Д.т.н. Г.А. Егоров (ОАО «ИНЭУМ им. И.С. Брука»)**, к.т.н. В.И. Шяудкулис,**
PhD М. Финотти, Д.А. Смотров (QStar Technologies, Inc.)

G. Egorov, V. Chiaoudkoulis, M. Finotti, D. Smotrov

Архитектурные особенности реализации облачных хранилищ для архивных применений

The specific architecture features of the cloud storage for the data archiving

Приводятся основные требования к облачным хранилищам для использования их в современных архивных хранилищах данных. Рассматриваются архитектурные особенности реализации облачных хранилищ данных для архивных применений на примере конкретной реализации проекта фирмы Qstar Technologies Inc.

Ключевые слова: облачные хранилища данных, архивация данных, архивные хранилища, кластеры.

The requirements to the cloud storage design used in the modern archiving systems are presented. The specific architecture features of the cloud storage for the data archiving are evaluated using QStar Technologies, Inc. product as an example.

Keywords: cloud storage, data archiving, data archiving systems, clusters.

Введение

Облачные хранилища, являясь примером высокого уровня виртуализации обычных хранилищ данных, удовлетворяют большинству требований, которые должны учитываться при архивации данных [1]. Технология облачных хранилищ, обеспечивая различные аспекты открытости, является основой для построения современных архивных систем.

Статья развивает тему архивации данных на базе технологий облачных хранилищ [1, 2]. Рассматриваются структура и принципы построения таких облачных хранилищ, обеспечивающих высокие эксплуатационные характеристики архивных систем с точки зрения

цены, производительности, энергопотребления и безопасности доступа к данным. Именно этим требованиям удовлетворяет программный продукт Qstar OSM (далее – система OSM), который по сути дела является высокопроизводительным, расширяемым и отказоустойчивым частным облачным хранилищем данных, разработанным специально для создания архивных систем [2, 3]. Несмотря на ориентацию для архивации данных, система OSM обеспечивает весь набор сервисов и возможностей облачных хранилищ. Система OSM поддерживает стандартный набор протоколов – TCP/IP, HTTP и OSD; при необходимости может быть достаточно просто добавлена поддержка других протоколов.

Сетевая архитектура OSM

Архитектура OSM строится на двух коммуникационных сетях для внешней и внутренней связи. Внутренняя сеть обеспечивает выполнение внутрикластерных операций и не влияет на производительность внешней сети. Эта сеть доступна только узлам кластера через отдельный интерфейс, который автоматически конфигурируется независимо от конфигурации OSM. Внешняя сеть реализуется на базе протоколов TCP/IP, и ее параметры полностью контролируются администратором системы. По умолчанию для конфигурации внешней сети используется протокол DHCP. Интерфейс управления системой OSM позволяет определять другие протоколы (например, статические IP-адреса), если это необходимо. Важно отметить, что протоколы TCP/IP используются во всех конфигурациях OSM. Это необходимо для обеспечения независимости управления кластером и его функционирования от способа взаимодействия с устройствами хранения во внутренней сети.

В зависимости от потребностей конфигурация системы OSM может быть построена для нижнего, среднего и высокого уровней производительности. Главным различием между этими уровнями является выбор типов устройств хранения и оборудования для построения внутренней сети кластера. Система нижнего уровня использует TCP/IP для всех

обменов данными, тогда как система среднего уровня основана на применении SAS-устройств хранения, а система верхнего уровня использует оптоволоконные технологии.

Система OSM нижнего уровня (рис. 1) строится с использованием необходимого количества узлов кластера, где каждый узел содержит жесткие диски для хранения данных. Пользовательские компьютеры подключаются к OSM через внешнюю сеть – обычно коммуникационный коммутатор. Минимальное количество узлов в данной конфигурации для предотвращения потери данных не должно быть меньше трех.

