All-flash для IaaS-сервисов?

Интервью с Максимом Березиным – руководителем направления облачных вычислений, директором Виртуального дата-центра КРОК.

Максим Березин – руководитель направления облачных вычислений, директор Виртуального дата-центра КРОК.

SN. В чем уникальность обновленной публичной облачной платформы, запущенной КРОК в конце прошлого года?

М.Б. Можно назвать несколько уникальных особенностей нашей облачной платформы. *Во-первых*, мы одними из первых в России для поддержки своих публичных IaaS-сервисов использовали all-flash массивы Violin, одновременно рассеяв миф об их меньшей конкурентоспособности в сравнении с традиционными и гибридными дисковыми массивами high-end класса. *Во-вторых*, нам удалось на базе одной платформы хранения обеспечить консолидацию и поддержку самой разнородной нагрузки с гарантированной производительностью и минимальными задержками при доступе к данным. Вообще, если говорить про гарантии производительности, то в облаке КРОК они тройные: обеспечиваются, помимо, флеш-СХД, мощными виртуальными дисками и сетью Infiniband. *В-третьих*, мы добились существенной минимизации администрирования платформы хранения Violin с точки зрения ее масштабируемости, управляемости (включая пиковые нагрузки), поддержки заданных SLA. *Наконец*, облако КРОК является собственной разработкой, большинство узлов которой построено на базе открытого ПО.

SN. Чем был обусловлен переход на all-flash платформу Violin?

М.Б. Мы запустили собственную публичную облачную платформу в России одними из первых еще в 2010 году. *Нам нужен был рост в геометрической прогрессии из года в год.* Чтобы обеспечить подобную динамику, архитектура облачной платформы должна была быть сверхгибкой и масштабируемой, а сама платформа – хорошо управляемой. В какой-то момент мы начали упираться в архитектурные ограничения, заложенные еще на заре развития облака, а именно: в подсистему хранения данных.

Одна из важнейших задач роста – обеспечение потребностей заказчиков не только в части объемов дисковой емкости, но и в части гарантированной производительности используемых дисков. Нужно было избежать взаимного влияния соседей по СХД, сделать ситуацию полностью управляемой, дать гарантии по производительности дисков в рамках SLA в пределах от 400 IOPS до 100 000 IOPS на диск.

Ко всему прочему, тема с гарантированными дисками в облаке была подогрета и законом о персональных данных, вступающим в силу с 1 сентября 2015 года. Многие заказчики размещали и размещают свои ИТ-сервисы за границей, зачастую – на выделенном физическом высокопроизводительном оборудовании в ЦОДе провайдеров. Данные теперь нужно перенести на территорию России, но от высокой гарантированной производительности дисков отказываться заказчики не могут, а ждать поставки оборудования в РФ уже нет возможности. Облако в этом случае является, пожалуй, единственным способом успеть перенести данные в РФ и получить аналогичные технические параметры производительности, как на используемом ранее физическом оборудовании провайдера.

Облако КРОК сейчас – это более 500 физических серверов и 12 хранилищ данных, распределенных между двумя удаленными площадками. Порядка 70% размещенных в облаке виртуальных серверов — это продуктивные среды крупных корпоративных клиентов и компаний средней величины. Тестовых сред и сред разработки всего порядка 30%. У нас размещен ряд довольно тяжелых инсталляций Oracle, SAP и различных высоконагруженных OLAP-систем заказчиков.

SN. Какие проблемы вам удалось решить с переходом на платформу Violin?

М.Б. На базе классических дисковых массивов с дисковыми полками и контроллерной парой мы запустились в 2010 г. и жили до недавнего времени. *Одна из основных проблем, с которой нам постоянно приходилось бороться, – на хранилищах данного типа практически отсутствовали внятные механизмы управления качеством предоставления услуг.* Классические дисковые СХД, как правило, позволяют только создать приоритеты выдаваемой производительности на определенные виртуальные диски. Когда количество таких дисков на каждом из массивов доходит до сотен и тысяч, о задании каких-либо приоритетов можно забыть. Подобные механизмы управления производительностью на СХД не только не работали, а просто ухудшали и без того сложную ситуацию с недостатком производительности и управляемости.

