Многоуровневые тома RAID для ЦОД и интернет-компаний

Обсуждение возможности создания многоуровневых томов (как по принципу кэширования данных, так и за счет объединения уровней) на базе анонсированного в сентябре 2013 года четвертого поколения решений Adaptec maxCache Plus.

Дмитрий Зотов – инженер компании PMC-Sierra, подразделение Adaptec by PMC.

Введение

В статье пойдет речь о так называемых уровневых томах RAID (RAID volume with tiering) или томах RAID, использующих tiering-схему. Возможность создания таких томов в полной мере появилась с приходом на рынок решений Adaptec 4-го поколения (с поддержкой SAS 12 Гбит/с, анонс – сентябрь 2013 г.) и, в частности, нового ПО – maxCache Plus для уровневого управления устройствами хранения.

Возможность назначать различные категории данных различным типам устройств хранения повышает производительность доступа к данным и снижает совокупные затраты. Решение Adaptec maxCache Plus обеспечивает возможности уровневого управления, позволяя использовать любые устройства хранения уровня блоков (RAID и HBA, флэш-устройства PCIe, чипсеты материнских плат) для создания виртуальных пулов хранения в серверной среде.

maxCache Plus управляет структурой хранилища посредством:

- *Volume Manager – управление вводом-выводом уровневого тома и маршрутизацией к правильной уровневой группе;*
- *Policy Engine – это интеллект продукта, который определяет, где разместить управляемые данные.*

Концепция

Уровневый (tiering) подход заметно усложняет концепцию RAID-тома. Понятие тома RAID (очень часто его называют виртуальным диском) становится более комплексным, более сложным объектом, иными словами, степень виртуализации усиливается и усложняется. Для непосвященных людей, выражаясь простым языком, то, что вы видите в вашей системе как устройство “C:”, в рамках концепции tiers – это уже не отдельный жесткий диск (или его часть) и не группа дисков, собранных в RAID-том, например RAID6. Это – работающая как единое целое группа томов RAID, например, RAID1 на дисках SSD и RAID6 на дисках HDD.

И это еще не все сложности. Сам контроллер RAID в данной схеме становится “виртуальным”, собранным методом объединения из ряда RAID-контроллеров и HBA-карточек. При этом интегрированные на материнской плате контроллеры RAID и HBA также могут быть включены в эту схему. Более того, такое объединение возможно и с контроллерами “дисков” NVRAM (которые по природе своей являются оперативной памятью) и с контроллерами PCIe.

Если опираться на теорию (людям непосвященным это может показаться довольно сложным, но необходимость в такого рода подходах уже назрела, поэтому, рискнем представить эти рассуждения в данной статье), то такие возможности заложены в базовом свойстве информационных систем: свойства “прямого и обратного” объединения.

Система хранения ведет себя как отдельная информационная система. В ней есть специфика. Но она полностью удовлетворяет определению информационной системы. Следовательно, к ней можно применить метод объединения или разделения.

Примеры такого объединения на более высоких уровнях, таких как отдельный сервер, – это кластерные решения. Берется один сервер как некая информационная система, затем другой сервер – как другая информационная система, и через специальное программное обеспечение (в качестве примера можно взять ПО компании Microsoft High Availability Cluster) они объединяются в один виртуальный сервер, т.е. два физических ведут себя как один виртуальный.

В качестве примера обратной операции (разъединения) можно взять сервер с VMware, где на одном физическом сервере генерируется несколько виртуальных машин.

С учетом абсолютной универсальности этого свойства можно задаться вопросом: почему на уровне систем это реализовано довольно давно, а на уровне подсистем мы только начинаем получать это как работающий и отлаженный механизм?

Проблема очевидна. Если взять контроллер Adaptec RAID, то он поддерживает большое количество операционных систем.

Возможность его объединения с другими контроллерами в один виртуальный, так, чтобы диски с разных контроллеров можно было бы использовать в одном томе RAID, не такая уж простая задача.

Неслучайно для линейки контроллеров Adaptec Series 8Q, которые поддерживаются почти всем классом операционных систем, такое объединение пока возможно только в операционных системах Linux и Windows (получить детальную информацию можно на сайте *www.adaptec.com/support* в разделе userguide для maxView Storage Manager контроллеров Series 8Q).

