NETAPP TECHNICAL REPORT

RAID-DP: Реализация NetApp схемы Double-Parity RAID для защиты данных

Jay White, Chris Lueth, и Jonathan Bell, NetApp

Март 2013 | TR-3298

Коротко о главном:

Этот документ предлагает углубленный обзор реализации RAID «с двойной контрольной суммой» – NetApp® RAID-DP®. Рассмотрены несколько аспектов устройства RAID-DP, включая принципы его работы, уровни надежности в сравнении с другими типами RAID, производительность и использование емкости дисков.

Оглавление

1 Введение

NetApp представила свою реализацию принципа «RAID с двумя дисками четности» под именем RAID-DP в 2003 году, когда она появилась в OS Data ONTAP® 6.5. С тех пор этот тип RAID стал типом RAID по умолчанию для всех систем хранения NetApp. Этот документ предлагает вашему вниманию обзор технологий RAID-DP, и то, как они существенно улучшают устойчивость к сбоям при различных сценариях отказа дисков. Другие затронутые области использования включают в себя ответы на вопросы, как дорого обходится RAID-DP (бесплатен), каковы специальные требования к оборудованию (отсутствуют), как осуществляется преобразование существующего тома RAID 4 в RAID-DP (просто, в обе стороны), и каково его влияние на показатели производительности. Этот документ также представляет сценарий восстановления после отказа двух дисков, и показывает, как RAID-DP позволяет сохранить доступность данных в случае отказа двух дисков одновременно.

2 Зачем нужен RAID-DP?

Как уже говорилось ранее, традиционный «single-parity RAID» (RAID типов 3, 4 и 5) предлагает удовлетворительную защиту от одного дискового отказа, это может быть полный отказ одного диска целиком, или ошибка считывания, так называемая «битовая ошибка». В обоих случаях считываемые данные восстанавливаются с использованием избыточности данных на сохранившихся дисках данных, плюс диске «контрольных сумм», *parity*. Если отказ это ошибка чтения, то воссоздание данных это почти моментальная процедура, и дисковый массив или том данных остается доступным. Однако, если проблема это отказ диска целиком, когда все данные, на нем хранившиеся, потеряны, все эти данные потребуется восстановить, и дисковый массив или том остается на длительное время в уязвимом состоянии *degraded mode* до тех пор, пока данные потерянного диска не будут реконструированы целиком на резервный диск. Именно существование *degraded mode* во время реконструкции данных RAID показывает, что классический *single-parity RAID* и его возможности по защите данных на сегодня не удовлетворяют возросшим требованиям защиты данных на современных дисках.

2.1 Как современные диски большой емкости работают в RAID

Современные диски продолжают развиваться, как и другие компьютерные технологии. Жесткие диски стали на порядки емче, чем во времена, когда появилась технология RAID. Несмотря на то, что жесткие диски становятся все емче, их надежность не растет пропорционально увеличению их емкости, и, что даже более важно, уровень так называемых «битовых ошибок» диска, увеличивается пропорционально увеличению их емкости. Эти три фактора — увеличение емкости отдельного диска, не улучшающаяся пропорционально росту их емкости надежность, и увеличивающийся уровень битовых ошибок на больших объемах современных дисков — все это ведет к серьезному несоответствию для классических структур «RAID с четностью» с использованием одного диска четности, требуемому пользователем уровню защиты данных на дисках.

Поскольку сегодня жесткие диски отказывают также как и во времена появления технологии RAID, эти технологии сегодня столь же востребованы, как и тогда. При отказе одного диска, RAID воссоздает потерянные данные за счет избыточности информации на оставшихся дисках и данных контрольной суммы на диске parity, и записывает ее на диск из резерва, так называемый *spare disk*. Однако, с тех времен, когда RAID впервые появился, увеличившийся объем дисков сильно удлинил процесс реконструкции RAID. При той же скорости вращения, время, необходимое на

восстановление потерянных данных, хранимых на диске емкостью 2TB, значительно больше, чем для диска 600GB и менее. Сложности с увеличенным временем реконструкции усугубляются еще и тем фактом, что все чаще используемые для хранения данных диски с интерфейсом SATA не только более медленные в процессе считывания-записи данных, но и несколько менее надежные, чем диски FC или SAS.

2.2 Защита данных в схеме RAID с одним parity disk на больших дисках

Существуют различные способы улучшить ситуацию с защитой данных на схеме RAID с одним *parity disk*, по мере того, как емкость дисков растет, но все они не лишены недостатков. Первый вариант, это продолжать покупать и использовать в системах хранения диски минимально доступной на рынке емкости, что позволит сохранять меньшее время восстановления после сбоя. Однако такой поход не практичен с любой точки зрения. Плотность хранения это критически важная величина в любом датацентре, и использование дисков меньшей емкости ведет к снижению объемов хранения на квадратный метр площади датацентра, и удорожанию эксплуатации системы хранения. Кроме того, производители систем хранения вынуждены использовать те модели дисков, которые им готовы поставлять производители в достаточных количествах, то есть наиболее массовые в производстве, а диски малой емкости часто могут быть просто недоступны в нужных количествах.

Второй способ использовать диски большой емкости в традиционном *single-parity RAID* чуть более практичен, но, с появлением double-parity RAID и, в частности, RAID-DP, также малопривлекателен по ряду причин. Это способ делать RAID-группы из сравнительно небольшого числа таких емких дисков, что снижает время реконструкции RAID. Продолжая аналогию с тем, как диски большого объема требуют больше времени на проведение реконструкции RAID, чем меньшие по объему диски, так и RAID-массив, состоящий из большого числа дисков, восстанавливается после сбоя дольше, чем состоящий из меньшего их числа. Однако RAID из небольшого числа дисков имеет существенный недостаток: в случае RAID-групп из малого числа дисков растет «оверхед» для контроллера и уменьшается *usable space*.

