EMC: перспективы СХД

В мае 2014 г. корпорация EMC объявила о заключении окончательного соглашения о приобретении частной компании DSSD, занимающейся разработкой инновационной высокомасштабируемой архитектуры флэш-систем хранения для in-memory баз данных и рабочих нагрузок по обработке больших данных с высокой интенсивностью ввода-вывода (таких как SAP HANA и Hadoop). Продукты на базе новой высокомасштабируемой архитектуры флэш-систем хранения DSSD запланированы к выпуску в 2015 г.

Сергей Бочарников – системный инженер, EMC Россия и СНГ.

Введение

В статье речь пойдет не об уже существующем продукте, а о технологии ближайшего будущего, которая появится достаточно скоро.

Многие классифицируют СХД-платформу, основываясь на физическом способе подсоединения («У них Инфинибенд между узлами!») или протоколе («Это – блочный!», или «NAS», или «мультипротокольный»). Это рассуждение неверно, поскольку только архитектура определяет основные характеристики хранилища. Все прочие элементы – лишь детали того, как та или иная система хранения построена.

Еще одно замечание: многие также часто задают и такой вопрос: «Так какая же из этих типов СХД самая лучшая?». Ответ: «В каждой конкретной ситуации и для каждой конкретной задачи (то есть типа нагрузки) может быть лучшей та или иная архитектура, но ни одна не может быть просто «лучшей» среди всех, для любой задачи сегодня или завтра».

Нагрузка предопределяет выбор архитектуры, а не наоборот. Утверждать обратное будет большим заблуждением. Один из примеров подобного заблуждения – утверждение о том, что мы все и везде скоро перейдем на использование флэш-накопителей. Например, один из заказчиков EMC (услугами которого в интернете большинство из нас пользуется ежедневно) имеет инфраструктуру с емкостью 85 Пбайт на 20 772 жестких дисков, что транпортируется на 8 грузовиках. Можно такой объем обслужить лишь флэш-накопителями? Вряд ли.

Другой заказчик (он в 10 раз больше): 550 стоек с оборудованием, 4416 физических серверов, 5000 виртуальных машин, 10 000 инсталляций ОС, 200 000 дисков, 890 Пбайт пространства. Можно создать такую инфраструктуру только на флэш-накопителях? Нет.

Однако флэш-память становится все дешевле. Уже сейчас можно купить накопитель емкостью 1 Тбайт меньше чем за $400 – это серьезный прогресс, но и магнитные технологии не стоят на месте.

Фазовый переход, углеродные нанотрубки – вот области, в которых сейчас сосредоточены основные усилия исследований и разработок. Мы имеем все шансы быть свидетелями, как к 2017-2018 годам магнитные технологии выйдут на новый этап конкуренции с твердотельными накопителями.

Новый тип нагрузок

Традиционные типы нагрузок успешно живут и развиваются: и OLTP, и файловые хранилища и многое другое. Но в последнее время значительное развитие получили приложения типа аналитика in-memory и аналитика больших данных (или OLAP-нагрузки)[1)].

Первый тип предполагает обработку БД в оперативной памяти. Может возникнуть заблуждение, что это – то же самое, что и кэширование в традиционной БД. Однако это не так. Кэширование – процесс, при котором наиболее часто используемые данные хранятся в оперативной памяти, что дает определенный выигрыш для операций чтения, в то время как операции записи или обновления существующих записей все равно имеют дело с диском. К тому же обслуживание кэша также требует ресурсов процессора и оперативной памяти!

Специализированная же in-memory DB лишена большого количества промежуточных операций, ставших ненужными индексов и кэшей, упрощена и нацелена на наиболее полное использование доступных ресурсов.

Для чего применяются такие БД? Изначально они использовались для более быстрой аналитики данных и позволяли значительно сократить время получения отчетов и анализа данных. Но это, скорее, эволюционный подход.

Настоящая революция случилась, когда стало понятно, что с помощью этой технологии стало возможно получить аналитику в реальном времени – то, что было недоступно для таких больших объемов данных ранее. Основные потребители таких решений находятся в области электронной коммерции, биржевых торгов, телекома и др., где решения должны приниматься немедленно, а не спустя некоторое время (сфера применения стремительно расширяется).

А каков объем данных, который обрабатывают in-memory DB? Сегодня в крупных организациях используются традиционные промышленные БД размером под 100 Тбайт. Может такой объем быть размещен в оперативной памяти? С одной стороны – почему бы и нет? С другой – оперативная память серверов, хоть и увеличилась в объемах за последнее время, тем не менее, еще не может вместить в себя такой объем данных, и нам приходится пользоваться внешними СХД.

