -системе, основанной на платформе BOINC, организован следующим образом.
Вычислительные узлы, имеющие свободные ресурсы, обращаются к серверу для получения новых рабочих заданий.
Сервер BOINC рассылает клиентским приложениям экземпляры рабочих заданий, клиенты выполняют вычисления и отсылают обратно результаты.
После получения результатов сервер проверяет и обрабатывает их, например занося в базу данных или автоматически создавая на их основе новые рабочие задания.

Архитектура системы на базе BOINC

Платформа BOINC позволяет выполнять в *grid* -системе специализированные приложения, в которых, в частности, может быть реализован механизм сохранения промежуточных результатов вычислений, графическое отображение процесса выполнения приложения, автоматическое распараллеливание вычислительного процесса в зависимости от количества доступных на клиенте вычислительных ядер и т.п.
Но унаследованные (неадаптированные для работы в *Grid*) приложения с помощью специальных *приложений-"оберток"* также могут быть перенесены на платформу BOINC без необходимости изменения их исходного кода и перекомпиляции. Приложения-"обертки" берут на себя взаимодействие с ядром программы-клиента, запуская исходное приложение как свой дочерний процесс.

Архитектура системы на базе BOINC



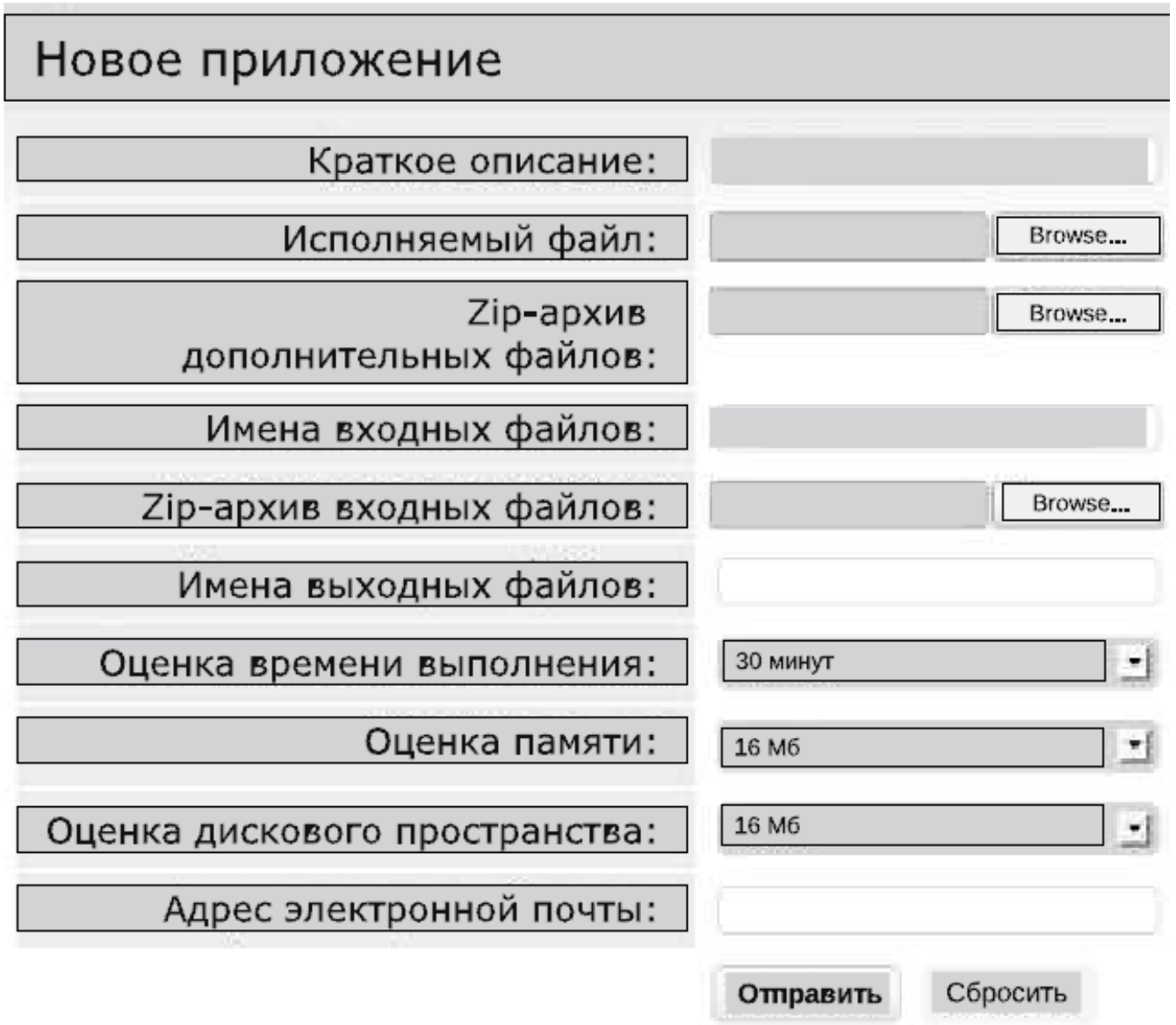
Web –интерфейс доступа к grid-сегменту

Для обеспечения удобного доступа пользователей к вычислительным ресурсам *Grid* был разработан специальный Web-интерфейс.

HTML-форма содержит:

- поля для загрузки на Web-сервер исполняемого файла приложения, дополнительных файлов приложения (при необходимости) и zip-архива входных файлов;
- поля для ввода имен входных и выходных файлов, которые требует исполняемый файл приложения;
- поля для ввода оценки ресурсов, требуемых для выполнения заданий;
- поле для ввода адреса электронной почты, на который требуется отправить результат.

Web –интерфейс доступа к Grid-сегменту



Web –интерфейс доступа к Grid-сегменту

Обязательными для заполнения являются поле для загрузки исполняемого файла приложения и поле для ввода адреса электронной почты.
Пользователь загружает исполняемый и дополнительные файлы приложения, указывает имена входных и выходных файлов и загружает zip-архив входных файлов.
Каждый из них послужит основой для создания отдельного рабочего задания.
Данные, введенные пользователем, обрабатываются на стороне Web-сервера и передаются BOINC-серверу.

