электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

**УДК 004.03/004.6**

# Тенденции развития систем хранения данных

**Зубарев А.В.**

*Студент, кафедра Компьютерные системы и сети»*
*МГТУ им. Н.Э. Баумана, г. Москва, Россия*
*Научный руководитель: Самарев Р.С., к.т.н.,доцент кафедры*
*«Компьютерные системы и сети» МГТУ им. Н.Э. Баумана, г. Москва, Россия*

МГТУ им. Н.Э. Баумана
strike@student.bmstu.ru

По статистике, количество данных, хранимых на цифровых носителях, увеличивается в два раза примерно за 1-1.5 года и будет сохранять эту тенденцию в течении нескольких лет. Так сегодня общий объем производимых данных составляет порядка 15 ПиБ в день[1]. В связи с этим возникает потребность эти данные хранить и иметь доступ к ним для последующей обработки и архивирования. Исследования показывают, что хранение данных составляет около 23% от всех расходов на информационную инфраструктуру в компаниях[2]. Поэтому наибольшую заинтересованность проявляют компании, которые обладают большим объемом хранимых данных, начиная с нескольких сотен терабайт, где абсолютное значение таких расходов весьма существенно. Именно такие компании являются основными заказчиками и потребителями передовых СХД.

В последние несколько лет сильно изменилось отношение компаний к данным. Если раньше данные, особенно собранные несколько лет назад, представляли лишь абстрактную ценность, то сейчас они неоспоримо являются одним из самых ценных активов компаний. Связано это, прежде всего, с заполнением «информационных» рынков и развитием аналитических систем, таких как автоматизированные системы управления (ERP), системы поддержки принятия решений и т. п.[2]. Для работы которых требуется большое количество разнообразных данных за длительный период.

Также продолжается постепенный переход с бумажных на цифровые носители в областях государственного сектора, что также предъявляет СХД свои запросы.

Обобщая запросы компаний можно выделить следующие потребности к СХД[2,3]:

- «вечное» хранение данных и надежность;
- быстродействие и доступность;
- масштабируемость и простота контроля;
- цена.

Под «вечным» хранением понимается то, что никакие данные, оказавшиеся в компании, никогда и ни при каких обстоятельствах не должны быть утрачены т. е. СХД должны обеспечивают соответствующую надежность хранимых данных. Вследствие наступления века информации, также важна скорость работы с данными, в случае СХД, скорость доступа к данным – ведь в современном информационном обществе фраза «время – деньги» как никогда актуальна, причем если раньше это время измерялось в днях, затем в часах, то теперь оно уже измеряется в минутах и секундах, а иногда и микросекундами, например, в биржевой торговле. В качестве примера можно взять случай января 2013 года, произошедший на американских биржах: некая компания получила данные отчета по запасам природного газа на 400 мс раньше официального выхода и за это время успела купить часть акций, которые после официального публикации отчета подорожали в цене.

Под масштабируемостью понимается возможность относительно простого увеличения пространства хранения до некоторых размеров, например, покупкой и подключением новых носителей. Конечно, идеалом считается бесконечно-масштабируемая система, но, к сожалению, такая система вряд ли появится из-за особенностей устройства нашего мира.

После того как мы определись с основными потребностями, можно перейти к тому как компании-производители СХД решают вставшие перед ними задачи.

Одним из самых перспективных и наиболее общих направлений развития СХД является виртуализация, которую в последние несколько лет стали внедрять и использовать практически во всех СХД. Виртуализация – это некоторое представление ресурсов, например, их логическое объединение. Применительно к теме СХД – это виртуализация дисковых ресурсов т. е. объединение какого-то числа этих ресурсов в единый или не единый дисковый пул. При этом происходит своего рода инпаксуляция физического носителя и его логического представления. Благодаря этому становится возможным использование таких технологий как[4]:

- динамическое выделение и распределение ресурсов (HDP);

- консолидация данных на всех носителях (твердотельных, дисковых, ленточных) и автоматическое распределение файлов по уровням доступа;
- практически не ограниченную масштабируемость (в том числе возможность географического разнесения данных);
- широкие возможности управления данными (миграция, резервное копирование данных).

Все эти технологии в некоторой мере связаны. Так без обеспечения динамического выделения ресурсов невозможно организовать хорошую масштабируемость, а удобное управление данными невозможно без грамотной консолидации данных.

Динамическое выделение ресурсов позволяет осуществлять более эффективное распределение памяти для сохранения данных: каждый виртуальный диск занимает в точности такую емкость в пуле хранения данных, состоящем из физических жестких дисков, какая ему действительно требуется, и при этом всегда пользуется всеми предоставляемыми ему ресурсами. Возможность объединения всех доступных жестких дисков в один массив повышает производительность и надежность т. к. появляется некоторая избыточность данных. В кластерных системах одни и те же виртуальные диски могут назначаться нескольким хостам, даже если у физических жестких дисков имеется только по одному порту. Также динамическое выделение памяти позволяет «увеличить» доступный размер дискового пула без остановки рабочего процесса, что сокращает издержки в работе СХД и дает возможность непрерывной работы. Стоит заметить, что иногда реализацию данной технологии называют «облаком» или «облачным хранилищем».