Наиболее надежной защитой до появления схемы *dual parity* обладал тип RAID 1, или «зеркалирование». В схеме RAID 1, процесс зеркалирования диска реплицировал точную копию всех данных на диске на другой диск. Хотя RAID 1 обеспечивал максимально возможный на тот момент уровень зашиты данных от дисковой ошибки, стоимость его использования была высока, так как требовалось в два раза больше дискового пространства для хранения того же объема данных. Как уже было отмечено раньше, использование RAID-групп меньшего размера, для обеспечения улучшенной отказоустойчивости, ведет к увеличению стоимости владения, так как уменьшает полезную емкость на вложенный доллар. С этой точки зрения, RAID 1 с его неприятными требованиями удвоить емкость массива для хранения того же объема данных, является наиболее дорогостоящим решением отказоустойчивого хранения, с наивысшей стоимостью владения.

2.3 Защита данных при использовании RAID-DP

Вскоре, с широким распространением в индустрии хранения дисков большого объема, создающих значительные проблемы для защиты хранимых на них данных, пользователи и аналитики стали нуждаться в новом решении для улучшенной надежности хранения на дисковых массивах. Для удовлетворения этих требований, NetApp выпустил в свет решение с новым типом RAID, под названием RAID-DP. RAID-DP означает *RAID with Double Parity*, и существенно увеличил отказоустойчивость при выходе из строя диска или ошибки чтения, в сравнении с традиционным

single-parity RAID. Если взять все соответствующие показатели типичного времени наработки на отказ для дисков, посчитать их по формуле для RAID-DP, и сравнить результаты с традиционным *single-parity RAID*, то вы увидите, что RAID-DP является в тысячи раз более надежным, на тех же самых дисках. С таким уровнем надежности, RAID-DP превосходит по отказоустойчивости даже RAID 10, но при этом за цену RAID 4. RAID-DP предлагает бизнесу наилучший уровень стоимости владения, без необходимости рисковать для получения этого результата сохранностью данных.

3 Как работает RAID-DP

Традиционные уровни существующих технологий RAID предлагают защиту данных несколькими способами. Вариант RAID, используемый NetApp, это вариант RAID 4, Данные сохраняются в «горизонтальных строках» блоков, данные контрольной суммы вычисляются для блоков, входящих в эту «строку», и сохраняются на специально выделенном для хранения этих данных диске. Однако, основной недостаток различных типов RAID с *single parity*, включая и модифицированный NetApp RAID 4, заключается в том, что они имеют возможность защитить данные только от одного дискового сбоя в группе дисков RAID.

3.1 RAID-DP с двумя parity-дисками

Что представляет собой RAID-DP и как в нем используется *double parity*?

На самом базовом уровне, RAID-DP добавляет второй диск *parity* в каждую RAID-группу, используемую *aggregate* или *traditional volume*. RAID-группа это структура, поверх которой находится *aggregate* и/или так называемый *traditional volume*. Каждая группа *NetApp RAID 4* имеет определенное количество дисков данных и один диск *parity*, а *aggregate* и/или *traditional volume* включают в себя одну или несколько таких групп. Если *parity disk* в дисковой группе RAID 4 хранит на себе данные «контрольной суммы», посчитанной «по горизонтали», проходящей по всем дискам с данными в группе RAID 4, дополнительный диск RAID-DP parity хранит на себе данные, рассчитанные для так называемой *«diagonal parity»*, по всем дискам группы RAID-DP. При использовании избыточных данных в этих двух страйпах *parity* в RAID-DP — одна традиционная, «горизонтальная», и одна «диагональная» — защита данных сохраняется даже в случае двойного отказа дисков в одной RAID-группе.

3.2 Пример работы RAID-DP

При использовании RAID-DP, традиционная, «горизонтальная» parity из RAID 4 также используется и становится частью механизма RAID-DP. Иначе говоря, то, как работает механизм parity в RAID 4, не изменяется с переходом системы хранения на RAID-DP. Процесс, при котором вычисляется контрольная сумма для «горизонтальной строки», по-прежнему работает в RAID-DP и вычисляет одну из двух контрольных сумм. На деле, если происходит одиночный отказ диска, или происходит ошибка чтения в результате *bad block* или битовой ошибки, то используется «горизонтальная» контрольная сумма RAID 4, и только с ее помощью восстанавливаются потерянные в этом случае данные, без использования дополнительных средств RAID-DP. В этом случае, компонент «diagonal parity» из RAID-DP это просто дополнительная защитная обертка вокруг «горизонтальной» parity.

RAID 4 с «горизонтальной» контрольной суммой

Рисунок 1 показывает пример, как вычисляется «горизонтальная» parity в традиционном NetApp RAID 4. Это первый шаг, необходимый для понимания работы механизма RAID-DP и *double parity*.

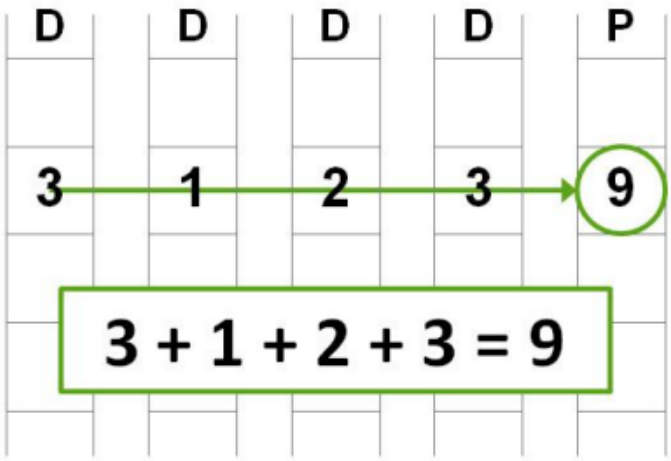

Рис. 1) Простая контрольная сумма в RAID 4.