Второй серьезной архитектурной проблемой стала жесткая привязка параметров емкости и производительности дисковых массивов. Проще говоря, есть 8 физических дисков по 150 ГБ со скоростью вращения 15K. Мы хотим собрать из них RAID10. Эффективная емкость логического диска получится на уровне 600 ГБ. Сам по себе каждый диск в отдельности выдает до 150 IOPs, а значит, логический диск выдаст порядка 1000 IOPs. В итоге получаем ситуацию: на каждый ТБ эффективной дисковой емкости приходится около1500 IOPs производительности.

Теперь посмотрим на наиболее распространенные задачи заказчиков: например, требуется разместить в облаке базу на 2 ТБ и получить производительность 10 000–50 000 IOPs. Чтобы получить такую производительность, нужно взять гораздо больше дисковой емкости, нежели на самом деле необходимо (10–50 ТБ). Вы можете возразить мне, что на массивах есть контроллерная пара с кэшом и что можно добавить полку с твердотельными дисками и т.д. Кэш контроллеров на случайном чтении/записи сильно ситуацию не улучшит. Твердотельные диски лишь отчасти спасают ситуацию с производительностью, а решение проблемы с управляемостью лишь отодвигают в будущее. К тому же, классический контроллер не может раскрыть всю производительность твердотельного диска.

Ситуация с производительностью дисков в облаке КРОК становилась критической, заказчики начинали жаловаться на непредсказуемые просадки по производительности дисковой подсистемы. Нужно было принимать решение о дальнейшем развитии облака. Рассматривали различные решения, в том числе и переход на high-end массивы. Но осознавали, что это не решение проблемы, а просто дорогой «костыль», который позволит прожить еще пару лет, пока мы не придем к аналогичной ситуации. Про «неуправляемость» подобных решений я уже сказал выше. *Также high-end массивы не предназначены для очень частого создания и удаления дисковых лунов* – данные операции занимают много времени и превращаются в длинные очереди из запросов, что не подходит для работы в облаке. В общем, задача не бралась малой кровью, и мы были вынуждены начать все с чистого листа, провести локальную революцию в части хранения данных в облаке.

SN. Какие основные потребности заказчиков вы удовлетворяли и какие проблемы вам пришлось решать при переводе своего облака на новую платформу хранения?

М.Б. Основные потребности заказчиков сводились к следующему:

Во-первых, нужно было обеспечить гарантированную производительность, за-

фиксированную в SLA, ввода/вывода без возможности взаимного влияния соседей по массиву друг на друга. *Во-вторых*, обеспечить прозрачное наращивание и уменьшение производительности дисковых ресурсов на лету с изменением тарифов за использование разных дисков. *В-третьих*, размещать в облаке высоконагруженные базы данных, требующие от 20 000 до 100 000 IOPs на диск. *В четвертых*, обеспечить высокую доступность данных. *В-пятых*, поддерживать самую разнородную нагрузку по соотношению операций чтения/записи при незначительном отклонении в производительности. *Наконец*, массив должен иметь уровень надежности корпоративного класса.

Мы давно смотрели в сторону использования all-flash массивов Violin. Так сложилось, что именно на базе данного класса решений наша облачная задача с хранением данных решалась наиболее красиво, а что самое интересное – финансово оправданно. При этом мы, тем не менее, протестировали практически все доступные на рынке решения корпоративного уровня, а также поработали с наиболее заметными стартапами. При выборе решения основными критериями были: высокая производительность и отсутствие зависимости дисковой емкости от количества IOPs; достаточное количество IOPs'ов для гарантии их выделения в любой момент времени; наличие Infiniband-адаптеров для подключения к существующей SAN; документированные и понятные средства управления массивами посредством API.

