Технический обзор

Функции RAID стека для динамического управления емкостью и производительностью

08 2016

Microsemi
Power Matters.™

1 Функции RAID Стека Для Динамического Управления Емкостью и Производительностью.

Функции RAID стека - Online Capacity Expansion (OCE) (динамическое расширение емкости RAID тома) и RAID Level Migration (RLM) (динамическое изменение уровня RAID для RAID тома) можно встретить в большинстве продуктов Microsemi, относящихся к системам хранения. Эти функции позволяют совершать динамические изменения в RAID томах без выключения системы или восстановления данных, что дает возможность динамически управлять емкостью и производительностью томов в случае возникновения потребности в более высоких значениях того и другого.

Важно заметить, однако, что современные системы хранения изменились очень значительно по сравнению с теми системами, во времена которых данные две функции появились в составе RAID стека. Поэтому, в ряде случаев использование OCE и RLM функций может являться нерациональным, а иногда и очень рискованным делом, по этой причине может быть предпочтительнее использовать другие опции (например, создание дополнительных томов или систем хранения).

Данная техническая статья обеспечивает обзор OCE и RLM функций и объясняет схемы их рационального использования.

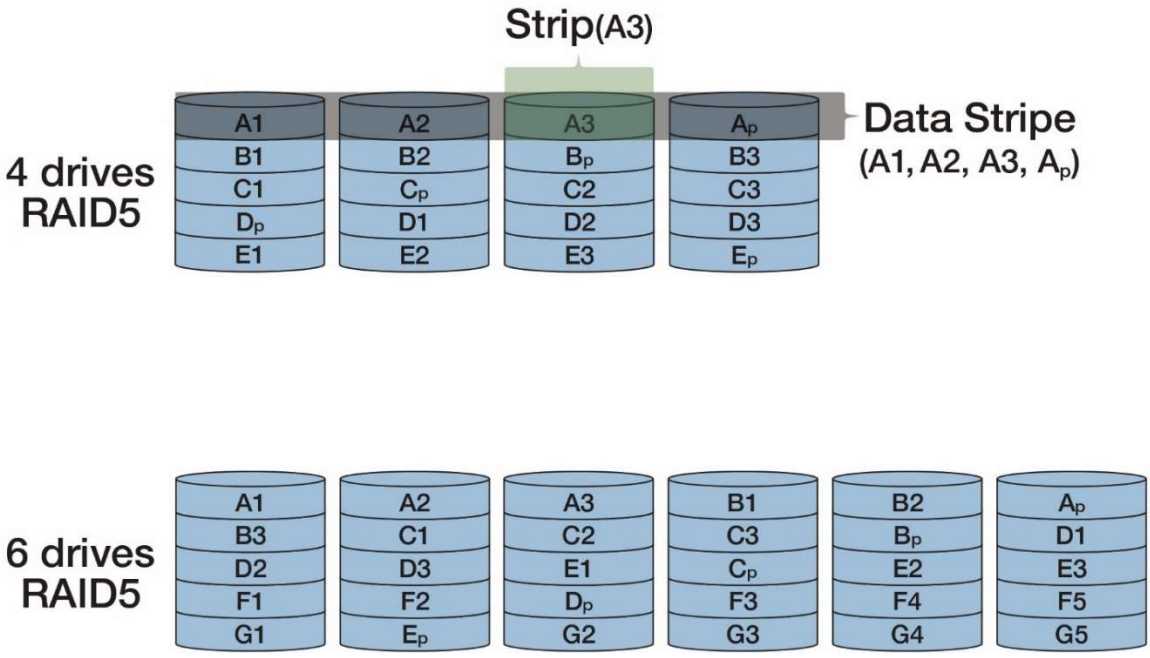
1.1 OCE и RLM

OCE предлагает пользователям при необходимости расширить емкость существующего RAID тома без резервного копирования и восстановления данных. Этот процесс работает в фоновом режиме. Он реорганизует существующие данные и создает избыточную информацию (например, Parity), а система хранения, точнее данный RAID том, в это же время, по-прежнему может отрабатывать запросы пользователей.
Исторически, когда RAID тома эксплуатировались без данной функции, замена дисков в томе на диски с большей емкостью или создание RAID тома с дополнительными дисками вынуждали пользователей к созданию резервной копии, удалению старого тома, созданию нового с нужными свойствами и переноса данных из резервной копии на новый том. Такой подход требовал перевода сервера и системы хранения в режим “offline”, пока создавался новый RAID том и данные резервировались и восстанавливались из копии.

