

Лекция 6-2. Дисковые массивы. Технология RAID

Избыточные массивы
недорогих дисков

Дисковые массивы

Дисковым массивом (Disc Array) называют набор жестких дисков, подключенных к одному многопортовому контроллеру. В простейшем случае контроллер интерпретирует их как независимые накопители, которые ОС может использовать для размещения логических разделов. Такой массив называется **JBOD (Just a Bunch of Discs)**.

Однако все **современные дисковые контроллеры** серверного назначения, а также большинство контроллеров настольных и мобильных (включая встроенные), **поддерживают** определенную логику для **объединения жестких дисков в один или несколько массивов**, каждый из которых представляется ОС единым диском. Это объединение преследует одну из двух целей (или обе вместе):

- **Повышение производительности;**
- **Повышение отказоустойчивости (надежности).**

Технология RAID

Технология объединения дисков в массив прорабатывалась в 70-х годах, однако название **RAID (Redundant Array of Inexpensive Discs)** было предложено в 1987 году (ун-т Беркли, США).

Суть идеи:

дорогостоящие серверные диски большого объема можно заменить набором дешевых и не столь надежных винчестеров настольного класса за счет усложнения логики доступа к ним со стороны контроллера.

Сейчас RAID расшифровывается как **Redundant Array of Independent Discs**,

т.к. задача снижения стоимости отошла на второй план. **Основной задачей стало обеспечение отказоустойчивости за счет введения избыточности** (дополнительных аппаратных ресурсов для хранения копий или контрольных кодов данных). При этом RAID может решать и задачу улучшения производительности.

ОЦЕНКА НАДЕЖНОСТНЫХ ХАРАКТЕРИСТИК RAID-МАССИВОВ

HDD считаются достаточно надежными устройствами – среднее время до выхода из строя (MTTF) жестких дисков корпоративного уровня составляет порядка 1,6 миллионов часов, а вероятность появления невосстановимой ошибки (UER) благодаря использованию кодов обнаружения ошибок (EDC), кодов коррекции ошибок (ECC) и различных проприетарных технологий поддержания целостности данных на носителе по оценкам производителей – не более чем 10^{-16} . Между тем в реальности частота ежегодных отказов (AFR) жестких дисков оценивается примерно в 0,75 %.

функциональный сбой и скрытая (или отложенная) ошибка

Функциональный сбой

Под функциональным сбоем, как правило, понимают выход из строя накопителя, который **может обнаружить управляющий им контроллер**, т.е. когда требуемые данные не могут быть прочитаны с накопителя.

- К основным причинам функциональных сбоев причисляют:
нарушение серворазметки,
- сбои в работе электроники накопителя,
- поломки считывающих головок,
- сбои системы позиционирования,
- превышение лимита критичных S.M.A.R.T. параметров.

Скрытые ошибки

Под скрытыми ошибками дисков (UDE) понимают **не обнаруживаемые электроникой накопителя ошибки** при записи данных (UWE), когда внешне нормальная операция записи влечет нарушение данных на соседних дорожках и/или не происходит модификация оригинальных данных, и ошибки при чтении данных (URE) при неправильной интерпретации кодов коррекции ошибок (в случае множественных ошибок) или считывании неверных данных из-за ошибок позиционирования.

К первопричинам отложенных ошибок относят

- производственные дефекты магнитного слоя,
- коррозионные и физические повреждения магнитного слоя в процессе эксплуатации,
- временные сбои в позиционировании магнитных головок, например из-за вибраций,
- ошибки позиционирования из-за термического расширения рабочей поверхности из-за нарушений температурного режима эксплуатации накопителя.

ОЦЕНКА НАДЕЖНОСТНЫХ ХАРАКТЕРИСТИК RAID-МАССИВОВ

Современные RAID-массивы обеспечивают надежное хранение данных только в случае функциональных сбоев накопителей, входящих в RAID. При появлении скрытых ошибок надежность хранения данных не гарантируется *Hafner J. L., Deenadhayalan V., Belluomini W. et al. // IBM Journal of Research and Development. 2008. V. 52. № 4/5. P. 431.*

Роль скрытых ошибок в процессах потери данных возрастает с увеличением информационной емкости HDD и приходом на рынок корпоративных систем хранения nearline-устройств [*Whittington W. // Desktop, Nearline & Enterprise HDDs [Электронный ресурс].* – Режим доступа: <http://www.snia.org/education/tutorials/2008/spring/storage/>

Для борьбы со скрытыми ошибками необходимы методы проактивного мониторинга на уровне систем хранения данных и исправления данных ошибок с использованием внешних кодов коррекции ошибок и/или обнаружения ошибок по отношению к стандартно реализуемым в RAID-массивах – «скрабинг» (scrubbing). Подобные технологии в настоящее время реализуются в ряде high-end систем хранения корпоративного класса (HP EVA, EMC2 Centera и т.п.).

Архитектура RAID

Технология RAID предполагает создание дисковой подсистемы, надежность и/или быстродействие которой в несколько раз выше, чем у каждого из входящих в ее состав жестких дисков.

Ядром RAID является многопортовый контроллер, который реализует определенную логику *распределения* (distribution) *данных* и их резервных копий/контрольных кодов *по* подключенным к нему жестким *дискам*. При этом для системного ПО один массив представляется одним **виртуальным диском**. Контроллер также может объединить в *массивы несколько массивов*, создав массив второго порядка. Как правило, массивы 3-го и более высокого порядка не реализовываются.

Архитектура RAID

Контроллер отвечает за распределение данных при записи (striping), сборку их при чтении (concatenating), контроль за целостностью (monitoring), восстановление массива при сбое диска/дисков (rebuilding).

Для оперативного и *прозрачного* восстановления к массиву может быть приписан резервный диск (Spare disc), который заменяет дефектный. При этом один резервный диск может приписываться к нескольким массивам. В обычном режиме, когда массив исправен, резервный диск не используется.

Обычно для массива RAID требуются диски идентичной емкости.

Для достижения высокой скорости они должны быть одной модели. При использовании разных дисков задействованный объем каждого будет равен объему меньшего среди дисков.

Уровни RAID

В рамках технологии RAID стандартно описано несколько **методов организации массивов**, получивших название «**уровни**». Чем выше уровень, тем больше для него требуется аппаратных ресурсов (в том числе самих дисков) и тем лучше его свойства (отказоустойчивость + производительность).

Каждый уровень обладает своими достоинствами и недостатками, ориентируясь на которые, следует выбирать уровень в зависимости от приоритетов выполняющихся на компьютере задач.

Уровни RAID-массивов

- Уровни, которые можно считать стандартизованными — RAID 0, RAID 1, RAID 2, RAID 3, RAID 4, RAID 5 и RAID 6.
- Применяются также различные комбинации RAID-уровней, что позволяет объединить их достоинства. Обычно это комбинация какого-либо отказоустойчивого уровня и нулевого уровня, применяемого для повышения производительности (RAID 1+0, RAID 0+1, RAID 50).

(Помимо стандартных, существует целый ряд проприетарных разработок, обычно – для серверных систем и систем хранения данных верхнего ценового класса)

- Встроенные контроллеры дешевых материнских плат поддерживают обычно уровни 0 и 1. На платах повыше классом реализованы также уровни 5 и 10 (или 0+1). Контроллеры серверов поддерживают также уровень 6, а также «улучшенные» уровни 1E, 5EE, 50, 60.
- Все современные RAID-контроллеры поддерживают функцию JBOD (не предназначена для создания массивов, а обеспечивает возможность подключения к RAID-контроллеру отдельных дисков).

Диаграммы уровней RAID

Далее будут приведены типовые диаграммы уровней RAID по возрастанию их технической сложности, требований к контроллеру и минимально необходимому числу жестких дисков.

Для примера будут изображены 4 диска, объединенные в 1 массив. На практике количество дисков и массивов бывает больше, но в общем случае 4 достаточно для создания массива любого стандартного уровня.

Буквами А, В, С и т.д. отмечены *стрипы* (strips) – последовательные блоки, на которые делится содержимое виртуального диска, сформированного контроллером из массива. Стрип – единица хранения данных на одном диске массива. Обычно размер стрипа можно задавать в настройках контроллера. От этого параметра зависят многие характеристики полученного массива.

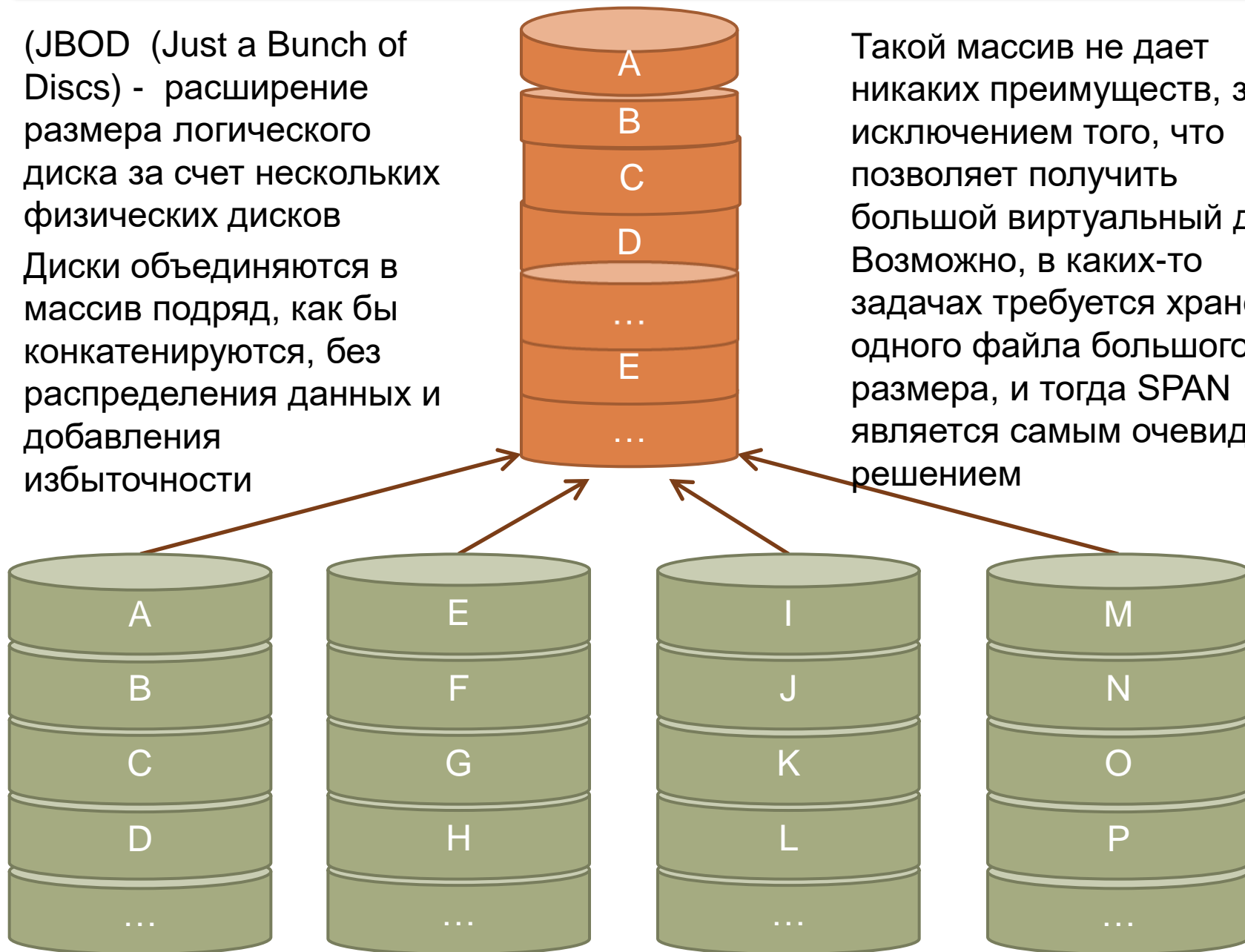
Страйп (stripe, по аналогии bit-byte) – это сумма всех стрипов с каждого из дисков массива. Размер страйпа также важен, т.к. он определяет, запрос какого размера может быть выполнен параллельно всеми дисками

SPAN (JBOD)

(JBOD (Just a Bunch of Discs) - расширение размера логического диска за счет нескольких физических дисков

