Содержание:

Введение

RAID — технология виртуализации данных, которая объединяет несколько дисков в логический элемент для избыточности и повышения производительности.

Базовые уровни RAID

RAID 0 — дисковый массив из двух или более жёстких дисков без резервирования. Информация разбивается на блоки данных фиксированной длины и записывается на оба/несколько дисков поочередно, то есть один блок на первый диск, а второй блок на второй диск соответственно.

(+): Скорость считывания файлов увеличивается в n раз, где n — количество дисков. При этом такая оптимальная производительность достигается только для больших запросов, когда фрагменты файла находятся на каждом из дисков.

(-): Увеличивается вероятность потери данных: если вероятность отказа 1 диска равна p, то вероятность выхода из строя массива RAID 0 из двух дисков равна 2p-p*p. Таким образом, если вероятность отказа одного диска за год равна 1 %, то вероятность отказа массива RAID0 из двух дисков составляет 1,99 %, то есть практически в два раза больше.

RAID 1 — массив из двух (или более) дисков, являющихся полными копиями друг друга. Не следует путать с массивами RAID 1+0 (RAID 10), RAID 0+1 (RAID 01), в которых используются более сложные механизмы зеркалирования.

(+): обеспечивает приемлемую скорость записи (такую же, как и без дублирования) и выигрыш по скорости чтения при распараллеливании запросов.

(+): имеет высокую надёжность — работает до тех пор, пока функционирует хотя бы один диск в массиве. Вероятность выхода из строя сразу двух дисков равна произведению вероятностей отказа каждого диска, то есть значительно ниже вероятности выхода из строя отдельного диска. На практике при выходе из строя одного из дисков следует срочно принимать меры — вновь восстанавливать

избыточность. Для этого с любым уровнем RAID (кроме нулевого) рекомендуют использовать диски горячего резерва.

(-): Недостаток RAID 1 в том, что по цене двух (и более) жестких дисков пользователь фактически получает объём лишь одного.

RAID 2 — Массивы такого типа основаны на использовании кода Хэмминга. Диски делятся на две группы: для данных и для кодов коррекции ошибок, причём если данные хранятся на 2 n − n − 1 {\displaystyle 2^{n}-n-1} дисках, то для хранения кодов коррекции необходимо n {\displaystyle n} дисков. Суммарное количество дисков при этом будет равняться 2 n − 1 {\displaystyle 2^{n}-1} . Данные распределяются по дискам, предназначенным для хранения информации, так же, как и в RAID 0, то есть они разбиваются на небольшие блоки по числу дисков. Оставшиеся диски хранят коды коррекции ошибок, по которым в случае выхода какого-либо жёсткого диска из строя возможно восстановление информации. Метод Хэмминга давно применяется в памяти типа ECC и позволяет на лету исправлять однократные и обнаруживать двукратные ошибки.

(+): массива RAID 2 является повышение скорости дисковых операций по сравнению с производительностью одного диска.

(-): массива RAID 2 является то, что минимальное количество дисков, при котором имеет смысл его использовать — 7, только начиная с этого количества для него требуется меньше дисков, чем для RAID 1 (4 диска с данными, 3 диска с кодами коррекции ошибок), в дальнейшем избыточность уменьшается по экспоненте.

RAID 3 — В массиве RAID 3 из n {\displaystyle n} дисков данные разбиваются на куски размером меньше сектора (разбиваются на байты или блоки) и распределяются по n − 1 {\displaystyle n-1} дискам. Ещё один диск используется для хранения блоков чётности. В RAID 2 для этой цели применялся n − 1 {\displaystyle n-1} диск, но большая часть информации на контрольных дисках использовалась для коррекции ошибок «на лету», в то же время большинство пользователей устраивает простое восстановление информации в случае её повреждения, для чего хватает данных, умещающихся на одном выделенном жёстком диске.

Отличия RAID 3 от RAID 2: невозможность коррекции ошибок на лету.

(+):

- высокая скорость чтения и записи данных;
- минимальное количество дисков для создания массива равно трём.

(-):

- массив этого типа хорош только для однозадачной работы с большими файлами, так как время доступа к отдельному сектору, разбитому по дискам, равно максимальному из интервалов доступа к секторам каждого из дисков. Для блоков малого размера время доступа намного больше времени чтения.
- большая нагрузка на контрольный диск, и, как следствие, его надёжность сильно падает по сравнению с дисками, хранящими данные.