Web –интерфейс доступа к grid-сегменту

Программа обработки введенных данных на стороне Web-сервера создает новое BOINC-приложение и шаблоны входных и выходных файлов.
В данном случае каждое приложение включает в себя программу-"обертку" в дополнение к исполняемому файлу приложения, загруженному пользователем.
Рабочие задания BOINC генерируются на основе загруженных входных файлов и созданных шаблонов.

Web –интерфейс доступа к grid-сегменту

Рабочие узлы *Grid* периодически отсылают на *Grid* -сервер запросы новых заданий. Программа-планировщик, входящая в состав BOINC, отвечает на запросы, распределяя неразосланные рабочие задания между узлами, которые запросили задания и которые соответствуют определенным критериям.

Например имеют достаточный объем оперативной памяти и свободного дискового пространства для выполнения заданий.

Ход выполнения рабочих заданий отображается на Web-сайте. Пользователю доступна информация об общем количестве рабочих заданий на сервере и их статусе (например, "в процессе выполнения" или "завершено").

Web –интерфейс доступа к grid-сегменту

По мере того как завершенные результаты поступают с рабочих узлов *Grid* на сервер, программа-валидатор, входящая в состав BOINC, проверяет выходные файлы и отмечает результат как правильный или ошибочный. Правильные результаты обрабатываются программой-ассимилятором, которая упаковывает их в zip-архив и отсылает пользователю по электронной почте после того, как все результаты рабочих заданий успешно обработаны.

Применение grid-сегмента в квантово-химических расчетах

Для этих расчетов использовался наиболее популярный программный пакет Firefly, частично основанный на исходном коде системы GAMESS (US). Firefly предназначен для выполнения расчетов ab initio и DFT (основе теории функционала плотности) на Intel-совместимых процессорах x86, AMD64 и EM64T.

Вычислительные методы ab initio, или неэмпирические, предназначены для расчета с максимально возможной точностью физических и химических свойств заданного химического соединения (как многоатомной системы) на основе представлений и методов квантовой механики.

Применение grid-сегмента в квантово-химических расчетах

Такие расчеты позволяют оценить вклад энергии, требующейся для перестройки электронных оболочек атомов, изменения геометрической структуры комплекса и сближения катионов между собой, в общую энергию образования частицы, состоящей из комплекса и катионов.
Исходные данные для проведения вычислительных экспериментов представляют собой наборы параметров, регулирующих работу программы, в том числе координаты атомов, метод расчета и базис волновых функций. Выходные файлы должны содержать информацию о полной энергии рассчитываемой системы, ее энергетических уровнях, параметрах молекулярных орбиталей и т.п.

Применение grid-сегмента в квантово-химических расчетах

При проведении вычислительных экспериментов в состав *Grid*-сегмента входили 14 вычислительных узлов с различными аппаратными и программными характеристиками, и разными настройками, связанными с организацией вычислений.

Два узла обслуживали проекты BOINC в монопольном режиме, а на десяти вычислительных узлах кластера расчеты в рамках *Grid* регулярно приостанавливались для выполнения расчетов пользовательских задач, запускаемых на кластере с помощью системы управления заданиями.

Суммарная пиковая производительность *Grid* составила 1.04 TFLOPS.

Система X-Com. http://x-com.parallel.ru/

Система метакомпьютинга *X-Com* предназначена для быстрого развертывания и проведения распределенных расчетов. Система представляет собой инструментарий для адаптации и поддержки выполнения программ в распределенных неоднородных средах. Система поддерживает среды с десятками тысяч вычислительных узлов (процессоров), обеспечивает корректную работу в условиях высокой динамичности состава среды, не требует административного доступа к ресурсам. Система X-Com написана на языке Perl. Это обеспечивает ее работоспособность на подавляющем большинстве современных программно-аппаратных платформ.

Система X-Com.

Ее архитектура основана на технологии клиент-сервер. Поэтому прикладная задача должна быть логически разделена на две части: *серверную* и *клиентскую*.

Серверная часть отвечает за разбиение задачи на множество независимых порций и объединение результатов. Для реализации серверной части задачи в системе *X-Com* имеется два прикладных интерфейса *API* – (Application Programming Interface) - интерфейс создания приложений.

В *простейшем случае*, если необходимо выполнить одну и ту же программу на множестве различных входных файлов, используется *API Files*, и тогда серверная часть задачи описывается в параметрах настройки сервера *X-Com* (указываются пути к входным и выходным каталогам и список файлов для обработки).

В менее тривиальных случаях применяется *API Perl*. Этот интерфейс предполагает написание модуля на языке Perl, реализующего заданный набор функций:

- ✓ инициализацию серверной части задачи,
- ✓ генерацию номера первой и последней порции,
- ✓ генерацию тела порции по нее номеру,
- ✓ обработку результатов готовой порции,
- ✓ условия завершения работы серверной части задачи,
- ✓ а также действия, выполняемые перед завершением.

Система X-Com.

Аналогичный подход применяется и для реализации *клиентской* части задачи.

В элементарном случае в параметрах настройки сервера *X-Com* указывается формат команды с именами входных и выходных файлов, которая будет запущена клиентом X-Com на вычислительном узле.

В более сложном случае клиентская часть задачи может быть описана двумя функциями на языке Perl:

- ✓ инициализация задачи на узле и
- ✓ обработка каждой порции данных.

После реализации клиентской и серверной части прикладной задачи формируется непосредственно вычислительная среда.

Система X-Com.

Этот процесс состоит из двух частей:

- ✓ запуск клиентов X-Com на вычислительных узлах и
- ✓ настройка и запуск сервера X-Com.

Клиент X-Com может быть запущен на вычислительном узле постоянно (в монопольном режиме) или запускаться в те моменты времени, когда узел не занят другими процессами (работа по занятости).

На высокопроизводительных вычислительных комплексах используются интерфейсы к штатным системам очередей.

В текущей версии X-Com поддерживается взаимодействие с *Cleo, Torque, LoadLeveler, Slurm*, а также *Unicore*.

Система X-Com.

Сервер X-Com запускается на специальной выделенной машине вручную либо с помощью подсистемы управления заданиями. Такой вариант предоставляет пользователям более удобный способ работы с распределенной средой, позволяя им оперировать привычными понятиями очереди заданий. При этом обеспечивается и последовательное выполнение задач на всех доступных ресурсах, и параллельное выполнение одновременно нескольких заданий. Возможен запуск заданий с учетом их требований к ресурсам, на которые они будут распределены.

