СХД для Web Scale IT

Рассматриваются особенности и реализации СХД для одного из самых быстро развивающихся трендов – Web Scale IT, который, по прогнозам, в 2017 г. получит отражение в архитектуре большинства глобальных компаний.

Сергей Платонов – руководитель исследовательской лаборатории RAIDIX.

Введение

Термин Web Scale IT впервые был предложен аналитическим агентством Gartner в середине мая 2013 г. Через год они предсказали, что данный подход к построению архитектуры будет использован более чем половиной глобальных компаний в 2017 г. Кроме того, Web Scale IT попал в топ-10 самых "горячих" технических трендов 2015 г.

Что же представляет собой Web Scale IT? Сегодня считается, что это – подход к построению инфраструктуры, принятый и опробованный такими гигантами, как Amazon, Facebook, Netflix, Rackspace и Google.

По прогнозам IDC, в 2017 г. значительный рост рынка Web Scale IT в денежном аспекте – $14 млрд и очень большой CAGR – 32%.

Gartner выделяет среди ключевых следующие элементы Web-scale IT:

- *промышленно-создаваемые датацентры с полной глобальной консолидацией всех компонент и централизованным управлением;*
- *веб-ориентированная или микросервисная архитектура;*
- *программируемое управление.*

Направление Web Scale IT связано с новым вызовами, прежде всего, это: 1) необходимость хранить, обрабатывать очень большой объем неструктурированной информации, получаемой от датчиков, мобильных устройств в результате сбора различной информации от web-порталов, web-магазинов, социальных сетей и др.; 2) необходимость обслуживания до сотен миллионов пользователей одновременно на базе нескольких георазпределенных сайтов, интегрированных в единое решение; 3) широкое распространение облачных сервисов.

В настоящее время эти тенденции объединяются термином "миграция от второй к т.н. "третьей платформе".

Web Scale IT инфраструктуры ориентированы на определенный класс задач (особенности которых отмечены выше) и в неполной мере соответствуют требованиям классических OLTP-/OLAP-приложений, которые будут существовать еще многие годы.

При реализации Web Scale IT инфраструктуры уровень хранения может быть полностью интегрирован в серверный узел или выделяться в отдельный "слой".

Решаемые проблемы

Web Scale IT используются для решения трех основных проблем современных ЦОД:

1) Масштабирование

Классические архитектуры ЦОД не позволяют предсказуемо расширяться, увеличивая производительность и объем эластично, без экспоненциального роста стоимости инфраструктуры.

2) Сложность обслуживания

- Десятки вендоров.
- Несколько разрозненных средств управления.
- Тысячи узлов.
- Старый подход к выделению ресурсов.
- Необходимость держать в штате специалистов по разным технологиям сильно увеличивают стоимость владения и время необходимое для выделение ресурсов и устранение проблем.

3) Недостаток места и высокий уровень энергопотребления

Проблемы с энергопотреблением и местом – бич современных центров обработки данных. Считается, что подход Web scale IT поможет значительно сократить расходы на электроэнергию.

Что Web Scale IT значит для рынка СХД?

Унификация оборудования приведет к тому, что доля традиционных систем хранения данных будет снижаться, и на их место будут приходить гиперконвергентные билд-блоки, которые в рамках одного шасси объединят систему хранения, вычислительные мощности, гипервизор и сетевой контроллер.

Билд-блоки должны объединяться в горизонтально масштабируемое решение, обеспечивая гранулярность расширения пространства и производительности.

Говоря о таких билд-блоках, нельзя не упомянуть про проект opencompute. Видимо, именно эти аппаратные платформы станут стандартом де-факто во многих ЦОД завтра.

Facebook периодически рапортует о значительном сокращении расходов после внедрения "открытых" серверов. Но автор статьи не обнаружил среди представленных блоков хранения действительно интересные решения.

На рынке уже есть несколько коммерческих продуктов, представляющих собой готовые решения для построения Web Scale IT инфраструктур.

В основе большинства из них можно найти корни из ПО с открытым исходным кодом, что несет только плюсы и хорошо зарекомендовало себя в датацентрах веб-гигантов.

Непонятно желание некоторых производителей работать только с DAS – ведь многие компании уже имеют инфраструктуру хранения и фабрики SAN и не готовы от них отказываться, а развертывание в параллель новых программно-ориентированных гиперконвергентных СХД только усложнит менеджмент. ПО должно иметь возможность не только объединять имеющиеся локальные ресурсы хранения серверов в пул, но и иметь возможность использования объема внешних СХД и управления ими через API.

Масштабирование СХД должно проводиться небольшими дозами, быть максимально автоматизировано и происходить в режиме plug-and-play. Рост емкости и производительности при этом должен быть предсказуем и линеен.

