Раздел 2. Лекция 2.

Магнитные и твердотельные накопители. Часть 2

Основные вопросы лекции

1. Накопители на твердотельных дисках. Гибридные жесткие диски. Накопители на базе флэш-памяти. Перспективы применения новых технологий энергонезависимой памяти для хранения данных.
2. Накопители на магнитной ленте, основные разновидности, характеристики, интерфейсы. Конструкция и принцип действия накопителей на магнитной ленте.
3. Дисковые массивы. Технология RAID.

1. Твердотельный накопитель (SSD)

SSD (Solid-state drive) — компьютерное немеханическое ЗУ на основе микросхем памяти. Кроме них, SSD содержит управляющий контроллер, который управляет процессом чтения / записи и структурой размещения данных, для кеш-памяти используется микросхема DDR DRAM.

Различают два вида твердотельных накопителей:

- SSD на основе памяти, подобной оперативной памяти компьютеров;
- SSD на основе флеш-памяти.

В настоящее время SSD используются в компактных устройствах: ноутбуках, нетбуках, коммуникаторах и смартфонах, но могут быть использованы и в системных блоках для повышения производительности. Некоторые известные производители переключились на выпуск твердотельных накопителей уже полностью, например Samsung продал бизнес по производству жѐстких дисков компании Seagate.

1. Принцип работы SSD

2. Гибридный жесткий диск (SSHD — Solid State Hybrid Drive

Такие устройства сочетают в одном устройстве накопитель на жестких магнитных дисках (HDD) и SSD относительно небольшого объема, в качестве кэша (для увеличения производительности и срока службы устройства, снижения энергопотребления).

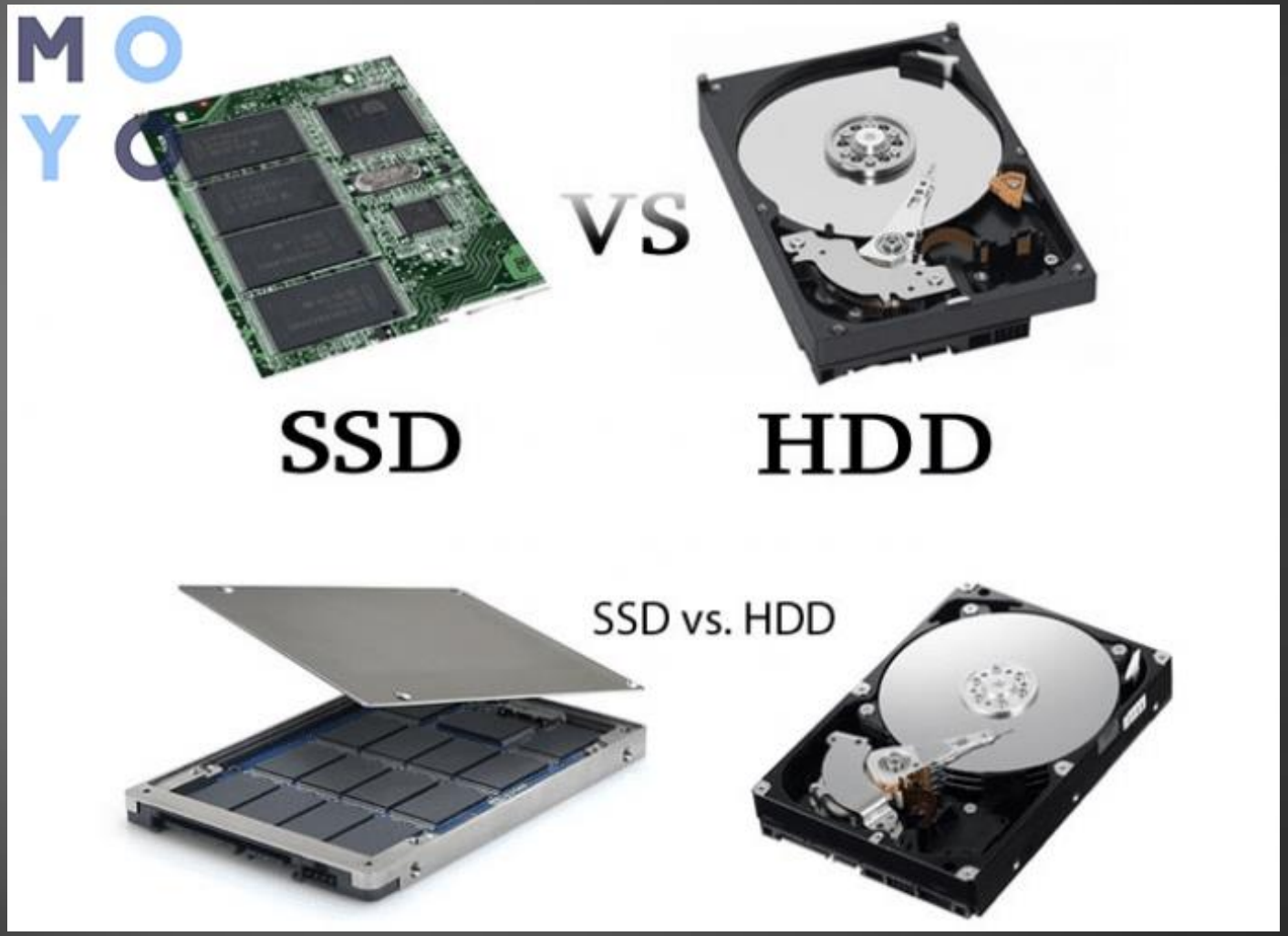
Используются, в основном, в переносных устройствах (ноутбуках, сотовых телефонах, планшетах и т. п.).

