CAStor: ПО для кластерного файлового CAS-хранилища

В статье дано представление об организации масштабируемого кластерного файлового хранилища для неструктурированных данных на основе ПО, позволяющего обращаться к файлам по их идентификаторам, формируемым по содержимому файла. Данный тип СХД получил название "контентно адресуемая архитектура хранения" (content addressed storage architecture – CAS).

Введение

Основатели компании Caringo были одними из первопроходцев, которые заложили основу CAS-хранилищ еще в конце 90-х годов прошлого века. Во многом рынок контентно адресуемых файловых хранилищ получил развитие за счет выхода на рынок в 2002 г. решения Centera компании EMC, разработанному на основе технологий компании FilePool (приобретенной EMC в 2001 г.).

Программное решение CAStor стало доступно на рынке в 2006 г. и позволяет организовывать высоко масштабируемые файловые CAS-хранилища на базе любых стандартных серверов, включая блэйд-серверы. По организации CAS-хранилище на базе ПО CAStor это просто множество (кластер) серверов, подключенных через NIC к LAN-сети. При этом они не обязательно должны быть гомогенными. К отдельному файлу такого CAS-хранилища можно обратиться по его уникальному идентификатору, который хранится у приложения (CAS-хранилище передает идентификор приложению при записи файла).

За счет того, что при адресации к файлу резко уменьшается объем информации вследствие отсутствия какой-либо информации о блоках данных (признак традиционных файловых систем), удается значительно расширить адресное пространство и довести его до петабайт.

Адресация по контенту требуется для организации файловых хранилищ с неизменяемым контентом или, например, для справочных данных, объем которых может доходить до сотен петабайт и более. Пропускная способность CAS-хранилищ на базе ПО CAStor ограничивается только пропускной способностью LAN или Ethernet-коммутатора.

Благодаря высокой степени самоуправляемости и самовосстановления CAS-хранилища, вносимого ПО CAStor, все манипуляции добавление/удаление с серверами (узлами) кластера можно производить в режиме онлайн.

В настоящее время на рынке существует достаточно много (см., например, SN № 3/36, 2008) горизонтально масштабируемых кластерных систем (необязательно с CAS-адресацией), однако в качестве одних из основных преимуществ рассматриваемого решения является его простота и доступность.

Где это работает?

Основная потребность в высокомасштабируемых файловых хранилищах связана с возрастанием потребности в хранении неструктурированного контента, под которым подразумевается файловая информация – неизменяемый контент, справочная информация и др., хранение которой необходимо в соответствии с регулирующими нормами, технологиями производства или требованиями бизнеса.

Приведем несколько цифр, свидетельствующих о тенденциях роста объемов информации.

По оценкам IDC, информация, генерируемая компаниями в соответствии с регулирующими требованиями и нормами, составляет 20% общего объема (что, по состоянию на 2007 г.[1)], равнялось около 60 экзабайт = 60 млн Пбайт).

Однако в настоящее время контент, генерируемый частными лицами, превосходит объемы корпоративного контента, большая часть которого хранится и предоставляется web 2.0 компаниями[2)].

Около 80% всей информации в компаниях хранится в виде файлов – документы, электронные таблицы, медиафайлы и др. с ежегодным ростом этой информации от 50% до 120%[3)].

В соответствии с данными другого исследования[4)], к 90% файлов никогда не обращаются после их создания, а из 10% оставшихся файлов к 65% обращаются только один раз (рис. 1).

Потребность в архивных системах возрастает прежде всего потому что, увеличение объемов информации происходит быстрее, чем развиваются технологии хранения, обеспечивающие улучшение характеристик традиционных (SAN/NAS) СХД.

Рис. 1. К 90% созданных файлов не обращаются никогда, а к 65% из оставшихся обращаются только один раз.

Архивные системы и решения, в частности, на базе ПО CAStor по большей части не заменяют высокопроизводительных (онлайновых) SAN/NAS систем, а лишь расширяют их возможности и позволяют в целом оптимизировать IT-инфраструктуру (рис. 2), позволяя хранить очень большие объемы данных в течение десятилетий, не задумываясь о каких-либо операциях миграции в течение всего срока хранения.