Остановились в итоге на массивах Violin Memory 6264. Они полностью удовлетворяли нашим требованиям. В итоге суммарно на 8 приобретенных массивов получили 0,56 ПБ сырой / 0,32 ПБ полезной флеш-емкости с суммарной производительностью 8 000 000 IOPs.

Это был безопасный выбор, учитывая, что Violin Memory создала All-Flash рынок в 2007 году и дальше всех продвинулась в совершенствовании продукта. Компания разработала систему с нуля совместно с Toshiba — изобретателем NAND-Flash и обладает ключевыми патентами на работу с флеш. В массиве нет никаких SAS-петель, батарей и прочих унаследованных от дисковых систем архаизмов, без компромиссов в виде SSD. Вместо этого — современная архитектура FFA на шине PCI. Используются носители собственной разработки VIMM емкостью 1,1 ТБ.

Гарантии производительности осуществляются путем выставления ограничений на уровне гипервизора на каждый диск в отдельности. Программное обеспечение по управлению массивами в автоматическом режиме определяет текущую утилизацию каждого из них и выбирает наименее занятые массивы для размещения новых дисков. Массив совсем компактный — 1 000 000 IOPs всего в 3 юнитах с номинальным временем отклика 250 мкс.

SN. Были ли какие-либо проблемы при миграции на новую платформу хранения?

М.Б. При миграции не все пошло гладко, и возникли интеграционные проблемы нашего облака и массивов. Суть их заключалась в том, что мы могли реализовать потенциал массивов примерно на 10% от их возможностей, а именно столкнулись с максимальным количеством создаваемых дисков на один массив. В облаке КРОК мы используем Infiniband как для сети хранения данных, так и для связи между виртуальными машинами. В массивах Violin Memory используется протокол SRP (SCSI over RDMA) для подключения LUN'ов к серверам по Infiniband. Этот протокол обладает следующей особенностью: контрольные команды используют subnet manager сети Infiniband. В обычной ситуации, когда количество LUN'ов и серверов не очень большое, как и подключений между ними, это не является проблемой. Но не в случае облака. Из-за того, что подключений, то есть путей между серверами и LUN'ами, очень много, subnet manager'ы сети Inifiniband уходили в себя при перестроении топологии сети. Просто не хватало процессорной мощности. Большое количество путей также создавало сложности на контроллерах доступа СХД — они начинали работать очень медленно, вызывая ошибки в драйверах, которые считали, что путь отвалился.

Как мы решили проблему? Вместе с компанией Violin Memory провели большую работу по оптимизации количества путей: каждый отдельный LUN мы стали подключать, то есть экспортировать, к отдельному физическому серверу. Сложность заключалась в том, что это нужно было делать при помощи ReST API массива, что тоже проходило нелегко. Если LUN экспортирован на все хосты и на еще один конкретный хост, то он экспортируется и с этого конкретного LUN'а. Это требует выполнения живой миграции всех LUN'ов, чтобы в итоге избавиться от экспорта на все хосты. Так как при каждом включении/выключении сервера проходил новый запрос на API массива, у нас повысились требования к производительности этого API. В результате потребовалось инициировать выпуск новых версий прошивок для оборудования.

С нашей стороны в решении этого вопроса на протяжении более полугода было занято три программиста и три разработчика. Мы написали огромное количество тестов для нашего программного кода, оптимизировали систему автоматизированной сборки и тестирования. В результате у нас добавились тысячи строк оттестированного программного кода. Может показаться, будто оптимизация количества путей — это простая задача, но на самом деле это не так: на ее выполнение нам потребовались серьезные трудозатраты и больше десяти месяцев. Но у нас была очень серьезная мотивация — мы работали на беспрецедентное по производительности и отказоустойчивости решение в интересах наших облачных заказчиков.