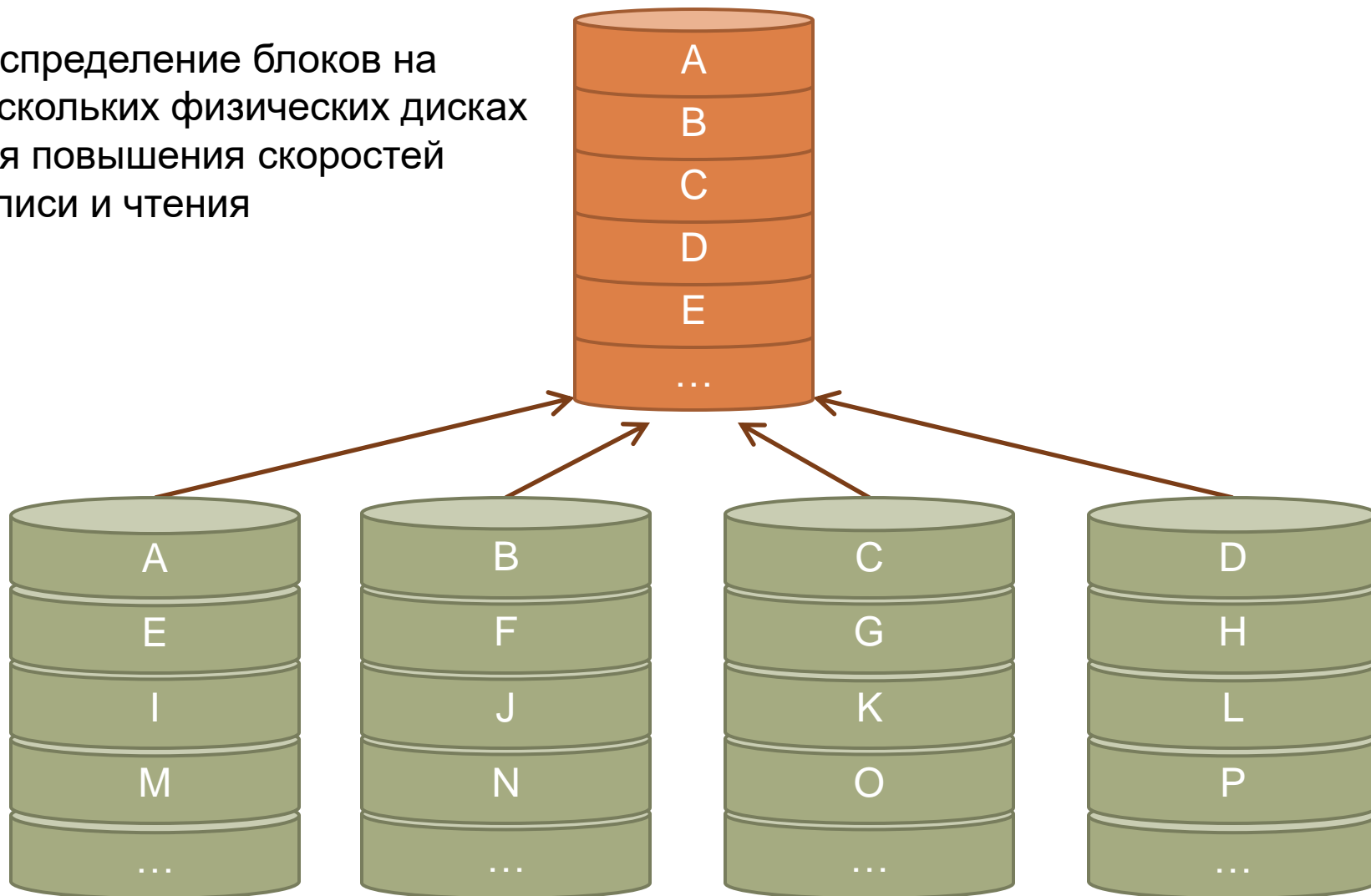
Диски объединяются в массив подряд, как бы конкатенируются, без распределения данных и добавления избыточности

Такой массив не дает никаких преимуществ, за исключением того, что позволяет получить большой виртуальный диск. Возможно, в каких-то задачах требуется хранение одного файла большого размера, и тогда SPAN является самым очевидным решением

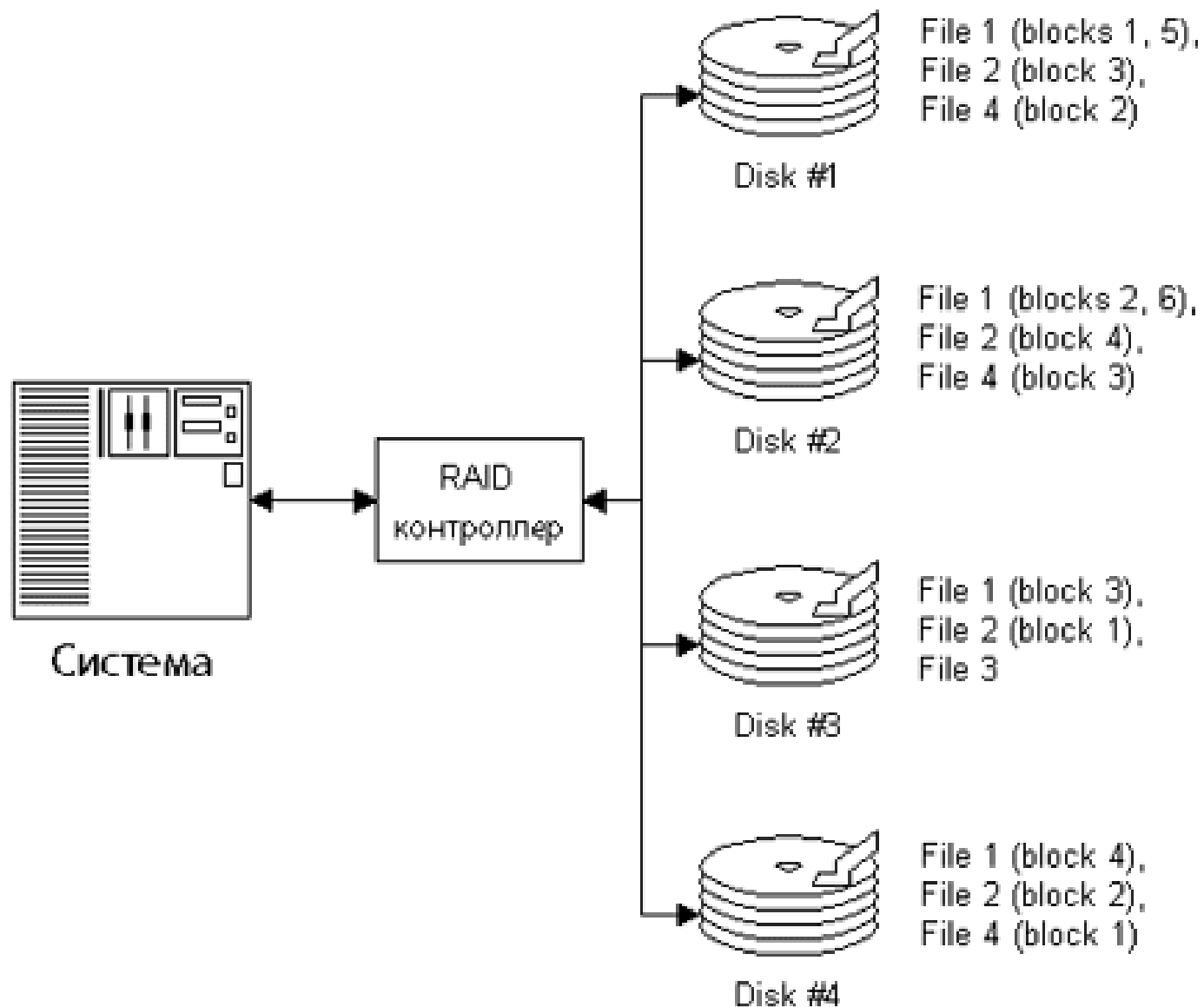


RAID 0 (Striping — «чередование»)

Распределение блоков на нескольких физических дисках для повышения скоростей записи и чтения



Пример RAID 0



RAID 0 – Плюсы и минусы

- + Самый простой и выгодный с точки зрения производительности массив. В нем присутствует распределение, но **нет избыточности – емкость массива равна сумме всех дисков**
- + Реализация RAID 0 очень проста, требует минимум аппаратных средств, а благодаря возможности параллельного чтения и записи может давать прирост, равный количеству дисков (при условии, что все запросы будут равны страйпу). Ускорение достигается в равной степени и для случайных, и для последовательных запросов.
- **отказоустойчивость** не только не повышается, но даже **снижается**, причем кратно количеству дисков (при условии равновероятного выхода из строя каждого). Для разрушения (без возможности восстановления) массива достаточно выхода из строя одного диска.

RAID 0 применяется в настольных машинах, а также в задачах, где данные могут быть легко восстановлены. На массиве RAID 0 обычно хранятся временные файлы при выполнении видеомонтажа, обработки изображений, 3D-графики, разного рода кэши, индексы баз данных, журналы работы и т.д.

Вероятность выхода из строя RAID 0 - 2 HDD

p - вероятность выхода из строя HDD

$$P(A_1) = P(A_2) = p \quad (1)$$

$q=1-p$ - вероятность работоспособного состояния .

$$P(\bar{A}_1) = P(\bar{A}_2) = q \quad (2)$$

A - событие выхода RAID0 из строя

$$1 = P(\bar{A}_1 \bar{A}_2) + P(A_1 \bar{A}_2) + P(\bar{A}_1 A_2) + P(A_1 A_2)$$

$$P(A) = P(A_1 \bar{A}_2) + P(\bar{A}_1 A_2) + P(A_1 A_2)$$

$$P(A) = 1 - P(\bar{A}_1 \bar{A}_2) \quad (3)$$

$$P(A) = 1 - P(\bar{A}_1)P(\bar{A}_2) = 1 - q^2$$

ПРИМЕР Найдем вероятность разрушения RAID 0 при $p = 0.03$

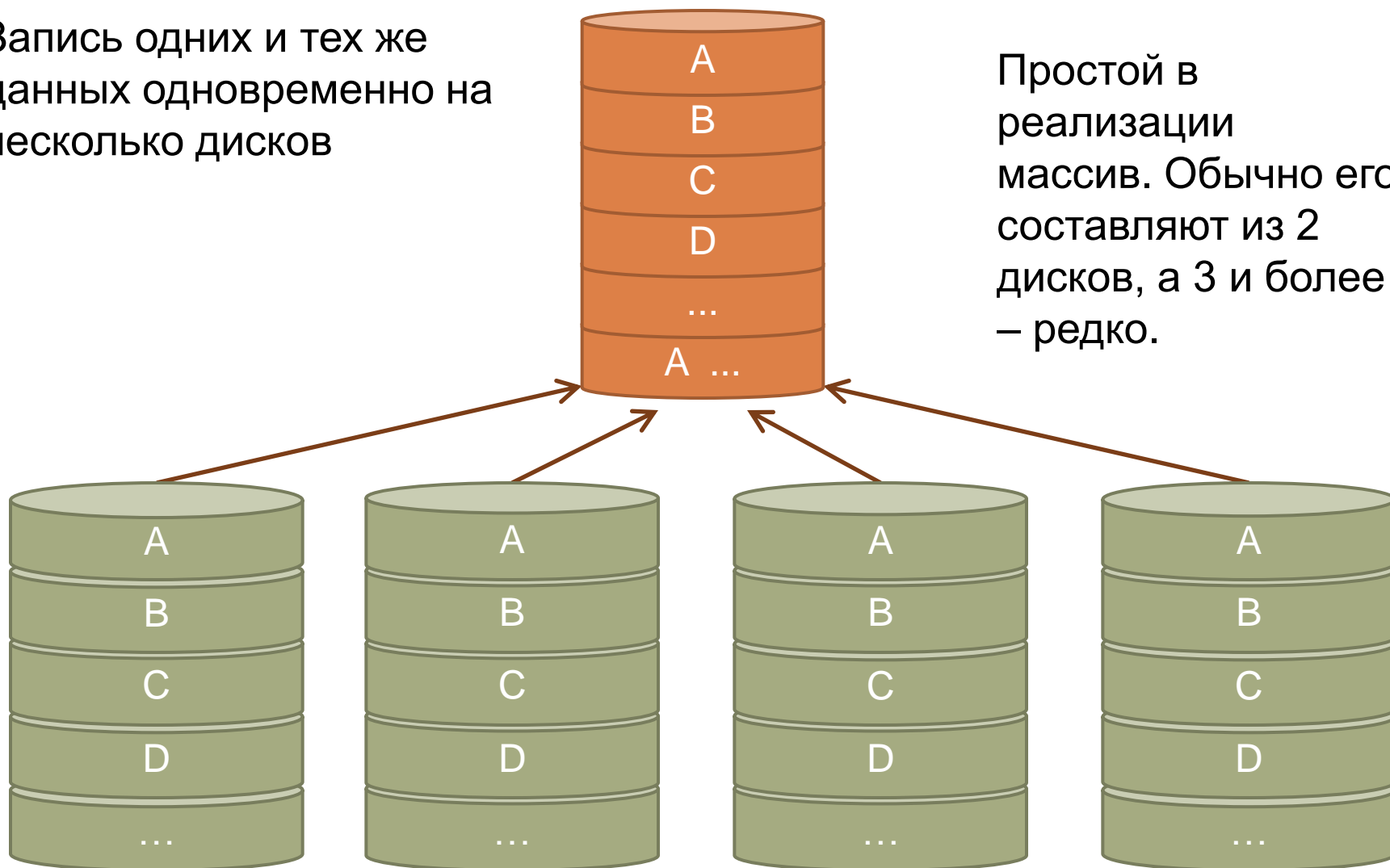
$$q = 1 - p = 0.97 \quad P(A) = 1 - q^2 = 0.0591$$

Вероятность разрушения RAID 0 равняется 5,91%.