Существенная особенность системы *X-Com* – возможность построения иерархических распределенных сред с произвольным количеством уровней.

Система X-Com.

Данная функциональность реализована с помощью промежуточных серверов X-Com.
Промежуточный сервер получает задания и необходимые данные от вышестоящего сервера X-Com (для этого сервера он представляется клиентом X-Com) и распределяет их внутри пула своих клиентов (для них он представляется единственным сервером X-Com).
Введение промежуточных серверов позволяет снизить нагрузку на центральный сервер распределенной среды, оптимизировать потоки данных и подключить к расчетам вычислительные ресурсы, находящиеся внутри закрытых сетей.

Система X-Com.

С помощью системы *X-Com* был решен целый ряд вычислительно емких задач из различных научных областей.
В каждом расчете отрабатывалась та или иная функциональность системы.
Система X-Com может применяться как основа для построения сервисов по распределению заданий на доступные вычислительные ресурсы. Совместно с компаниями Тесис и Сигма Технология в 2009-2010 гг. был реализован программный комплекс для решения оптимизационных гидродинамических задач. Программный комплекс объединял оптимизатор *IOSO*, решатель *FlowVision* и систему *X-Com*. Последняя отвечала за взаимодействие между оптимизатором, установленном на рабочем месте пользователя комплекса и генерирующим пакеты заданий для решателя, и системой очередей, через которую осуществлялся запуск пакета *FlowVision* на суперкомпьютерах МГУ.

Система X-Com.

На суперкомпьютере *СКИФ МГУ "Чебышев"* с помощью системы *X-Com* был реализован сервис выполнения пакетов однопроцессорных задач на суперкомпьютерах. Основная задача сервиса – дополнение системы управления прохождением задач - Cleo функциональностью по группировке нескольких однопроцессорных задач на один узел суперкомпьютера. Это повышало эффективность использования ресурсов вычислительной системы. Система X-Com применялась для исследования свойств прикладных задач на процессорном полигоне НИВЦ МГУ.

Система X-Com.

В рамках этой работы проводилось сравнение времени выполнения приложения, откомпилированного с использованием различных компиляторов, запускаемого по одному или более процессов на узел, на имеющихся в полигоне программно-аппаратных платформах.
Таким образом оценивалась эффективность работы приложения в зависимости от набора факторов (архитектура узла, компилятор и его опции, наличие или отсутствие параллельно работающих процессов).

Что такое Мета-компьютинг?

Термин возник вместе с развитием высокоскоростной сетевой инфраструктуры в начале 90-х годов.
Он относился к объединению нескольких разнородных вычислительных ресурсов в локальной сети организации для решения одной задачи.
Основная цель построения мета-компьютера заключалась в оптимальном распределении частей работы по вычислительным системам различной архитектуры и различной мощности.

Что такое Мета-компьютинг?

Компонентами "мета-компьютера" могут быть и простейшие ПК, и мощные массивно-параллельные системы.
Здесь важно то, мета-компьютер может не иметь постоянной конфигурации - отдельные компоненты могут включаться в его конфигурацию или отключаться от нее. При этом технологии мета-компьютинга обеспечивают непрерывное функционирование системы в целом.

Современные исследовательские проекты направлены на обеспечение прозрачного доступа пользователей через Интернет к необходимым распределенным вычислительным ресурсам, и прозрачного подключения простаивающих вычислительных систем к мета-компьютерам.

Что такое Мета-компьютинг?

Так предварительная обработка данных и генерация сеток для счета может производиться на пользовательской рабочей станции.
Основное моделирование на *векторно-конвейерном суперкомпьютере*.
Решение больших систем линейных уравнений - на *массивно-параллелльной* системе.
Визуализация результатов - на специальной *графической станции*.

В дальнейшем, исследования в области технологий мета-компьютинга были развиты в сторону однородного доступа к вычислительным ресурсам большого числа (вплоть до нескольких тысяч) компьютеров в локальной или глобальной сети.

Что такое Мета-компьютинг?

Наилучшим образом для решения на мета-компьютерах подходят задачи переборного и поискового типа. Здесь вычислительные узлы практически не взаимодействуют друг с другом, а основную часть работы производят в автономном режиме. Основная схема работы примерно такая: специальный агент, расположенный на вычислительном узле (компьютере пользователя), определяет факт простоя этого компьютера, соединяется с управляющим узлом мета-компьютера и получает от него очередную порцию работы (область в пространстве перебора). По окончании счета по данной порции вычислительный узел передает обратно отчет о фактически проделанном переборе или сигнал о достижении цели поиска.

Наиболее известные проекты по мета-компьютингу и распределенным вычислениям в Интернет

Distributed.net http://www.distributed.net/

GIMPS - Great Internet Mersenne Prime Search

SETI@home http://lnfm1.sai.msu.ru/SETI/koi/

http://ru.wikipedia.org/wiki/SETI@home

TERRA ONE

Globus

The Metacomputing Project

PACX-MPI

Condor

Distributed.net http://www.distributed.net/

Одно из самых больших объединений пользователей Интернет, предоставляющих свои компьютеры для решения крупных переборных задач. Основные проекты связаны с задачами взлома шифров (RSA Challenges). Так, 19 января 1999 года была решена предложенная RSA Data Security задача расшифровки фразы, закодированной с помощью шифра DES-III. В настоящее время в distributed.net идет работа по расшифровке фразы, закодированной с 64-битным ключом (RC5-64). С момента начала проекта в нем зарегистрировались 191 тыс. человек. Достигнута скорость перебора, равная 75 млрд. ключей в секунду (всего требуется проверить 2^{64} ключей). За решение этой задачи RSA предлагает приз в $10 тыс.

GIMPS http://www.mersenne.org/

Поиск простых чисел Мерсенна (т.е. простых чисел вида 2^P-1). С начала проекта было найдено 4 таких простых числа. Организация Electronic Frontier Foundation предлагает приз в $100 тыс. за нахождение простого числа Мерсенна с числом цифр 10 млн.

SETI@home.
http://ru.wikipedia.org/wiki/SETI@home

SETI@home (*Search for Extra-Terrestrial Intelligence at Home* — поиск внеземного разума на дому) — научный некоммерческий проект добровольных вычислений на платформе BOINC, использующий свободные вычислительные ресурсы на компьютерах добровольцев для поиска радиосигналов внеземных цивилизаций.