Такой подход не только вызывает остановку работы с системой хранения, но и риск потери данных, если что-то пойдет не так с резервной копией. До появления функции OCE другой способ состоял в том, чтобы добавить дополнительную емкость к RAID тому, что увеличивает стоимость и, кроме этого, крайне сложно предсказать, какая емкость будет нужна.
Существует два метода OCE, которые могут быть использованы, чтобы расширить емкость существующего RAID тома. Первый метод – заменить каждый диск в составе тома на диск с большей емкостью, и когда все устройства будут иметь более высокую емкость, использовать появившуюся дополнительную емкость для расширения существующего тома или создания в этой области другого тома, если это позволяет RAID стек. Второй метод – физически добавить еще дисков к существующему тому и произвести размещение данных в структуре RAID тома так, чтобы в состав тома вошли эти новые диски.

Данные рисунки показывают пример работы OCE. В этом примере все bx, cx и dx блоки данных были по-новому размещены на дисках при расширении с 4 до 6 дисков. Новая область для новых данных появляется на RAID томе из 6 дисков.

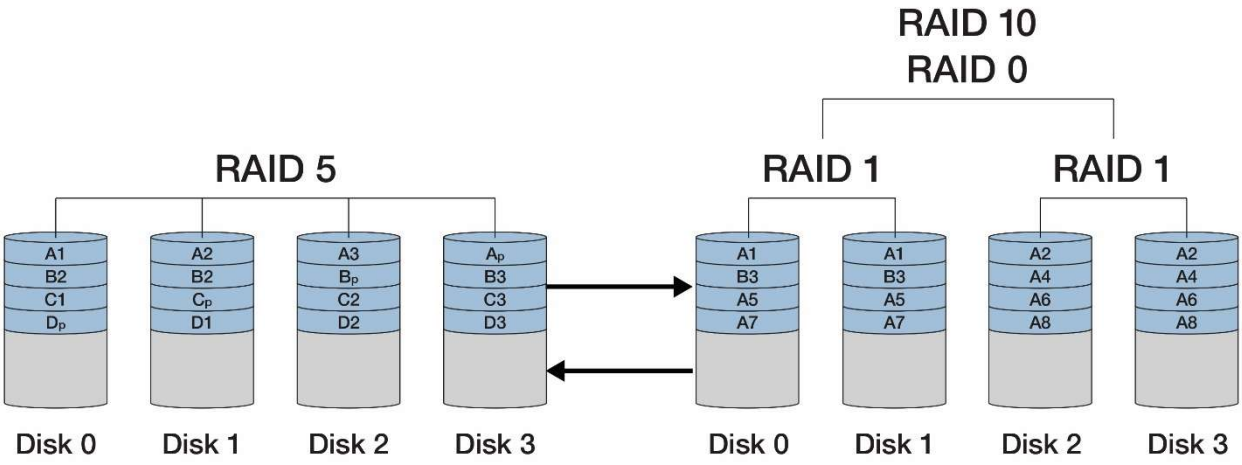
Рисунок 1. Пример использования OCE



RLM позволяет пользователю изменить уровень RAID на существующем наборе дисков (например, перейти от использования RAID5 уровня к RAID10 на каком-то одном RAID томе). Изменение размера блоков данных, которые распределены по дискам в составе RAID тома, также рассматривается, как подфункция или часть RLM функции. RLM обычно используется, когда существующий RAID уровень не обеспечивает тех характеристик производительности, которые требуются, или другой уровень RAID более предпочтительный с точки зрения стоимости за Gb. Основное преимущество данной функции заключается в том, что уровень RAID или размер блока можно изменить без потери доступа к системе хранения и пользовательских данных.