На рисунке вы видите группу дисков в RAID 4, состоящую из четырех дисков данных (первые четыре колонки, помеченные «D») и одного диска контрольной суммы (правая колонка, помеченная «P»). Строки на рисунке 1 представляют стандартные блоки 4KB, используемые в традиционной реализации NetApp RAID 4 . В рассматриваемом случае наша «контрольная сумма» вычисляется путем сложения всех значений в каждом из блоков по горизонтали между собой, и записи получившейся суммы в блок parity (3 + 1 + 2 + 3 = 9). На практике, в реальной системе, контрольная сумма вычисляется с помощью операции «Исключающее ИЛИ» (exclusive OR, XOR), но для иллюстрации, мы воспользуемся простым сложением. Если нам понадобится восстановить данные, потерянные в результате дискового сбоя, мы выполняем эти операции в обратном порядке. Например, если первый диск выйдет из строя, то RAID 4 восстановит хранившиеся на нем данные (3, в первой колонке рисунка 1), вычтя значения блоков оставшихся дисков из значения «контрольной суммы» (9 – 3 – 2 – 1 = 3). Этот пример восстановления данных помогает понять, каким образом контрольная сумма RAID типа *single parity* защищает от сбоя диска, но только одного.

Добавление страйпов diagonal parity в RAID-DP

Рисунок 2 добавляет в структуру одну диагональную полосу-страйп, отмеченную блоками синего цвета, и второй диск для контрольных сумм, помеченный «DP», в существующую структуру группы RAID 4, описанную ранее. Это показывает, что структура RAID-DP является надстройкой над лежащей «ниже» структурой RAID 4 с «горизонтальной» контрольной суммой.

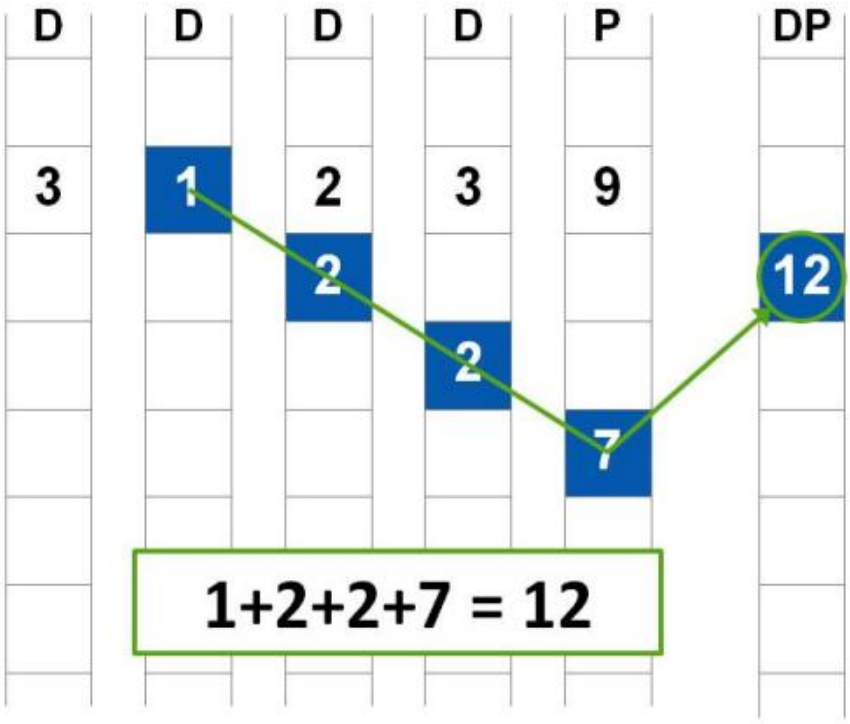

Рис. 2) Добавление *diagonal parity*.

Диагональный страйп в реальной системе вычисляется с использованием операции XOR, но, как и в примере выше, для упрощения мы на иллюстрации используем обычное суммирование, (1 + 2 + 2 + 7 = 12).

Одним из наиболее важных моментов является то, что диагональный страйп включает в себя и блок обычной, «горизонтальной» контрольной суммы. RAID-DP включает в себя все диски оригинальной структуры RAID 4, как диски данных, так и диск, хранящий данные *horizontal parity* RAID-4. Рисунок 3 дополняет картину остальными блоками на дисках данных, плюс страйпами горизонтальной и диагональной контрольной суммы.