Логотип SETI@Home	
Тип	Распределённые вычисления
Разработчик	Калифорнийский университет в Беркли
Языки интерфейса	Мультиязычная, включаяРусский
Первый выпуск	17 мая 1999
Аппаратная платформа	Кроссплатформенное программное обеспечение
Последняя версия	7.4.42 (10 марта 2015)
Тестовая версия	7.2.42 (28 февраля 2014)
Лицензия	LGPL (как часть BOINC)
Сайт	setiathome.ssl.berkeley.edu

Обсерватория Аресибо — крупнейший радиотелескоп в мире. «Тарелка» телескопа, имеющая в диаметре 305 м и сконструированная в воронке на земле, фокусирует радиоволны на подвижной антенне. Конструкция антенны подвешена в центре на 18 тросах, крепящихся на трех башнях высотой 110 м каждая.

SETI HOME

Антенна установлена внутри защитного купола, отсеивающего радиосигналы от помех.

Как работает проект.

Проект заключается в обработке данных радиотелескопа обсерватории Аресибо на предмет поиска сигналов, которые можно интерпретировать как искусственные.
Данные, получаемые с облучателя радиотелескопа, записываются с высокой плотностью на магнитную ленту (заполняя примерно одну 35-гигабайтную DLT плёнку в день). При обработке данные с каждой ленты разбиваются на 33000 блоков по 1049600 байт, что составляет 1,7 сек времени записи с телескопа. Затем 48 блоков конвертируются в 256 заданий на расчёт, которые рассылаются не менее чем на 1024 компьютера участников проекта.

После обработки результаты передаются компьютером участника проекта в Space Sciences Laboratory (SSL) Калифорнийского университета, Беркли (США), с помощью программного обеспечения BOINC.
Каждый пользователь персонального компьютера, имеющий доступ к Интернету, может подключиться к проекту (такой подход даёт беспрецедентную вычислительную мощность, обусловленную большим количеством компьютеров, участвующих в обработке данных).

Новаторский подход к проблеме продемонстрировали астрономы из Университета Калифорнии в Беркли: в 1999 г. они запустили в действие проект*SETI@home*. Идея проекта — привлечь к работе миллионы владельцев персональных компьютеров, чьи машины большую часть времени просто бездействуют. Те, кто участвует в проекте, скачивают из Интернета и устанавливают на своем компьютере пакет программ, которые работают в режиме скринсейвера, а потому не доставляют владельцу никаких неудобств. Эти программы участвуют в расшифровке сигналов, принятых радиотелескопом.

До настоящего момента к проекту присоединились 5 млн пользователей в 200 с лишним странах мира; вместе они потратили электричества больше чем на миллиард долларов, но каждому пользователю участие в проекте стоило недорого. Это самый масштабный коллективный компьютерный проект в истории; он мог бы послужить образцом для других проектов, где требуются большие вычислительные мощности. Тем не менее до сих пор проект *SETI@home* также не обнаружил ни одного разумного сигнала.

SETI@home (Search for ExtraTerrestrial Intelligence at home.
Поиск ВнеЗемного Разума на дому) - это научный эксперимент, который использует соединяемые через сеть Интернет компьютеры в Поиске ВнеЗемного Разума (SETI).
Вы можете принять участие запустив бесплатную программу, которая загружает и анализирует данные с радиотелескопа.

SETI@home http://setiathome.ssl.berkeley.edu

Проект Search for Extraterrestrial Intelligence - поиск внеземных цивилизаций с помощью распределенной обработки данных, поступающих с радиотелескопа Arecibo в Пуэрто-Рико. Калифорнийский университет Беркли.
Основа - распределенные вычисления через Интернет. Это теоретически позволяет получить виртуальный суперкомпьютер с производительностью в сотни раз большей, чем у самого мощного ныне существующего компьютера.
Присоединится может любой желающий. Доступны клиентские программы для Windows, Mac, UNIX, OS/2 (клиент Windows срабатывает в качестве screen-saver'a).
Для участия в проекте зарегистрировались около 2 млн. человек.

SETI@home Развитие проекта

На 25 марта 2012 г. проект является наиболее популярным на платформе BOINC — общее число участников проекта составляет более 1,2 млн. По объёму вычислений в день по состоянию на 25 марта 2012 года проект занимает пятую позицию с результатом 1,6 петафлопс.
Результаты используются также и для исследования других астрономических объектов.
Дальнейшее продолжение и дополнение к проекту SETI@Home — проект AstroPulse (Beta (астрономические исследования).
Для AstroPulse (Beta) существуют клиенты для GNU/Linux (в том числе и для 64-разрядных версий) и Microsoft Windows.
27 января 2009 года было объявлено о создании нового открытого проекта — *setiQuest*.
В его основу лягут исходные коды SETI@Home.

Развитие проекта

Астроном Джилл Тартер запустила в интернете проект SETIQuest , который призван вступить в контакт с внеземными мирами при помощи специальных сигналов. Всем желающим предлагается улучшить существующий алгоритм обработки сигналов для поисков внеземной жизни, для этого на сайте будут раскрыты его исходные коды.

SETIQuest выложил большое количество данных SETI (Search for Extraterrestrial Intelligence) по поиску инопланетян.
Сейчас организаторы предлагают принять более активное участие в поисках внеземной жизни.

Развитие проекта

Зарегистрировавшись на сайте, участник SETIQuest может совершенствовать код программ, которые используются для расшифровки и обработки цифрового сигнала с системы телескопов Аллена (Allen Telescope Array) - основного инструмента работы SETI.

Новички могут присоединяться к SETIQuest по мере его развития и изучать обработанные данные в коллективном поиске возможных сигналов искусственного происхождения. Достаточно скачать программу, которая будет расшифровывать сигналы, собранные массивом радиотелескопов Аллена.

Цель проекта - вычленить из общего шума периодические сигналы, источником которых могут быть приборы обитателей иных миров.

The DIMES project

Distributed Internet MEasurements & Simulations.
Распределённый научный проект, нацеленный на изучение структуры и топологии Интернета.

Интернет построен таким образом, что единственным эффективным способом построить его карту является сделать это распределённо.

Поэтому организаторы просят вас принять участие в проекте. Здесь важны не столько ваши циклы CPU или сетевой трафик (который наш клиент почти не потребляет), сколько ваше местонахождение. Чем в большем количестве мест работают клиенты проекта, тем точнее будут составленные им карты.

