ScaleIO: заоблачные перспективы

С момента своего появления в 2010 г. облако КРОК претерпело две значительные модернизации. Первая — в 2015 г. — связана с началом использования AFA-массивов Violin и продвижением сервисов с гарантированной производительностью. Вторая модернизация пришлась на 2016 г., когда был осуществлен переход с СХД на базе корпоративной файловой системы на гиперконвергентные SDS с использованием ПО ScaleIO, а также развертыванием файловой системы Ceph для объектного доступа вместо "рукописной".

Антон Семчишен – менеджер по продвижению комплексных решений департамента вычислительных систем компании КРОК.

Предыстория развития облака

Публичные облачные инфраструктуры по мере своего развития нуждаются в модернизации программно-аппаратной базы. Это может быть обусловлено, например, стартом крупных проектов, реализация которых требует более гибкого и быстрого масштабирования инфраструктуры.

Как показал наш опыт, работоспособным инструментом для эффективного апгрейда облака в текущих условиях может стать внедрение программно-определяемых хранилищ. В частности, такое решение на базе свободно распространяемой системы Ceph было использовано нами при модернизации объектного хранилища, применяемого для хранения файлов, образов виртуальных машин, статистического контента. Это позволило упростить управление хранилищем, обеспечить бОльшую надежность и гибкость работы с данными.

Несмотря на то, что теоретически Ceph позволяет поддерживать блочный, объектный и файловый доступ к данным, после тестирования решения мы выявили два важных для нас недостатка: Ceph дает меньшую производительность при работе с высокопроизводительными SSD дисками в сравнении с обычными СХД. Кроме того, чем сильнее нагружается общее кластерное решение, тем больше утилизируются серверы, на которых развернута Ceph. В результате при достижении определенного уровня нагрузки целесообразность использования Ceph сводится на нет. Поэтому в настоящее время в составе облака КРОК Ceph используется только в качестве объектного хранилища.

По мере увеличения количества облачных заказчиков и усложнения сервисов, предоставляемых на базе облака, появляется острая необходимость усовершенствовать блочное хранилище, которое у нас развернуто с использованием кластерной системы GPFS. Ее эксплуатация стала слишком дорогостоящим удовольствием, т.к. отнимает очень много ресурсов на поддержание надежной работы и тестирование. В настоящее время на рынке наблюдается рост количества гиперконвергентных решений, в которых подсистема хранения данных представляет собой программно-определяемую СХД (SDS), распределенную по вычислительным узлам. При этом в качестве носителей используются встроенные диски. Такой подход к построению подсистемы хранения отлично вписывается в концепцию облачного сервиса – с ростом количества вычислительных узлов растет объем и производительность СХД. Чтобы окончательно определиться с поставщиком требуемого для нас решения для блочного хранилища, мы приступили к масштабному тестированию гиперконвергентных SDS-решений.

Табл. 1. Требования к системе хранения, предъявленные в рамках тестирования КРОК.

Требования	Пояснение
Производительность	Система должна иметь примерно следующую производительность с одного хоста (3 сервера с 4 флеш-дисками на каждом): • последовательное чтение —100к IOPs; • рандомное чтение — 40к IOPs; • последовательная запись — 20к IOPs; • рандомная запись — 10к IOPs. *Данные взяты из максимума, полученного при тестировании
Отказоустойчивость	Наиболее важные компоненты системы (серверы метаданных, серверы мониторинга и управления кластером) должны быть задублированы (active-active, active-standby). В случае active-standby важно время переключения на резервный сервис, просадка производительности (примерно в %) и время недоступности системы, если таковое будет
Надежность	Все данные и метаданные должны иметь реплику и находиться на разных хостах кластера
Документированность	Система должна иметь полную, подробную и поддерживаемую в актуальном виде документацию по архитектуре и управлению.
Управление	Кластер должен управляться либо через web-интерфейс, либо через cli
Мониторинг	Желательно, чтобы система имела свои собственные утилиты мониторинга, которые можно привязать к существующим системам мониторинга. В случае отсутствия таковых необходимо проработать возможность и методы мониторинга кластера самописными скриптами и привязки их к существующим системам мониторинга
Support or community	Решение должно иметь профессиональный саппорт для заведения заявок или большое комьюнити для выяснения вопросов в случае необходимости
Масштабируемость	Система должна понятно и прозрачно масштабироваться. При этом скорость ребалансинга системы должна контролироваться и изменяться в зависимости от нужд
Затраты ресурсов	Система должна затрачивать минимальное количество ресурсов как с серверной стороны, так и клиентской: • затраты CPU на серверной стороне; • затраты memory на серверной стороне; • затраты CPU на клиентской стороне; • затраты memory на клиентской стороне.
Поддержка с клиентской стороны	Решение на клиентской стороне должно поддерживаться нативно, без использования fuse.