Необходимо заметить, что рынок хранения неструктурированного контента очень большой, прежде всего, с точки

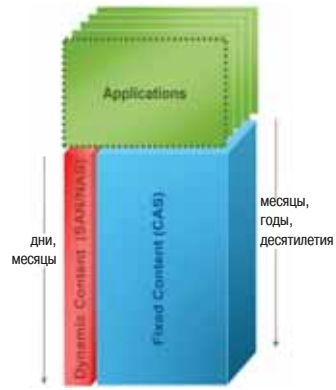

Рис. 2. Архивные системы позволяют оптимизировать онлайновые СХД и обеспечить надежное хранение неизменяемого контента в течение десятилетий, не задумываясь об операциях миграции данных в течение всего срока хранения.

[1)] The Expanding Digital Universe, IDC, David Reinsel et al, March 2007

[2)] The Diverse and Exploding Digital Universe, IDC, Christopher Chute et al, March 2008

[3)] The Economic Impact of File Virtualization, IDC, Richard Villars, May 2007

[4)] Measurement and Analysis of Large-Scale Network File System Workloads, UC Santa Cruz

зрения особенностей применения этих систем. Поэтому приоритезация показателей архивных систем в каждом случае может быть уникальной.

Как это работает?

Архивные решения на базе ПО CAStor это прежде всего простота первоначального развертывания, дальнейшего масштабирования и управления.

ПО CAStor – аппаратно-независимое решение. Оно работает на любых стандартных серверах с любым форм-фактором – блэйдах, серверах, ПК и др. Обычно это сервер архитектуры X86 с более чем 1 Гбайт оперативной памяти, одним или более жестким диском и сетевой картой Gigabit Ethernet. ПО не устанавливается на диск, так что вся емкость может быть использована для хранения контента. Для инициации узла кластера в сервер вставляется USB с ПО CAStor и через 60 секунд сервер CAStor готов к работе.

CAStor масштабируется от нескольких серверов и 1,5 Тбайт до сотен узлов и свыше 1 Пбайт. Кластер организуется и поддерживается самостоятельно при подключении к LAN (рис. 3). Для достижения максимальной производительности каждый узел (и клиент) кластера должен подключаться непосредственно к Ethernet-коммутатору с пропускной способностью 1 GE и выше.

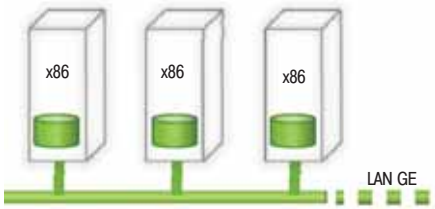

Рис. 3. Архивная CAS-система на базе ПО CAStor это множество любых стандартных серверов, подключенных к LAN.

Оборудование можно смешивать и оно может быть разных поколений и от разных производителей. При добавлении нового сервера он просто подключается к сети/кластеру CAStor. При этом доступный объем кластера CAStor увеличивается динамически при работающей системе без необходимости настройки хранилища. После загрузки нового узла (или несколько узлов) в кластере, при необходимости его удаления он выводится из эксплуатации, все содержимое на "списываемых" серверах будет перемещено на другие узлы кластера, диски будут зачищены от остатков информации, после чего оборудование можно удалить. Все это выполняется во время работы кластера CAStor и без воздействия на приложения или доступность данных.

CAStor органично оптимизирует использование узлов кластера, позволяя лучше использовать все доступное оборудование. Небольшой кластер размером 10 Тбайт в департаменте такой же жизнеспособный, как и больший кластер на 100 Тбайт в ЦОД.

В CAStor все файлы реплицируются по умолчанию. Если один диск выходит из

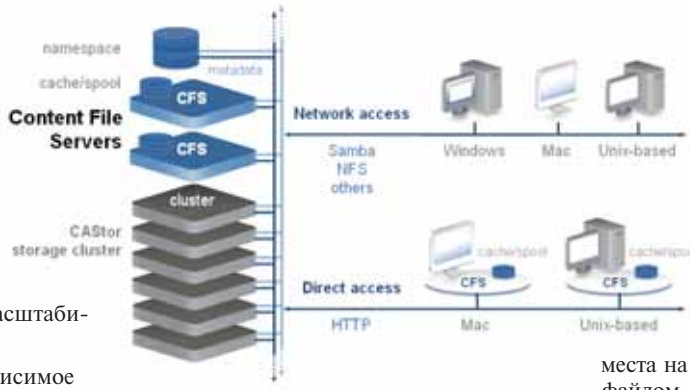

Рис. 4. Кластер CAStor поддерживает все основные интерфейсы по доступу к объектам: HTTP 1.1, CIFS, NFS, FTP, WebDAV, клиентов Mac.