Понимание структуры Интернета — важная исследовательская задача. Её решение сможет позволить сделать Интернет более приятным местом для каждого из нас.

The DIMES project

Distributed Internet MEasurements & Simulations.
Распределённый научный проект, нацеленный на изучение структуры и топологии Интернета.

Клиент DIMES выполняет измерения состояния сети, такие как TRACEROUTE или PING. Делает это он с низкой частотой, потребляя в пике до 1 килобайта в секунду. Агент никуда не отсылает ни информацию о действиях, выполняемых на компьютере, ни вашу личную информацию — он отсылает ТОЛЬКО результаты измерений.

Проект DIMES http://www.evergrow.org/ является частью EVERGROW — исследовательского консорциума, состоящего из более чем 20 университетов и исследовательских институтов в различных странах.

Центр проекта находится на факультете EE-систем университета Тель-Авива.

The DIMES project

Distributed Internet MEasurements & Simulations.
Распределённый научный проект, нацеленный на изучение структуры и топологии Интернета.

Для участия в проекте требуется постоянное подключение к Интернету. Скорость подключения не важна.
Клиент написан на Java, и требует установленной Java версии 1.4 или выше. Распространяется вместе с исходным текстом, под лицензией GNU LGPL. Дистрибутив клиента весит около 4.5 мегабайт.

У проекта есть огромный плюс: в свёрнутом состоянии, клиент абсолютно не потребляет CPU. А это значит, что можно одновременно участвовать в DIMES, и в каком-нибудь другом проекте.

TERRA ONE

Коммерческий проект компании Cerentis ставит своей целью объединение множества персональных компьютеров, подключенных (или периодически подключаемых) к Интернет, для решения задач анализа информации, предоставляемой различными заказчиками.

Клиентские компьютеры (TerraProcessor), подключенные к TERRA ONE, используются во время простаивания с помощью screen-saver'a.

За обработку информации владельцы ПК получают возможность покупки в Интернет-магазинах - им начисляются "кредиты" (TerraPoints) за каждую единицу обработанной информации.

Globus http://www.globus.org.

Проект реализуется в Argonne National Lab. Цель The Globus Project - построение т.н. "computational grids", включающих в себя вычислительные системы, системы визуализации, экспериментальные установки.

В рамках проекта проводятся исследовании по построению распределенных алгоритмов, обеспечению безопасности и отказоустойчивости мета-компьютеров.

Программное обеспечение Globus 1.0.0 доступно бесплатно. Доступна реализация MPI (MPICH-G) поверх Globus.

Для тестирования Globus был создан реальный метакомпьютер GUSTO (testbed environment), который включает около 40 компонент с суммарной пиковой производительностью 2.5 TFLOPS.

Globus http://www.globus.org.

В рамках проекта разработан ряд программных средств:

•Globus Resource Allocation Manager - единообразный интерфейс к различным "локальным" системам распределения нагрузки (LSF, NQE, LoadLeveler). Для описания требований приложения к ресурсам разработан специальный язык RSL (Resource Specification Language).

•Globus Security Infrastructure - система аутентификации на базе открытого ключа и X.09-сертификатов.

•Metacomputing Directory Service (MDS) - репозиторий информации о вычислительных ресурсах, входящих в метакомпьютер.

•Nexus - коммуникационная библиотека.

•Heartbeat Monitor (HBM) - средство мониторинга, позволяющее определить сбой некоторых машин и процессов, входящих в метакомпьютер.

•Globus Access to Secondary Storage (GASS) - средство доступа к удаленным данным через URL (Uniform Resource Locator).

Legion http://legion.virginia.edu/

A Worldwide Virtual Computer университета Вирджинии.
Цель - разработка объектно-ориентированного ПО для построения виртуальных мета-компьютеров, включающих до нескольких миллионов индивидуальных хостов, объединенных высокоскоростными сетями.
Пользователь, работающий на своем домашнем компьютере, должен иметь абсолютно прозрачный доступ ко всем ресурсам мета-компьютера.
В рамках Legion возможно исполнение параллельных приложений - поддерживаются библиотеки MPI и PVM, а также язык Mentat.
Программное обеспечение проекта доступно бесплатно (поддерживаются платформы SGI IRIX, Linux, Alpha/OSF1, RS/6000).

Распределенные вычисления Grid computing

Распределенные вычисления являются технологией, которая с одной стороны оказала влияние на появление концепции облачных вычислений, а с другой стороны имеет ряд существенных отличий.
Речь идет о коллективных, или распределённых вычислениях (grid computing) - когда большая ресурсоёмкая вычислительная задача распределяется для выполнения между множеством компьютеров, объединённых в мощный вычислительный кластер сетью в общем случае или интернетом в частности.

Распределенные вычисления Grid computing

Установление общего протокола в сети Интернет непосредственно привело к быстрому росту онлайн пользователей. Это привело к необходимости выполнять больше изменений в текущих протоколах и к созданию новых. На текущий момент обширно используется протокол Ipv4 (четвёртая версия IP протокола), но ограничение адресного пространства, заданного ipv4, неизбежно приведет к использованию протокола ipv6.
В течение долгого времени усовершенствовалось аппаратное и программное обеспечение, в результате чего удалось построить общий интерфейс в Интернет.
Использование веб-браузеров привело к использованию модели "Облака", взамен традиционной модели информационного центра.

Распределенные вычисления Grid computing

В начале 1990-ых, Иэн Фостер и Карл Кесселмен представили их понятие *Грид* вычислений. Они использовали аналогию с электрической сетью, где пользователи могли подключаться и использовать услугу. *Грид* вычисления во многом опираются на методах, используемых в кластерных вычислительных моделях, где многократные независимые группы, действуют, как сеть просто потому, что они не все расположены в пределах той же области. В частности, развитие Грид технологий позволило создать так называемые GRID-сети, в которых группа участников могла общими усилиями решать сложные задачи.

Распределенные вычисления Grid computing

Так, сотрудники IBM создали интернациональную команду grid-вычислений, позволившую существенно продвинуться в области борьбы с вирусом иммунного дефицита.
Целые команды из разных стран присоединяли свои вычислительные мощности и помогли "обсчитать" и смоделировать наиболее перспективные формы для создания лекарства от СПИДа..."