RAID 4 — RAID 4 похож на RAID 3, но отличается от него тем, что данные разбиваются на блоки, а не на байты. Таким образом, удалось отчасти «победить» проблему низкой скорости передачи данных небольшого объёма. Запись же производится медленно из-за того, что чётность для блока генерируется при записи и записывается на единственный диск.

Из широко распространённых систем хранения RAID-4 применяется на устройствах компании NetApp (NetApp FAS), где его недостатки успешно устранены за счет работы дисков в специальном режиме групповой записи, определяемом используемой на устройствах внутренней файловой системой WAFL.

RAID 5 — Основным недостатком уровней RAID от 2-го до 4-го является невозможность производить параллельные операции записи, так как для хранения информации о чётности используется отдельный контрольный диск. RAID 5 не имеет этого недостатка. Блоки данных и контрольные суммы циклически записываются на все диски массива, нет асимметричности конфигурации дисков. Под контрольными суммами подразумевается результат операции XOR (исключающее или). Этот метод, по сути, обеспечивает отказоустойчивость 5 версии. Для хранения результата xor требуется всего 1 диск, размер которого равен размеру любого другого диска в RAID.

Минимальное количество используемых дисков равно трём.

(+): RAID 5 получил широкое распространение, в первую очередь благодаря своей экономичности. Объём дискового массива RAID 5 рассчитывается по формуле (n-1) *hddsize, где n — число дисков в массиве, а hddsize — размер диска (наименьшего, если диски имеют разный размер). Например, для массива из четырёх дисков по 80 гигабайт общий объём будет (4 − 1) * 80 = 240 гигабайт, то есть «потеряется»

всего 25 % против 50 % RAID 10. И с увеличением количества дисков в массиве экономия (по сравнению с другими уровнями RAID, обладающими отказоустойчивостью) продолжает увеличиваться.

RAID 5 обеспечивает высокую скорость чтения — выигрыш достигается за счёт независимых потоков данных с нескольких дисков массива, которые могут обрабатываться параллельно.

(-): Производительность RAID 5 заметно ниже на операциях типа Random Write (записи в произвольном порядке), при которых производительность падает на 10-25 % от производительности RAID 0 (или RAID 10), так как требует большего количества операций с дисками (каждая операция записи, за исключением так называемых full-stripe write-ов, заменяется на контроллере RAID на четыре — две операции чтения и две операции записи).

При выходе из строя одного диска надёжность тома сразу снижается до уровня RAID 0 с соответствующим количеством дисков n-1 — то есть в n-1 раз ниже надёжности одного диска — данное состояние называется критическим (degrade или critical). Для возвращения массива к нормальной работе требуется длительный процесс восстановления, связанный с ощутимой потерей производительности и повышенным риском.

В ходе восстановления (rebuild или reconstruction) контроллер осуществляет длительное интенсивное чтение, которое может спровоцировать выход из строя ещё одного или нескольких дисков массива. Кроме того, в ходе чтения могут выявляться ранее не обнаруженные сбои чтения в массивах cold data (данных, к которым не обращаются при обычной работе массива, архивные и малоактивные данные), препятствующие восстановлению. Если до полного восстановления массива произойдет выход из строя, или возникнет невосстановимая ошибка чтения хотя бы на ещё одном диске, то массив разрушается и данные на нём восстановлению обычными методами не подлежат. Для предотвращения таких ситуаций в RAID-контроллерах может применяться анализ атрибутов S.M.A.R.T.

RAID 6 — похож на RAID 5, но имеет более высокую степень надёжности — три диска данных и два диска контроля чётности. Основан на кодах Рида — Соломона и обеспечивает работоспособность после одновременного выхода из строя любых двух дисков. Обычно использование RAID-6 вызывает примерно 10-15 % падение производительности дисковой группы, относительно RAID 5, что вызвано большим объёмом работы для контроллера (более сложный алгоритм расчёта контрольных

сумм), а также необходимостью читать и перезаписывать больше дисковых блоков при записи каждого блока.

Программный RAID - Для реализации RAID можно применять не только аппаратные средства, но и полностью программные компоненты (драйверы). Например, в системах на ядре Linux поддержка существует непосредственно на уровне ядра. Управлять RAID-устройствами в Linux можно с помощью утилиты mdadm. Программный RAID имеет свои достоинства и недостатки. С одной стороны, он ничего не стоит (в отличие от аппаратных RAID-контроллеров, цена которых от $150). С другой стороны, программный RAID использует некоторое количество ресурсов центрального процессора.