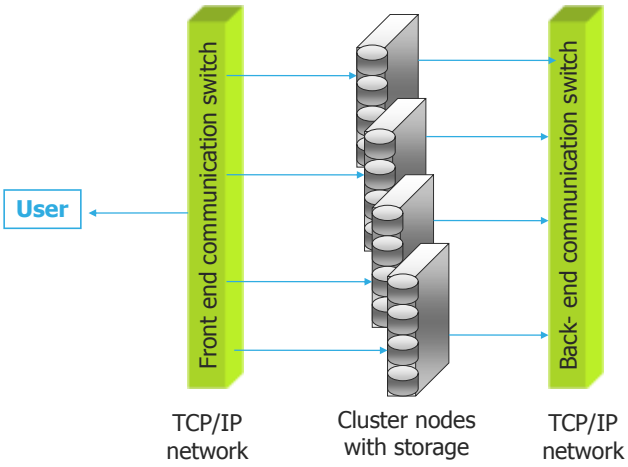


Рис. 1

Конфигурация OSM нижнего уровня

Система нижнего уровня ограничена с точки зрения производительности используемыми протоколами TCP/IP. Другое ограничение связано с доступностью внутренних дисков только самому узлу. В такой конфигурации при сбое одного из узлов недоступными становятся все устройства хранения, принадлежащие этому узлу. Конечно, диски могут быть перемещены в другие узлы, но это требует некоторого ручного вмешательства. Сбой любого узла не приводит, однако, к потере данных или их недоступности, т.к. копии всех объектов автоматически сохраняются в других узлах системы.

Система среднего уровня (рис. 2) строится с использованием дисков типа SAS, подключенных к внутренней сети кластера. Технология SAS предоставляет новый уровень сервиса с минимальным увеличением конечной стоимости системы – возможность досту-

па к любому устройству хранения из любого узла системы внутри кластера. В этом случае при сбое узла обеспечивается переназначение его диска на другой работающий узел в системе, что увеличивает надежность доступа к данным и улучшает эксплуатационные возможности.

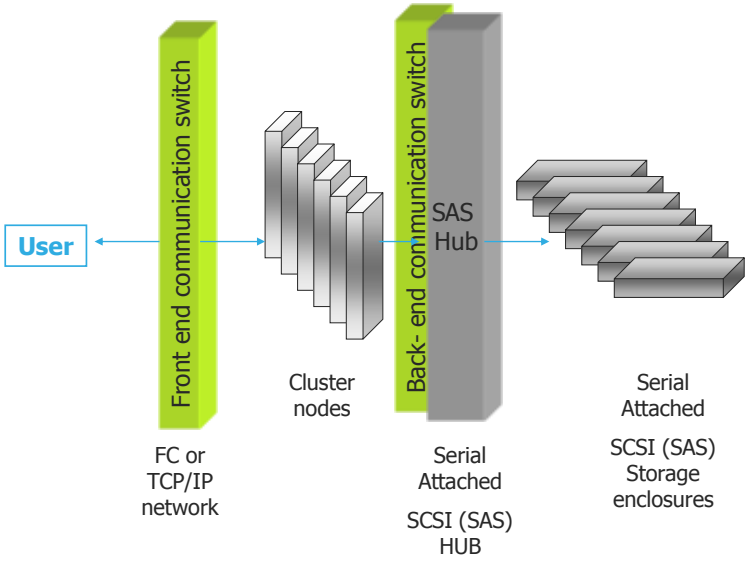


Рис. 2

Конфигурация OSM среднего уровня

Система, построенная на базе технологии SAS, дешевле системы с использованием оптоволоконных технологий, но в то же время она обеспечивает высокие уровни производительности и доступности данных, а также возможность разделения доступа к дисковым устройствам между узлами кластера.

Высокопроизводительная конфигурация OSM (рис. 3) строится с использованием оптоволоконных технологий, которые используются для внутренней сети кластера. В этом случае любой узел кластера имеет доступ к любому диску и может управлять им в случае сбоя любого узла.