Обратите внимание, как это усложняет жизнь. Зададимся вопросом: какие операционные системы поддерживает RAID-контроллер?

Если вы его не виртуализируете – это один список. Если виртуализируете методом объединения – другой.

Если предположить, что поддерживается и метод разъединения (один физический контроллер распадается на несколько виртуальных) – это уже может быть и третий список. Вы скажете: научная фантастика. Нет, практически уже реальная жизнь.

Использование

Вполне логично возникают вопросы: для чего разработана такая степень виртуализации, нужна ли такая сложная структура, оправдывает ли она свое существование, и почему еще вчера в ней не было столь большой потребности, а сегодня на подобные решения уже имеется устойчивый спрос?

Не вдаваясь в детальный анализ, перечислим несколько очевидных причин.

Создание систем хранения для большого числа пользователей

Здесь подразумеваются системы хранения, используемые в датацентрах. При таком подходе решения типа tiering позволяют более эффективно и с меньшими затратами получить нужные требования к производительности, емкости и надежности хранилища.

Создание виртуальных машин

Создание большого количества виртуальных машин на одной “железной”

(hadware) платформе приводит к новому подходу к системе хранения. Выделять "целый" том RAID для каждой виртуальной машины слишком нерационально. Как решение создается один или несколько виртуальных томов по схеме tiering, где, опять-таки, свойства производительности, надежности и емкости максимально оптимизированы и по цене, и по другим параметрам (например, количество SSD и HDD во внутренних дисковых корзинах серверной платформы), и полезная емкость раздается нужным образом виртуальным машинам. Такой подход позволяет не задумываться об управлении производительностью для отдельных приложений на виртуальных машинах, через настройки отдельных RAID-томов – это задача специальных алгоритмов внутри уровневой схемы tiering.

Совмещение наиболее сильных свойств дисков SSD и HDD и других типов памяти в одном решении

Как правило, для этого используют более частные или производные решения для схемы tiering: такие, как гибридные тома (hybrid volumes) и SSD-кэширование (SSD-cache). Для проектов высокого класса tiering в чистом виде дает наиболее высокий экономический эффект.

Здесь указаны в качестве примера несколько основных причин – полный список занял бы слишком много места. Из вышесказанного понятно, что данное решение востребовано в проектах класса enterprise, в проектах hosting- и internet-площадок. Это достаточно популярный вид проектов, из чего можно смело сделать заключение, что механизм, о котором мы собираемся поговорить подробно, будет находить высокий спрос в среде проектировщиков и потребителей решений для систем хранения.

Рис. 1. Контроллер Adaptec 81605ZQ.

Рис. 2. SAS3-экспандер Adaptec 82885T.

Практическая реализация

Описанную выше схему можно реализовать с помощью контроллеров Adaptec Series 8Q (рис. 1) и специального дополнительного пакета программного обеспечения, прилагаемого к утилите управления – maxCache Plus.

Контроллеры Adaptec Series 8Q по праву можно включить в число продуктов для датацентров и интернет-компаний. Впервые в линейку добавлен отдельный 36-портовый SAS-экспандер для более удобного подключения большого количества дисков HDD и SSD (рис. 2).

На примере Adaptec Series 8Q мы видим появление в линейке современных контроллеров RAID не только новых подходов к самой линейке (почти 50% моделей имеют опцию Q, появились отдельные SAS-экспандеры и т.д.), но и новых функций, реализованных через настройки контроллера. Большей частью все эти особенности продиктованы необходимостью в создании подсистем хранения для серверов корпоративных и коммерческих центров обработки данных, а также интернет-компаний. Практическая реализация таких схем, как "виртуальный контроллер" и "уровневые тома (tiering)" призвана дать пользователям возможность оценить их реальную необходимость и преимущества в таких решениях.

Этапы реализации

На первом шаге в сервер необходимо установновить хотя бы один контроллер Adaptec Series 8Q. Установить драйверы всех устройств и начать установку утилиты управления maxView Storage Manager.

В утилите, помимо традиционных компонентов для управления, выбрать установку пакета maxCache Plus, который, по сути, является драйвером виртуального контроллера (рис. 3).