Гибридные жесткие диски являются промежуточным решением между SSD и HDD. За счет наличия дополнительной Flash-памяти они способны буферизовать наиболее часто используемые файлы и впоследствии обращаться к ним с гораздо более высокой скоростью. Алгоритм Adaptive Memory, записывающий наиболее часто используемые данные в твердотельный кэш объемом 8 Гбайт. - позволяет заметно ускорить работу операционной системы и часто используемых приложений.

Производители:

- Seagate Laptop Thin SSHD (2,5”)
- Desktop SSHD (3,5”)
- TOSHIBA

1. Физическое исполнение SSD и HDD



1. Плюсы и минусы SSD

+ Высокая скорость чтения и записи (в разы по сравнению с HDD).
+ Относительно низкое энергопотребление.
+ Полное отсутствие шума и вибрации.
+ Менее чувствительны к механическим воздействиям и внешним электромагнитным полям.
+ Более широким диапазоном рабочих температур.
+ Низкое тепловыделение, что способствует улучшению производительности.

- Ограниченное количество циклов перезаписи (10 000 - 100 000).
- Высокая стоимость.
- Проблемы с восстановлением данных после резкого скачка напряжения и др.

1. RAM drive

RAM drive, RAM disk (диск в памяти), электронный диск — компьютерная технология, позволяющая хранить данные в быстродействующей оперативной памяти как на блочном устройстве (диске). Может быть реализована как программно, так и аппаратно.

Основные достоинства:

- Крайне высокая скорость чтения (измеряется гигабайтами в секунду);
- Крайне высокая скорость (IOPS - операций ввода-вывода в секунду). Некоторые образцы оперативной памяти типа DDR3 позволяют достигать более 1 000 000 IOPS. Для сравнения IOPS современных жестких дисков составляет 20-300. IOPS NAND SSD накопителей 700-100 000.
- Отсутствие задержек при произвольном доступе;
- Реализация без использования дополнительных аппаратных компонентов;
- Цена за гигабайт сопоставима с ценой за гигабайт NAND SSD накопителя.

Основные недостатки:

- Потребление крайне ценного ресурса (оперативной памяти);
- Малые ёмкости (при наличии на рынке жестких дисков в 500—4000Гб, модули оперативной памяти исчисляются гигабайтами);
- Потеря содержимого при отключении подачи напряжения (решается сохранением содержимого на диске при выключении, однако риск есть).

1. Примеры реализаций RAM drive

MS-DOS - RAMDRIVE.SYS — драйвер операционной системы

- COMBI.SYS — драйвер, созданный для реализации максимально эффективного использования памяти, задействованной для электронного диска. Свободное пространство электронного диска, созданного этой программой, использовалось как кэш для жѐсткого диска.

RAMDisk — от Dataram для Windows 9x, 2000, XP, Vista, Seven, Server 2000, 2003, 2008. Поддержка 32-х и 64-битных версий.

Linux реализует три вида ram-disk:

- Специализированный архив в формате cpio для размещения модулей для начальной загрузки (initrd).
- Файловая система, размещающаяся в памяти tmpfs (используется чаще всего для хранения временных данных, сохранение которых не актуально между перезагрузками и к которым нужен быстрый доступ).
- Блочный ramdisk (модуль brd), позволяющий создавать блочные устройства (вида /dev/ram0).

FreeBSD - Поддержка RAM-диска встроена в базовую систему, реализуется драйвером md(4), настраивается программой mdconfig(8).

1. Flash-память

Flash – «быстрый, мгновенный» при описании своих новых микросхем.
Изобретателем считается **Intel**, представившая в 1988 году флэш-память с архитектурой NOR.
Годом позже **Toshiba** разработала архитектуру NAND, которая и сегодня используется наряду с той же NOR в микросхемах флэш.