Что сдерживает высокопроизводительную аналитику

Давайте посмотрим, какой путь проделывают данные в традиционной, диск-ориентированной БД (рис. 1):

- приложение хочет изменить какую-то запись в БД и запрашивает эту запись у СУБД;
- СУБД смотрит в свой кэш, не находит там запись и обращается к файловой системе;
- файловая система смотрит в кэш файловой системы, также не находит там нужных данных и обращается к диску;
- получив нужные данные, ФС кладет запись в свой кэш и передает их в СУБД;

Рис. 1. Цепочка обращений при доступе приложения к данным.

1) Термин СХД исторически связывался с хранением данных для OLTP-приложений. Параллельно ему существовал термин – хранилище данных (data warehouse, DW), который исключительно ассоциировался с хранением БД для OLAP-приложений. В настоящее время эти два понятия конвергируются, т.е. все чаще единая аппаратная платформа (и все чаще одна БД с возможностью построчного и поколоночного хранения таблиц) используются для поддержки как OLTP-, так и OLAP-приложений. Тому пример – DSSD-решение. Однако специализированные DW с расширенными возможностями, прежде всего, в части интеграции неструктурированных данных и поддержки высокопроизводительной аналитики еще длительное время будут представлены на рынке.

– СУБД также размещает запись в своем кэше и отдает запись приложению;
– приложение, обработав данные, хочет записать их и передает в СУБД;
– СУБД обновляет информацию в своем кэше и передает данные файловой системе;
– файловая система актуализирует кэш ФС и производит запись на диск.

И только после этого операция выполнена.

Много ли времени тратится на это? Было проведено тестирование, в ходе которого для начала взяли традиционную БД и разместили ее данные в RAMDISK – т.е., по сути, в той же оперативной памяти. Выигрыш в скорости по сравнению с диском составил чуть менее 3 раз (рис. 2).

А затем взяли специализированную in-memory DB и повторили тестирование. Выигрыш в скорости по сравнению с прошлым разом составил невероятные... 420 раз! И это – за счет того, что нам больше не нужны эти промежуточные стеки кэшей, файловых систем, SCSI-обращений[2)].

Остается только скорость работы с памятью. Кстати, какова она?

Рассмотрим на примере какого-нибудь современного процессора, скажем Xeon E7. У нас есть регистр процессора, но его емкость очень небольшая (буквально десяток килобайт). Поэтому мы пользуемся кэш-памятью самого процессора, скорость обращений к ней – единицы наносекунд, а емкость – порядка 10 Мбайт. Когда они заканчиваются, нам приходится использовать память DRAM через шину DMI. Задержка возрастает до десятков наносекунд, но объем увеличивается до единиц терабайт. Однако и этого объема может стать недостаточно. И в этом случае самый быстродействующий накопитель – SSD-плата (например, PCIe) в самом сервере. Скорость обращения к ней увеличивается уже примерно до 50... микросекунд (то есть мы переходим на следующий порядок величин), зато емкость начинает измеряться десятками терабайт. Когда нам понадобится больше, мы вынуждены будем обратиться к внешней СХД – и там задержки могут составить до 1 миллисекунды (даже если у нас самая быстрая СХД XtremIO), зато и доступная емкость – сотни терабайт и даже намного больше (рис. 3).

Итак, мы выяснили, что мешает действительно быстрой работе нового типа нагрузки – in-memory DB:

Рис. 2. Сравнение времени доступа к данным при разных схемах их хранения.

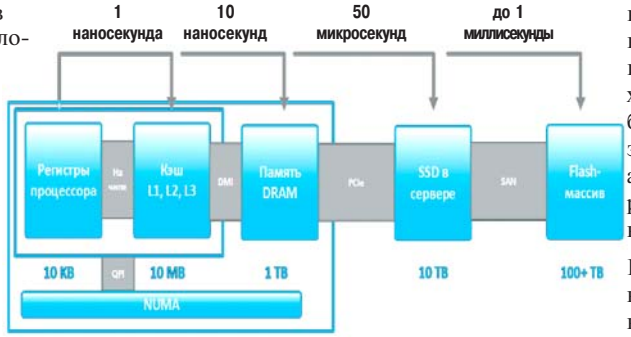

Рис. 3. Сравнение уровней хранения данных по объему и задержкам.

– многочисленные программные «прослойки», такие как файловая система, SCSI-стек, каждый со своими кэшами и буферами;
– значительная удаленность хранилища от процессора, то есть от работающего там приложения.