Кластеризация

Система OSM строится на стандартном оборудовании, объединенном в многомашинный комплекс в виде кластера. Физический кластер, включая узлы, сетевое оборудо-

вание, жесткие диски, является единицей управления с точки зрения администратора системы. На базе физического кластера определяются логические кластеры, при этом каждый логический кластер может содержать определенное количество узлов, входящих в физический кластер. Максимальное количество логических кластеров ограничено значением 32, где нулевой логический кластер имеет специальное назначение (в него включаются все новые узлы и устройства хранения, установленные на них).

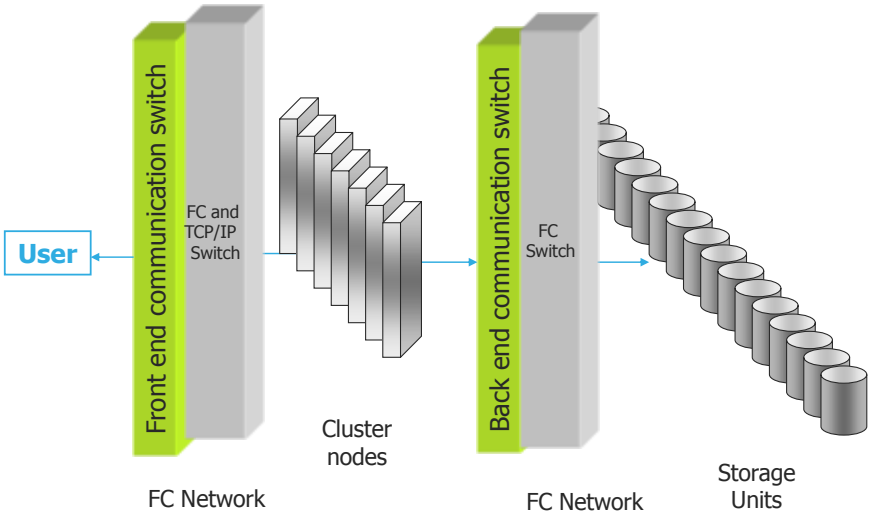


Рис. 3

Высокопроизводительная конфигурация OSM

С точки зрения администрирования каждый новый узел автоматически добавляется в нулевой логический кластер и становится ресурсом системы. Узлы в нулевом кластере не могут использоваться, но доступны для распределения как члены нулевого кластера. Это позволяет администратору включать эти ресурсы в конкретный логический кластер (с ненулевым номером). В большинстве случаев решение о логическом разбиении системы на кластеры принимается в зависимости от требований по защите и разделению данных.

Узлами кластера являются обычные персональные компьютеры, к которым предъявляются некоторые специфичные требования, в соответствии с которыми узел должен:

• быть совместим с архитектурой Intel x64_86. Это требуется, потому что узлы используют одно и то же программное обеспечение (ПО), которое скомпилировано для этих

архитектур;

- предоставлять как минимум два TCP/IP-интерфейса со скоростью 1 Гбит (один – для внутренней сети, второй – для внешней);
- поддерживать объем оперативной памяти от 8 ГБайт и выше;
- предоставлять некоторую минимальную производительность процессора и желательно более высокую, если требуется поддержка компрессии данных и их шифрование;
- предоставлять возможность управления частотой центрального процессора для обеспечения выбранной политики энергосбережения. Данная возможность позволяет адаптировать потребляемую мощность к реальной нагрузке системы;
- предоставлять поддержку специальных устройств (например, USB или SSD-диски) для загрузки и поддержки операционной системы и ПО. Следует иметь в виду, что диски для хранения данных не содержат файлов, специфичных для конкретного узла, что позволяет заменять оборудование узла без сложностей, связанных с переконфигурацией;
- иметь поддержку протокола IPMI. На практике это означает, что данный узел должен иметь специальный модуль, который позволяет управлять энергопитанием узла удаленно через интерфейс TCP/IP. Эта опция важна, если пользователь хочет осуществлять агрессивное управление энергопотреблением кластера;
- иметь соответствующие адаптеры в случае использования SAS или оптоволоконных устройств.

Большинство современных персональных компьютеров предоставляют все указанные опции, и с этой точки зрения можно констатировать, что OSM может работать практически на любом современном Intel-компьютере.