Вызовы

Вызов № 1: уже в ближайшей перспективе большая доля СХД будет строиться на стандартных серверах с открытым или/и проприетарным ПО, обладая минимальными издержками на управление, петабайтной/экзабайтной масштабируемостью, поддержкой, как правило, нескольких основных способов доступа к данным. При этом будут наблюдаться три тенденции:

– наполнение корпоративной функциональностью (поддержка хранения структурированных данных, повышение уровня надежности и производительности, георазпределенность и др.), например, hadoop-кластеров, которые изначально были ориентирова-

ны только на неструктурированные данные с низкими требованиями по надежности;

– реализация корпоративных СХД (поддержка, прежде всего, классических OLTP-/OLAP-приложений) в бОльшей своей доле на стандартных серверах (за исключением all-flash СХД, в которых компонента на базе проприетарных технологий будет иметь решающее значение);

– упрощение средств управления сервисами данных, с одновременной интеграцией всех СХД ЦОД под единым интерфейсом и консолью.

Масштабирование до огромного количества узлов (с возможностью разнесения в разные ЦОД) требует пересмотра подходов к обеспечению доступности и консистентности данных.

Репликация данных "стоит" дорого и уже не всегда справляется с обеспечением должного уровня доступности. Возрастающая популярность flash, наличие локальности IO и возможная распределенность кластера не позволяет использовать классические методы кодирования.

Также не нужно забывать и про необходимость масштабироваться на лету и исправлять ошибки в полностью автоматическом режиме – "самоизлечиваться".

Распределенные системы также подвержены так называемой "византийской" ошибке, для ее обнаружения используются алгоритмы обеспечения консенсуса, но, если посмотреть на существующие решения, мы заметим, что из этого небогатого арсенала используется только один инструмент – Paxos.

Вызов № 2: разработчикам решений необходимо проделать большую работу по адаптации ключевых технологий хранения и обеспечения доступности данных.

Гипермасштабируемые системы хранения и обеспечение доступа к данным тысяч узлов потребуют пересмотра подхода к интерфейсам. Про то, что POSIX-совместимые файловые системы не справляются со своими задачами в гипермасштабируемых архитектурах, известно всем. Блочные интерфейсы тоже близки к пику масштабируемости (или уже достигли его, если говорить о flash и SCSI).

Также гранулярность в виде таких объектов, как LUN, в блочных СХД, не позволяет эффективно использовать возможности СХД для того, чтобы в достаточной мере удовлетворять изменяющимся требованиям от приложений: мы можем изменить параметры, сделать мгновенную копию только целого LUN, на котором могут находиться десятки виртуальных машин.

Автоматизация выделения ресурсов и управления СХД нуждается в согласованном интерфейсе между приложениями и системой хранения.

Системы и их узлы должны управляться на основе заранее предопределенных политик и абстрагировать от группы администрирования все аппаратные ресурсы. Классическое выделение ресурсов неприменимо в средах с тысячами узлов.

СХД должны в режиме реального времени реагировать на изменения бизнес-требований и выполнять предоставление необходимых параметров производительности доступа без вмешательства со стороны группы администрирования.

Сейчас наиболее богатыми по возможностям API обладают гипервизор и система управления ЦОД от компании VMware. Эта же компания предложила изменить подход к хранению и реализовала в шестой версии продуктов технологию VVol. Технологии поддерживаются большинством вендоров на рынке, но их нет в других гипервизорах.

Рабочая группа SNIA по формированию единого видения программно-конфигурируемых СХД для взаимодействия между приложениями и системой хранения предлагает использовать для взаимодействия CDMI.

А ведь есть и другие предложения: opencompute, например, разрабатывает собственное API для взаимодействия с flash-памятью, производители жестких дисков предлагают новые интерфейсы взаимодействия диска с контроллером или приложением для того, чтобы типичные ошибки или недостатки не привели к катастрофе в инфраструктурах, содержащих тысячи дисков.

Вызов № 3: производители должны договориться и создать единый стандарт взаимодействия приложений и гипервизоров с системами хранения данных. Управление всеми ресурсами должно происходить из единого интерфейса.

Гипермасштабируемые ЦОД действительно должны использовать подход Software Defined Everything, и программно-конфигурируемые СХД являются одним из основных компонентов этого подхода.

Реализации Web Scale IT

Концепция Web scale IT получила отражение в следующих классах решений.

Мега-ЦОДы от глобальных web-провайдеров

Это классическое направление развития Web Scale IT, с которого “все началось”. К настоящему времени разработано второе поколение Мега-ЦОДов, летом – ожидается третье. И хотя их архитектура была существенно доработана, тем не менее, они во многом не соответствуют корпоративным требованиям. Мега-ЦОДы строятся на базе распределенных файловых систем и кластерной архитектуры (сотни тысяч серверов, основная ориентация – web-приложения) и развиваются самостоятельно несколькими мировыми глобальными сервис-провайдерами (Google, Amazon и др.). Компонента хранения – неотъемлемая часть узла кластера.