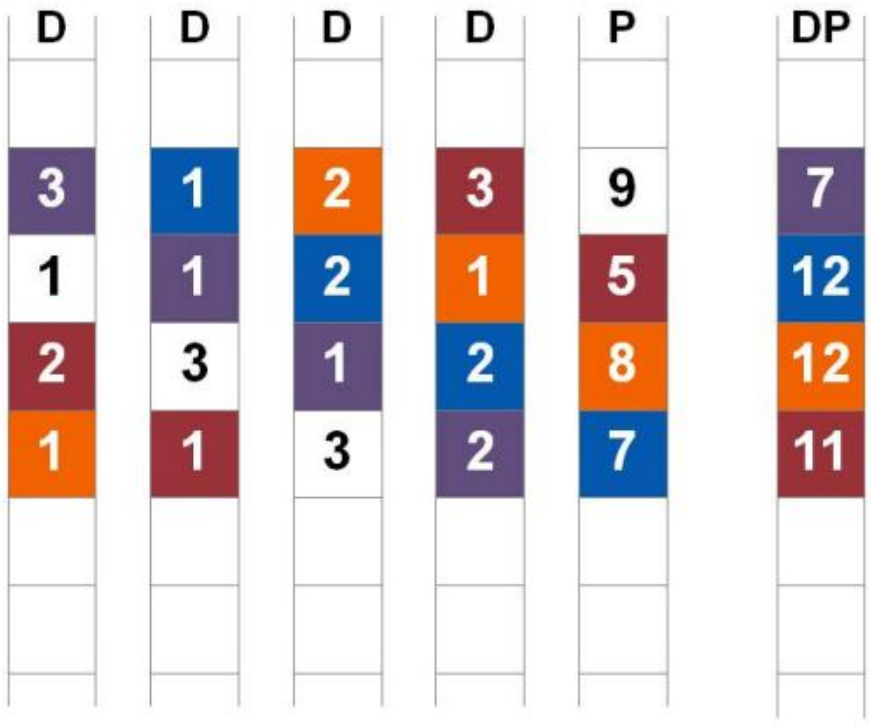

Рис. 3) Пример вычисления *diagonal parity*.

Один из аспектов устройства RAID-DP, показанный на рисунке 3, это то, что диагональный страйп «оборачивается» вокруг края «горизонтального» страйпа контрольной суммы RAID-4. Два важных условия, определяющих возможность RAID-DP восстановить данные после двойного дискового сбоя, могут быть не вполне ясны из приведенного примера. Первое условие заключается в том, что каждый диагональный страйп не включает в себя один, и только один диск группы, но каждый последующий страйп не включает в себя при этом разные диски. Это ведет нас ко второму условию, что один диагональный страйп не получает сгенерированного блока контрольной суммы. В данном примере это страйп из белых блоков. В приведенном далее примере работы реконструкции RAID станет очевидно, что отсутствие одного полного диагонального страйпа не влияет на способность RAID-DP восстановить данные при двойном отказе дисков группы.

Важно отметить, что устройство RAID-DP diagonal parity, рассмотренное в этом примере, в реальных системах хранения работает на десятках дисков в одной RAID-группе и на миллионах «горизонтальных строк» блоков данных, по всей ширине группы RAID 4. Хотя проиллюстрировать работу алгоритмов RAID-DP проще на небольшой картинке, восстановление RAID-группы гораздо большего размера работает точно также, как показано на этом небольшом примере, вне зависимости от числа дисков в RAID-группе.

Доказательство того, что RAID-DP в самом деле восстанавливает данные после одновременного отказа двух дисков может быть проведено двумя путями. Один – используя математические теоремы и доказательства, и другой – пройдя на примере все процедуры, которые проводит система для восстановления данных после сбоя. В этом документе мы воспользуемся вторым способом показать, как работает защита данных RAID-DP. Для углубленного рассмотрения математического аппарата, стоящего за механизмами RAID-DP, смотрите работу Row-Diagonal Parity for Double Disk Failure Correction, размещенную на сайте USENIX Organization.

Процесс реконструкции данных в RAID-DP

Возьмем рисунок, приведенный выше, и представим себе, что в иллюстрируемой структуре вышло из строя сразу два диска. Это будет означать, что все данные блоков в первых двух колонках слева – потеряны, как изображено на рисунке 4.

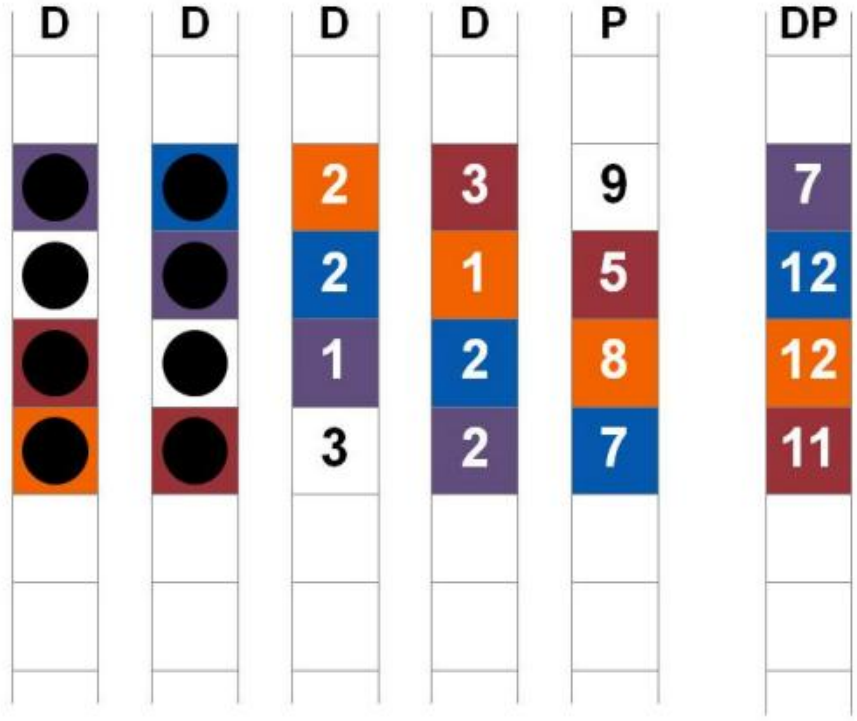

Рис. 4) Два отказавших диска.