Следующий рисунок показывает, как используется RLM для перехода с RAID5 уровня на RAID10.

Рисунок 2. RAID5 и RAID10



1.2 Проблемы При Использовании OCE и RLM

RLM и OCE выполняются в фоновом режиме, и, таким образом, они могут быть использованы в том случае, пока система находится в режиме online. Однако, эти операции обычно требуют перемещения всех блоков данных и, поскольку современные диски могут быть очень большого размера (ТБ), это требует перемещения, копирования и перестановки очень большого количества информации и блоков данных.

В то время, когда данные функции были разработаны и внедрены, размер дисков enterprise класса был в пределах нескольких десятков ГБ, а размер near-line дисков в пределах нескольких сотен ГБ. При таких значениях емкости для дисков со скоростями от 10 kRPM до 15 kRPM enterprise класса и дисков 7.2 kRPM near-line класса начальный RLM или OCE процессы завершались бы в течении нескольких часов (если система хранения не была нагружена пользовательскими запросами), а в общем – начиная, с примерно, нескольких часов (для абсолютно ненагруженных пользовательской работой дисков) до нескольких дней максимум (наихудший сценарий, когда система хранения находится под максимальной нагрузкой). Исходя из вышесказанного, существовали рекомендации запускать RLM или OCE процесс в то время, когда диски не сильно загружены или совсем не загружены (например, ночью), чтобы уменьшить влияние производительности на пользователей и приложения.

На сегодняшний день скорость вращения и, как следствие, производительность таких устройств не изменилась значительно (для задач RLM и OCE), но зато емкость выросла больше чем в десятки раз. Для современных дисков с емкостью в несколько ТБ, при использовании RLM и OCE функций, весь процесс может занять от нескольких дней для томов без пользовательской нагрузки до десятков недель, чтобы достичь завершения при наличии нагрузки. Последние технологические решения, такие как представление SMR (Shingled Magnetic Recording), еще более замедляет перемещение данных, поскольку такие диски не поддерживают блочную запись (для подобных операций используется контейнер гораздо большего размера, называемый chunk, что очень сильно сказывается на производительности).

Не только значительное увеличение емкости дисков приносит проблемы для процессов RLM и OCE. Кроме этого, стоит обратить внимание на количество данных, которые должны быть прочитаны и записаны. Процесс, который перемещает данные, должен тщательно мониториться, поскольку потеря питания или другие проблемы (бэд блоки или выход дисков из строя) должны отрабатываться без потери данных и доступа. Для выполнения такого мониторинга необходимо отслеживание транзакции перемещения данных. Процесс мониторинга влияет и на само перемещение данных и на общую производительность, когда выполняется параллельно с пользовательскими операциями. Для управления общей производительностью в настройкак контроллера есть опция, которая позволяет управлять приоритетом фоновой задачи, чтобы уменьшить ее влияние на производительность, но это заметно увеличивает время выполнения самих RLM и OCE процессов.

Поскольку RLM и OCE процессы пытаются минимизировать влияние на время пользовательских транзакций и производительность RAID тома, их рекомендуется запускать с низшим (low) приоритетом для фонового режима. Для современных жестких дисков с большой емкостью этот процесс не сможет быть завершен за типичное время неактивности (например, выходные дни или ночь). Кроме этого, многие приложения и шаблоны трафика не имеют времени с низкой или нулевой активностью пользователей. Как следствие, RLM и OCE процессы становятся чрезвычайно длинными по времени и приносят очень высокий риск, связанный с потерей данных. Это должно приниматься во внимание, когда создаются сценарии использования подобных функций.