SN. Несколько слов о конечных результатах, которые вам удалось достичь при развертывании IaaS-сервисов.

М.Б. Диапазон доступной производительности дисков, которые можно создать на массивах, — до 100 000 IOPs на диск. Причем, производительность эта гарантированная, а не плавающая как обычно принято на рынке публичных облачных платформ.

Мы предоставляем заказчикам по умолчанию диски со следующей производительностью: 400, 1000, 3000, 5000, 10000 IOPs, соответственно. Заказчики через портал самообслуживания имеют возможность запуска дисков разной производительности, а также смены параметров их количества IOPs на лету. А диски производительностью от 10 000 IOPs до 100 000 IOPs добавляются на портал самообслуживания по запросу в службу технической поддержки. Как правило, это индивидуальные уровни хранения, параметры которых определяются для каждого заказчика в отдельности по итогам нагрузочного тестирования. Мы их не прячем, просто действительно далеко не всем нужны такие высокопроизводительные уровни хранения. Нам не жалко, ведь среднее количество IOPs на 1ТБ емкости в нашем облаке – 25 000. Это действительно фантастический показатель.

Массивы распределены между двумя облачными платформами КРОК в распределенных дата-центрах на территории Москвы: в одном – 5 хранилищ, в другом – 3 (рис. 1). Это позволяет заказчикам строить Disaster recovery решения с использованием высоконагруженных систем как на одной площадке, так и на другой. Причем управление массивами на обеих площадках выполняется посредством единого портала самообслуживания.

У КРОК есть отдельный SLA на производительность дисков, помимо стандартного SLA на доступность виртуальных серверов. Определение недоступности производительности гарантированных дисков в нем звучит следующим образом: «Недоступность производительности флеш-диска – состояние флеш-диска, когда в течение пяти минут процессор виртуальной машины, к которой он подключен, ожидает данные от дисковой подсистемы более 10% времени, или задержка

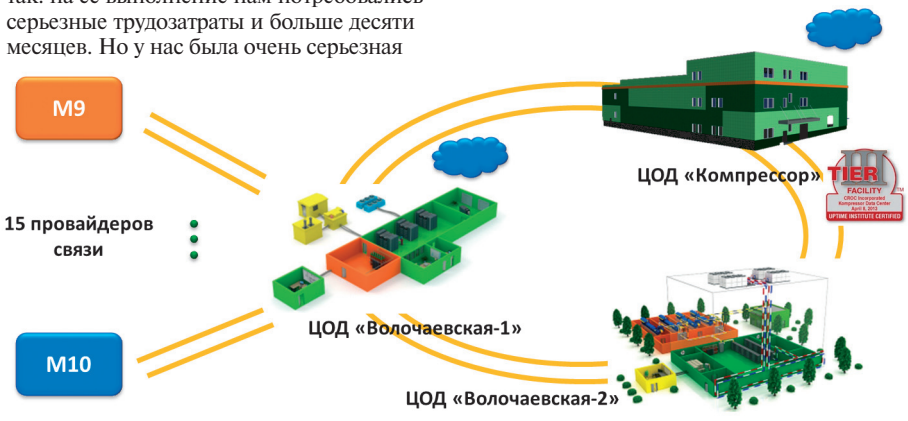

Рис. 1. Массивы распределены между двумя облачными платформами КРОК.

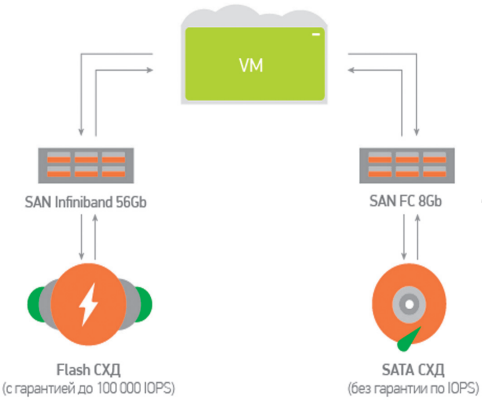

Рис. 2. В облаке также есть и SATA-хранилище данных, на котором можно разместить наименее горячие данные по оптимизированным тарифам.