Когда возникает событие двойного дискового отказа, RAID-DP первым делом начинает искать цепочку, с которой она начнет реконструкцию. В нашем случае, допустим, это будет диагональный страйп, блоки которого помечены на рисунке голубым цветом. Помните, что восстановление данных после одиночного отказа диска в RAID-4 возможно только в том случае, если потерян не более, чем один элемент группы. Помня об этом проследим, что для голубого диагонального страйпа на рисунке 4 отсутствует всего один из пяти имеющихся блоков. Имея четыре из пяти доступных элементов, RAID-DP имеет достаточно данных для реконструкции утраченных данных в голубом блоке. Рисунок 5 показывает то, как эти данные восстанавливаются на доступном системе *hot spare disk* .

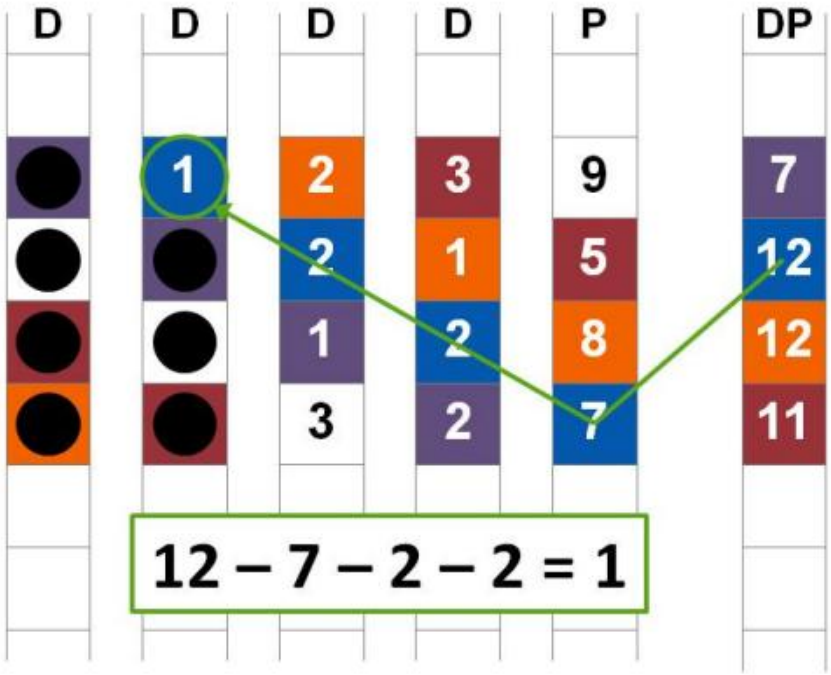

Рис. 5) Восстановление значения «голубого» блока с помощью данных диагонали блоков.

Данные голубого блока восстановлены с помощью простой арифметической процедуры со значениями диагональной строки блоков, рассмотренной ранее (12 – 7 – 2 - 2 = 1). Итак, одна голубая диагональ, и значение входившего в нее потерянного блока - восстановлены, процесс восстановления данных теперь берет горизонтальную строку. Так как благодаря восстановлению данных в диагональной строке, мы восстановили один из блоков, появилась возможность восстановить соседний по горизонтали блок с помощью простой «горизонтальной» контрольной суммы (9 – 3 – 2 – 1 = 3). Результат вы видите на рисунке 6.

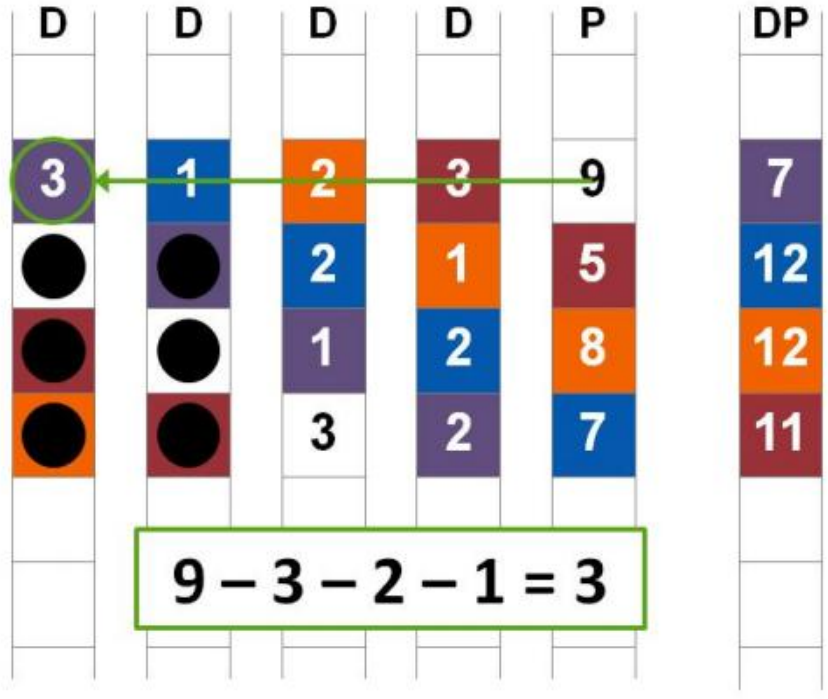

Рис. 6) Восстановление потерянного блока с помощью данных строки блоков.

RAID-DP теперь может воспользоваться восстановленными данными для того, чтобы восстановить блок, входящий в следующий диагональный страйп. Начиная с верхнего левого блока, воссозданного из «горизонтальной» контрольной суммы, RAID-DP восстановит значение потерянного «фиолетового» блока в диагональном страйпе, как показано на рисунке 7.