1.3 Оптимальное Использование OCE и RLM

Устройства хранения на основе flash памяти создают новый уровень (tier) в системах хранения с различными шаблонами использования, с меньшим влиянием на производительность для операций случайного доступа. Как результат, такие устройства на сегодняшний день гораздо лучше подходят для RLM и OCE процессов. В то же время, данные устройства имеют более высокую цену из расчета отношения доллар/GB, что в конечном счете делает функции типа OCE более ценными, чем раньше, поскольку пользователи теперь могут использовать тома по принципу“ растем по мере надобности” и добавлять более дорогие SSD диски только тогда, когда в этом возникает необходимость. Поскольку устройства хранения, основанные на flash памяти становятся более и более эффективными с точки зрения цены с каждым новым поколением, возможность добавить дополнительную емкость позже экономит деньги.

Ситуация противоположна для HDD. Эти устройства лучше оптимизированы для систем с большой емкостью и, таким образом, дешевле изначально добавить дополнительную емкость под будущие потребности.

1.4 Заключение

Вместо того, чтобы использовать устройства терабайтной емкости, которые имеют производительность с низкой выборкой, мы советуем реализовывать системы хранения на SSD дисках. Для наиболее оптимальных результатов OCE и RLM функции не должны использоваться с HHD дисками. Таким образом, тестирование RLM и OCE функций в сегодняшней продуктовой линейке контроллеров Microsemi большей частью фокусируется на SSD дисках, нежели чем на HDD.

Пользователи должны следовать рекомендациям, представленным в этом техническом руководстве, и выбирать уровни (tiers) хранения так, чтобы избегать использования RLM и OCE на втором уровне (tier), где используются жесткие диски (HDDs). Функции RLM и OCE наиболее эффективны для систем хранения enterprise класса, которые обеспечивают высокую производительность и надежность. Такие высокопроизводительные и надежные продукты обеспечивают минимальное влияние на производительность для пользователей и приложений, и продолжительность процесса перемещения данных, используемых данными функциями, завершается за приемлемый промежуток времени. Это обеспечивает рациональное размещение систем хранения на площадках заказчиков данных решений.

Современные условия для систем хранения данных являются более сложными, чем во времена, когда были представлены функции OCE и RLM. Хотя RLM и OCE процессы могут занимать очень значительное время (это время больше не вмещается в традиционные “окна” обслуживания систем хранения из-за очень большой емкости) и не могут быть инвертированы или оставновлены, данные функции могут быть крайне удобны в ряде ситуаций. Обращайтесь в компанию Microsemi, чтобы убедиться в правильности выбора технических решений и получить консультации относительно Ваших проектов.

Microsemi Corporate главный офис:
One Enterprise, Aliso Viejo,
CA 92656 USA.
Для звонков внутри США:
+1 (800) 713-4113.
За пределами США: +1 (949) 380-6100
Факс: +1 (949) 215-4996.
Email: **sales.support@microsemi.com.**
www.microsemi.com

О компании Microsemi

Microsemi Corporation (Nasdaq: MSCC) предлагает полный набор полупроводниковых и системных решений для аэрокосмической и оборонной отраслей, телекоммуникаций, центров обработки данных и промышленных рынков. Компания предлагает следующие продукты: высокопроизводительные радиационно-устойчивые комбинированные интегральные схемы; программируемые логические интегральные схемы; однокристальные схемы; специализированные заказные интегральные схемы; системы управления электропитанием; устройства для хронометража и синхронизации; системы точного времени, задающие мировой стандарт времени; устройства для обработки голоса; радиочастотные системы; дискретные элементы; системы хранения и связи корпоративного уровня; технологии безопасности и масштабируемые противовзломные системы; решения Ethernet; интегральные схемы и промежуточные устройства с питанием через Ethernet; а также услуги индивидуального проектирования. Главный офис компании Microsemi расположен в городе Алисо-Вьехо (штат Калифорния, США). В подразделениях компании во всем мире работают около 4 800 сотрудников. Подробнее на сайте www.microsemi.com.

ESC-2161475