получения данных от дисковой подсистемы более 5 мс и при этом количество запросов ввода-вывода (IOPs) на флеш-диск меньше, чем заявленная производительность флеш-диска на 3%».

Под заявленной производительностью флеш-диска здесь подразумеваются как раз стандартные уровни хранения 400, 1000, 3000, 5000, 10000 IOPs, соответственно, или индивидуальные уровни хранения от 10 000 IOPs до 100 000 IOPs. Как только мы выходим за заявленные параметры производительности, начинают отсчитываться минуты недоступности, а как только мы выходим за SLA 99,9, то сразу же попадаем на штрафы. Одним словом, работать без соответствия SLA совсем не в наших интересах.

Что же делать, если гарантированная производительность дисков не нужна, а нужно просто хранить большой объем данных по умеренным ценам? В облаке также есть и SATA-хранилище данных, на котором можно разместить наименее горячие данные по оптимизированным тарифам (рис. 2). А миграция между разными типами хранилищ производится на лету.

Услуга оказалась настолько востребованной и своевременной, что первые восемь флеш-массивов в нашем облаке были заполнены на 60% за 4 месяца. В настоящий момент мы устанавливаем еще восемь таких же массивов. Это будет первое облако в России с более чем 1 ПБ флеш-памяти.

(Материал подготовлен на основе статьи в корпоративном блоге КРОК на Хабрахабр)

Платформа Violin Memory получила статус VMware Ready

Май 2015 г. – Компания Violin Memory®, Inc. объявили о присвоении флэш-платформе хранения данных Violin FSP 7300 статуса VMware Ready™ и завершении сертификации VMware Horizon Fast Track 2.0 Proven Storage. В результате всесторонней проверки Violin получила высшую оценку со стороны VMware.

«Богатый функционал платформы хранения Violin FSP7300 нацелен на решение задач производительности, снижения затрат и масштабируемости в проектах построения инфраструктуры виртуальных станций (VDI), – заявила директор по маркетингу Violin Memory Эми Лав. – И позволяет пользователям управлять VDI-средами и расширять их по мере необходимости с минимальными затратами».

«Мы рады тому, что Violin FSP 7300 может заслуженно носить эмблему VMware Ready™, которая, на деле означает совместимость и применимость решения Violin Memory в инфраструктурах VMware, упрощая клиентам реализацию проектов», – отметил старший директор подразделения Global Technology Partnering Organization компании VMware Ховард Холл.

«На фоне экономического спада в России проекты виртуальных станций активно реализуются. Флэш позволяет поймать двух зайцев сразу: снизить стоимость проекта и обеспечить масштабирование станций до нескольких тысяч с сохранением высокой производительности и низкого времени отклика. Сертификация Violin с VMware открывает рынку возможность выбирать флэш-платформу, которая адресно решает обе задачи», – сообщил глава представительства Violin Memory в России и СНГ Максим Зубарев.

Программа VMware Ready™ технологических альянс-партнеров (TAP) позволяет клиентам выбирать продукты третьих сторон, сертифицированные для работы в инфраструктуре VMware. Именно такие продукты позволяют клиентам снижать риски и затраты проекта по сравнению с кустарными решениями. Программа VMware TAP насчитывает тысячи участников по всему миру, включает лучших в отрасли технологических партнеров с обязательством предоставлять проверенные практики и бизнес-решения, в соответствии с потребностями клиентов. *Больше информации о платформе хранения Violin 7300* FSP: *http://www.vmware.com/files/pdf/partners/violin/VMware-Horizon-Violin-Memory-Solution-Brief.pdf.*