Используя традиционную схему управления (в качестве ведущего контроллера выбирается именно Adaptec Series 8Q, но диски можно будет выбирать и назначать на уровни с разных контроллеров, доступных в системе, см. рис. 4.), произвести назначение томов RAID из дисков на разные уровни (tiers).

Пока наша схема имеет всего два уровня иерархии хранения, но в будущем их станет больше.

Традиционно уровень 0 (tier 0) предназначен для высокопроизводительных дисков (таких, как устройства NVMe

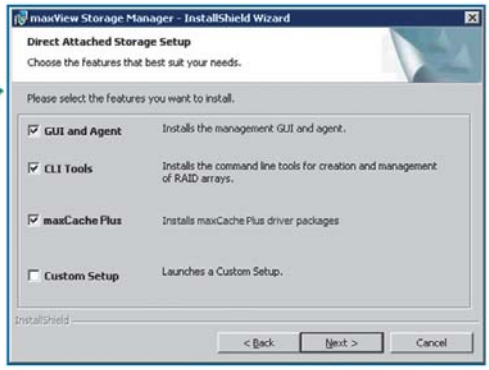

Рис. 3. Схема установки Adaptec MaxCache Plus.

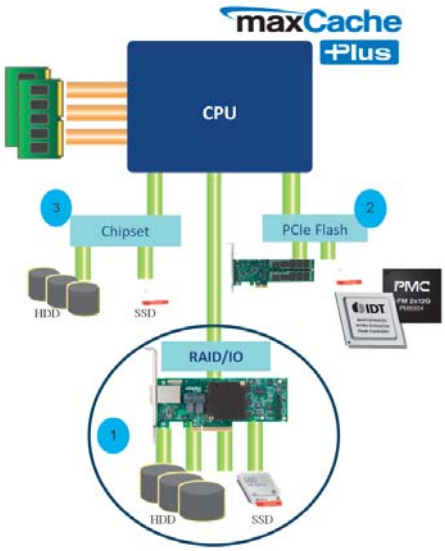

Рис. 4. Структурное представление многоуровнего RAID-тома.

Рис. 5. Объединение в рамках одного тома двух режимов: кэширования и объединения данных.

RAM, PCIe SSD или традиционные диски SAS/SATA SSD). Например, для уровня 0 можно использовать диски PCIe SSD контроллера 2 или диски SSD контроллера 1 (рис. 4).

Уровень 1 (tier 1) используется для относительно менее производительных устройств, таких, как классические диски SAS/SATA HDD (с контроллера 3 и контроллера 1).

Из SSD создаем том RAID, например, RAID1, и назначаем его к уровню tier 0, из HDD создаем том, например, RAID6 и назначаем его к уровню tier 1.

Далее объединяем оба уровня, указывая нужную нам емкость от RAID6 и RAID1 и режим работы уровневого тома. По большому счету, есть два режима работы: режим кэширования и режим объединения (рис. 5).

При режиме кэширования (изначально все данные размещаются на уровне tier 1) "горячие" данные только копируются на уровень tier 0. "Горячие" данные – это данные, к которым пользователи обращаются чаще всего (определяются они динамически внутренним алгоритмом контроллера). Такой режим не является чем-то необычным. Он поддерживается контроллерами Adaptec Series 5Q, 6Q и 7Q и пользуется большой популярностью среди решений для высокопроизводительных систем хранения. При этом полезная емкость уровневого тома равна емкости тома на уровне tier 1.

Второй режим принципиально новый. В нем горячие данные физически перемещаются с уровня tier 1 на уровень tier 0. Полезная емкость равна сумме полезных емкостей уровней tier 0 и tier 1. Такой режим возможен только на контроллерах Adaptec Series 8Q. Его преимущество перед SSD-кэшированием как раз и заключается в том, что полезная емкость дисков SSD добавляется к общей емкости. Недостаток заключается в том, что "горячие" данные находятся только на уровне tier 0, поэтому уровень tier 0 должен обладать высокой надежностью. Использование томов без избыточности, таких как JBOD и RAID0, на нем неприемлемо.

Заключение

Использование многоуровневых томов RAID, несмотря на их высокую эффективность для многих типов приложений, требует высокой степени проработки проектов, хорошего понимания принципов работы и в ряде случаев тщательнейшего тестирования.

