

"[ЗА]ОБЛАЧНЫЕ" БАЗЫ ДАННЫХ ДЛЯ НАУКИ

М.Е. Прохоров, О.С. Бартунов

В последние годы в Сети стали развиваться публичные мегапроекты, подобные Amazon, Google и Yahoo, для поддержания которых используются горизонтально масштабируемые системы содержащие десятки и сотни тысяч компьютеров. Для обеспечения надежности подобных систем требуется наличие достаточно большого резерва, т.е. готовых к работе, но не используемых в системе компьютеров.

Некоторое время назад у разработчиков подобных проектов возникла идея, что эти простаивающие мощности можно сдавать в аренду нуждающимся в них клиентам из интернет-сообщества.

Немного позднее эта идея дополнилась созданием на базе неиспользуемых ресурсов мегапроектов публичных сервисов достаточно высокого уровня, обладающих обширной функциональностью и готовых клиентскому использованию.

Подобный подход получил название "Cloud computing" [1], что обычно переводится на русский как "облачные" или "заоблачные вычисления" (а также "облачные" сервисы, службы, базы данных, архивы и т.д.).

Можно привести следующие примеры подобных сервисов.

- Онлайн-приложения Google: GMail, Docs, Calendar, Google Base и т.д. [2]
- Сервисы студии 37signals: средство управления проектами Basecamp, средство совместной работы над текстами Writeboard, онлайн-бизнес-чат Campfire и т.д. [3]
- Сервисы от Amazon.com: Elastic Computing Cloud (EC2) – виртуальные вычислительные мощности; Simple Storage Service (S3) – место для хранения данных; SimpleDB – база данных, предназначенная для хранения коротких таблиц и эффективного выполнения сложных запросов к ним [4].

Не вдаваясь в этимологию появления в составе этого термина слова "облако" следует заметить обоснованность самого возникновения нового термина, поскольку cloud computing имеет ряд характерных особенностей, отличающих его и от "распределенных вычислений", и от системы "обычных" web-сервисов, и от "web-2.0" в целом. К этим особенностям относятся:

- максимальная простота и удобство доступа к ресурсу (обычно с помощью браузера через web-интерфейс);
 - возможность сильного (почти бесконечного) масштабирования;
 - отсутствие минимальных ограничений на объем или интенсивность использования ресурса.
- Есть три причины, по которым идея "заоблачных сервисов" возникла именно сейчас. Первая – количественное развитие технологий на стороне пользователя:
- доступные широкополосные и безлимитные каналы, позволяющие оперативно передавать достаточно большие объемы информации;
 - персональные компьютеры с мощными процессорами и большими объемами оперативной памяти на которых с приемлемой скоростью выполняются даже приложения написанные на "медленных" языках (например на интерпретируемом JavaScript);
 - рост возможностей, а также стандартизация и взаимная совместимость современных браузеров.

Второй фактор – появление крупных игроков в интернет-индустрии. Из приведенных выше примеров Google и Amazon с самого начала занимались бизнесом, другие пришли в сеть, когда распознали в ней выгодное место для приложения своих разработок. Все они "вдруг" обнаружили, что простаивающим серверам можно найти интересное для клиентов и выгодное для себя применение. Причем именно для разработчиков сетевых мегапроектов стоимость одного "лишнего" сервера оказывается минимальной.

Третья причина – коммерческая – стоимость внимания клиентов к ресурсу растет гораздо быстрее, чем трафик, диски, сервера, разработка приложений.

Рассмотрим какие достоинства и недостатки имеют "облачные технологии" с точки зрения их пользователей.

Достоинства.

Простота использования.

Для работы с cloud-сервисом достаточно браузера, что делает эту технологию независимой от используемой операционной системы и позволяет пользователю работать в привычном ему удобном окружении.

Быстрота начала работы.

Cloud-сервисы готовы к работе. Отсутствует этап создания инфраструктуры. Необходимо только определить необходимую структуру данных и загрузить их в cloud-систему и сервис (приложение) готов к работе.

Высокий пороговый уровень.

Даже минимальные начальные предложения в cloud-приложениях имеют очень высокие ресурсные параметры. Так объем почты Gmail по-умолчанию вырос с 1 до 6 Гб (на сегодняшний день). Стандартные объемы cloud-DB и cloud-архивов составляют десятки и сотни гигабайт.

Возможности масштабирования.

Практически все cloud-сервисы предоставляют возможность очень сильного увеличения предоставляемых ими ресурсов. При этом стоимость услуг растет практически линейно, а эффективность функционирования системы снижается незначительно.

Дешевизна.

Вероятно это основное преимущество данного подхода. Для индивидуальных пользователей использование многих cloud-сервисов с начальным объемом предоставляемых ресурсов оказывается бесплатным (в большинстве случаев при этом владелец сервиса получает доходы от рекламы). Стоимость платных услуг при использовании больших ресурсов оказывается близка к их себестоимости. Такие расценки на арендуемые ресурсы и услуги возможны только в рамках мегапроектов, при этом они остаются для их разработчиков выгодными (или, по крайней мере, не очень затратными). Для пользователей использование cloud-сервисов оказывается заметно (иногда в десятки раз) выгоднее создания собственной инфраструктуры и разработки приложений.

Недостатки.

Ограниченность функциональности при работе через браузер.

Этот недостаток преодолевается путем написания собственных приложений используя данные об API cloud-системы, предоставляемых ее разработчиками. В научной среде такой подход широко распространен (вне cloud-контекста) и не вызывает особых трудностей.

Системы постоянной готовности.

Следствием использования резервных мощностей гораздо более крупных мегапроектов является невозможность использования cloud-технологий в системах критичных к потерям входных данных, а также в системах с гарантированным временем ответа (тем более в системах реального времени). Этот недостаток заложен в самой природе cloud-технологии и на сегодняшний день устранен быть не может.

Типы баз данных.

Большинство cloud-DB, предлагаемых сегодня клиентам, являются не реляционными, а имеют структуру ключ – значение. Это связано с резким снижением эффективности работы реляционных баз данных в распределенной среде. Для систем ключ – значение это проблема выражена гораздо слабее.

Нерешенные вопросы.

К ним относятся вопросы длительного (многолетнего) хранения данных в cloud-системах и ряд юридических и законодательных ограничений на использование cloud-систем и технологий.

В докладе будет рассмотрена практика и перспективы использования "облачных" технологий в науке на примерах дистанционного зондирования Земли, космических и астрономических экспериментов.

ЛИТЕРАТУРА:

1. Wikipedia "Cloud computing" // http://en.wikipedia.org/wiki/Cloud_computing
2. <http://www.google.com>
3. <http://www.37signals.com/>
4. <http://www.amazon.com>