С точки зрения управления оборудованием устройства типа Blade с поддержкой оптоволоконных соединений являются идеальным решением для OSM. Преимущество таких систем заключается в возможности при необходимости добавления нового оборудования с целью увеличения производительности или эффективного управления энергопо-

треблением.

Узлы в логическом кластере OSM функционируют независимо друг от друга. Такая независимость – одна из отличительных особенностей системы OSM, что делает её более гибкой, чем другие кластерные системы. Например, OpenStack-архитектура предполагает наличие специальных «proxy»-компьютеров для доступа к данным [4]. За счет этого обеспечивается взаимозаменяемость всех узлов системы, т.к. все они имеют одно и то же ПО и возможность включения в логические кластеры новых узлов для расширения системы.

Несмотря на универсальность используемого ПО имеется ряд операций внутри кластера, которые нуждаются в правильной координации исполнения. Для обеспечения такой внутрикластерной синхронизации выбирается один мастер-узел. Процесс выбора основан на широковещательной рассылке сообщений и позволяет надежно выбирать узел на роль мастер-узла внутри кластера. В случае если мастер-узел по каким то причинам отключается, то оставшиеся узлы повторяют процесс выбора нового мастер-узла в кластере. С архитектурной точки зрения мастер-узел ничем не отличается от обычного узла. Такой узел можно рассматривать как узел с назначенной специальной функцией внутри кластера.

Одной из функций мастер-узла является назначение устройств хранения узлам кластера. Эта операция должна выполняться в последовательном режиме с целью обеспечения целостности ресурсов кластера, что особенно важно, когда в кластере используются разделяемые устройства хранения (рис. 2 и 3). Мастер-узел также отвечает за планирование управления энергоснабжением. Управление энергоснабжением реализует каждый узел, но мастер-узел принимает решения о том, какую политику энергосбережения следует использовать.

В любой момент функционирования кластера его конфигурация доступна для любого узла. Узлы кластера через определенные интервалы времени передают свою конфигурацию (при помощи broadcast-сообщений) всем доступным узлам кластера. Каждый узел ответственен за получение таких сообщений и поддержание таблицы конфигурации кла-

стера (достаточно сложной структуры данных), которая содержит информацию о состоянии узла (количество процессоров, текущая частота процессоров, загрузка процессоров, объем использования памяти, состояние модулей хранения и т.д.), необходимую для управления и наблюдения за состоянием системы и энергопотреблением. Таблица конфигурации кластера также включает в себя информацию, относящуюся к логическому статусу системы (количество сохраненных объектов, объем использованного дискового пространства, информация о состоянии устройств хранения и параметры загрузки узла), которая активно используется для балансировки нагрузки на узел и для реализации политик балансировки при записи данных.

Управление системой OSM построено на основе использования таблицы конфигурации кластера. Пользователь осуществляет управление OSM через WEB-ориентированный графический интерфейс пользователя с любого авторизованного WEB-клиента. Администратор при помощи браузера может соединяться с любым узлом кластера, т.к. таблица конфигурации кластера доступна любому узлу.

Для удобства администрирования пользователю предоставляется графический интерфейс, отображающий физическую модель узлов кластера с присоединенными модулями хранилища. Каждый элемент интерфейса может обрабатывать события от «мышки» и позволяет пользователю получать информацию по общему состоянию кластера, а также детализированную информацию по состоянию каждого элемента.