RAID 1 (Mirroring - «зеркалирование»)

Запись одних и тех же данных одновременно на несколько дисков

Простой в реализации массив. Обычно его составляют из 2 дисков, а 3 и более – редко.



RAID 1 – Плюсы и минусы

- + **высокая степень отказоустойчивости** при минимальном использовании аппаратных средств. Для работы массива достаточно, чтобы оставался рабочим хотя бы один (причем любой) из дисков.
- + при организации параллельного **доступа возможно ускорение всех операций чтения**, как у массива RAID 0. Операция чтения по времени выполнения ограничена быстродействием самого медленного диска в массиве.
- + простота реализации
- + дает **наивысшую скорость восстановления массива**, причем эта операция легко выполняется в фоновом режиме.
- **потери дисковой емкости**: фактически емкость массива равна емкости одного диска.

В чистом виде применяется редко, в основном для задач, где требуется наивысшее сочетание быстродействия и отказоустойчивости, пусть и за счет повышения стоимости: финансовая отчетность, банковские системы, различные корпоративные базы данных и т.д.

Вероятность выхода из строя RAID 1

RAID1 - 2 HDD

p - вероятность выхода из строя HDD

$q=1-p$ - вероятность работоспособного состояния .

A - событие выхода RAID1 из строя

$$P(A_1) = P(A_2) = p \quad (1)$$

$$P(\bar{A}_1) = P(\bar{A}_2) = q \quad (2)$$

$$\bar{A}_1 \bar{A}_2 \quad A_1 \bar{A}_2 \quad A_1 \bar{A}_2 \quad A_1 A_2$$

$$P(A) = P(A_1 A_2) = P(A_1)P(A_2) = p^2 \quad (3)$$

Пример: Пусть вероятность выхода из строя HDD в течение года равняется 3%. Найдем вероятность разрушения RAID 1

$$p = 0.03$$

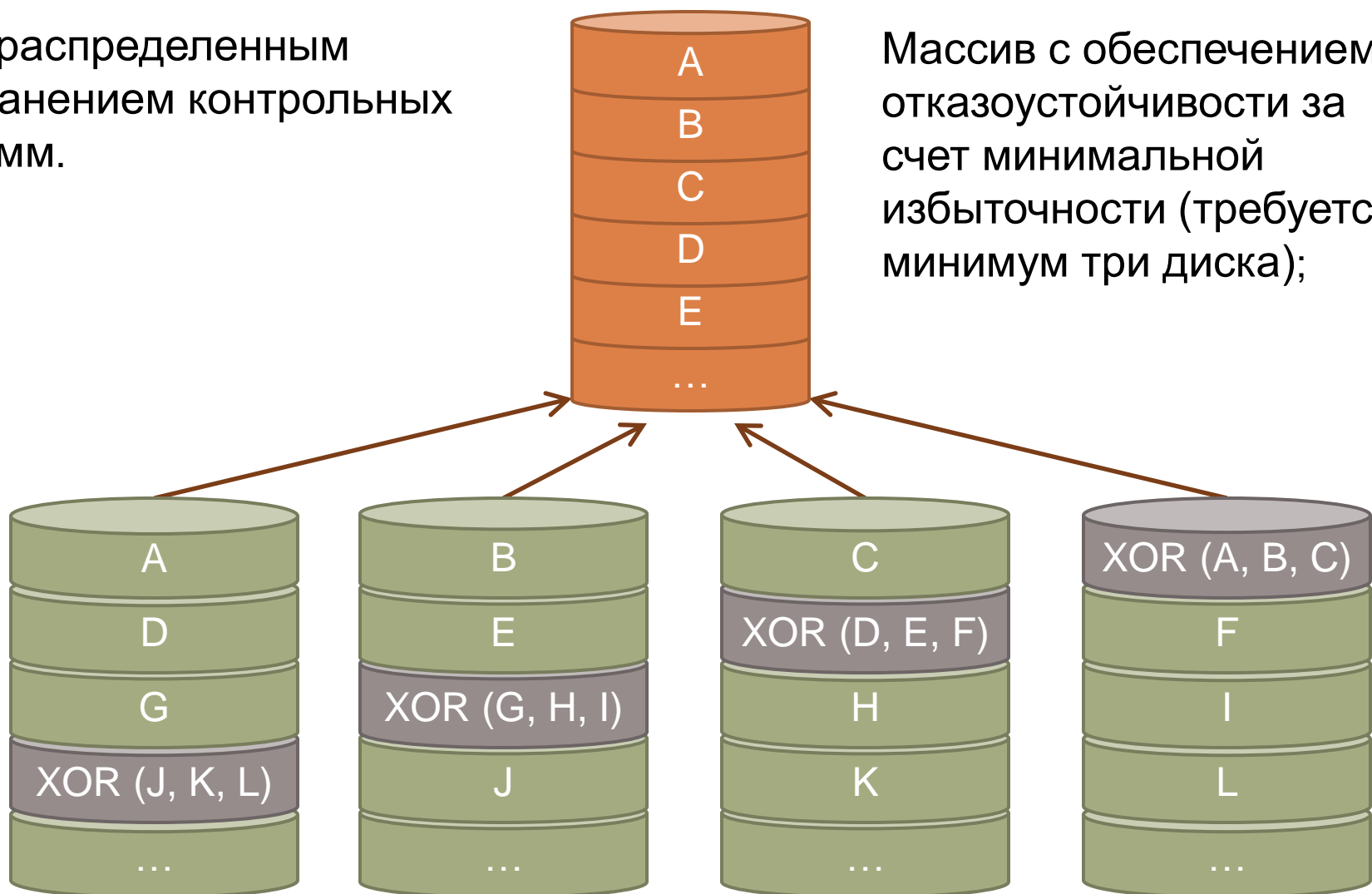
$$P(A) = p^2 = 0.0009$$

Ответ: Вероятность разрушения RAID 1 равняется 0,09%.

RAID 5 (Striping with parity)

Отказоустойчивый массив независимых дисков с распределенной четностью (Independent Data disks with distributed parity blocks)

С распределенным хранением контрольных сумм.



Массив с обеспечением отказоустойчивости за счет минимальной избыточности (требуется минимум три диска);

Для каждого страйпа вычисляется блок четности методом побитного XOR и записывается на один из дисков по очереди (в циклическом порядке), что позволяет равномерно нагружать массив и осуществлять конкурентные запросы параллельно.

Пример. Массив содержит n дисков, размер страйпа d .

Для каждой порции из $n-1$ страйпов рассчитывается контрольная сумма $p = d_1 \text{ XOR } d_2 \text{ XOR } \dots \text{ XOR } d_{n-1}$.

Страйп d_1 записывается на первый диск, страйп d_2 — на второй и так далее вплоть до страйпа d_{n-1} , который записывается на $(n-1)$ -диск. Далее на n -й диск записывается контрольная сумма p_n , и процесс циклически повторяется с первого диска, на который записывается страйп d_n .

Если один из дисков, например второй, вышел из строя, то блок d_2 окажется недоступным при считывании. Однако его значение легко восстановить по контрольной сумме и по значениям остальных блоков с помощью все той же операции XOR: $d_2 = d_1 \text{ XOR } p \text{ XOR } \dots \text{ XOR } d_{n-1}$.

Массив RAID 5 **защищает только от выхода из строя одного диска** и способен, пусть и со значительным снижением скорости, работать без него до той поры, пока не будет установлен новый винчестер. Потери на избыточность составляют ровно один диск, но в случае большого количества дисков эти потери незначительны.

RAID 5 – Плюсы и минусы

- + **Скорость работы RAID 5 при чтении так же высока, как и у RAID 0 и RAID 1.**
- + **предоставляет компромисс между отказоустойчивостью и избыточностью при возможности достижения высокого быстродействия** при наличии эффективного контроллера.
(является наиболее часто используемым).
- скорость записи, особенно случайной, может существенно снижаться, т.к. для записи хотя бы одного стрипа приходится прочитать весь страйп и обновить блок четности. Контроллеры с достаточным объемом кэш-памяти и функцией отложенной записи могут компенсировать этот недостаток, но не до конца.
- сложность восстановления массива. К тому же в этот момент массив подвержен разрушению при порче второго диска.

RAID 5 применяется для большинства серверных задач, кроме хранения баз данных, для которых требуется высокое быстродействие при случайной записи.

Вероятность выхода из строя RAID 5

RAID5 - 3 HDD

p - вероятность выхода из строя HDD

$q=1-p$ - вероятность работоспособного состояния .

A - событие выхода RAID5 из строя

$$P(A_1) = P(A_2) = p \quad (1)$$

$$P(\bar{A}_1) = P(\bar{A}_2) = q \quad (2)$$

$$\begin{array}{ccccccc} \bar{A}_1 \bar{A}_2 \bar{A}_3 & A_1 \bar{A}_2 \bar{A}_3 & \bar{A}_1 A_2 \bar{A}_3 & \bar{A}_1 \bar{A}_2 A_3 & & & \\ A_1 A_2 \bar{A}_3 & A_1 \bar{A}_2 A_3 & \bar{A}_1 A_2 A_3 & & A_1 A_2 A_3 & & \\ C_3^2 & & & & C_3^3 & & \end{array} \quad (3)$$

Два диска из
трех

три из трех

$$P(A_1 A_2 \bar{A}_3) = P(A_1 \bar{A}_2 A_3) = P(\bar{A}_1 A_2 A_3) = p^2 q$$

$$P(A_1 A_2 A_3) = p^3$$

$$P(A) = C_3^2 p^2 q + C_3^3 p^3 = 3p^2 q + p^3$$

Вероятность выхода из строя RAID 5

RAID5 - ***N*** HDD

p - вероятность выхода из строя HDD

q=1-p - вероятность работоспособного состояния .

A- событие выхода RAID5 из строя

$$P(A_1) = P(A_2) = p \quad (1)$$

$$P(\bar{A}_1) = P(\bar{A}_2) = q \quad (2)$$

$$P(A) = C_N^2 p^2 q^{N-2} + C_N^3 p^3 q^{N-3} + \dots + C_N^N p^N q^0$$

$$P(A) = \sum_{i=2}^N C_N^i p^i q^{N-i} \quad (4)$$

$$\sum_{i=0}^N C_N^i p^i q^{N-i} = 1 \quad , \text{ где } q = 1 - p \quad (5)$$

$$P(A) = 1 - q^N - N p q^{N-1}$$

Пример

Пусть вероятность выхода из строя HDD в течение года равняется 3%.