Распределенные вычисления Grid computing

Вообще говоря, " *Грид* -вычисления" — это форма распределённых вычислений, в которой «виртуальный суперкомпьютер» представлен в виде кластера соединённых с помощью сети, слабосвязанных, гетерогенных компьютеров, работающих вместе для выполнения огромного количества заданий (операций, работ). Эта технология применяется для решения научных, математических задач, требующих значительных вычислительных ресурсов. Грид-вычисления используются также в коммерческой инфраструктуре для решения таких трудоёмких задач, как экономическое прогнозирование, сейсмоанализ, разработка и изучение свойств новых лекарств.

Распределенные вычисления Grid computing

Грид с точки зрения сетевой организации представляет собой согласованную, открытую и стандартизованную среду, которая обеспечивает гибкое, безопасное, скоординированное разделение вычислительных ресурсов и ресурсов хранения информации, которые являются частью этой среды, в рамках одной виртуальной организации.

Грид - географически распределенная инфраструктура, объединяющая множество ресурсов разных типов (процессоры, долговременная и оперативная память, хранилища и базы данных, сети), доступ к которым пользователь может получить из любой точки, независимо от места их расположения.

Распределенные вычисления Grid computing

Грид предполагает коллективный разделяемый режим доступа к ресурсам и к связанным с ними услугам в рамках глобально распределенных виртуальных организаций,
состоящих из предприятий и отдельных специалистов, совместно использующих общие ресурсы. В каждой виртуальной организации имеется своя собственная политика поведения ее участников, которые должны соблюдать установленные правила. Виртуальная организация может образовываться динамически и иметь ограниченное время существования.

Распределенные вычисления Grid computing

Потенциал технологий *грид* уже сейчас оценивается очень высоко: он имеет стратегический характер, и в близкой перспективе *грид* должен стать вычислительным инструментарием для развития высоких технологий в различных сферах человеческой деятельности, подобно тому, как подобным инструментарием стали персональный компьютер и интернет. Такие высокие оценки можно объяснить способностью *грид* на основе безопасного и надежного удаленного доступа к ресурсам глобально распределенной инфраструктуры решить две проблемы:

Распределенные вычисления Grid computing

1. создания распределенных вычислительных систем сверхвысокой пропускной способности из серийно выпускаемого оборудования при одновременном повышении эффективности (до 100%) имеющегося парка вычислительной техники путем предоставления в *грид* временно простаивающих ресурсов;
2. создания широкомасштабных систем мониторинга, управления, комплексного анализа и обслуживания с глобально распределенными источниками данных, способных поддерживать жизнедеятельность государственных структур, организаций и корпораций.

Распределенные вычисления Grid computing

На практике границы между этими (*grid* и *cloud*) типами вычислений достаточно размыты.
Сегодня с успехом можно встретить "облачные" системы на базе модели распределённых вычислений, и наоборот.
Однако будущее облачных вычислений всё же значительно масштабнее распределённых систем, к тому же не каждый "облачный сервис" требует больших вычислительных мощностей с единой управляющей инфраструктурой или централизованным пунктом обработки платежей.

Проблема «NUG30»

Из группы задач QAP (quadratic assignment problem) типа задачи коммивояжера, возникшей еще в 1968 году при решении задачи тестирования в прикладной теории размещений.

Имеется множество точек куда должно быть доставлено какое-либо воздействие и требуется определить стоимость передачи этих воздействий для каждой пары точек.

Поток воздействий между каждой парой умножается на расстояние между точками и так для всех пар. В задаче «nug30» имеется 30 воздействий, каждое из которых должно быть доставлено в 30 фиксированных точек.

Требуется минимизировать стоимость передачи воздействий между точками, иначе говоря, найти оптимальный маршрут.

Проблема «NUG30»

Задача имеет массу практических применений от планирования сети госпиталей или заводов до проектирования микропроцессоров. Однако несмотря на простоту постановки ее оптимальное решение весьма нетривиально.

В рамках проекта PACI (Advanced Computational Infrastructure) над решением задачи «nug30» работало в среднем 650 процессоров (в пиковые периоды их число составляло 1009), установленных на компьютерах пяти различных платформ, физически расположенных в восьми разных местах по всему миру (несколько штатов США плюс компьютерный центр в Италии).

Проблема «NUG30»

Счет шел в течение семи дней, что составило 96 тыс. часов процессорного времени.
Без использования метакомпьютера процесс решения затянулся бы на многие годы.
Инструментальным средством, позволившим выполнить такой огромный объем работ был тандем из системы Condor и Globus.

Кластерная система Condor

(*www.cs.wisc.edu/condor/downloads*).

В университете Wisconsin-Madison была развернута первая конфигурация.

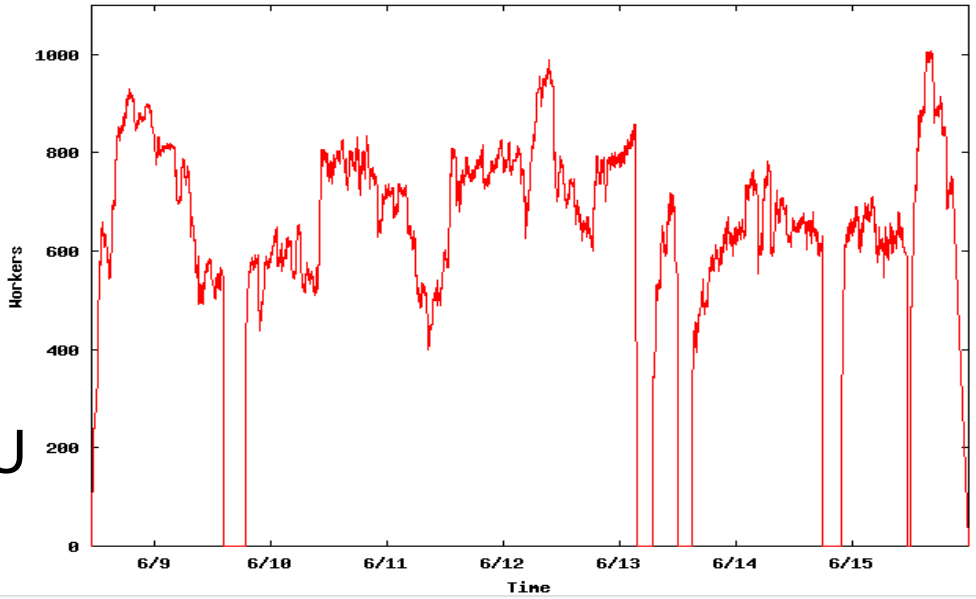
На сегодняшний день в университете имеется 350 настольных UNIX станций, которые включены в сеть Condor и предоставляют доступ для работы пользователям со всего мира.