Табл. 2. Сравнение результатов тестирования ScaleIO и Ceph при нагрузке тестом FIO (3000IOPS 4k) по загрузке системных ресурсов узлов.

	ScaleIO			Ceph		
	write	read	mix (70R/30W)	write	read	mix (70R/30W)
mem_serv	750M	750M	750M	300M	300M	350M
cpu_serv	<10%	<10%	<10%	25-45%	25-47%	20-60%

Сравнительное тестирование SDS-решений

Назначением всех представленных на рынке программно-определяемых решений является оптимизация инфраструктуры хранения. Вопрос состоял только в выборе наиболее подходящей системы, отвечающей нашим техническим требованиям (табл. 1). В частности, для нас было важно, чтобы решение, *во-первых*, поддерживало нагрузку, генерируемую виртуальными машинами. *Во-вторых*, данное ПО не должно сильно влиять на производительность узлов, на которых оно размещается.

В поиске оптимального варианта мы протестировали возможности восьми кластерных файловых систем. Для этого развернули стенд из 6 серверов (6 SAS + 6 SSD дисков), 36 SATA и 36 SSD дисков по 12 дисков на сервер, использующих среду передачи данных на базе Infiniband (IPoIB).

Некоторые из тестируемых решений были «отметены» сразу, другие проверялись достаточно долго в условиях разных нагрузок. Особое внимание уделялось также масштабируемости и отказоустойчивости. Из всего набора тестов особое внимание было уделено производительности. СХД, созданная на базе локальных дисков физичес-

ких серверов под управлением EMC ScaleIO, «выдала» наилучшие результаты. Она была отказоустойчивой, в наименьшей мере нагружала серверы, что для нас было крайне важным, т.к. это предоставляло возможность запустить на тех же серверах максимальное число виртуальных машин.

Как видно из табл. 2, работа Ceph занимает в подсистеме виртуализации от четверти до половины ресурсов CPU, что может быть приемлемо только при использовании этого решения на выделенных под нужды хранения серверах. Оба решения использовали в тестировании менее 1 Гбайт памяти, что является приемлемым при использовании современных конфигураций узлов кластера (256+ГБ).

Важные архитектурные особенности решения на базе EMC ScaleIO

В настоящее время дедупликация, компрессия, шифрование в составе решения EMC ScaleIO не поддерживается и не используется (при этом не следует забывать, что поддержка и криптографии, и дедупликации – это ресурсоемкие операции, *прим. ред.*). Одним из важных преимуществ программно-определяемых СХД является возможность получения новых функциональных возможностей – в классических СХД, как правило, появление нового функционала, в лучшем случае, требует замены контроллеров. Например, во второй версии ScaleIO была добавлена возможность использования SSD дисков для кэширования данных, хотя это можно было реализовать и ранее с помощью аппаратной поддержки функции кэширования в контроллерах. Заказчики, которые сильно беспокоятся за безопасность данных, по-прежнему могут воспользоваться программными средствами шифрования на уровне ОС.

Все процедуры/сервисы данных, связанные с поддержанием доступности данных (например, репликация) реализуются в рамках одного кластера на одной площадке с использованием коммутаторов 56 Гбит/с Infiniband. Защита данных в кластере осуществляется путем дублирования данных между узлами кластера с фактором репликации 2 (единственный доступный вариант). При этом процесс репликации, очистки и равномерного перераспределения данных осуществляется автоматически, без вмешательства администратора. Роль управляющего кластера распределена как минимум между тремя серверами (выделенные серверы метаданных не используются). Настройки ScaleIO позволяют создавать так называемые домены доступности – мы можем настроить репликацию данных внутри кластера таким образом, чтобы блок данных и его копия никогда не находились в одной серверной стойке или одном машинном зале, что позволяет дополнительно увеличить доступность данных. Теоретически данный функционал может быть использован для репликации между ЦОДами, однако такая модель использования приведет к катастрофическому росту задержек внутри кластера, поэтому производитель рекомендует использовать для такой репликации специальное решение – RecoverPoint.

Антон Семчишен,
компания КРОК

Эволюция Oracle SPARC

Июнь 2016 г. – Корпорация Oracle представила новые серверы на базе процессоров SPARC S7. Эти процессоры были впервые анонсированы на конференции Hot Chips 2015. Тогда же Oracle официально объявила о своих планах относительно будущих SPARC, предназначенных для использования в горизонтально масштабируемых системах. Теперь эти планы становятся реальностью.

Наряду с новейшим поколением процессоров S7 компания выпускает новые серверы, оптимизированные программно-аппаратные комплексы и облачные сервисы Oracle Cloud – полный набор решений, развертываемых на площадке заказчика и в облаке.