Ядро Linux 2.6.28 (последнее из вышедших в 2008 году) поддерживает программные RAID следующих уровней: 0, 1, 4, 5, 6, 10. Реализация позволяет создавать RAID на отдельных разделах дисков, что аналогично описанному ниже Matrix RAID. Поддерживается загрузка с RAID.

ОС семейства Windows NT, такие как Windows NT 3.1/3.5/3.51/NT4/2000/XP/2003 изначально, с момента проектирования данного семейства, поддерживают программный RAID 0, RAID 1 и RAID 5 (см. Dynamic Disk). Более точно, Windows XP Pro поддерживает RAID 0. Поддержка RAID 1 и RAID 5 заблокирована разработчиками, но, тем не менее, может быть включена, путём редактирования системных бинарных файлов ОС, что запрещено лицензионным соглашением. Windows 7 поддерживает программный RAID 0 и RAID 1, Windows Server 2003 — 0, 1 и 5. Windows XP Home не поддерживает RAID.

В ОС FreeBSD есть несколько реализаций программного RAID. Так, atacontrol, может как полностью строить программный RAID, так и может поддерживать полуаппаратный RAID на таких чипах, как ICH5R. Во FreeBSD, начиная с версии 5.0, дисковая подсистема управляется встроенным в ядро механизмом GEOM. GEOM предоставляет модульную дисковую структуру, благодаря которой родились такие модули как gstripe (RAID 0), gmirror (RAID 1), graid3 (RAID 3), gconcat (объединение нескольких дисков в единый дисковый раздел). Также существуют устаревшие классы ccd (RAID 0, RAID 1) и gvinum (менеджер логических томов vinum). Начиная с FreeBSD 7.2 поддерживается файловая система ZFS, в которой можно собирать следующие уровни RAID: 0, 1, 5, 6, а также комбинируемые уровни.

OpenSolaris и Solaris 10 используют Solaris Volume Manager, который поддерживает RAID 0, RAID 1, RAID 5 и любые их комбинации, как 1+0. Поддержка RAID 6

осуществляется в файловой системе ZFS.

Hybrid RAID - это некоторые из обычных уровней RAID, но в сочетании с дополнительным ПО и твердотельными накопителями (SSD), которые используются как кэш для чтения. В результате производительность системы повышается, так как SSD обладают значительно лучшими скоростными характеристиками по сравнению с HDD. Существует несколько реализаций, например Crucial Adrenaline, либо некоторые контроллеры Adaptec бюджетного класса. На данный момент Hybrid RAID не рекомендуется использовать в серверах ввиду малого ресурса SSD, исключение составляют специальные серверные SSD с повышенным ресурсом.

Дальнейшее развитие идеи RAID

Идея RAID-массивов — в объединении дисков, каждый из которых рассматривается как набор секторов, и в результате драйвер файловой системы «видит» как бы единый диск и работает с ним, не обращая внимания на его внутреннюю структуру. Однако, можно добиться существенного повышения производительности и надёжности дисковой системы, если драйвер файловой системы будет «знать» о том, что работает не с одним диском, а с набором дисков.

Более того, при разрушении любого из дисков в составе RAID 0 вся информация в массиве окажется потерянной. Но если драйвер файловой системы разместил каждый файл на одном диске, и при этом правильно организована структура каталогов, то при разрушении любого из дисков будут потеряны только файлы, находившиеся на этом диске; а файлы, целиком находящиеся на сохранившихся дисках, останутся доступными. Схожая идея «повышения надёжности» реализована в массивах JBOD.

Размещение файлов по принципу «каждый файл целиком находится на одном диске» сложным/неоднозначным образом влияет на производительность дисковой системы. Для мелких файлов латентность (время позиционирования головки над нужным треком + время ожидания прихода нужного сектора под головку) важнее, чем время собственно чтения/записи; поэтому если мелкий файл целиком находится на одном диске, доступ к нему будет быстрее, чем если он разнесён на два диска (структура RAID-массивов такова, что мелкий файл не может оказаться на трёх и более дисках). Для крупных файлов размещение строго на одном диске может оказаться хуже, чем размещение на нескольких дисках; однако, это проявится только если обмен данными производится большими блоками; либо если

к файлу делается много мелких обращений в асинхронном режиме, что позволяет работать сразу со всеми дисками, на которых размещён этот файл.

Источники литературы

- https://ru.wikipedia.org/wiki/RAID
- https://integrus.ru/blog/tipy-raid-massivov.html
- https://beginpc.ru/raznoe/what-is-raid
- https://www.ixbt.com/storage/raids.html