Вообще, разработчики ПО не хотели бы иметь дело со всеми этими промежуточными уровнями доступа. В идеальном случае, было бы здорово, если бы хранилище данных напрямую общалось бы с API приложения или его библиотеками.

СХД для высокопроизводительной аналитики

В мае 2014 г. EMC приобрела компанию DSSD и, соответственно, ее разработки флэш-систем для хранения и обработки БД в ОП (типа SAP HANA) и для аналитики больших данных в реальном времени (в интеграции, например, с Hadoop).

Продукты на базе новой высокомасштабируемая архитектура флэш-систем хранения DSSD запланированы к выпуску в 2015 году и будут предлагать следующие возможности оптимизации (рис. 4):

– оптимизация для баз данных в памяти (SAP HANA , GemFire и т.д.);
– оптимизация для аналитики в реальном времени (управление рисками, обнаружение мошенничества, часто используемые приложения, Pivotal HD и т.д.);
– оптимизация для высокопроизводительных приложений, используемых в исследовательских организациях и правительственных учреждениях (секвенирование генома, распознавание лиц, анализ климата и т.п.).

Высокомасштабируемую архитектуру флэш-систем хранения DSSD выберут заказчики, которым необходима платформа, обеспечивающая беспрецедентную производительность для размещенных в памяти приложений и приложений для больших данных с высокими требованиями к вводу-выводу (таких как SAP HANA и Hadoop). В такой конфигурации флэш-система DSSD будет использоваться как самый быстродействующий уровень многоуровневой архитектуры хранения.

Архитектура DSSD дана на рис. 4. Она представляет собой большое количество флэш-памяти (для увеличения пропускной способности), сгруппированное с небольшими контроллерами и дополнительно снабженное сверху некоторым количеством более быстрой памяти для сглаживания задержек флэш. Добавьте сюда ПО автоматизации, объединение в пул ресурсов, возможность работать напрямую с API приложения.

На сегодняшний день продукт еще настолько новый, что не объявлены практически никакие его характеристики – ни задержка, ни емкость, ни полный перечень сервисов по обработке данных. По предварительным данным цифры будут весьма впечатляющими.

Однако дело не столько в цифрах, сколько в самой концепции – разместить хранилище данных максимально близко к приложению, убрать все лишние прослойки и стеки, полностью использовать все преимущества флэш-памяти (рис. 5).

Раньше были попытки называть серверной памятью твердотельные массивы, подключенные к серверу. Но, если обращение к данным происходит через scsi_write... или file.open() – это не серверная память.

Первый экземпляр нового типа СХД, предназначенный для тестирования, уже существует.

Сергей Бочарников,
EMC Россия и СНГ.

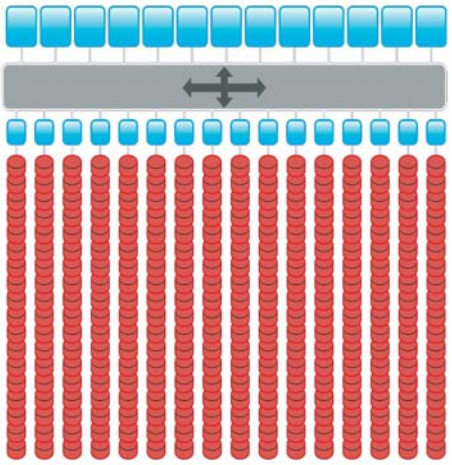

Рис. 4. Архитектура DSSD.

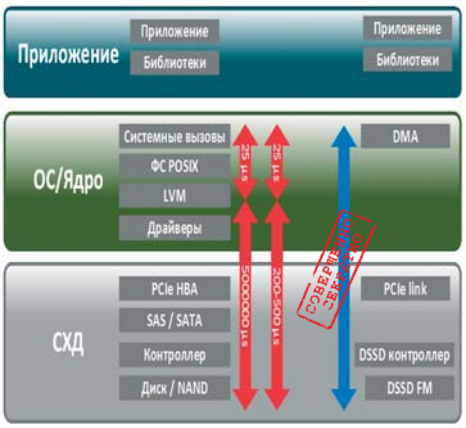

Рис. 5. Сранение задержек при доступе к данным по стандартным протоколам для локального серверного HDD и локального SSD (подключение по PCIe) с DSSD *(http://virtualgeek.typepad.com/virtual_geek/2014/05/a-new-5th-branch-in-the-storage-tree-of-life.html).*

2) Полный отчет о результатах тестирования доступен по ссылкам:
– *virtualgeek.typepad.com;*
– *http://www.mcobject.com/in_memory_database.*