Мега-ЦОДы от ведущих разработчиков ПО

Это второе направление развития глобальных мега-ЦОДов, которые разрабатываются на базе модульной кластерной архитектуры (с использованием гиперконвергентных билд-блоков), но с повышенной надежностью и функциональностью (например, наличие сжатия, дедупликации, репликации и др.) – в отличие от предыдущего класса (пример – разработки от Nutanix). Ориентация – на сервис-провайдеров, которым требуются сотни тысяч серверов. Помимо web, поддерживается широкий класс офисных приложений, а также основные гипервизоры. Основной метод доступа – файловый. Компонента хранения – неотъемлемая часть узла кластера. Один администратор может управлять тысячами серверов.

Гипермасштабируемые СХД от глобальных вендоров

Данный класс представляет выделенный уровень гипермасштабируемых СХД, который разрабатывается рядом ключевых мировых вендоров (EMC, Fujitsu, IBM и др.). Поддерживаются только уровень хранения и сервисы данных. Возможны все типы доступа к данным, включая SQL-запросы к hadoop-кластерам. Примеры: EMC Elastic Cloud Storage, EMC ScaleIO, Fujitsu Eternus CD10000 (см. статью в данном SN, *прим. ред.*), IBM Spectrum Storage, IBM Elastic Cloud Storage (см. SN № 3/59, 2014, *прим. ред.*) и др.

Если сравнивать удельную стоимость решений отмеченных выше классов, то она возрастает от первого класса к третьему.

ПО управления единым пулом гетерогенных СХД

Виртуализация управления СХД с одновременным упрощением и автоматизацией ряда сервисов (поддержание уровней обслуживания), а также интеграция “классических” и “новых” СХД – одна из ключевых задач при реализации Web Scale IT. В качестве примера может служить EMC ViPR Cotroller, который позиционируется как ПО для автоматизации управления гетерогенными ресурсами хранения (см. статью в данном SN, *прим. ред.*) с разными способами доступа к данным.

Отдельно остановимся на продукте, реализующим подход SDS, – EMC ScaleIO, который, как и основные новинки компании, является приобретенным стартапом. В отличие от ViPR, ScaleIO, будучи программно-определяемой СХД, реализует все функции программным образом и создает пул хранения из локальных дисков серверов, обеспечивая пользователя функциями СХД корпоративного класса.

ScaleIO по своей сути является распределенным блочным устройством и имеет асимметричную архитектуру. Поддерживаются многие операционные системы, физические и виртуальные серверы, все типы носителей – HDD, SSD, PCIe flasf card, а также автоматическая перенастройка и перебалансировка. Есть мгновенные снимки, шифрование данных на лету и поддержка quality of service (ограничение полосы пропускания для различных приложений). Для защиты от по-

тери данных используется репликация, что достаточно дорого при современных объемах данных. При отказе узла начинается автоматическое восстановление на свободное пространство имеющихся серверов. Масштабируемость может достигать тысяч серверов.

ScaleIO поддерживает только Ethernet-соединение и только блочный доступ, что сильно ограничивает схему применения. Продукту явно недостает таких функций, как дедупликация и сжатие на лету.

Год назад компания VMware представила свой продукт класса SDS, называемый vSAN (в феврале 2015 г. появилась его последняя версия).

Его можно смело назвать наиболее полноценной реализацией программно-определяемой СХД. VMware "знает цену" автоматизации и давно уже создает и оттачивает API взаимодействия СХД и гипервизора.

vSAN обладает функциями создания мгновенных снимков и копий, репликации данных. Поддерживается кэширование на SSD, а начиная с версии 6.0 – и полноценная работу all-flash массивов. При подключении новых узлов и заполненности пространства более чем на 80% выполняется проактивная ребалансировка кластера. vSAN обладает богатыми возможностями по автоматизации выделения ресурсов и обеспечению QoS. Эти функции обеспечивает Storage Policy Based Management.

Недостатками решения можно назвать отсутствие функций data reduction и поддержку только технологий репликации для защиты от сбоя узлов. Кластер vSAN масштабируется до 64 узлов, что значительно ниже, чем у других продуктов. Кроме того, нельзя не отметить, что vSAN работает только в инфраструктуре виртуализации VMware.

Не так давно компания Citrix приобрела компанию Sanbolic, являющуюся сильным игроком на рынке SDS. Sanbolic Scale-Out Platform позволяет построить масштабируемую до тысяч узлов инфраструктуру хранения. В качестве ресурсов хранения Sanbolic использует SAN и NAS СХД, локальные диски серверов и облачные хранилища (такие как AWS).

Функциональность Sanbolic включает:

- поддержку файлового и блочного уровней хранения;
- распределенный RAID;
- мгновенные копии и клоны;
- репликацию, в том числе в режиме active-active на большие расстояния;
- дедупликацию;
- автоматизированное многоуровневое хранение;
- динамический QoS.

Программно-определяемые СХД также имеются и у оставшихся двух лидеров виртуализации: RedHat приобрел Gluster и Inktank, а Parrallels разработал собственную Cloud File System.

Сергей Платонов,
исследовательская лаборатория RAIDIX