Концепция модуля хранилища

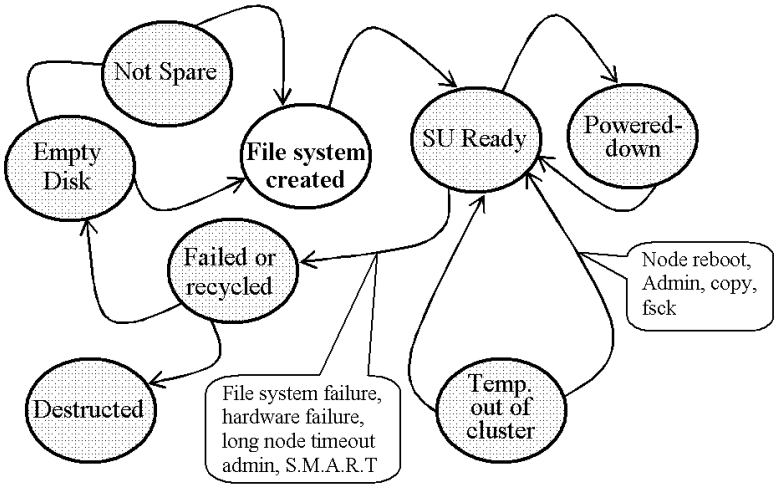
Одной из важных особенностей архитектуры OSM является концепция модуля хранилища (Storage Unit – SU). Модуль хранилища – элемент хранилища для хранения данных с использованием объектно-ориентированного подхода. В самом простом виде SU – это отдельный жесткий диск (подсоединенный через интерфейсы ATA, SATA, SCSI, FC, NAS), отформатированный специальным способом и способный хранить объекты OSM

(практически, JBOD – просто группа дисков). В качестве модуля хранилища также могут выступать ресурсы NFS, CIFS или любой другой файловой системы. Важно отметить, что OSM также поддерживает диски RAID, но это не является необходимостью. Архитектура RAID имеет много преимуществ, однако работа с такими устройствами имеет и свои проблемы. Например, устройства RAID не обеспечивают управление энергоснабжением и имеют большое время восстановления после сбоя. С точки зрения управления энергоснабжением в кластере более предпочтительным является JBOD, а не RAID.

Новые узлы включаются в OSM с неотформатированными дисками. Когда новый узел из нулевого кластера включается в конкретный логический кластер, эти диски форматируются, и им назначаются уникальные числовые идентификаторы внутри данного логического кластера. Это делается под управлением мастер-узла, т.к. повторение этих идентификаторов недопустимо. Архитектура системы OSM в большей степени ориентирована на модули хранилища, чем собственно на узлы, которые выступают просто как элементы обработки данных, хранящихся на модулях хранилища. Чем больше узлов в кластере, тем выше общая производительность этого кластера. Чем больше модулей хранилища в кластере, тем выше надежность хранения данных внутри этого кластера и тем больше его ёмкость. С другой стороны, узлы могут передавать друг другу управление дисками (модулями хранилища), что очень важно в системах с разделяемым доступом к дискам.

Мастер-узел является тем узлом, который распределяет ресурсы между узлами кластера, например, в случае отказа какого-то узла, при подсоединении нового модуля хранилища или при отказе какого-то диска, для перебалансировки нагрузки узла. Использование JBOD позволяет реализовать управление энергоснабжением, т.к. логически модуль хранилища не разделяется и контролируется только операционной системой узла. Файловая система модуля хранилища может быть демонтирована, модуль может быть отключен, перезапущен и смонтирован заново, когда нагрузка на систему возрастает. Управление

состоянием модуля хранилища осуществляется владеющим им узлом. Жизненный цикл



модуля хранилища представлен на рис. 4.

Рис. 4

Жизненный цикл модуля хранилища OSM

Если модуль хранилища начинает отказывать (это определяется с использованием SMART-технологии для дисков), он может быть выведен из эксплуатации, а данные с него перемещены на другие модули хранилища. Пустой диск становится резервным (hot spare) или форматированным модулем хранилища (file system created) с назначенным уникальным идентификатором в результате добавления его в кластер (SU-ready). В состоянии SU-Ready можно отключить питание диска (Powered-Down), или он может быть временно исключен из кластера, например, для обслуживания. Если модуль хранилища дал сбой, то он исключается из кластера полностью, и его номер не может быть назначен другому модулю хранилища. Данные с дефектного устройства могут быть сканированы на другое устройство, пока в кластере существуют копии этих данных. Система OSM не пытается каким то образом «оживить» дефектное устройство. Практика показывает, что если диск начал работать неправильно, то практически невозможно восстановить все данные с такого устройства. Архитектура OSM предоставляет простые технологии миграции данных. Когда новое устройство добавляется в кластер, система восстанавливает на нем объекты с

оставшихся дисков, и, таким образом, ни одна копия объекта не будет потеряна. Такой подход значительно отличается от технологии RAID, т.к. OSM будет копировать только блоки, которые реально используются, а в RAID будет восстанавливаться каждый блок вне зависимости от того, используется он или нет.