В будущем такие схемы будут совершенствоваться. Как пример: количество уровней tier будет увеличиваться, алгоритмы переноса данных между уровнями tier будут усложняться.

О степени популярности подобных решений можно будет судить уже в самом ближайшем будущем.

Дмитрий Зотов,
компания PMC-Sierra,
подразделение Adaptec by PMC

Oracle: новые разработки

Октябрь 2014 г. – Важнейшими премьерами прошедшего Oracle OpenWorld стали технология Oracle Software in Silicon Cloud, новая флэш-система хранения данных Oracle FS1 Series, опция Oracle Database In-Memory Option для СУБД Oracle 12.7, Oracle Big Data SQL, Oracle Zero Data Loss Recovery Appliance и др.

Функции Software in Silicon в процессоре Oracle SPARC M7

Application Data Integrity — это первая в истории полная реализация на аппаратном уровне проверки обращений к оперативной памяти. Призванная помочь снизить риски работоспособности систем, вызванные ошибками в системе безопасности, такими как HeartBleed, она обеспечивает аппаратный мониторинг в реальном времени обращения к памяти программных процессов и пресекает неразрешенный доступ к памяти, связан ли он ошибкой программирования или с атаками, использующими переполнение буфера. Кроме того, она помогает ускорить разработку кода, а также обеспечивать качество, надежность и безопасность программного обеспечения.

Query Acceleration повышает производительность обработки запросов к базе данных в оперативной памяти, работая с потоком данных непосредственно из памяти через интерфейсы с пропускной способностью – до 160 Гбайт/с, что обеспечивает огромный прирост производительности. Для ускорения запросов в процессоре SPARC M7 реализовано несколько блоков.

Блоки декомпрессии в аппаратных ускорителях Software in Silicon значительно увеличивают эффективный объем доступной для использования памяти. Эти блоки выполняют декомпрессию данных на одном процессоре со скоростью, эквивалентной использованию 16 PCI-модулей декомпрессии или 60 процессорных ядер. Это позволяет хранить в оперативной памяти базы данных в сжатом виде, при этом обеспечивая доступ и выполнение операций с ними без потери производительности.

Средства ускорения запросов и декомпрессии могут сочетаться для повышения производительности и емкости, обеспечивая максимально эффективное использование ресурсов памяти, пропускной способности и процессорных ядер для революционного повышения производительности. Автоматический контроль целостности данных может работать постоянно для повышения надежности и безопасности.

После максимального ускорения и обеспечения безопасности работы приложений с использованием технологии Software in Silicon разработчики могут также улучшить процесс установки и развертывания своих программных продуктов, интегрируя, создавая и тестируя шаблоны виртуальной машины для Oracle Solaris.

Oracle SPARC M7 является первым процессором с революционной технологией Software in Silicon. Выпуск систем на базе процессора SPARC M7 запланирован на 2015 г.

Новая флэш-система хранения данных Oracle FS1 Series обладает возможностью масштабирования флэш-памяти до петабайтов. Новое ПО QoS Plus для управления системой обеспечивает оптимальный баланс между производительностью, затратами и преимуществами для бизнеса с помощью функций точного автоматического распределения данных по уровням хранения, безопасной мультиарендности для приложений и изоляции данных пользователей – и все это на единой масштабируемой платформе хранения данных.

Oracle Database In-Memory Option дает возможность работать с таблицей или с ее частью по прямым адресам в ОП. Это позволяет ускорять аналитические запросы от 100 до 3000 раз. Помимо этого, в 2 раза ускоряется обработка транзакций OLTP-приложений.

Oracle Big Data SQL – ПО, доступное только в составе Oracle Big Data Appliance. Эта разработка позволяет на основе SQL-запросов интегрировать данные в Hadoop-кластерах и на NoSQL-хранилищах с данными в Oracle Database.

Oracle Zero Data Loss Recovery Appliance (Recovery Appliance) значительно уменьшает влияние операций резервного копирования на производительность серверов и сетей, практически исключая необходимость приостанавливать или замедлять работу приложений на период резервного копирования. Облачная архитектура позволяет использовать Recovery Appliance для защиты тысяч баз данных, сокращая затраты и упрощая использование разрозненных систем резервного копирования.