Найдем вероятность разрушения RAID5 на трех HDD

$$p = 0.03$$

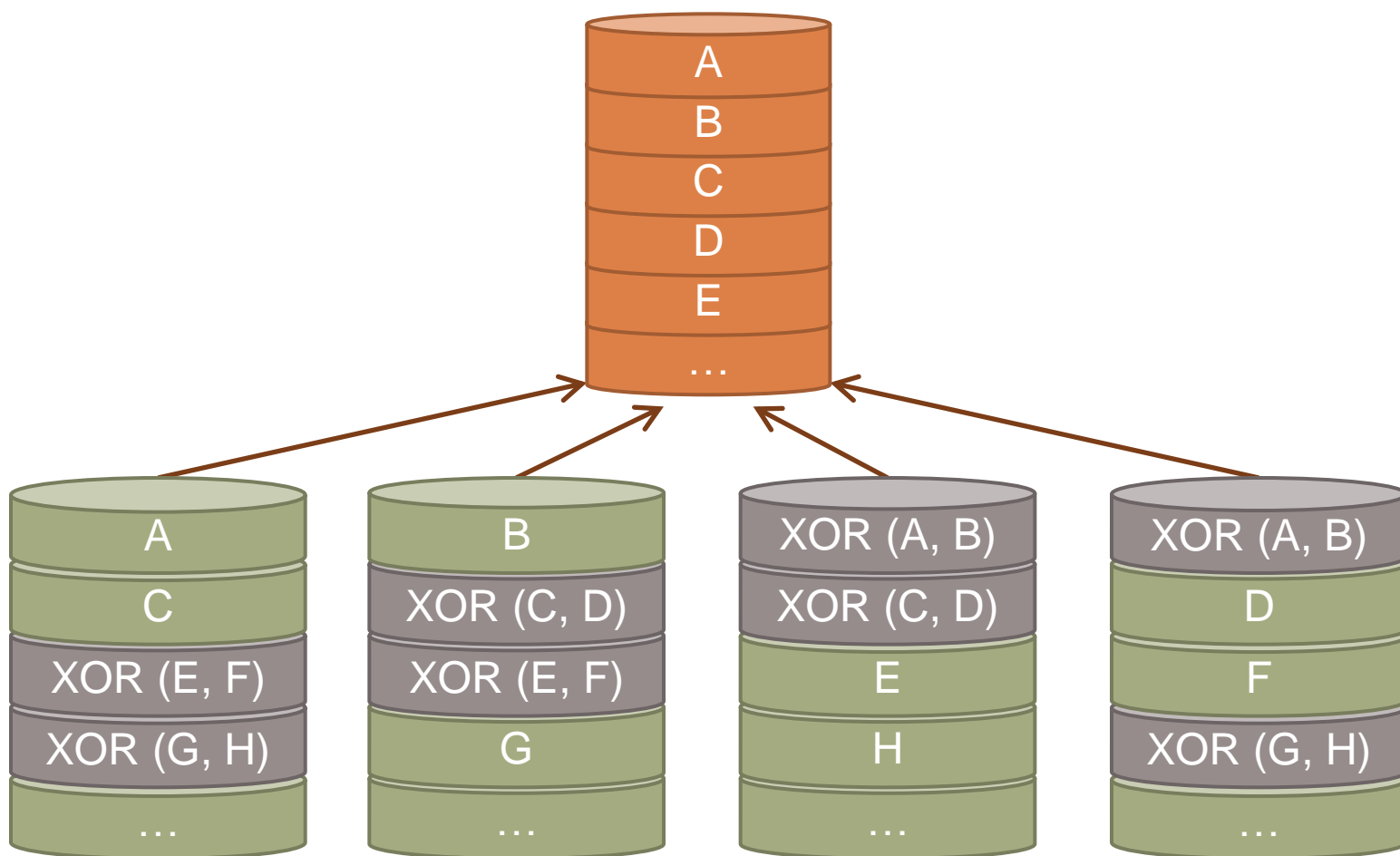
$$q = 1 - p = 0.97$$

$$P(A) = 1 - 0,97^3 - 3 * 0,03 * 0,97^2 = 0.002646$$

Ответ: Вероятность разрушения RAID 5 равняется ,0,26%.

RAID 6 (Striping with dual parity)

RAID 6. Отказоустойчивый массив независимых дисков с двумя независимыми распределенными схемами четности (Independent Data disks with two independent distributed parity schemes)



Похож на RAID 5, но имеет более высокую степень надежности — под контрольные суммы выделяется емкость 2-х дисков, рассчитываются 2 суммы по разным алгоритмам, причем с чередованием. Благодаря этому решается задача устойчивости к отказу сразу двух дисков, а также ускорения и защиты процесса восстановления массива при отказе одного диска.

Вместе с тем скорость чтения и особенно записи снижается, если не применяется достаточно мощный контроллер. Кроме того, затраты на избыточность составляют уже два диска, а для массива необходимо не менее 4 дисков.

RAID 6 реализуется только в мощных контроллерах и применяется для общих серверных задач, где высок риск потери ценных данных. Не теряет производительности при потере одного из дисков, что важно для некоторых задач.

Восстановление двух дисков

Будем вычислять два синдрома

$$\begin{cases} P = \sum_{i=0}^{N-1} D_i \\ Q = \sum_{i=0}^{N-1} q_i D_i \end{cases}$$

Тогда для двух утраченных (неизвестных) D_α, D_β имеем систему уравнений

$$\begin{cases} D_\alpha + D_\beta = P - \sum_{i \neq \alpha, \beta} D_i \\ q_\alpha D_\alpha + q_\beta D_\beta = Q - \sum_{i \neq \alpha, \beta} q_i D_i \end{cases}, i \neq \alpha, \beta; \alpha \neq \beta$$

Если система однозначно разрешима для любых α, β , то в страйпе можно восстановить два любых утраченных диска

Пусть $P = D_0 + D_1 + \dots + D_\alpha + \dots + D_{N-1} = \sum_{i=0}^{N-1} D_i$

Тогда $D_\alpha = P - \sum_{i=0, i \neq \alpha}^{N-1} D_i$

Вероятность выхода из строя RAID 6

RAID6 - ***N*** HDD

p - вероятность выхода из строя HDD

q=1-p - вероятность работоспособного состояния .

$$P(A_1) = P(A_2) = p \quad (1)$$

$$P(\bar{A}_1) = P(\bar{A}_2) = q \quad (2)$$

A- событие выхода RAID5 из строя

Справедливы те же, рассуждения что и для RAID 5. Отличие составляет только то, что RAID 6 считается разрушенным при выходе трех HDD. Следовательно, искомая вероятность равна:

$$P(A) = \sum_{i=3}^N C_N^i p^i q^{N-i} \quad (7)$$

$$\sum_{i=0}^N C_N^i p^i q^{N-i} = 1 \quad , \text{ где } q = 1 - p \quad (5)$$

$$P(A) = 1 - q^N - Npq^{N-1} - 0.5N(N-1)p^2q^{N-2}$$

Пример

Пусть вероятность выхода из строя HDD в течение года равняется 3%.

Найдем вероятность разрушения RAID6 из четырех HDD

$$N = 4$$

$$p = 0.03$$

$$q = 1 - p = 0.97$$

$$P(A) = 1 - 0,97^4 - 4 * 0,03 * 0,97^3 - 6 * 0,03^2 * 0,97^2 = 0.000105$$

Ответ: Вероятность разрушения RAID6 равняется ,0,01%.

Восстановление одного диска (P,Q – диски с контр. суммами)

Если отказал диск P или Q, то восстанавливать данные не нужно

Если отказал один диск данных, то восстанавливать его можно с использованием синдрома P (как RAID-5)

Если отказал диск данных и диск с P, то восстановить данные можно с помощью синдрома Q решив уравнение

$$q_{\alpha} D_{\alpha} = Q - \sum q_i D_i, i \neq \alpha$$

Если изменился один диск данных, то пересчитать синдромы можно имея старое и новое значение диска

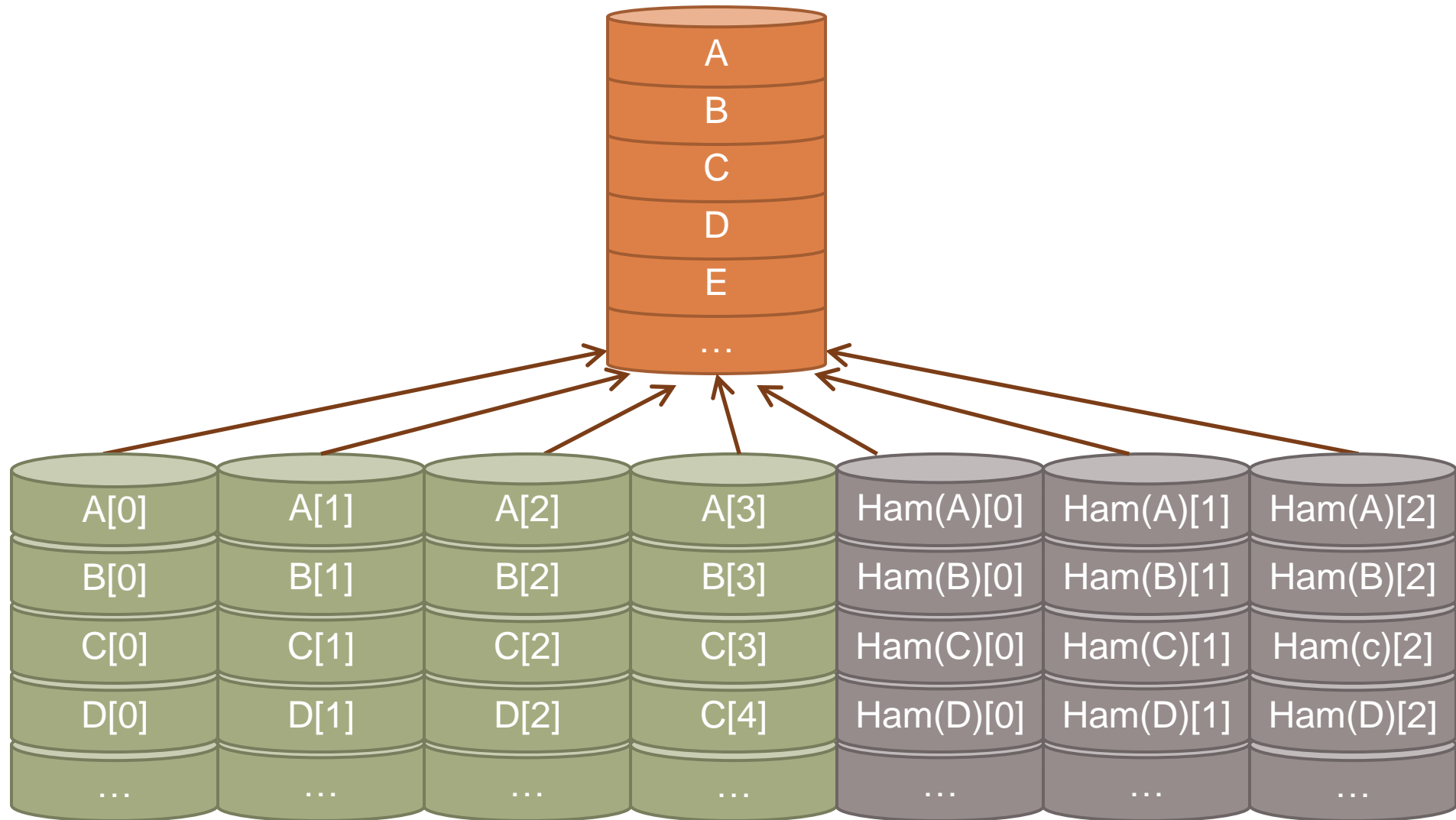
• тривиально

Уровни RAID 2-4

Создатели концепции RAID предусмотрели еще несколько вариантов реализации массивов, которые позволяют уменьшить избыточность при сохранении высокого уровня отказоустойчивости. К сожалению, разработчики устройств не поддержали эти уровни.


Реализация RAID 2, RAID 3 и RAID 4 практически не встречается в современных контроллерах жестких дисков ввиду высокой технической сложности и отсутствии явных преимуществ перед уровнем RAID 5.

RAID 2 (Bit-striping with Hamming code)



- В массивах такого типа диски делятся на две группы — для данных и для кодов коррекции ошибок, причем если данные хранятся на n дисках, то для хранения кодов коррекции необходимо n дисков.

Поток данных разбивается на слова таким образом, что количество бит в слове равно количеству дисков и при записи слова каждый отдельный бит записывается на свой диск. Для каждого слова вычисляется код коррекции ошибок, который записывается на выделенные диски для хранения контрольной информации. Их число равно количеству бит в слове контрольной суммы.



В данном массиве распределение выполняется не поблочно, а побитно, размер страйпа – 4 бита. Для каждого страйпа вычисляется 3-битный код Хэмминга, который представляет собой инверсию поразрядного XOR номеров позиций, в которых находятся «1», взятых в двоичном коде. Для 4-битных слов разрядность номеров позиций составляет 3 бита (нумерация с 1), поэтому классический массив RAID 2 состоит из 7 дисков – 4 диска с данными и 3 диска с кодами Хэмминга. Допустимы и варианты с иным количеством дисков.