Система свободно распространяется в загрузочных модулях для следующих платформ:

- HP PA-RISC (PA7000 и PA8000) HPUX 10.20;
- Sun SPARC Sun4m,c, Sun UltraSPARC Solaris 2.5.x, 2.6, Solaris 2.7 («с ограничениями»);
- Silicon Graphics MIPS (R4400, R4600, R8000, R10000) IRIX 6.2, 6.3, 6.4, 6.5;
- Intel x86, Pentium Linux 2.0.x, 2.2.x, glibc20, glibc21, libc5, Solaris 2.5.x, 2.6, Windows NT 4.0 («с ограничениями»);
- Digital ALPHA OSF/1 (Digital Unix) 4.x, Linux 2.0.x, Linux 2.2.x («с ограничениями»).

Математики решили задачу NUG30

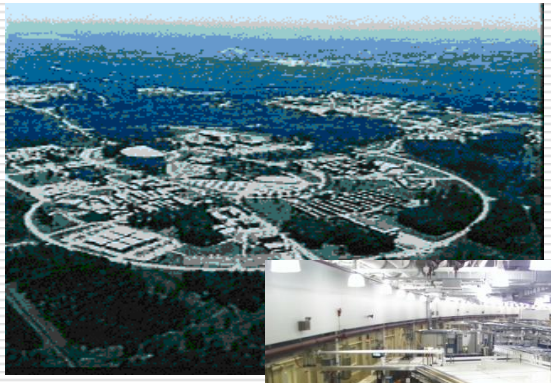
- Поиск решения NUG30 quadratic assignment problem
- Совместная работа математиков и компьютерных специалистов
- Condor-G произвёл 3.46E8 CPU секунд за 7 дней (max 1009 процессоров) в США и Италии (8 организаций)



14,5,28,24,1,3,16,15,
10,9,21,2,4,29,25,22,
13,26,17,30,6,20,19,
8,18,7,27,12,11,23