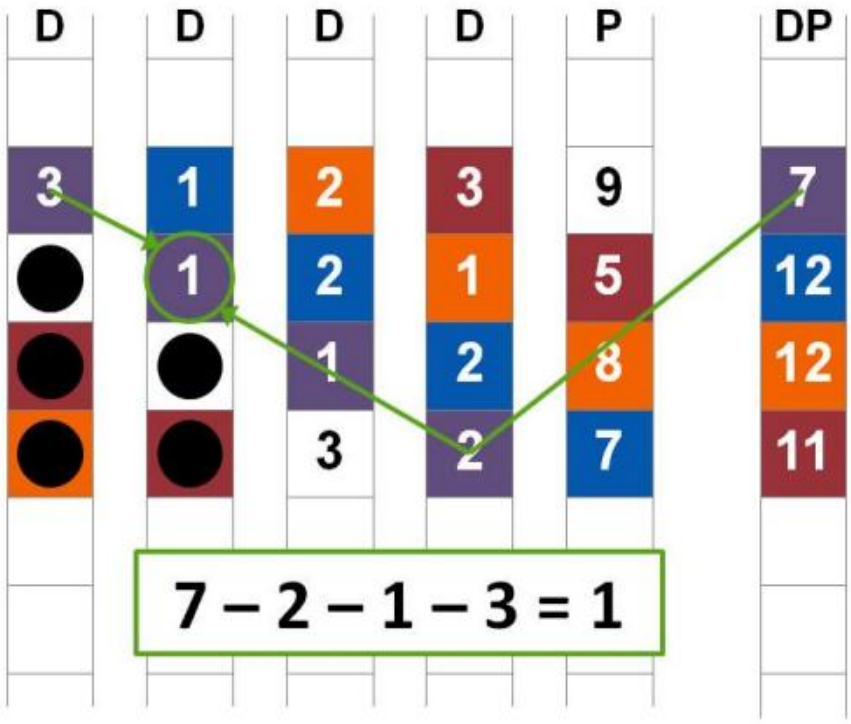

Рис. 7) Восстановление «фиолетового» блока с помощью данных диагонали блоков.

И вновь, как только RAID-DP восстановил нам блок «диагональной» контрольной суммы, появляется достаточно информации для восстановления значения следующего блока из данных «горизонтальной» контрольной суммы в первой колонке, как показано на рисунке 8.

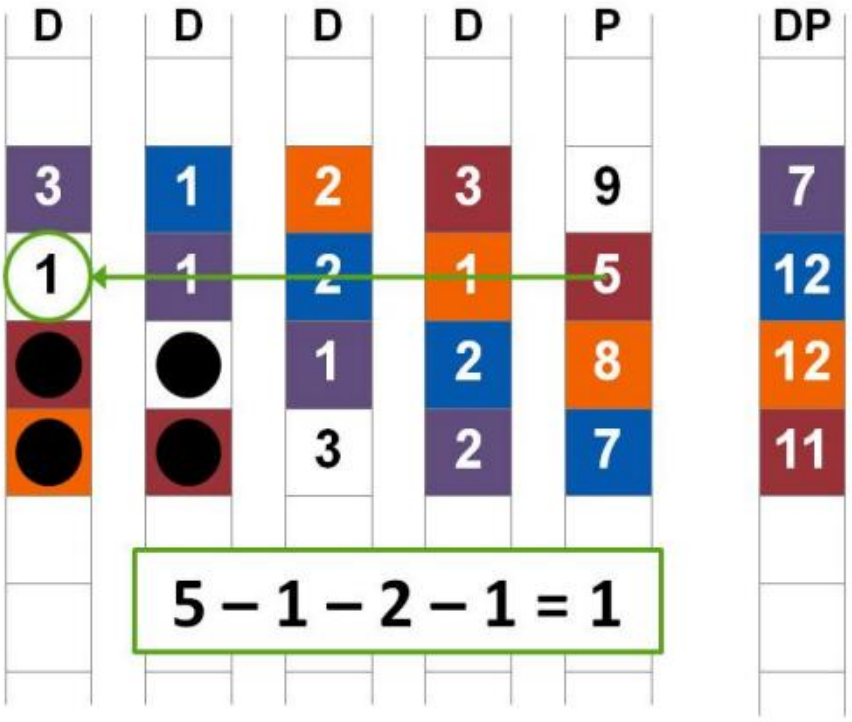

Рис. 8) Восстановление потерянного блока с помощью данных строки блоков.

Как мы уже отметили ранее, в нашем примере белый диагональный страйп неполон, в нем нет необходимого диагонального блока parity для восстановления соответствующего блока данных в нем. Для восстановления RAID-DP просматривает оставшиеся блоки, и находит, что может восстановить блок в «оранжевом» страйпе, как показано на рисунке 9.

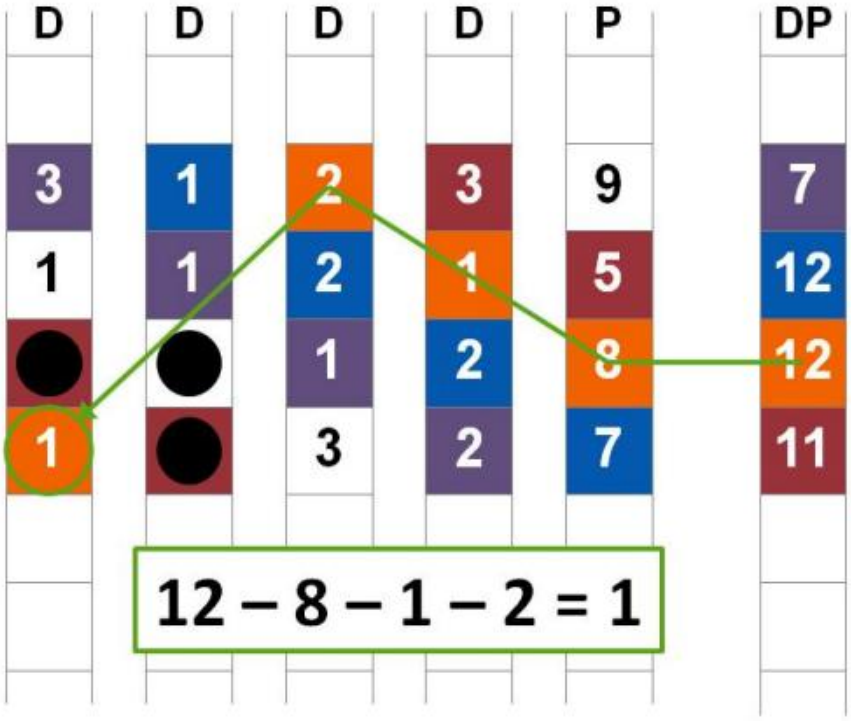

Рис. 9) Восстановление «оранжевого» блока с помощью данных диагонали блоков.