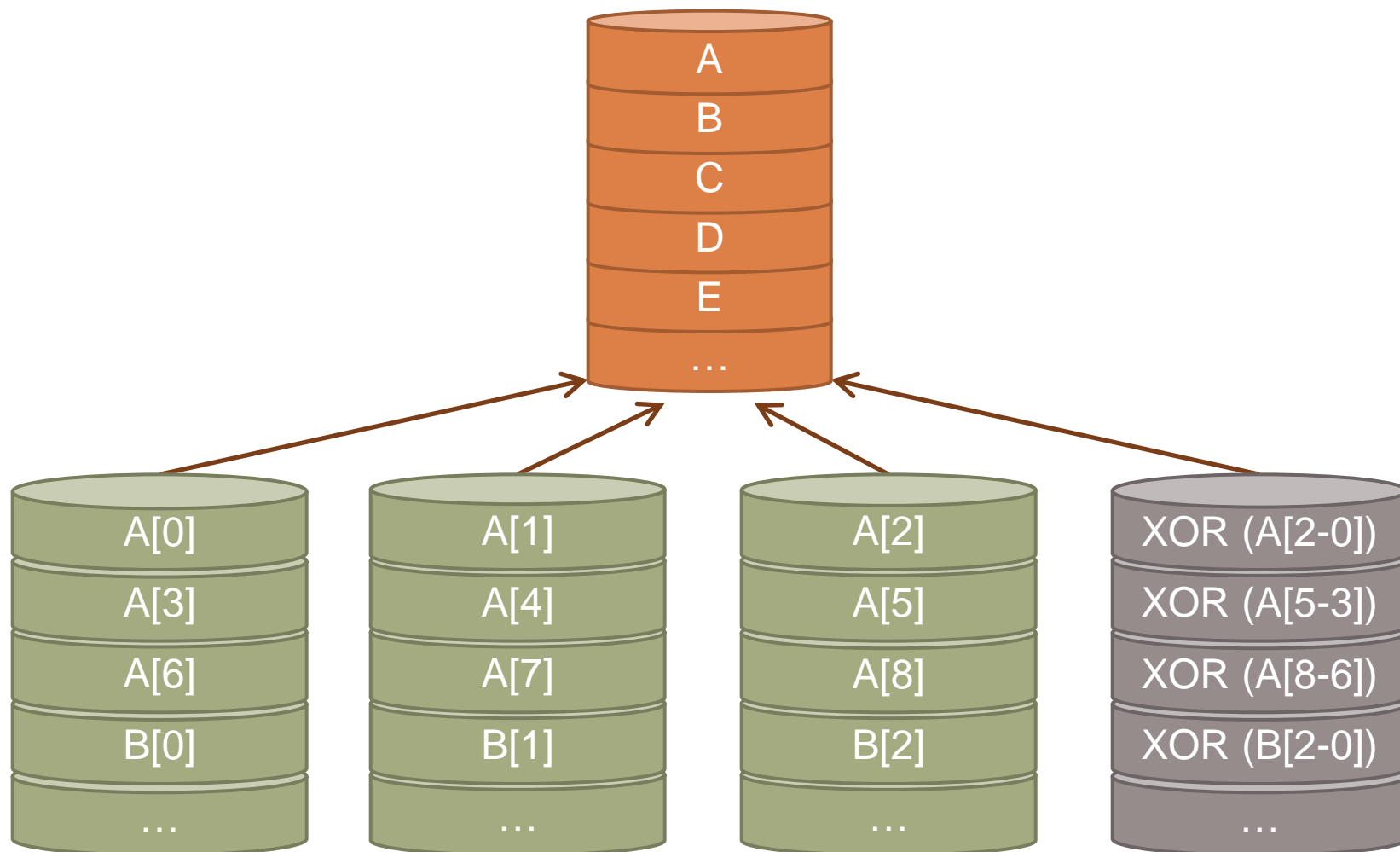
Метод Хемминга давно применяется в памяти типа ECC и позволяет на лету исправлять однократные и обнаруживать двукратные ошибки.

Массив RAID 2 предполагает строго синхронную и одновременную работу всех дисков сразу, за счет чего обеспечивается мгновенное обнаружение до 4 бит ошибки в каждом байте и исправление до 2 бит. При этом мгновенная скорость чтения/записи равна сумме скоростей всех дисков с данными.

Данный массив не реализован на практике, т.к. технически очень сложно обеспечить полную синхронность всех дисков, а иначе все преимущества этого массива теряются. Кроме того, RAID 2 обеспечивает защиту данных на исправных дисках, а это диски «умеют» делать и сами, за счет хранения кодов ECC в секторах.

RAID 3 (Parallel transfer with parity)

Отказоустойчивый массив с параллельной передачей данных и четностью
(Parallel Transfer Disks with Parity)



Данный массив использует побайтное распределение данных по дискам с дополнительным диском, используемым для хранения байтов четности. При этом предполагается строго синхронное обращение ко всем дискам, как в случае с RAID 2, однако потери на избыточность равны только одному диску вне зависимости от состава массива.

Соответственно массив защищен от потери только одного диска.

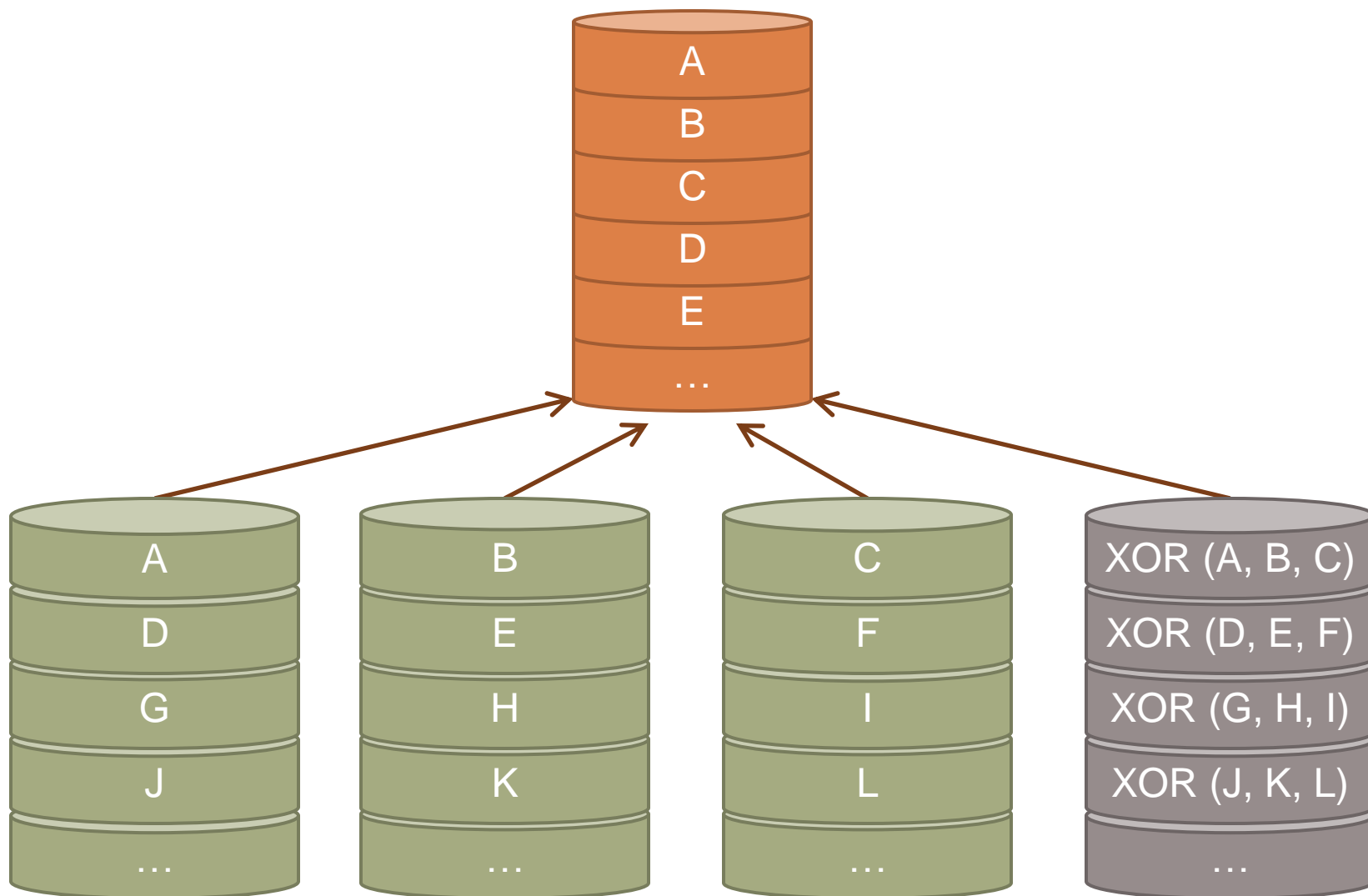
Преимущество RAID 3 – умножение скорости чтения на число дисков с данными, поскольку работают они параллельно и одновременно.

Недостаток RAID 3 – невозможность параллельного выполнения нескольких запросов, т.к. для каждого блока данных (который имеет размер не менее сектора) используются сразу все винчестеры.

Из-за высокой технической сложности RAID 3 используется редко и только в задачах, предполагающих доступ к файлам большого размера: системы видеонаблюдения, потокового воспроизведения, архивирования, видеомонтажа и т.п.

RAID 4 (Striping with dedicated parity)

Отказоустойчивый массив независимых дисков с разделяемым диском четности Independent Data disks with shared Parity disk)



RAID 4 похож на RAID 3, но отличается от него тем, что данные разбиваются на блоки, а не на байты; Данный массив идентичен RAID 5, однако для хранения блоков четности используется один и тот же диск. К обычным преимуществам RAID 5 добавляется ускорение доступа при чтении за счет того, что диск четности в этом случае вообще не используется. Потери на избыточность – один диск на массив, прирост скорости чтения равен числу дисков минус один.

При записи, как и в RAID 5, требуется вычитывать весь страйп и обновлять блок четности. А поскольку все блоки четности находятся на одном диске (в RAID 5 - на разных), этот диск сразу становится узким местом, т.к. к нему приходится обращаться во время всех операций записи без исключения. дисках, что повышает вероятность параллельного выполнения нескольких коротких запросов записи.

Ввиду имеющегося недостатка RAID 4 практически не используется.

RAID 7

Отказоустойчивый массив, оптимизированный для повышения производительности (Optimized Asynchrony for High I/O Rates as well as High Data Transfer Rates)

Является зарегистрированной торговой маркой корпорации Storage Computer. Во многом он похож на RAID 4 с возможностью кэширования данных. В состав RAID 7 входит контроллер с встроенным микропроцессором под управлением операционной системы реального времени (SOS). Она позволяет обрабатывать все запросы на передачу данных (как между отдельными дисками, так и между массивом и компьютером) асинхронно и независимо.

Блок вычисления контрольных сумм интегрирован с блоком буферизации, для хранения информации о четности используется отдельный диск, который может быть размещен на любом канале. RAID 7 имеет высокую скорость передачи данных и обработки запросов, хорошее масштабирование (при увеличении числа дисков повышается скорость записи). Самым большим недостатком этого уровня является стоимость его реализации.

Расширенные уровни RAID

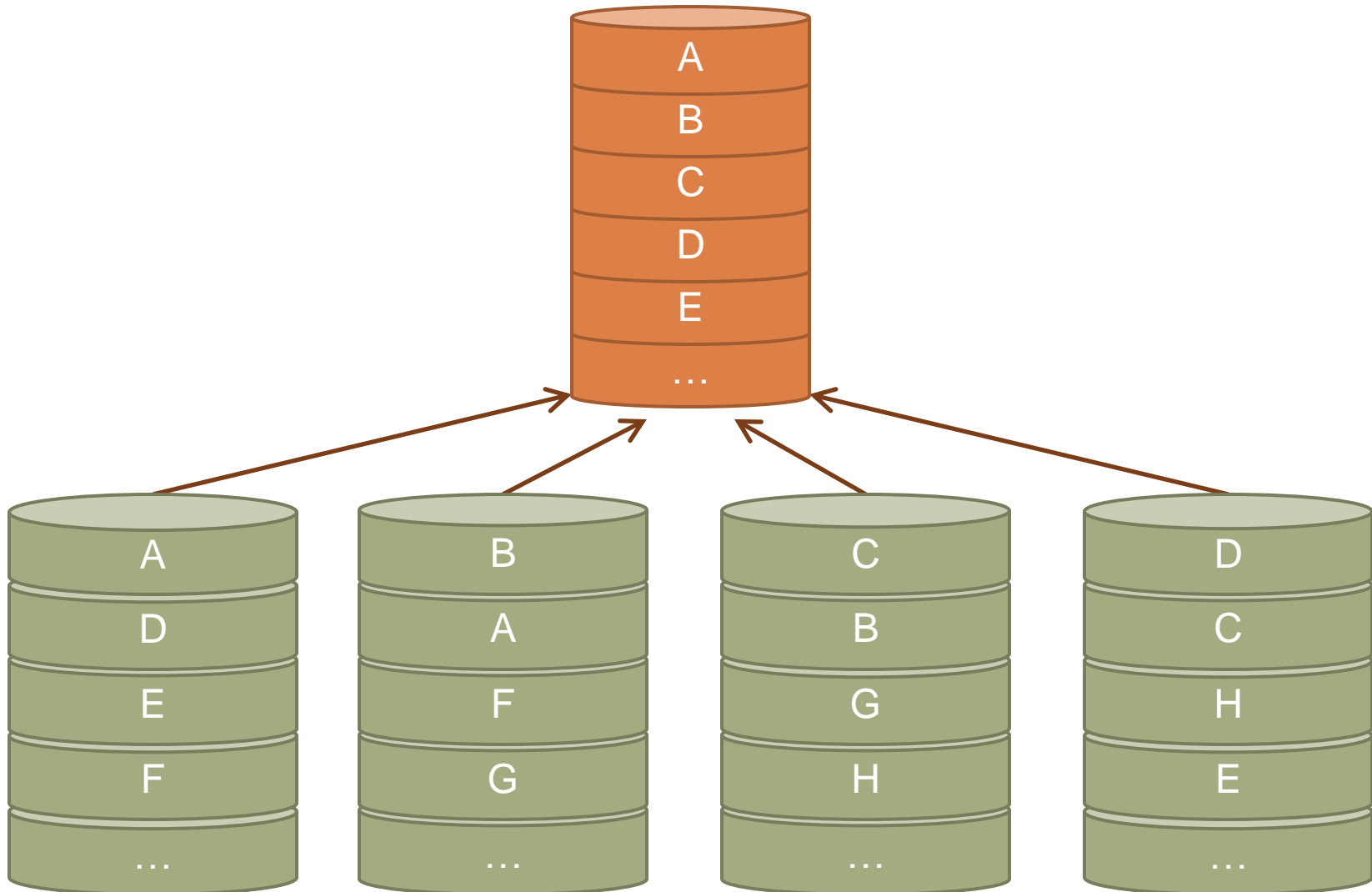
Некоторые изготовители систем хранения данных не были удовлетворены стандартными уровнями RAID, которые имеют определенные недостатки. Чтобы их устранить, были разработаны **«расширенные» уровни RAID, которые модифицируют исходные путем внесения более сложных правил распределения и дублирования данных.**