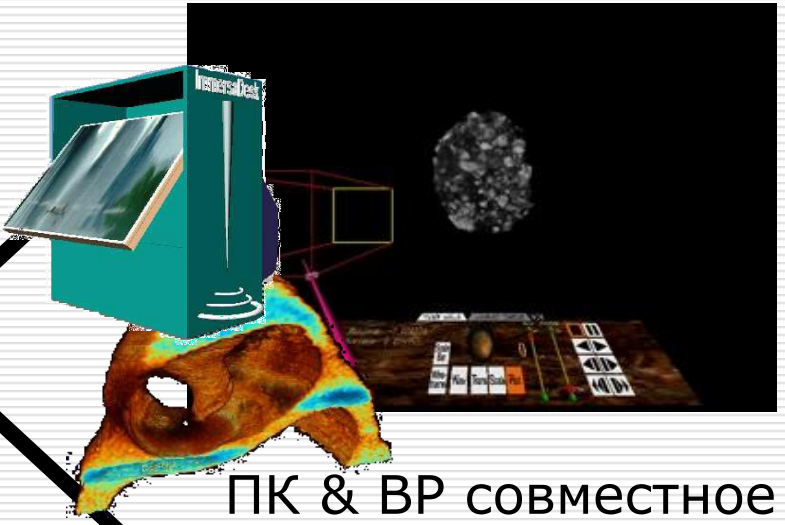
Доступ в сети к научным инструментам

Advanced Photon Source



Сбор данных в режиме реального времени

wide-area dissemination



ПК & ВР совместное управление

архивы

Томографическая реконструкция

Grids в Физике Высоких энергий

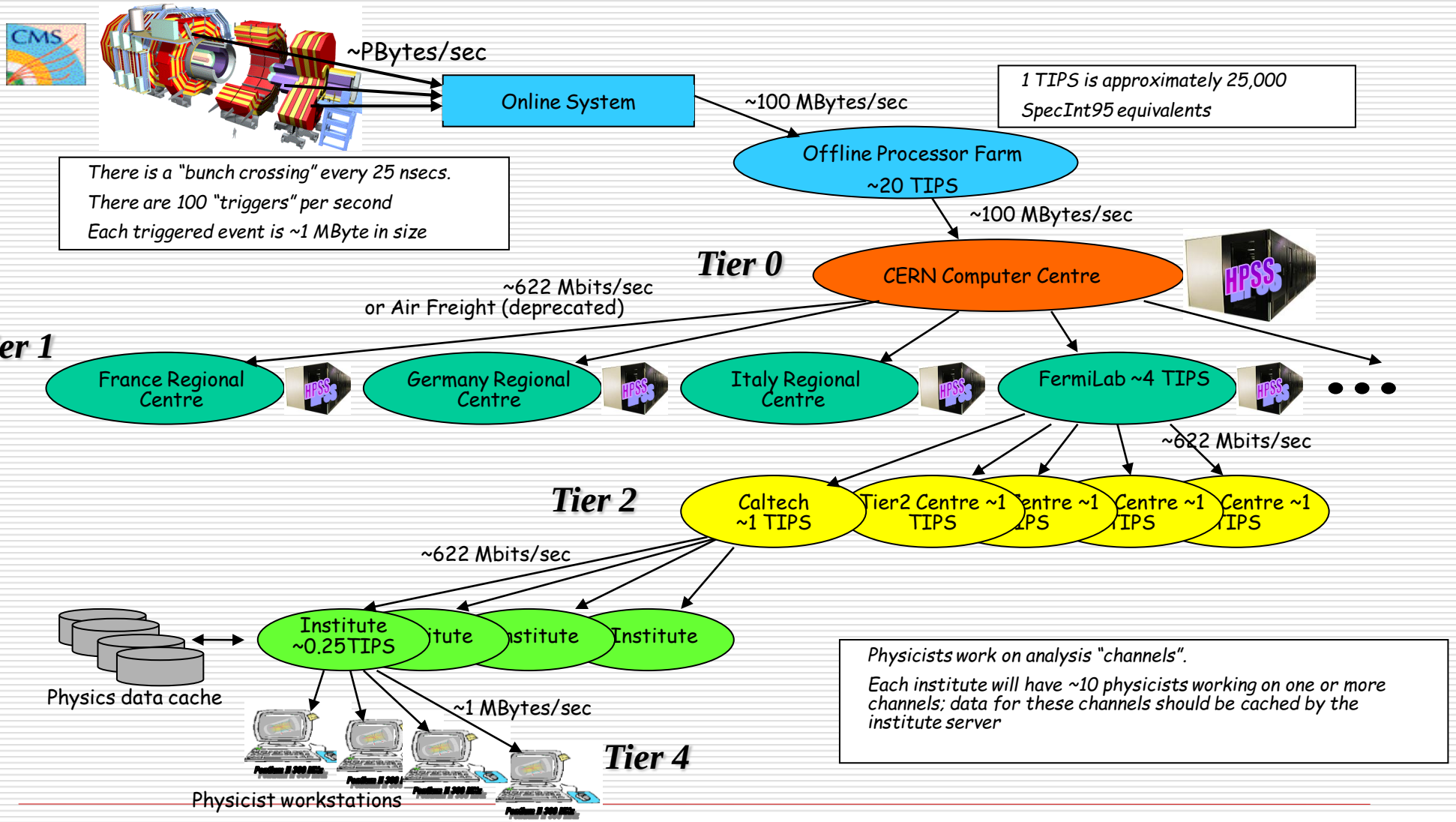


Image courtesy Harvey Newman, Caltech

Домашние компьютеры тестируют лекарства от СПИДа

- Кто =
  - 1000s домашних ПК
  - компания Entropia
  - Научно-исследовательская компания Scripps
- Единая Цель = ускорить исследования в области СПИДа

Технология MapReduce

Основными недостатками Grid-систем являются — хранение всех данных на одном жестком диске в каждой ноде, вместо распределения данных между нодами кластера, низкая скорость чтения данных с жесткого диска, которая обычно ограничена 75 Mb/sec, а также проблемы синхронизации данных между нодами. Мотивацией к созданию новой технологии послужила необходимость быстрой обработки большого количества данных и генерации на основе результатов обработки других выходных данных, таких как статистика и результаты поиска. При работе с таким количеством данных возрастает потребность в надежности, простоте поддержки и отказоустойчивости системы.

Технология MapReduce

Технология Map/Reduce была разработана исходя именно из этих потребностей. Соответственно, область применения Map/Reduce лежит там, где обрабатываемое количество данных исчисляется в терабайтах. Типичные задачи, решаемые с помощью Map/Reduce:

- Индексация большого количества данных
- Анализ логов или генерация статистики
- Data mining
- Сортировка

Технология MapReduce

MapReduce — программный фреймворк, представленный компанией Google, используемый для параллельных вычислений над очень большими, несколько петабайт, наборами данных в компьютерных кластерах.
MapReduce — это фреймворк для вычисления некоторых наборов распределенных задач с использованием большого количества компьютеров (называемых «нодами»), образующих кластер.

Работа MapReduce состоит из двух шагов: Map и Reduce.

Технология MapReduce

На *Map*-шаге происходит предварительная обработка входных данных.
Для этого один из компьютеров (называемый главным узлом — master node) получает входные данные задачи, разделяет их на части и передает другим компьютерам (рабочим узлам — worker node) для предварительной обработки.
Название данный шаг получил от одноименной функции высшего порядка.

На *Reduce*-шаге происходит свёртка предварительно обработанных данных.
Главный узел получает ответы от рабочих узлов и на их основе формирует результат — решение задачи, которая изначально формулировалась.

Технология MapReduce

Преимущество *MapReduce* заключается в том, что он позволяет распределенно производить операции предварительной обработки и свертки.
Операции предварительной обработки работают независимо друг от друга и могут производиться параллельно (хотя на практике это ограничено источником входных данных и/или количеством используемых процессоров).
Аналогично, множество рабочих узлов могут осуществлять свертку — для этого необходимо только чтобы все результаты предварительной обработки с одним конкретным значением ключа обрабатывались одним рабочим узлом в один момент времени.

Технология MapReduce

Хотя этот процесс может быть менее эффективным по сравнению с более последовательными алгоритмами, MapReduce может быть применен к большим объемам данных, которые могут обрабатываться большим количеством серверов. Так, MapReduce может быть использован для сортировки петабайта данных, что займет всего лишь несколько часов. Параллелизм также дает некоторые возможности восстановления после частичных сбоев серверов: если в рабочем узле, производящем операцию предварительной обработки или свертки, возникает сбой, то его работа может быть передана другому рабочему узлу (при условии, что входные данные для проводимой операции доступны).

Заключение

В мировой экономике происходят процесса глобализации и информационной интеграции. Они затронули и нашу страну, которая в силу географического положения и размеров вынуждена применять распределенные информационные системы. Распределенные информационные системы обеспечивают работу с данными, расположенными на разных серверах, различных аппаратно-программных платформах и хранящимися в различных форматах. Они легко расширяются, основаны на открытых стандартах и протоколах, обеспечивают интеграцию своих ресурсов с другими ИС, предоставляют пользователям простые интерфейсы.

Заключение

Распределенные вычисления и облачные технологии, технология MapReduce близки по своей сути и каждый имеют свои преимущества и недостатки.
На данный момент идет активная разработка и совершенствование данных технологи. Но речь идет именно о разработке, а не об использовании. На данный момент многие бояться именно самого факта, что информацию будут хранить сторонние люди. И хотя почти невозможность утери либо кражи данных уже доказана, немногие готовы довериться подобным сервисам.
Так же сказывается недостаточное на данный период времени качество, стабильность и скорость Интернет-соединений, что создает ощутимые трудности для разработчиков.

Заключение

Однако, несмотря на эти существенные недостатки, плюсы от внедрения данной технологии ясны всем.

Ведь это экономия для потребителей, борьба с пиратством для разработчиков, минимизация затрат в IT сфере для бизнеса, унификация сетевых стандартов для всех пользователей.

Так же следует помнить, что облако из тысяч машин способно решать действительно большие вычислительные задачи, необходимые современным ученым всех отраслей.

Лекция окончена!

Благодарю за внимание!