Еще одной перспективной технологией является автоматическое распределение данных в зависимости от их использования, которая является важным компонентом обеспечения эффективной работы в облачной среде[1]. Так как она делает возможным перемещать активно используемые файлы в область хранения с высокой производительностью операций ввода/вывода в зависимости от требований виртуальных дисков. А данные, которые запрашиваются реже определенного количества раз, переносятся в зону с меньшей производительностью. Так с ее помощью можно предотвратить возникновение пиковых нагрузок. В результате обеспечивается эффективное использование и слаженная работа разнородных ресурсов хранения.

Эту же технологию используют и в гибридных жестких дисках, только в случае СХД работают в больших масштабах.

В последнее время стала использоваться технология географического разнесения данных, которая позволяет создавать катастрофоустойчивые решения на случай прекращения

работы одного дата-центра и снижать задержки при работе с данными т. к. данные становятся физические ближе к пользователю.

В целов, технология виртуализации сделала управление данными более простым для конечного потребителя благодаря тому, что теперь клиент может работать с самими данными, еще больше абстрагировавшись от физического представления.

Одной из самых «медленных» систем компьютерных систем был и остается жесткий диск. В какой-то мере данную проблему должны решить твердотельные накопители (SSD), которые обладают гораздо более высокими показателями чтения/записи (при произвольном доступе к данным), а также обладают меньшим потреблением электроэнергии. Но обладают существенным минусами в виде большой цены за мегабайт (цена прямо пропорциональная емкости), меньшей максимальной емкости одного носителя, относительно малым количеством перезаписей и меньшей надежностью, чем у HDD. Поэтому СХД не торопятся переходить на данную технология, а стараются ускорить текущую инфраструктуру, лишь точечно внедряя SSD-носители[1].

В настоящее время наиболее часто твердотельные носители используют как компоненты многоуровневых СХД и как кеш или буферную память т. е. в СХД реализуются различные сочетания дисковых и твердотельных носителей.

Стоит заменить, что некоторые производители выпускают СХД, целиком основанные на SSD-носителях. Наиболее логично использование таких систем в суперкомпьютерах, где не требуется хранение особо большого количества данных, но требуется очень быстрый доступ к ним, например, в суперкомпьютере СКИФ Аврора используются только твердотельные диски. Также перспективным является использование таких систем в кеширующих прокси-серверах и системах биллинга.

С увеличением размера данных также встала проблема пропускной способности сетей. Поэтому СХД постепенно развиваются и эту область. Так для подключения носителей внутри одного хранилища продолжает свое развитие технология Fibre Channel: 16GFC. А как внешние интерфейсы СХД в большинстве переходят на iSCSI, что позволяет строить системы на основе, в большинстве случаев уже существующей, инфраструктуры Ethernet, либо на Fibre Channel over Ethernet (FCoE переносит фреймы Fibre Channel через Ethernet)[.

Традиционный подход к хранилищам данных состоял в непосредственном подключении серверов к системе хранения (Direct-attached storage - DAS). Помимо DAS, существуют устройства хранения данных, подключаемые к сети — NAS (Network-attached storage), а также компоненты сетей хранения данных — SAN (Storage area network). И NAS, и SAN системы появились в качестве альтернативы архитектуре DAS. Причем каждое решение

разрабатывалось как ответ на растущие требования к СХД и основывалось на использовании развитых в то технологиях.

Архитектуры сетевых систем хранения были разработаны в 1990-х гг., и их задачей было устранение основных недостатков систем DAS. В общем случае сетевые решения в области систем хранения должны были реализовать три задачи: снизить затраты и сложность управления данными, уменьшить трафик локальных сетей, повысить степень готовности данных и общую производительность. При этом архитектуры NAS и SAN решают различные аспекты общей проблемы. Результатом стало одновременное сосуществование двух сетевых архитектур, каждая из которых имеет свои преимущества и функциональные возможности[2].

В настоящее время почти все СХД пр актически полностью отказались от DAS в пользу NAS и SAN. То есть происходит разделение хранения данных от места их обработки.

Одной из важных задач, касающейся самого хранения данных, является оптимизация хранения данных. Исторический способ оптимизации занимаемого данными пространства сводится к их сжатию, позднее получила распространение дедупликация данных – технология, при помощи которой обнаруживаются и исключаются избыточные (повторяющиеся) данные. Например, путем замены повторов ссылками на первую копию. Главная проблема данной технологии заключается в её больший ресурсоемкости из-за чего она не получила широкого распространения в основных хранилищах, но получила распространение в системах архивации (например, семейство HP D2D Backup Systems) и в программно-ориентированных системах.