После того, как RAID-DP восстановил необходимый диагональный блок, процесс вновь переключается на горизонтальный ряд. Как уже рассмотрено на примерах выше, повторяется операция восстановления блока по данным горизонтального ряда блоков, показанного на рисунке 10.

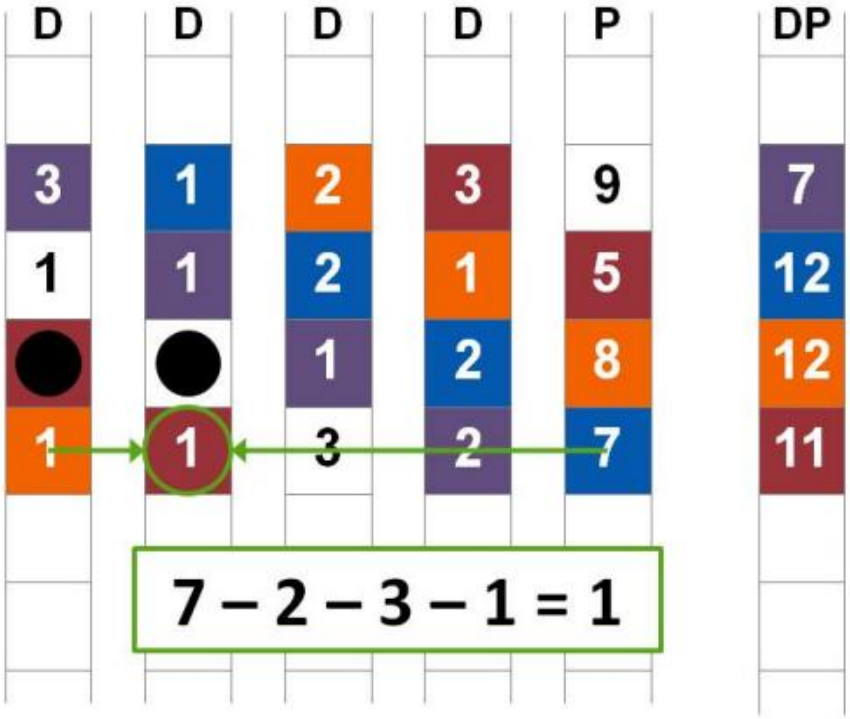

Рис. 10) Восстановление содержимого блока с помощью данных строки блоков.

Получив данные в «красном» восстановленном блоке, реконструкция RAID использует его вновь, для восстановления данных в отсутствующем блоке красного диагонального страйпа, что показано на рисунке 11.

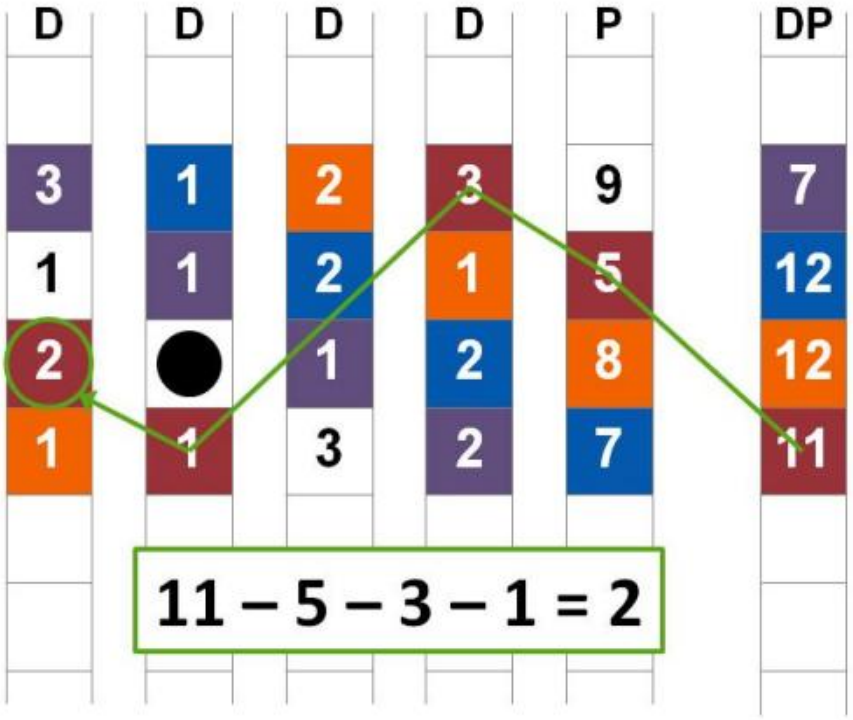

Рис. 11) Восстановление «красного» блока с помощью данных диагонали блоков.

И наконец, теперь у нас есть остаточно информации, чтобы восстановить блок данных в неполном «белом» страйпе, который мы восстановим с использованием данных «горизонтальной» контрольной суммы. На финальной схеме рисунка 12 мы видим, что все потерянные в результате сбоя двух дисков блоки успешно восстановлены с использованием технологии RAID-DP.