Каждый объект, хранящийся в системе OSM, обрабатывается с целью обеспечения целостности данных и эффективности его хранения. В этом, пожалуй, и состоит самое большое преимущество облачных хранилищ перед простыми дисковыми устройствами, когда единицей хранения является не блок, а объект переменной длины, который можно защитить, проверить, реплицировать, зашифровать.

Целостность данных поддерживается за счет вычисления контрольных сумм данных объекта с использованием методов XOR8, SHA1 или SHA256. Таким образом, каждый раз проверяя реплики объектов, мы можем установить, какая реплика повреждена, и восстановить ее из хороших копий. Для более эффективного хранения объекты могут быть сжаты перед передачей их реплик на другие узлы, что существенно уменьшает необходимую для хранения этих объектов память и увеличивает производительность передачи данных по внутренней сети. Когда контрольная сумма данных объекта подсчитана, формируется уникальный описатель для каждого объекта, который может быть использован для поиска объекта с таким же содержимым. Каждый раз, когда обнаруживается аналогичный по содержанию объект, система принимает решение о создании ссылки на уже существующий объект (дедупликация данных) вместо того, чтобы сохранить аналогичные данные еще раз. Этот метод однократного хранения применим для работы на уровне модуля хранилища, но не на уровне кластера. Поиск описателя по всему кластеру связан с уменьшением производительности из-за необходимости постоянного поддержания базы данных контрольных сумм на уровне кластера. В то же время необходимость репликации данных с целью обеспечения надежности хранения входит в противоречие с дедупликацией.

С целью обеспечения безопасности объекты могут шифроваться дополнительно. Для

поддержки шифрования используется система управления ключами шифрования. База данных системы управления ключами шифрования может быть скопирована на съемный носитель для безопасного хранения. Ключи шифрования также шифруются в целях безопасности.

Пользовательские интерфейсы OSM

С точки зрения пользователя, OSM является облачным хранилищем с объектно-ориентированным интерфейсом доступа. Внутреннее функционирование системы OSM полностью скрыто от пользователя. В настоящее время взаимодействие пользователя с OSM осуществляется через протоколы TCP/IP с использованием программного протокола OSD. Протокол HTTP не нуждается в дополнительных программных интерфейсах, т.к. все возможности системы OSM (включая балансировку нагрузки) являются его частью. Несмотря на это OSM предоставляет специальную библиотеку HTTP для облегчения процесса разработки приложений.

Для пользовательских приложений кластер OSM представляется как единая система хранения, доступная через пул IP-адресов и портов. Пользователь может вручную сконфигурировать эти пулы адресов или использовать протоколы LDAP или SLP для того, чтобы определить адреса узлов OSM автоматически. Другие методы определения адресации, такие как DNS или NIS, также могут быть использованы. Когда адрес (или пул адресов узлов) в сети OSM определен, пользователь может открыть сессию, предоставив информацию по авторизации в кластере. В большинстве случаев для определения полного набора адресов узлов кластера необходимо указать IP-адрес и порт хотя бы для одного узла.

Наиболее важными запросами системы являются запросы, относящиеся к работе с объектами. Объекты создаются по запросам «CREATE», «CREATE AND WRITE», «WRITE» и «READ» или их HTTP-эквивалентам. Запросы в кластер поступают к одному

из узлов кластера через интерфейсы TCP/IP со стороны внешней сети. В этот момент система OSM должна принять решение о том, где следует создавать новый объект. Для этого узел использует таблицу конфигурации кластера. Если данный узел загружен достаточно сильно, то он может вернуть код 301 или 307 по протоколу HTTP, что будет означать, что данный запрос должен быть перенаправлен к другому узлу. Если запрос принимается, то узел начинает создание объекта локально, а также активизирует процесс создания реплик этого объекта.