Среди множества таких расширений есть варианты, которые поддерживаются практически все изготовители контроллеров для серверов. Впрочем, расширенные уровни обычно сложнее в реализации, поэтому их поддержка ограничена.

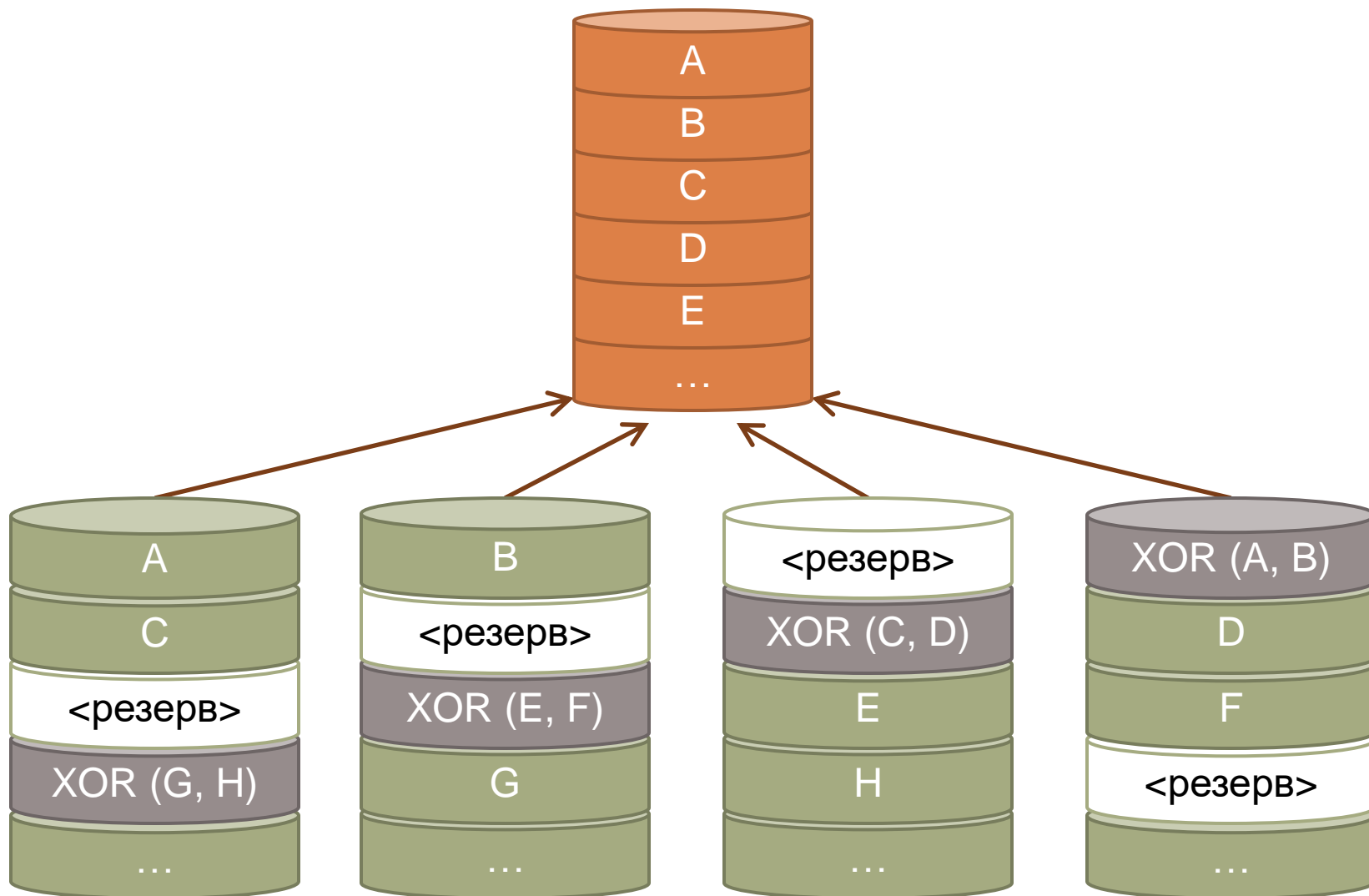
RAID 1E существует в различных вариантах, которые объединяет общий принцип – **использование распределения данных одновременно с дублированием на тех же дисках**. Обычно выполняется дублирование по страйпам (каждый страйп записывается дважды), еще лучше – с переменной порядка стрипов (как на диаграмме), что дает возможность восстановить данные, если выйдут из строя более одного диска, расположенные в массиве не рядом друг с другом. Также применяется **двукратная запись одного стрипа**, что по сути мало отличается от первого способа.

Основное преимущество RAID 1E – возможность реализации на массивах с нечетным количеством дисков, что невозможно для RAID 0. Сохранены все преимущества RAID 1, равно как и недостатки, главный из которых – потеря 50% полезной емкости дисков на избыточность.

RAID 1E (Striping with mirroring)



RAID 5EE (Hot space striping)



Массивы RAID 5E и 6E отличаются тем, что включают в свой состав на один диск больше, чем требуется. При этом на каждом диске создается свободная зона, которая будет использована при выходе из строя одного из дисков (виртуальный spare disc). В процессе восстановления выполняется «компрессия» массива (с заполнением резервных зон), и на выходе получается обычный RAID 5/6. Чтобы он снова стал 5E/6E, требуется «декомпрессия» с применением нового диска.

Массив RAID 5EE используется чаще, поскольку он использует распределение по дискам резервных зон в том же порядке, что и блоков четности. В итоге нагрузка на диски становится более равномерной, а процесс «компрессии» проходит значительно быстрее (восстановленный блок пропавшего диска записывается вместо предусмотренной в каждом страйпе резервной зоны).

Данный массив обладает большей избыточностью (на один диск), но позволяет обойтись без резервного диска, который обычно простаивает. С другой стороны, при наличии нескольких массивов резервный диск отводится только один, что позволяет сэкономить.

Гибридные, или комбинированные массивы RAID

Если в массиве RAID имеется более 3 дисков, то появляется возможность построения гибридного, или многоуровневого массива, в котором сочетаются структуры сразу двух массивов различного типа. Это позволяет получить сумму преимуществ двух типов, но за счет усложнения логики работы и затрат на диски.

Как правило, в многоуровневых массивах стараются сочетать высокое быстродействие массива типа RAID 0 и отказоустойчивостью массивов других типов. За счет применения большого количества дисков это удастся сделать.

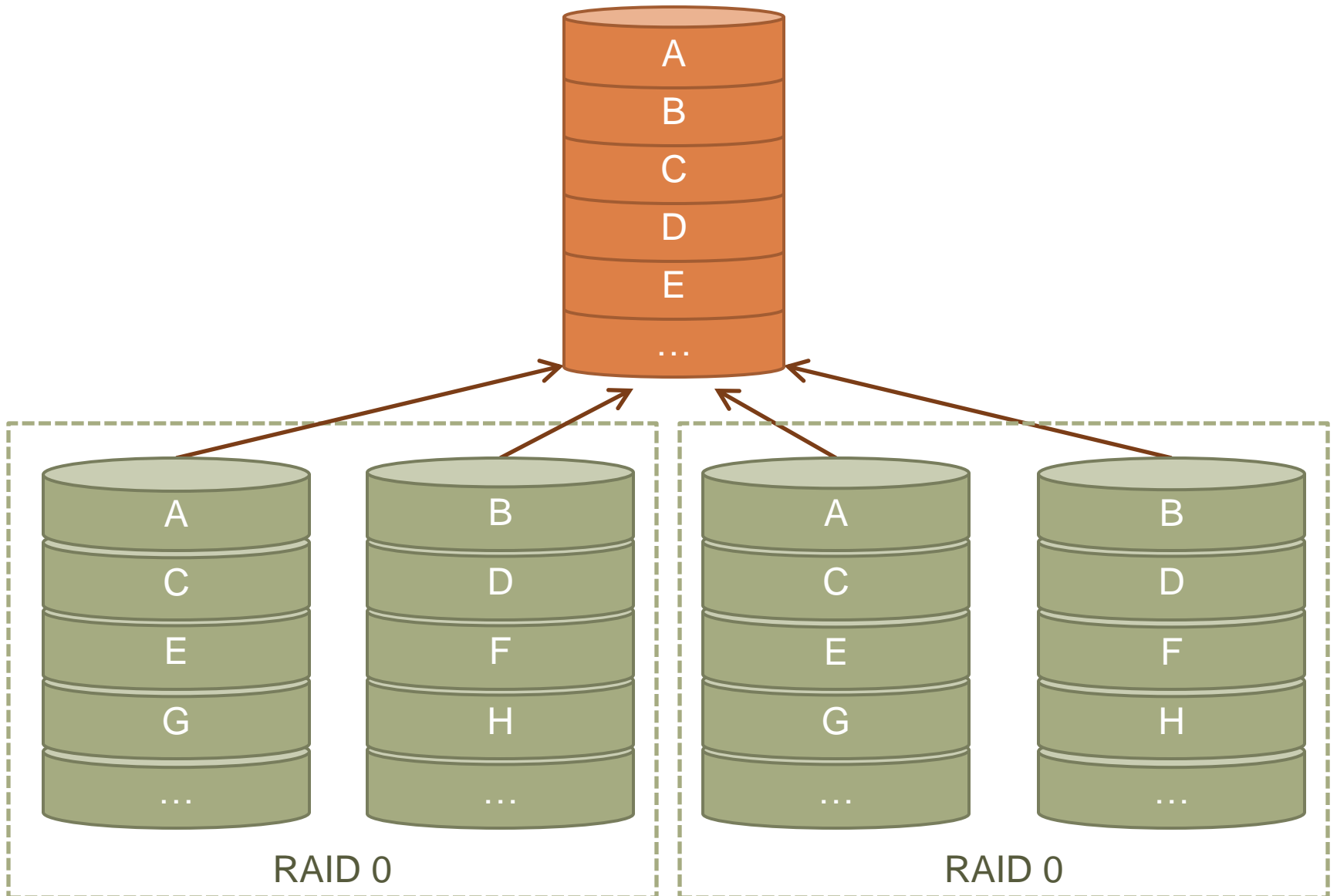
Массивы типа 10 или 0+1 часто поддерживаются настольными системами, другие массивы реализуются в серверных контроллерах.

Зеркало на многих дисках

RAID 1+0 или RAID 0+1.

- Под RAID 10 (RAID 1+0) имеют в виду вариант, когда два или более RAID 1 объединяются в RAID 0.
- Вариант, когда два RAID 0 объединяются в RAID 1, называется RAID 0+1. Достоинства и недостатки такие же, как и у уровня RAID 0. Как и в других случаях, рекомендуется включать в массив диски горячего резерва из расчёта один резервный на пять рабочих.