AMD: графика с 32 Гбайт для HPC

Июль 2015 г. – Компания AMD представила новую серверную графику AMD FirePro™ S9170 – первую в мире одночиповую видеокарту с 32 ГБ памяти на борту. Она предназначена для обработки ресурсоемких задач с двойной точностью (DGEMM) и поддерживает библиотеки OpenCL™ 2.0 (она также готова к работе с инструментами разработчиков OpenMP и OpenACC, которые будут представлены в 3 кв. 2015 г.). Новая графика, основанная на базе архитектуры AMD Graphics Core Next (GCN), обеспечивает пиковую производительность, до 5,24 терафлопс одинарной точности, а также предоставляет самые широкие возможности для вычислений с двойной точностью при пиковой производительности на уровне до 2,62 терафлопс[1)].

«AMD занимает первую строчку в списке Green500 по состоянию на ноябрь 2014 г. Сегодня портфолио AMD для производительных вычислительных систем дополнила новая видеокарта AMD FirePro S9170», – сказал Шон Бёрк *(Sean Burke)*, вице-президент и генеральный менеджер подразделения профессиональной графики AMD. – Серверная графика AMD FirePro S9170 может ускорить обработку комплексных нагрузок в области научных изысканий, анализа данных, сейсмологии – все это за счет максимально доступного количества памяти в 32 Гбайт. Мы разработали наше решение специально для суперкомпьютеров, и оно позволяет добиться высокой производительности при скромных затратах электроэнергии».

«Существуют такие задачи HPC, которые требуют наличия максимального количества данных на устройстве постоянно, и поэтому 32 Гбайт памяти на борту AMD FirePro S9170 – рекордное количество для одночипового графического адаптера – поможет ускорить научные расчеты, реализация которых раньше была просто невозможна, – говорит Саймон МакИнтош-Смит *(Simon McIntosh-Smith)*, глава Microelectronics Research Group в Бристольском Университете. – Например, наша новая версия транспортного кода SNAP на базе OpenCL из Национальной Лаборатории Лос-Аламоса требует хранить максимально возможный объем данных на устройстве, и 32 Гбайт поможет нам исследовать проблемы значительно большего масштаба и делать это быстрее, чем когда-либо. Большой объем памяти вместе с пропускной способностью в 320 Гб/с и двойной вычислительной точностью делают серверный графический процессор AMD FirePro S9170 «убойным» решением для множества высокопроизводительных вычислений».

1) Самая высокопроизводительная серверная графическая карта Nvidia с одним процессором на май 2015 г – Tesla K40 – демонстрирует пиковую производительность при двойной точности на уровне 1,43 терафлопс. Самый высокопроизводительный адаптер Nvidia с двумя GPU на май 2015 г – Tesla K80 – демонстрирует пиковую производительность при двойной точности 1,87 терафлопс. Спецификации продуктов Nvidia можно уточнить на сайте *http://www.nvidia.com/object/tesla-servers.html.*

Gemalto: 100 Гбит/с сетевой шифратор

Июнь 2015 г. – Компания Gemalto объявила о выпуске нового многоканального высокоскоростного устройства шифрования сети – SafeNet CN8000, скорость шифрования которого составляет 100 Гбит/с в одном блоке. SafeNet CN8000 позволяет зашифровывать множество сетевых подключений и еще больше трафика, обеспечивая требуемую производительность и безопасность для организаций, которые работают с крупномасштабными сетями с высокой пропускной способностью.

SafeNet CN8000 поддерживает мультиарендность, что является значительным преимуществом и обеспечивает гибкость в обработке информации для тех организаций, которым требуется разделять определенные конфиденциальные данные и сетевые подключения.

SafeNet CN8000 работает на основе квантового формирования случайных чисел, что позволяет гарантировать высококачественную произвольность, обеспечивающую защиту от целенаправленных криптоатак.