ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

Лекция 5. Грид-технологии и Большие данные

Курс лекций

Что такое грид-вычисления?

Грид-вычисления (от grid – решетка, сеть):

- форма распределенных вычислений, в которой «виртуальный суперкомпьютер» представлен в виде кластеров соединённых с помощью сети, слабосвязанных, гетерогенных компьютеров, серверов, СХД, ЦОД, работающих вместе для выполнения очень большого объема вычислений (заданий, операций, работ).
- географически распределённая 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.

Преимущество распределённых вычислений - отдельная ячейка вычислительной системы может быть приобретена как обычный неспециализированный компьютер, т.е. вычислительные мощности суперкомпьютера, можно получить с гораздо меньшей стоимостью.

Распределенный характер грид-систем роднит их с распределенными информационными системами.

История и примеры

Обосновали Я.Фостер, К.Киссельман, С.Тики в начале 90-х годов. Их определение: «грид-компьютинг - это скоординированное разделение ресурсов и решение задач в динамически меняющихся виртуальных организациях со многими участниками».

Метафора, обозначавшая возможность простого доступа к вычислительным ресурсам, как к электрической сети (аналог power grid)

Применяется преимущественно для научных исследований, требующих громадных вычислительных ресурсов.

Примеры:

- большой адронный коллайдер самый большой в мире ускоритель элементарных частиц (85 % всех вычислительных задач сейчас выполняется вне ЦЕРНа);
- проект Fusion - разработка метода получения электроэнергии с помощью термоядерного синтеза на экспериментальном реакторе (ТОКАМАК).

В России: Дубна (ОИЯИ), Москва (НИИЯФ МГУ, ФИАН, ИТЭФ), Протвино (ИФВЭ), Гатчина (ПИЯФ). В единую сеть с этими центрами связаны и центры других стран-участниц ОИЯИ — в Харькове, Минске, Ереване, Софии, Баку, Тбилиси.

Критерии грид-системы

Отцы-основатели видят 3 критерия. Система называется грид, если она:

1. координирует использование ресурсов при отсутствии централизованного управления этими ресурсами (если это не так, мы имеем дело с локальной системой управления);
2. использует стандартные, открытые, универсальные протоколы и интерфейсы (если это не так, мы имеем дело со специализированной прикладной системой);
3. нетривиальным (точнее, неаддитивным) образом обеспечивает высококачественное обслуживание (выгода от использования комбинированной системы значительно выше, чем от суммы ее отдельных частей) .

Грид-системы и суперкомпьютер

В обоих случаях используется принцип распараллеливания вычислений и имеется некоторое управляющее ПО.

В суперкомпьютерах – большое число процессоров объединяется локальной высокоскоростной шиной.

В грид-системах – вычислительные ресурсы, сконцентрированные в различных ЦОД (серверы со стандартными процессорами, СХД, ИБП и т. д. объединяются через сети (локальные и/или глобальные) при помощи стандартных протоколов.

Грид-системы и облако (grid & cloud)

Я.Фостер: «Облака выросли из грид-вычислений и основываются на концепции инфраструктуры грид. Эволюция подхода заключается в том, что вместо предоставления "сырых" вычислительных ресурсов и ресурсов хранения данных, в облаках обеспечивается предоставление более абстрактных ресурсов в виде сервисов».

Различия

Грид-системы:	Облачные вычисления
- одна сложная задача распределяется на несколько вычислительных узлов, что обеспечивает высокую загрузку вычислительных ресурсов;	- нескольких задач выполняются на одном физическом сервере, разделенном на виртуальные машины;
- используются для исполнения задач за ограниченный промежуток времени;	- ориентированы на предоставление "долгоживущих" сервисов;
- ориентированы на решение отдельных научных задач посредством суперкомпьютерных систем;	- ориентированы на непрерывное предоставление определенных сервисов конечным пользователям;
- строятся на базе нескольких компаний с четкими правилами взаимодействия и предоставления программно-аппаратных ресурсов;	- позволяют любой компании использовать сервисы, оплачивая только те ресурсы, которые необходимы для решения ее собственных задач;
- предоставляют программно-аппаратную базу для развертывания вычислительной инфраструктуры;	- предоставляют интегрированный подход для всех моделей информационных услуг: IaaS, PaaS, SaaS;
- интерфейсы ориентированы на взаимодействие посредством специального интерфейса, которым может воспользоваться только профессиональный программист.	- для каждой модели (IaaS, PaaS, SaaS) предоставляется свой интерфейс, что позволяет удовлетворить потребности, как отдельных пользователей, так и корпоративных клиентов.

Big Data: Новое слово в ИТ?

Большие Данные - серия подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов и значительного многообразия для получения результатов, доступных для восприятия человека. Этот подход актуален в условиях непрерывного прироста информации, увеличения ее неоднородности и распределенности по узлам вычислительной сети. Данный подход сформировался в конце 2000-х годов, как альтернативных традиционным СУБД и решениями класса бизнес аналитики (BI - Business Intelligence).