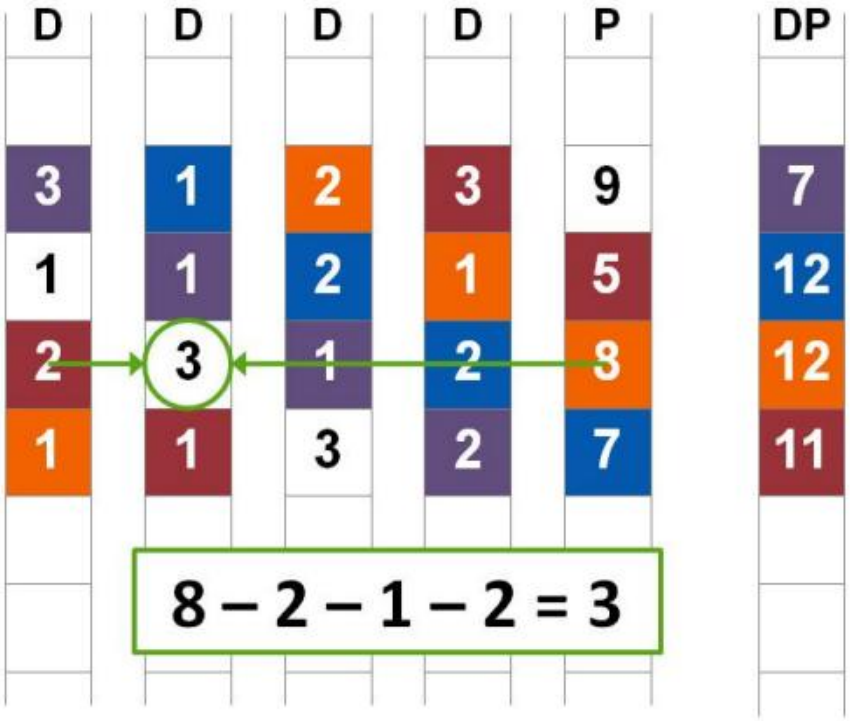

Рис. 12) Восстановление потерянного блока с помощью данных строки блоков.

Особенности работы RAID-DP

Рассмотренный выше пример восстановления данных на RAID-DP выглядит довольно длинным и сложным. Но следует упомянуть еще несколько особенностей работы RAID-DP, которые не являются очевидными и требуют отдельного уточнения. В случае двойного отказа дисков, RAID-DP автоматически поднимает приоритет процесса реконструкции, поэтому процесс восстановления RAID выполняется быстрее. В результате, время реконструкции RAID с двумя отказавшими дисками может быть даже меньше, чем для одного отказавшего диска. Вторая существенная особенность RAID-DP при двойном отказе диска заключается в том, что когда второй диск отказывает неодновременно с первым, то существенная часть информации к тому времени уже бывает восстановлена традиционным способом, с помощью простой «горизонтальной» parity. RAID-DP автоматически учитывает это, начиная работать только в том месте массива, где утрачены сразу два элемента группы.

4 Обзор RAID-DP

RAID-DP доступен на системах хранения NetApp без дополнительных затрат или особых аппаратных требований. Некоторые дополнительные сведения, не рассмотренные выше, приведены в этой главе.

4.1 Уровень защиты при использовании RAID-DP

На низшем уровне, RAID-DP обеспечивает защиту против отказа двух дисков в одной группе RAID одновременно, или от отказа одного диска и несчитавшегося блока или битовой ошибки до того, как реконструкция RAID завершилась. Наивысший уровень защиты будет достигнут при использовании RAID-DP совместно с технологией SyncMirror®, при этом защита обеспечивается от одновременного отказа пяти дисков в группе, или четырех и несчитавшегося блока и/или битовой ошибки в процессе реконструкции RAID.

4.2 Управление томами с RAID-DP

С позиции управления, когда aggregate создан (или сконвертирован) с использованием RAID-DP, он работает также как и в случае использования RAID 4. Все рекомендации, практики и руководства идентичны, вне зависимости от того, используется RAID 4 или RAID-DP, таким образом, для типовых процедур управления администраторам NetApp не требуется каких-либо изменений. Даже если контроллер системы хранения работает с миксом из aggregates или томов традиционного типа, сделанных поверх RAID 4 и RAID-DP, все операции при этом остаются одними и теми же.

4.3 Производительность RAID-DP

Производительность томов, использующих RAID-DP, сравнима с таковой при использовании RAID 4. Производительность операций чтения идентична для RAID обоих типов. При операциях записи, в зависимости от типа этих операций, производительность RAID-DP снижается примерно на 2-3%, в сравнении с RAID 4. Причиной этого небольшого снижения является дополнительная операция записи на второй диск с данными «диагональной четности» тома RAID-DP. Нет никакого заметного эффекта увеличения загрузки CPU при использовании RAID-DP в сравнении с RAID 4.

5 Выводы

RAID-DP обеспечивает существенное улучшение защиты данных, что особенно существенно в условиях современного резкого роста объемов жестких дисков. Особенно приятно то, что эти

улучшения, предлагаемые NetApp, не стоят дополнительных денег для пользователей. Для инсталляций, использующих Data ONTAP 6.5 и новее, преобразование существующих томов RAID 4 в новый тип RAID – это простая операция. В отличие от других реализаций RAID с защитой от сбоя двух дисков одновременно, производительность RAID-DP не страдает, и сравнима с производительностью RAID 4, поэтому RAID-DP не требует для своего использования дополнительных ресурсов, как это обычно требуется в решениях других производителей.