строя, то всегда есть другая копия. Система автоматически обнаружит и повторно реплицирует файл в фоновом режиме. В действительности, чем больше кластер, тем быстрее произойдет восстановление. Нет необходимости тратить время и деньги на настройку RAID.

Особенности работы

Интеграция с CAStor

Приложения интегрируются с CAStor с помощью стандартного HTTP 1.1. CAStor использует в качестве собственного интерфейса доступа упрощенный набор стандарта HTTP 1.1, называемый Simple Content Storage Protocol (SCSP). Это широко используемый протокол который никогда не устареет и не потребует портирования. Важно то, что отсутствует закрытый API, и любое приложение или веб-сервис могут быть связаны с CAStor за считанные часы. Для приложений, у которых нет доступа к исходному коду или они базируются на традиционных файловых протоколах, Caringo предлагает CAStor Content File Server (CFS) — файловую систему на базе Linux, которая поддерживает основные файловые протоколы, включая CIFS, NFS, FTP, WebDAV, и клиентов Mac, а также может работать обычная файловая система Linux (рис. 4). CAStor CFS не является классической файловой системой. Это тонкий связующий слой, который выглядит для приложений как файловая система и говорит на языке HTTP с CAStor. Со стороны доступа он представляет стандартный интерфейс файловой системы, а на стороне хранения дает огромное плоское адресное пространство, крупномасштабное, высокопроизводительное и надежное.

Отличие CAStor от традиционных файловых систем, использующих блочное хранение

В блочном методе файлы разбиваются на кусочки и хранятся в виде множества «блоков». Адрес каждого блока должен быть управляемым. Этот метод не только добавляет сложности но и ограничивает количество хранимых файлов и имеет ограничения общего объема хранилища.

Представьте файловую систему с 4 млн файлов и каждый из них разбит на 5 блоков, это означает, что файловая система должна управлять 20 млн адресов блоков. Как только требование вырастет до 100 миллионов или миллиарда файлов становится очевидным, что эта файловая система столкнется с проблемой масштабируемости.

В отличие от файловых систем, которые работают поверх блочного хранения, CAStor предоставляет единое адресное пространство для хранения контента и не имеет сложной файловой иерархии, имен папок или физического места на диске, связанного с каждым файлом. Он не разбивает файл на блоки (куски) и ему не нужно потом управлять множеством отдельных блоков, связанных с каждым файлом. CAStor хранит файлы как единые объекты в непрерывном дисковом пространстве и ему нужно управлять только одним UUID (Universally Unique ID) для каждого элемента контента. Это означает, что CAStor CFS просто управляет UUID, давая возможность масштабироваться сверх ограничений традиционных файловых систем в количестве файлов и поддерживаемом объеме.

Когда приложение/клиент первый раз записывает файл в CAS, ему возвращается ключ или уникальный идентификатор для будущего доступа к нему. Когда файл запрашивается для загрузки, приложение передает ключ на CAS, и файл возвращается. Нет никакой иерархии файлов, имен папок или места на диске, связанного с файлом. При перемещении файла в хранилище ключ не меняется. UUID могут храниться в приложениях, документах или в базах данных. Нет никаких ограничений, где или как хранится UUID.

Каждый UUID, идентифицирующий объект, имеет размер 128 бит. Файловое пространство, которое можно адресовать с помощью такой длины идентификатора, колоссально. Caringo для иллюстрации этого приводит следующий пример: если поверхность всей Земли (включая сушу и воду) разделить на элементы площадью 1 мм^2, а затем каждый из этих элементов разделить еще на 670 квадриллионов (10^{12}) более мелких элементов – получим адресное пространство CAStor.

На настоящий момент вследствие текущего технического уровня развития данный потенциал можно использовать только в самой незначительной степени, но по мере технологического прогресса (уже в ближайшие годы появятся диски на десятки терабайт, или настольные системы петабайтной емкости + 100 GE) он будет востребован все в большей степени.

Как CAStor гарантирует целостность данных?