Количество реплик объекта является конфигурируемым параметром, минимальное количество реплик – 2. Реплики создаются на разных узлах кластера; если количество узлов меньше числа реплик, то реплики создаются на разных дисках кластера. Определяются понятия минимального и максимального количества реплик, а также понятие специальной реплики. Узел возвращает положительный ответ пользователю при успешном создании минимального количества реплик, после чего начинает асинхронный процесс по созданию дополнительных реплик объекта до достижения их максимального количества, заданного для этого применения. Для каждого сохраненного объекта генерируется уникальный идентификатор, использующий идентификатор модуля хранилища (диска) в качестве ключа. Когда поступает запрос на чтение объекта, номер модуля хранилища используется для определения местоположения объекта с помощью таблицы конфигурации кластера. Если такой модуль хранилища не доступен, осуществляется поиск объекта по всему кластеру.

Особенность системы OSM состоит в поддержке специальных реплик объектов. Специальная реплика – это реплика объекта, хранимая на удаленной системе. Такой удаленной системой может быть QStar ASM на базе сменных носителей [2]. Например, политика хранения реплик может утверждать о том, что объект должен быть немедленно реплицирован на два модуля хранения, а одна реплика должна храниться в удаленной системе на базе съемных носителей (например, ленте или BluRay-диске). Если удаленные узлы

доступны, политика может позволить создавать две локальные реплики перед возвратом результата пользователю и после этого создать две специальные реплики для этого объекта, но уже асинхронно.

Балансировка загрузки и управление энергопотреблением

Управление энергопотреблением основано на алгоритме балансировки нагрузки. Например, при низкой нагрузке частота на процессоре может быть снижена, модули хранения остановлены или выключены целые узлы. Если нагрузка узлов кластера достигает определенного уровня, выключенный узел может быть перезапущен. Для управления энергопотреблением в OSM используется встроенный модуль статистики, который хранит временн***ы***е отметки для различных режимов работы системы. Такой подход позволяет создавать политики, основанные на условиях, подобных следующему – «если загрузка системы меньше 10% в течение пяти минут, выключи узел, где большая часть свободного места уже израсходована».

В самом общем виде интерфейс управления энергопотреблением в OSM представляется как простой слайдер с 10-ю позициями. Внутри каждого из значений слайдера имеется набор регулируемых параметров, доступных через кнопку Advanced, где может производиться точная настройка параметров управления энергопотреблением (в основном это набор различных временн***ы***х уставок). Это позволяет администратору определить до 10 стандартных настроек управления энергопотреблением либо для обеспечения наибольшей производительности (когда практически ни один из ресурсов кластера не отключается), либо для получения наилучших параметров энергосбережения (когда многие подсистемы могут быть отключены).

Заключение

Облачные хранилища в достаточной степени удовлетворяют требованиям архивации

данных. Необходимо отметить, что с точки зрения обеспечения безопасности доступа к данным крупные организации и предприятия отдают предпочтение использованию частных облачных хранилищ. Использование частных облачных хранилищ для архивации данных связано с высокими требованиями к эксплуатационным характеристикам (эффективность энергопотребления, цена эксплуатации и производительность. Этим требованиям в полной мере удовлетворяет продукт OSM фирмы QStar Technologies, Inc., архитектурные особенности которого рассмотрены в данной статье.

Литература

1. Егоров Г.А., Шяудкулис В.И., Финотти М. Проблемы построения современных архивных хранилищ данных. – «Информационные технологии», 2012, № 12.

2. Егоров Г.А., Шяудкулис В.И., Финотти М., Беляков М.И. Принципы практической реализации современных архивных хранилищ данных. – «Информационные технологии и вычислительные системы», 2013, № 1.

3. [Qstar ASM] http//www.qstar.com/

4. [OpenStack] http://swiftstack.com/openstack-swift/architecture/