RAID 0+1 (Mirrored RAID 0)

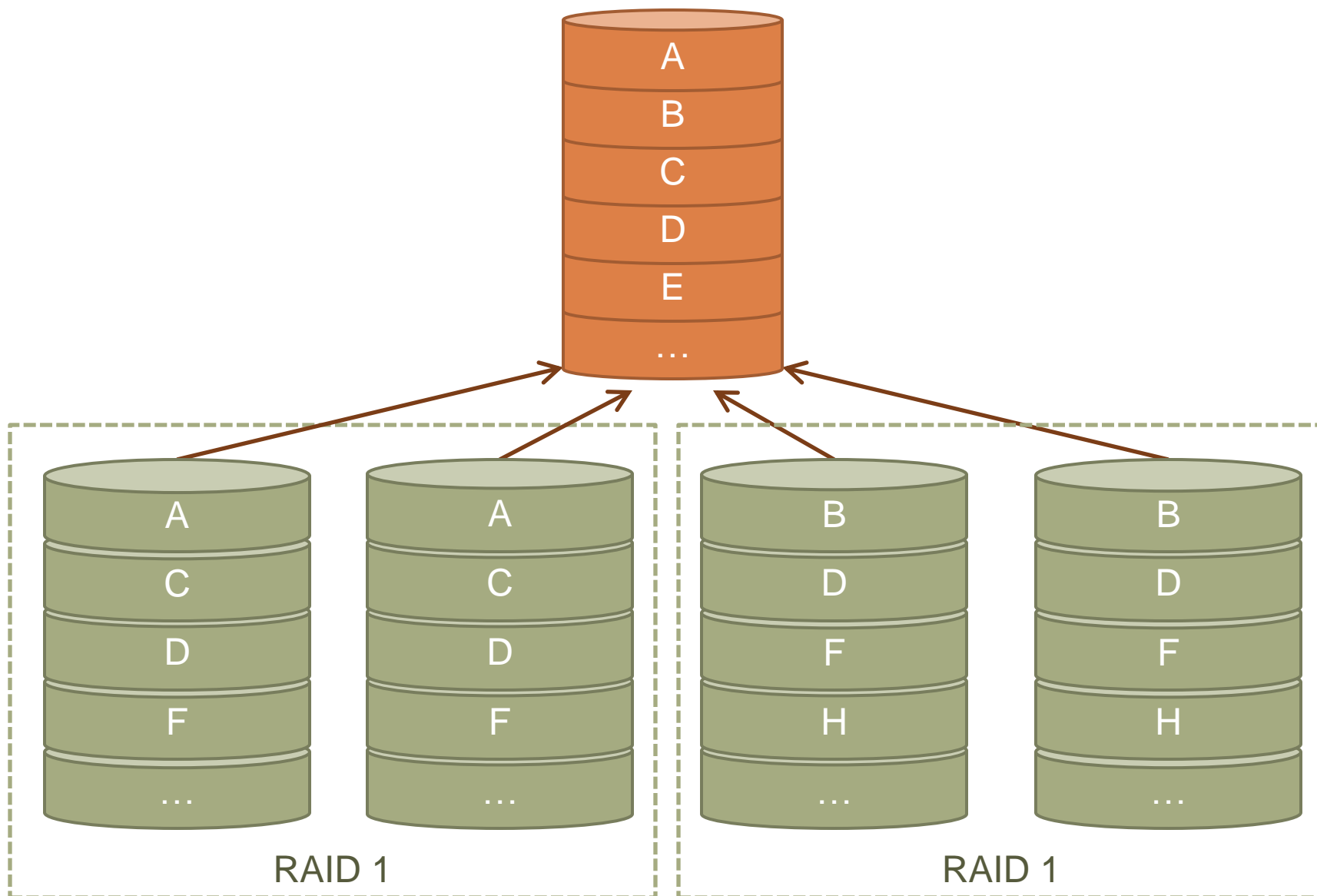


Данный массив представляет собой две (или более) копии массивов типа RAID 0. За счет этого добавляется отсутствующая у RAID 0 возможность отказоустойчивости. Кроме того, RAID 0+1 крайне прост в реализации и не требует серьезных аппаратных ресурсов, в том числе для процесса восстановления.

Недостатки

- затраты на избыточность (как и в RAID 1, теряется ровно половина дискового пространства), требуется не менее 4 дисков. При этом по сравнению с RAID 1 обеспечивается всего лишь возрастание скорости записи.
- резкое снижение отказоустойчивости при выходе из строя хотя бы одного (или нескольких при многократном зеркалировании) диска. Массив сразу превращается в RAID 0 (дубликат сразу разрушается), надежность которого в несколько раз ниже, чем надежность каждого диска.

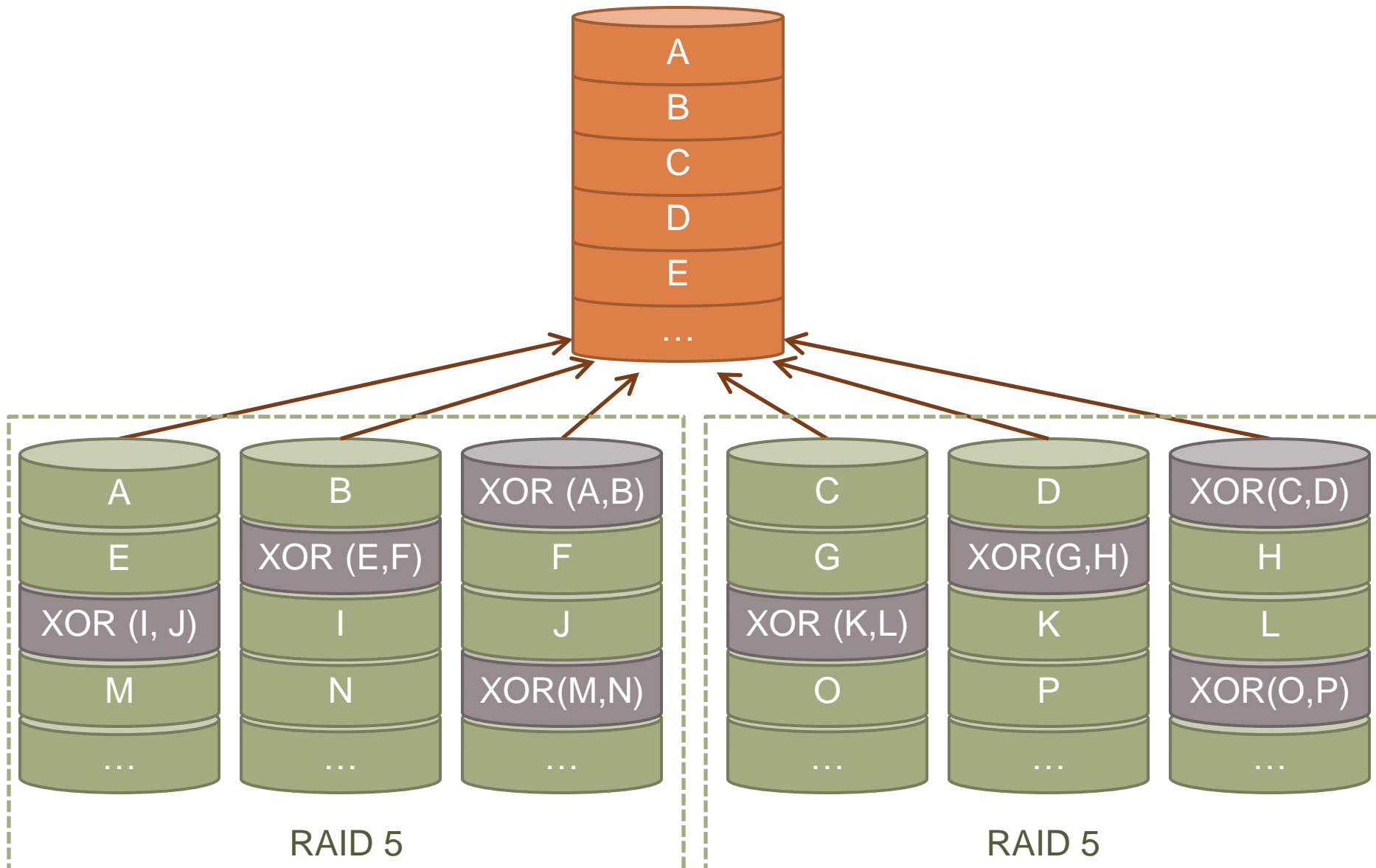
RAID 10 (Striped RAID 1)



RAID 10 – Плюсы и минусы

- На первый взгляд, данный массив кажется идентичным RAID 0+1. И действительно, диски всего лишь переставлены, а принцип размещения информации не изменился.
- + Тем не менее, у RAID 10 имеется одно серьезное преимущество: выход из строя одного из дисков приводит всего лишь к ослаблению (за счет исчезновения дубликата) одного из массивов RAID 1, в то время как остальные продолжают оставаться устойчивыми к отказам. Вероятность того, что второй отказавший диск будет из того же RAID 1, что и первый, невысока, если RAID 10 состоит из большого количества RAID 1.
 - Фактически у RAID 10 один недостаток – большие затраты на избыточность. Поэтому этот вариант используют для массивов, состоящих из большого количества дисков. Задачи те же, что и для RAID 1 – повышенные требования к отказоустойчивости при необходимости высокого быстродействия.

RAID 50 (Striped RAID 5)



Данный массив состоит из набора массивов RAID 5, данные между которыми распределены, как в массиве RAID 0. Этот способ является более предпочтительным, чем создание одного массива RAID 5 из такого же количества дисков. При этом потери на избыточность можно регулировать, подбирая количество дисков, входящих в состав RAID 5.

Данный массив обеспечивает такую же производительность по чтению и такую же отказоустойчивость, что и RAID 10, но с меньшими затратами на избыточность. Операции записи требуют все так же много времени, особенно при отсутствии эффективного кэширования (а при большом числе дисков так оно и есть).

Еще одно преимущество RAID 50 – более высокая скорость восстановления при выходе из строя одного из дисков (восстанавливать требуется только один из компонентов массива RAID 0, остальные продолжают работать) по сравнению с одним RAID 5.

Существует также вариант RAID 60, аналогичный RAID 50.

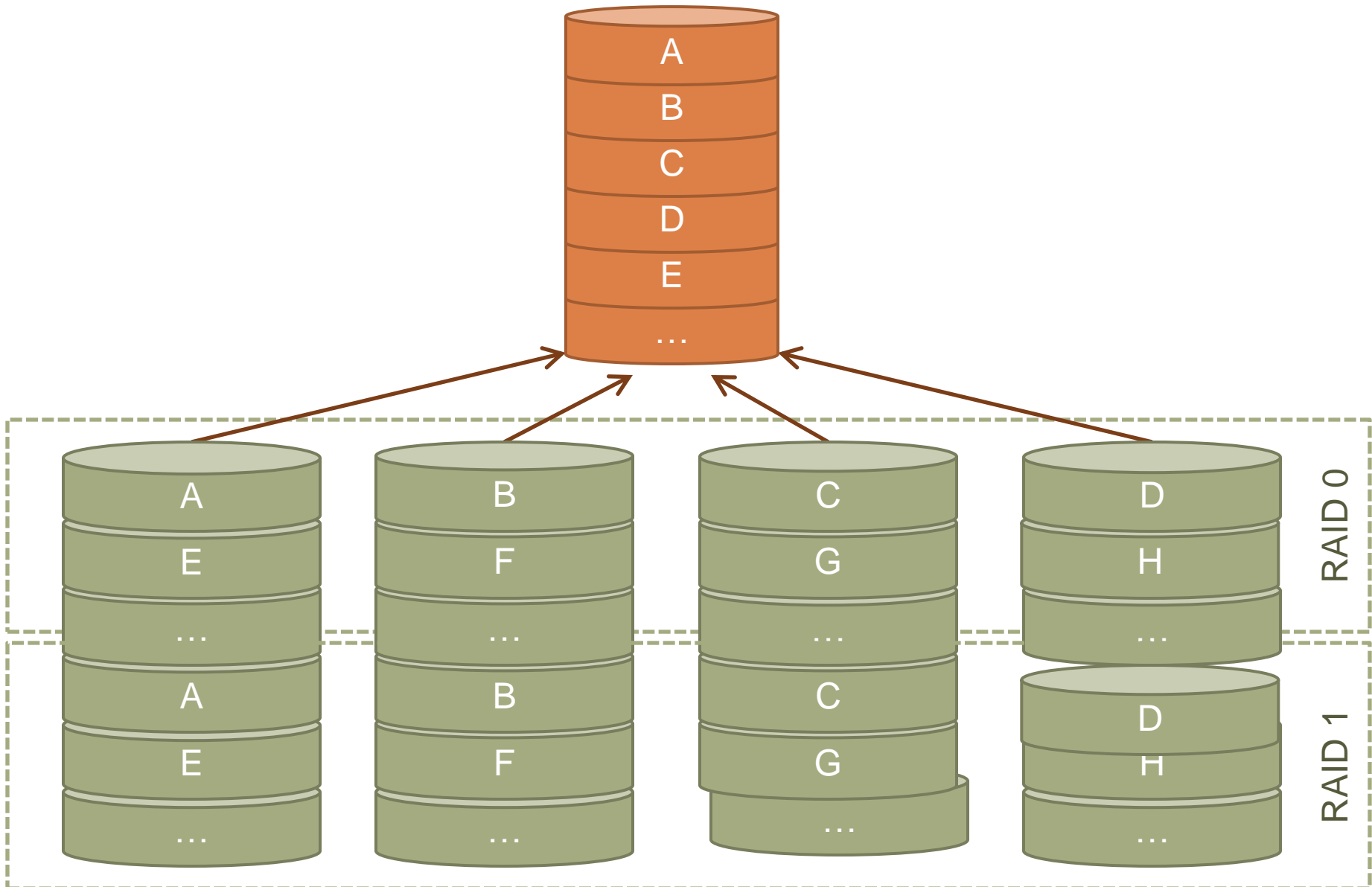
RAID 0+3 (03) и 3+0 (30)

- По идее сочетание чередования и RAID 3 дает выигрыш в скорости, но он довольно мал. Зато система заметно усложняется. Наиболее простой уровень 3+0. Из двух массивов RAID 3 строится страйп, и минимальное количество требуемых дисков – 6. Получившийся RAID 3+0 с точки зрения надежности лучше, чем 0+3.
- Достоинства этих комбинаций в довольно высоком проценте использования емкости дисков и высокой скорости чтения данных. Недостатки – высокая цена, сложность системы.