CAStor использует алгоритм хеширования, который вычисляет дайджест, иногда называемый цифровой отпечаток, основанный на последовательности бит для каждого контентного объекта (файла). Дайджест используется CAStor Health Processor (HP), который работает в фоне и непрерывно проверяет целостность контента для определения каких-

либо повреждений на диске. Если объект определяется как поврежденный, то из другой целой реплики хранимой в системе генерируется новая реплика. Это гарантирует, что в CAStor всегда присутствует нужное число доступных "чистых" реплик.

Хеш-дайджест также используется как Content Integrity Seal – печать целостности контента, которая является методом подтверждения аутентичности объекта контента для целей регулирования и доказательства в открытой, проверяемой пользователем структуре данных. В отличие от других систем CAS, CAStor отделяет адрес контента (UUID) от цифрового отпечатка (дайджеста), позволяя просто обновить хеш-алгоритм, если исходный был скомпрометирован. Это уже произошло с алгоритмами MD5 и SHA-1 и прозрачно обновляемый запатентованный хеш гарантирует долгосрочную целостность контента.

Поддержка ILM-процедур в CAStor на основе пользовательских метаданных

После записи объекта на CAStor он уже не может быть изменен. Это отвечает регулирующим требованиям, таким как SEC17a4 и Sarbanes-Oxley Act. При этом приложения и/или пользователи могут определить атрибуты метаданных для уникального описания контента объектов и управления им в течение жизненного цикла.

CAStor хранит все метаданные вместе с контентом и сохраняет их на протяжении всего их жизненного цикла. Другие элементы метаданных включают в себя число обслуживаемых реплик, срок хранения, тип контента, имя файла, исходное приложение и многое другое. CAStor также поддерживает специальный элемент метаданных LifePoint, который позволяет приложениям описывать, как файл должен управляться в течение его жизненного цикла в CAStor.

Администрирование CAStor

Кластер CAStor просто администрировать. Он устраняет необходимость в настройке хранилищ при добавлении новой емкости. Способность к самовосстановлению позволяет кластеру CAStor прозрачно восстанавливать данные со сбойного узла или диска без влияния на доступность данных. Если узел вышел из строя, кластер незамедлительно распознает потери и оставшиеся узлы кластера начинают вместе реплицировать контент поврежденного узла. Это происходит без участия администратора или влияния на приложения и доступность данных. Также кластер CAStor самобалансируется, автоматически выравнивая хранимый контент между узлами кластера для оптимальной производительности и устранения узких мест. Все эти действия требуют минимальной административной нагрузки, кластер CAStor может одинаково управляться из центрального веб-интерфейса как для 3 узлов в кластере, так и для 3000.

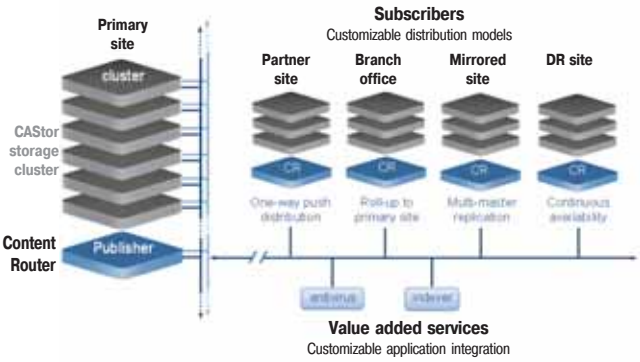

Рис. 5. Отдельный программный модуль CAStor Content Router позволяет компаниям географичеки (до 1000 км) распределять контент, основываясь на бизнес-правилах.

Поддерживается протокол SNMP. Наблюдение и управление кластерами может осуществляться с помощью общей платформы управления системами, например Novell ZENworks или HP OpenView.

Поддержка дедупликации данных

В CAStor применена архитектура асинхронного однократного хранения, таким образом, фоновые процессы устраняют дублированный контент через какое-то время с минимальным или нулевым воздействием на производительность.

Вместе с тем, такие функции как шифрование, индексация, поиск – полностью возлагаются на приложения и CAStor не поддерживаются.

Поддержка катастрофоустойчивости

Отдельный программный модуль CAStor Content Router позволяет компаниям географически (до 1000 км) распределять контент, основываясь на определенных бизнес-правилах. CAStor Content Router дает возможность установить значимость метаданных для зеркалирования контента в кластере, реплицирования в аварийный ЦОД, а также интеллектуально распределять специфичный контент на удаленные кластеры, например в филиалах (рис. 5). Поддерживаются конфигурации 1:1, 1:много и много:1.