Новые процессоры – новая «экономика»

Новые процессоры построены на ядре SPARC M7, но вместо 32 вычислительных ядер имеют только восемь. Благодаря этому серверы на базе S7 удалось сделать более компактными, энергоэффективными и доступными по цене. Они будут привлекательны для компаний любого размера, которые из экономических соображений подумывают о миграции на архитектуру x86, а также тех, для кого мощности платформ на M7 избыточны, а это, прежде всего, большое количество организаций, продолжающих эксплуатировать SPARC-системы предыдущих поколений.

Теперь платформы на базе S7 могут конкурировать по цене с серверами x86 при производительности, до двух раз превышающей показатели сопоставимых Intel-систем в расчете на вычислительное ядро. Важно также отметить, что поскольку в S7 используется микропроцессорная архитектура M7, новые серверы предлагают те же функциональные возможности, включая реализуемые в «ПО на кристалле» (Software in Silicon), такие как ускорение SQL-запросов, защита памяти, поддержка выделенных страниц и др. Эта функциональность превосходит все, что могут сегодня предложить производители в данном сегменте. Новые серверы доступны и в облаке – в платформенном сервисе Oracle Cloud Compute.

S7 в «железе»

Oracle представила новые продукты – два двухпроцессорных стоечных сервера S7-2 и S7-2L в корпусе 1U и 2U на базе 8-ядерных процессоров S7 с тактовой частотой 4,27 ГГц и функциональностью «ПО на кристалле», включая контроль целостности данных, обеспечение информационной безопасности и аппаратную поддержку аналитических операций. Впрочем, данные системы подходят не только для задач аналитики, но и для самого широкого спектра нагрузок. Их поставки начнутся в течение ближайшего месяца.

«Эффективная производительность на ядро у новых систем Oracle на 50%–100% превосходит показатели типовых серверов x86, при выполнении Java-приложений серверы на процессорах S7 опережают конкурентов в 1,7 раза, а для баз данных этот коэффициент составляет 1,6», – сообщил Маршал Чой (Marshall Choy), вице-президент по развитию системных продуктов. Что касается аналитических задач, то в этом им вовсе нет равных – они решаются до 10 раз быстрее благодаря позаимствованному из M7 ускорителю операций. Наличие открытых интерфейсов программирования (API) делает эти возможности доступными и сторонним разработчикам. А встроенные средства безопасности позволяют заказчикам создавать полностью защищенное облако со сквозным шифрованием данных.

Анонсированы и новые оптимизированные программно-аппаратные комплексы на базе S7. В их числе: простой в развертывании Oracle MiniCluster S7-2. Эта платформа предназначена для создания конфигураций высокой доступности или с повышенными требованиями к безопасности, использования в качестве UNIX-сервера или сервера баз данных. Экономически такое решение более выигрышное, чем самостоятельная интеграция всех требуемых компонентов программно-аппаратного стека. Таким образом, Oracle теперь предлагает широкий выбор кластерных конфигураций – от старшей модели (Oracle SuperCluster) до систем среднего класса.

Новый этап облачных вычислений

По словам Маршала Чоя, новая линейка систем на базе процессоров S7 позволит Oracle расширить спектр своих предложений, предоставить заказчикам системы для самых разных задач – от высокопроизводительных платформ на базе M7 до серверов среднего класса на процессорах S7. Последние станут оптимальным решением для таких приложений, как ERP-системы, приложения бэк-офиса, различные облачные нагрузки и др. Новинки также представляют интерес для провайдеров облачных сервисов, могут с выгодой использоваться для построения ЦОД высокой плотности.

Intel Core™ i7 Extreme Edition

Июнь 2016 г. — Корпорация Intel представила в России первый 10-ядерный процессор Intel® Core™ i7 Extreme Edition для настольных систем. Это самый мощный чип для десктопов за всю историю отрасли демонстрирует беспрецедентную производительность и поддерживает ряд уникальных возможностей.

Наличие кэш-памяти Smart Cache емкостью 25 МБ непосредственно на процессоре, поддержка быстрой ОЗУ, работающей на частоте до 2400 МГц, а также возможность подключения до 4 дискретных графических карт — все это даст серьезное преимущество при запуске самых современных компьютерных игр.

Новые процессоры серии Extreme Edition позволяют работать сразу в нескольких требовательных к вычислительным ресурсам приложениях без заметного снижения производительности системы, легко создавая видео с разрешением 4K или монтировать панорамные кадры с углом обзора 360 градусов.

Среди уникальных функций: возможность оверклокинга каждого ядра в отдельности, управление коэффициентом AVX для повышения стабильности системы и др.