Характеризуются тремя большими «V»:

1. volume – объем (терабайты - 2^{40}, петабайты - 2^{50}, экзабайты - 2^{60});
2. velocity – скорость (и прироста данных, и их обработки, и выдачи по запросу; в идеале – в реальном масштабе времени);

3. variety - многообразие (возможность одновременной обработки различных типов структурированных и неструктурированных данных - информации с сенсоров, поисковых систем, социальных сетей, медицинская и финансовая информация, SMS, мультимедиа: фотографии, презентации с графикой, музыкой, аудио и видео).

Есть еще две характеристики «Больших данных» — их ценность (принятие верного решения в нужный момент времени) и возможность работы с ними без предварительной подготовки данных.

История: статья Клиффорда Линча «Как могут повлиять на будущее науки технологии, открывающие возможности работы с большими объёмами данных?» в журнале «Nature» от 03.09.2008г.

Метафоры: аналог «Большой нефти», «Большой руде» и т.д.

По итогам 2011 года - **явление номер два** в информационной инфраструктуре после виртуализации.

Актуальность: в 2013 году объем мировых данных превысил 1,2 зеттабайт (2^{70}), в 2015 ожидалось уже 8 зеттабайт [для справки, есть еще 1 йотабайт = 2^{80}]. Т.е. почти удвоение по закону Мура. Если записать 8 ZB на диски, то это будет примерно 20 стопок высотой от Земли до Луны.

Google - 31 миллиард запросов в месяц, в день обрабатывает более 1 петабайта.

Facebook – 750 миллионов пользователей,10 млн. загрузок фотографий ежечасно. «Нравится» – 3 млрд. раз в день.

Twitter – 400 млн. обращений в день в 2012г. С увеличением в год на 200%.

Области применения

- военные применения (космическая и аэроразведка, мониторинг ситуации)
- научные исследования (мониторинг среды, зондирование атмосферы, расшифровка генома человека);
- медицина (обследование организма в целом, анализ аномалий генов конкретного человека, статистика);
- коммерция (анализ влияния большого числа факторов на объемы продаж большого числа товаров).

Основные тренды развития

1. Сложность данных (данные могут быть как структурированными, так и неструктурированными).
2. Сложность анализа (могут анализироваться одновременно изображения, видео, тексты, производиться распознавании образов т. д.).
3. Растущие требования к бизнес аналитике (прогнозирование в реальном масштабе времени).
4. Меняющаяся экономика вычислений (облачные вычисления снижают стоимость хранения и обработки данных).
5. Легкость и дешевизна распараллеливания обработки.

Особенности работы с Большими Данными

1. Анализируются все данные, а не статистические выборки

Для определения зоны распространения гриппа N1H1 специалисты Google выявили 45 из 50 миллионов условий поиска определенных лекарств и сравнив их с зонами распределения гриппа за 2003-2008 годы. Точность определения территорий распространения заболевания составила 97%.

Стив Джобс продлил себе жизнь на несколько лет проанализировав свою ДНК полностью, что позволило врачам менять лекарства при мутациях его раковой опухоли.

Компания Xoom, специализирующаяся на денежных переводах, проанализировав все данные по операциям с кредитными картами, обнаружила действия преступной группировки.

Анализ результатов всех боев в борьбе сумо позволил выявить наиболее вероятные договорные бои.

2. Отсутствие точности

В мире БД высокая точность невозможна – данные постоянно меняются, неупорядочены, разного качества, разбросаны по разным серверам иногда по всему миру.

Переводы Google охватывали миллионы страниц переводимых документов различного качества, взятых из интернет-контента. Система содержала триллион слов в 95 миллиардах англоязычных предложений сомнительного качества. К середине 2012 года служба охватила более 60 языков и способна принимать голосовой ввод с 14 языков для моментального перевода.

Индекс потребительских цен – опрос по ценам на 23 000 товаров в 90 городах США. Сканирование Web-страниц позволяет учесть стоимость 5 млн. товаров, хотя точность сведений гораздо ниже, чем при опросах.

Отсутствие жесткого структурирования записей в базах данных.

3. Корреляция, а не причинность

Отход от поиска причинностей: вместо причинностей – корреляции. Если мы знаем, что сочетание двух веществ излечивает определенную болезнь, то нам не так важно, почему это происходит.

Amazon - предложение книг не по тому, что покупал данный человек ранее, а по схожести самих книг, т.е. по корреляции содержания.

Walmart – повышенный спрос на тосты Pop-Tarts в период приближения стихийных бедствий.

Российские особенности

- основные потребители – банковский сектор (работа с клиентскими базами) и телеком (анализ абонентской базы);

- перспективны госсектор (электронное правительство) и медицина (быстрый анализ общего состояния пациента);

- ИТ-компании типа Google и Amazon, держатели больших объемов данных, пока отсутствуют, но перспективны «Яндекс», «Mail.ru»;

- научно-исследовательские организации могли бы использовать, но бюджеты маловаты;

- МСБ не имеет бюджетов для работы с БД;

- исследовательские центры EMC в Санкт-Петербурге и Сколково (биомедицина и энергоэффективность).