Поддерживается служба подписки, с помощью которой подписчик получает от публикаторов данные для обработки. Это позволяет работать другим приложениям обработки контента, таким как антивирус, индексация, сжатие и шифрование.

Тестирование производительности

Результаты тестирования CAStor, проведенные самой компанией Caringo при условии, что все узлы подключены к Gigabit Ethernet-коммутатору, каждый узел представляет сервер Capricorn Petabox PS3000, который имеет 4 HDD с интерфейсом 3 Гбит/с/s SATA с общей емкостью 3,0 Тбайт на узел, показали потоковую производительность свыше 1 Гбайт/с при 32 узлах и размерах файла 40 Мбайт/с. Результаты тестирования для файлов размером 32 Кбайт и от 350 Кбайт до 40 Мбайт для операций чтения и записи показаны на рис. 6, 7.

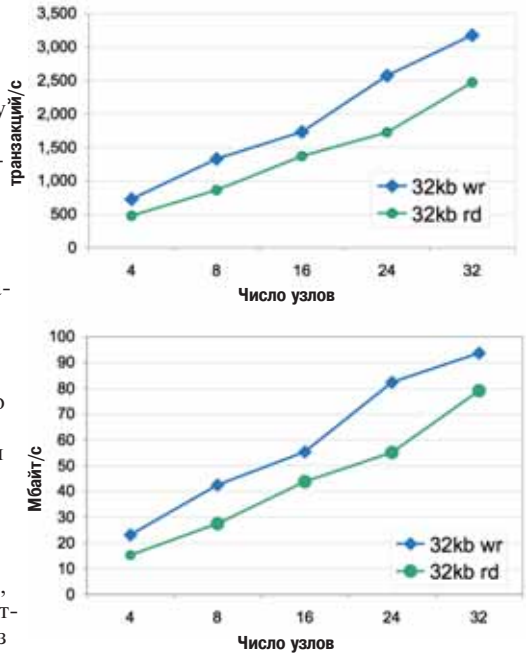

Рис. 6. Производительность кластера CAStor в зависимости от числа узлов для файлов размером 32 Кбайт в транзакциях в сек. (вверху) и в мегабайтах в сек. (внизу).

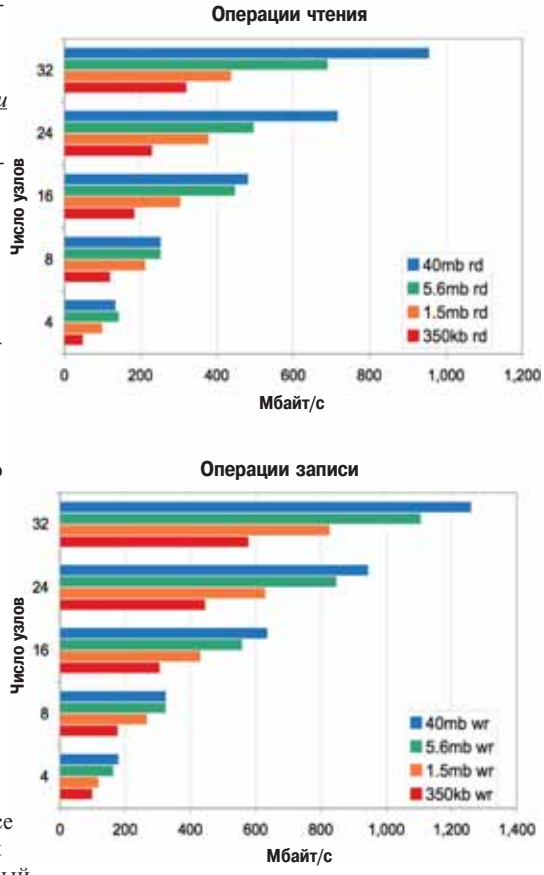

Рис. 7. Производительность кластера CAStor в зависимости от числа узлов для файлов размером 350 Кбайт до 40 Мбайт для операций чтения (вверху) и записи (внизу) в мегабайтах в сек.

Заключение

Доля архивных систем в мире возрастает постоянно. По оценкам IDC, 2009 год отмечен как год начала массового использования архивных хранилищ во всех секторах бизнеса.

Виталий Сайфуллин,
компания "Москит"