В настоящее время ведущие компании начинают налаживать производство систем дедупликации. Уже существуют и внедряются специализированные файловые системы (ФС) с возможностью дедупликации: XFS, WALF (NetApp), HUMMER, Fossil. Основным отличием таких ФС является подход к нахождению совпадающих частей данных. Так XFS осуществляет их нахождения «на лету», а ФС WALP компании NetApp делает это во время наименьшей нагрузки сервера, что позволяет экономить на вычислительных ресурсах системы, но требует дополнительную «буферную» емкость[5].

Наиболее перспективным является считается программно-независимая (блочная) дедупликация, позволяющая работать с любыми данными т. е. без потребности внесения каких-либо модификаций в программу. Не смотря на явные плюсы такого подхода, такие решения являются весьма ресурсоемкими и дорогостоящими, поэтому владельцы СХД предпочитают продолжать внедрять проверенные и надежные системы программно-ориентированнной дедупликации[5]. Стоит отметить, что программно-ориентированных

подход выгоден еще тем, что происходит постепенный переход к сервис-ориентированному хранению данных.

С совершенствованием технологии виртуализации и стремлением потребителей не вникать в тонкости работы СХД, а пользоваться хранилищем как некоторой услугой, начался постепенный переход к динамическим моделям обработки данных[4]:

- сервис-ориентированной архитектуры (SOA);
- сервис-ориентированной инфраструктуры (SOI);
- сервис-ориентированного решения для хранения данных (SOS).

Так некоторого времени производители приложений самостоятельно разрабатывали разные версии своих решений для разных серверных платформ или исползьовали открытые решения. Важной технологической тенденцией стало создание адаптируемых платформ для решения различных аналитических задач, которые включают аппаратную составляющую СУБД. Пользователей уже не волнует, кто сделал для их компьютера процессор, оперативную память или накопитель, — они рассматривают хранилище данных как некую услугу.

В связи с этим не очень хорошие перспективы у поставщиков, специализирующихся исключительно на ПО. Изготовители оборудования приобретают программо-технические компании и сами внедряют их в свои продукты.

Сама архитектура сервис-ориентированных решений для хранения данных предполагает следующие услуги:

- на уровне объектов: индексация, поиск, классификация данных;
- на уровне файлов: визуализация репликация, перенос, устранение, дублирование, защита, шифрование, архивирование данных;
- на уровне блоков: создание разделов, выделение ресурсов, управление томами, репликация, перенос, защита данных.

Так же сервис-ориентированная архитектура дает возможность преодолеть разрыв между информационными технологиями и бизнесом. Сегодня SOA стала реальной методологией построения информационных систем. Следует отметить, что SOA – это именно методология, а не продукт. Архитектура SOA создается под конкретное предприятие с учетом специфики его деятельности. Преимущество подхода, основанного на SOA, состоит в том, что он предполагает создание модульной архитектуры, где определенные компоненты систем поставлены в соответствие автоматизируемым функциям.

Обязательным условием внедрения SOA является наличие бизнес-проекта, т. е. модели бизнеса, описывающей выполняемые процессы, организационные структуры, стратегические

и тактические задачи, политики ведения бизнеса. Где SOA обеспечивает адаптацию СХД к текущему состоянию проекта.

Подводя итоги можно сказать, что массивы хранения данных, созданные на основе архитектуре двадцатилетней давности, не способны справиться с представляемыми им требованиями, что заставляет производитель СХД разрабатывать и внедрять новые решения для хранения данных, которые сильно отличаются от старых систем. Поэтому в настоящее время происходит очередная трансформация модели хранения данных. Активно внедряется виртуализация, благодаря которой стало возможным использование таких возможностей как динамическое выделение ресурсов и упрощенное управление данными.

Если раньше СХД представляли собой raid-контроллер с расширенными возможностями и некоторым количеством cache-памяти, то сейчас это уже сложное комплексное решение, позволяющее решать намного более широкий спектр задач. Все больше систем хранения данных будет ориентироваться на работу в «облачных» средах и соответствовать требованиям, предъявляемым к таким новым инфраструктурам.

**Список литературы**

1. Орлов С. Системы хранения: эффективность и стоимость // Журнал сетевых решений/ LAN. 2012, №1.
2. Система хранения данных // Tadvieser: государство, бизнес, ИТ.2013.URL.www.tadviser.ru/index.php/Статья:Система_хранения_данных (дата обращения: 17.02.2013).
3. Литвинов И., Ройфман Р., Шовкун Р. Тенденции в области систем хранения данных в России и мире.2013.URL.http://www.itsec.ru/articles2/Oborandteh/tendencii-v-oblasti-sistem-hraneniya-dannih-v-rossii-i-mire/ (дата обращения:17.02.2013).
4. Тенденции развития систем хранения данных // Оборудование. Технический Альманах 2008, №2. С. 26-29.
5. Обзор наиболее быстрорастущего сектора рынка - решений по дедупликации данных // Storage News 2008, №2, С. 2-7.