RAID 1+5 (15) и 5+1 (51)

- Этот уровень построен на сочетании зеркалирования или дуплекса и чередования с распределенной четностью. Основная цель RAID 15 и 51 – значительное повышение надежности. Массив 1+5 продолжает работать при отказе трех накопителей, а 5+1 - даже при потере пяти из восьми жестких дисков! Платить приходится большим количеством неиспользуемой емкости дисков и общим удорожанием системы.
- Чаще всего для построения RAID 5+1 используют два контроллера RAID 5, которые зеркалируют на программном уровне, что позволяет снизить затраты.

Matrix RAID



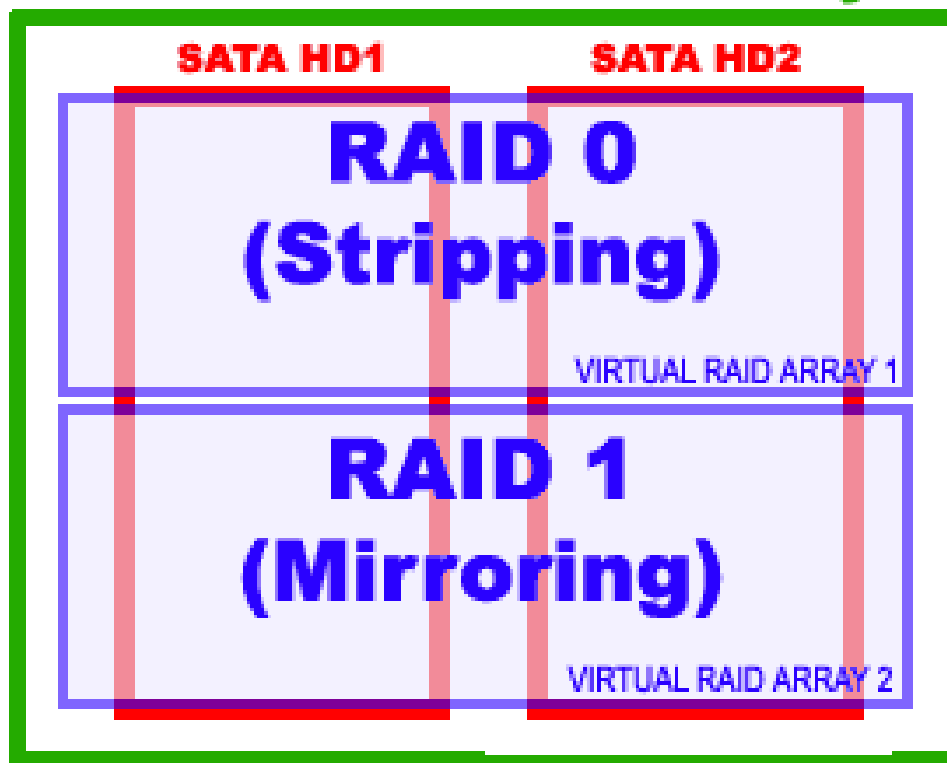
Данный вариант не является стандартным уровнем RAID, но упомянут тут в связи с тем, что систем с его поддержкой существует огромное множество.

Matrix RAID разработан Intel и реализован во всех настольных чипсетах с южными мостами с суффиксом “R” (“DH”, “OH”, “M”). Суть этого решения – в реализации нескольких сегментов RAID-массива на одном и том же диске. В данном случае мы наблюдаем RAID 0+1 на четырех дисках, но с распределением не по 2, а по 4 дискам сразу, что дает в идеальном случае увеличение быстродействия в 2 раза.

Matrix позволяет реализовать несколько вариантов RAID, а не только 0+1. Основное его преимущество – возможность экономии физических дисков (потери дискового пространства все равно остаются). Это актуально прежде всего для настольных систем или серверов высокой плотности.

Matrix RAID - одновременно массив уровней RAID 0 и RAID 1 при использовании только двух дисков с интерфейсом Serial ATA.

Intel ICH6 Matrix RAID Array



EXAMPLE :

SATA HD1 = 80 GB

SATA HD2 = 80 GB

C:\ - VIRTUAL RAID ARRAY 1 = 80 GB / RAID 0

D:\ - VIRTUAL RAID ARRAY 2 = 40 GB / RAID 1

Таблица характеристик RAID

Уровень RAID	Мин. кол-во дисков	Возможность отказа	Скорость чтения	Скорость записи	Ёмкость, % от суммы всех дисков
RAID 0	2	нет	x N	x N	100%
RAID 1	2	число зеркал	x N	как у диска	50%
RAID 1E	3	число зеркал	x N	почти x N	50%
RAID 5	3	1	x (N-1)	низкая	67-94%
RAID 5EE	4	1	x (N-2)	низкая	50-88%
RAID 6	4	2	x (N-2)	низкая	50-88%
RAID 6EE	5	2	x (N-3)	низкая	40-75%
RAID 0+1	4	1	x (N/2)	x N/2	50%
RAID 10	4	1 на RAID 1	x M (N/2)	x N/2	50%
RAID 50	6	1 на RAID 5	x (N-M)	низкая	67-94%
RAID 60	8	2 на RAID 6	x (N-2M)	низкая	50-88%

N – общее число дисков в массиве, M – число подмассивов

Контролер RAID

Существует несколько вариантов реализации поддержки RAID со стороны контроллера жестких дисков:

- Программная реализация.
- Встроенный в чипсет контроллер PCI IDE.
- Отдельная плата расширения с интерфейсом PCI/PCI Express.
- Внешний модуль (может входить в состав системы хранения данных типа NAS, SAN и др.).

Программная реализация обычно имеет вид надстройки (промежуточного слоя) для драйверов, реализующих работу с хост-контроллером жестких дисков. Уровни RAID 0 и RAID 1 реализуют практически все современные ОС, другие уровни требуют усложнения логики доступа к дискам, а потому программно не реализуются.

Встроенные контроллеры ограничены в ресурсах, а потому реализуют уровни 0, 1, реже 5, 0+1 и 10. При этом не обеспечивается эффективное кэширование, из-за чего производительность невысока.

Контроллер RAID

Классический RAID-контроллер представляет собой либо плату расширения, либо модуль во внешнем исполнении (напр., встроенный в систему хранения данных).

В состав типичного контроллера входят:

- Микроконтроллер или микропроцессор (зачастую универсального назначения), который выполняет основные функции управления и поддержки массивов.
- Интерфейсная часть (системной шины, интерфейсов жестких дисков), может входить в состав микроконтроллера.
- Память большого объема для реализации кэширования (может иметь модульную конструкцию).
- Флэш-ПЗУ для хранения микропрограммы, BIOS, Setup.
- Индикаторы и/или динамики, разъемы для подключения средств индикации.
- Разъем для подключения аккумулятора.

Функции RAID-контролера

- Обслуживание массивов: создание, поддержка работы, мониторинг, восстановление с применением резервного диска, конвертирование в другой уровень реального времени.
- Настройка параметров при помощи Setup (доступен на этапе POST) и/или графического интерфейса (обычно – с веб-интерфейсом).
- Уведомление администратора об ошибках (звуком, светом, по email и т.д.), индикация порта, к которому подключен сбойный диск.
- Ведение журналов работы.

К основным операциям, с помощью которого RAID-контроллер оптимизирует доступ к дискам массива, относятся упреждающее чтение, отложенная запись и кэширование всех операций. При этом для защиты массивов от повреждений при пропадании питания RAID-контроллер может комплектоваться аккумулятором, который обеспечивает сохранность содержимого памяти (кэша) в течение нескольких суток – до восстановления энергоснабжения дисков.

Заключение —

псевдо RAID-массивы

- **MULTIPATH** — массив, позволяющий создавать разные псевдо-дисковые устройства для одного физического диска;
- **FAULTY** — псевдо RAID-массив.

Уменьшение роли RAID5 в области использования жестких дисков HDD.

Современные RAID-контроллеры: новые возможности. Автор:

Дмитрий Зотов, инженер, PMC-Sierra 24.02.2013

<http://www.bytemag.ru/articles/detail.php?ID=21362>


- интегрированные на материнские платы для домашних ПК RAID-контроллеры поддерживают далеко не все RAID-уровни. Двухпортовые RAID-контроллеры поддерживают только уровни 0 и 1, а RAID-контроллеры с большим количеством портов (например, 6-портовый RAID-контроллер, интегрированный в южный мост чипсета ICH9R/ICH10R) — также уровни 10 и 5.
- Кроме того, если говорить о материнских платах на чипсетах Intel, то в них тоже реализована функция Intel Matrix RAID, которая позволяет создать на нескольких жестких дисках одновременно RAID-матрицы нескольких уровней, выделив для каждой из них часть дискового пространства.

- **Рост популярности гибридных томов.** В широком смысле гибридный том – это любой том, где одновременно используются и традиционные жесткие диски (HDD), и твердотельные SSD. В силу этого такое решение, как SSD-кэширование, тоже является одним из вариантов реализации гибридного тома. В RAID-контроллерах Adaptec функция «гибридный том» (Hybrid Volume) подразумевает специальный режим для томов RAID1, 10, где используются как HDD-, так и SSD-диски.
- Сам рост популярности гибридных томов объясняется довольно просто. SSD-диски в чистом виде не находят широкого применения, поскольку ряд их серверных свойств пока находится в разработке. Цена SSD-дисков довольно высока. Но в то же время SSD-решения обладают уникальной производительностью. Гибридные тома позволяют добавить надежности и емкости со стороны HDD-дисков, производительности со стороны SSD-дисков и оптимизировать цену такого решения.

Возможность использовать RAID-тома SSD

Хорошим примером уже имеющегося на рынке контроллера, который разработан с учетом широкого использования SSD-дисков, является 7 серия контроллеров Adaptec. Ядро контроллера имеет показатель на уровне больше 500 тыс. IOPS для случайного трафика и около 6 Гбайт/с для последовательных шаблонов трафика. Такие показатели позволяют использовать SSD-диски средней категории производительности (наиболее популярных моделей) в количестве, по крайней мере равном количеству портов на контроллере (8–16 SSD).

- Вполне очевидно, что следующим шагом к поддержке SSD-томов будет использование технологии SAS3 – 12 Гбит/с.
- Подчеркнем, что есть ряд факторов, которые пока препятствуют широкому распространению SSD RAID-томов и вытеснению HDD RAID-томов и гибридных томов. К этим факторам относятся высокая стоимость SSD, их низкая емкость, гарантированная остановка записи при записи определенного количества информации, оптимизация алгоритмов кэширования контроллеров для работы с SSD, самое начало внедрения поддержки TRIM-команд для RAID-контроллеров и т.п.

- 
- **Дополнительные функции RAID-томов**
 - Рассмотрим функции RAID-томов, влияющие на их производительность или надежность.
 - **Copy Back Hot Spare**
 - **Защита кэширования через оперативную память с помощью модуля AFM**
 - **SSD-кэширование**
 - **Функция Power Management**
 - **Управление**

RAID TP

обеспечивает сохранность и доступность информации при одновременном выходе из строя трех любых дисков;