Собственно, сейчас можно сказать, что это два различных вида памяти, имеющие в чем-то схожую технологию производства.

1. Особенности Flash-памяти

Среди главных достоинств можно назвать следующие:

- энергонезависимость, т.е. способность хранить информацию при выключенном питании (энергия расходуется только в момент записи данных);
- информация может храниться очень длительное время (десятки лет);
- сравнительно небольшие размеры;
- высокая надежность хранения данных, в том числе устойчивость к механическим нагрузкам;
- не содержит движущихся деталей (как в жестких дисках).

Основные недостатки флэш-памяти:

- невысокая скорость передачи данных (в сравнении с динамической оперативной памятью);
- незначительный объем (по сравнению с жесткими дисками);
- ограничение по количеству циклов перезаписи (хотя эта цифра в современных разработках очень высока – более миллиона циклов).

1. Базовые элементы Flash-памяти

Флэш-память строится на однотранзисторных элементах памяти с «плавающим» затвором, что обеспечивает высокую плотность хранения информации.

Существуют различные технологии построения базовых элементов флэш-памяти, разработанные ее основными производителями. Эти технологии отличаются количеством слоев, методами стирания и записи данных, а также структурной организацией, что отражается в их названии.

Наиболее широко известны NOR и NAND типы флэш-памяти, запоминающие транзисторы в которых подключены к разрядным шинам, соответственно, параллельно и последовательно.

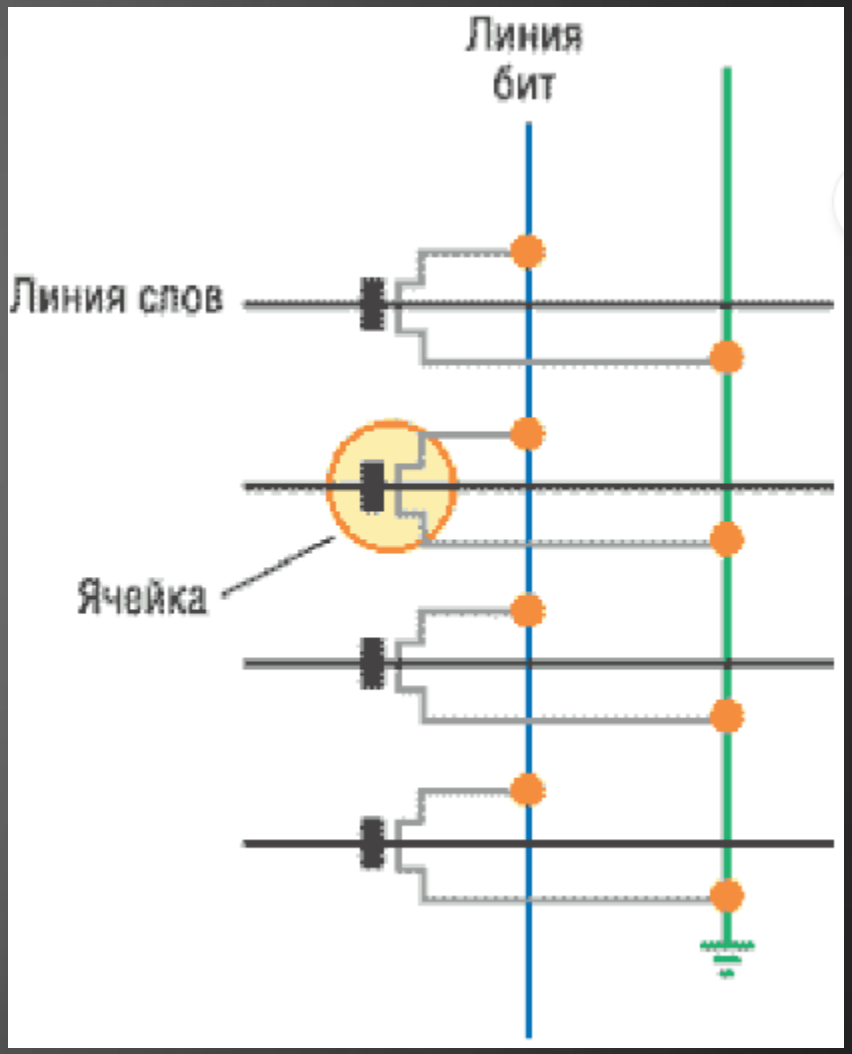
1. Архитектура Flash-памяти (NOR)

память на основе ячеек NOR (логическая функция ИЛИ-НЕ).

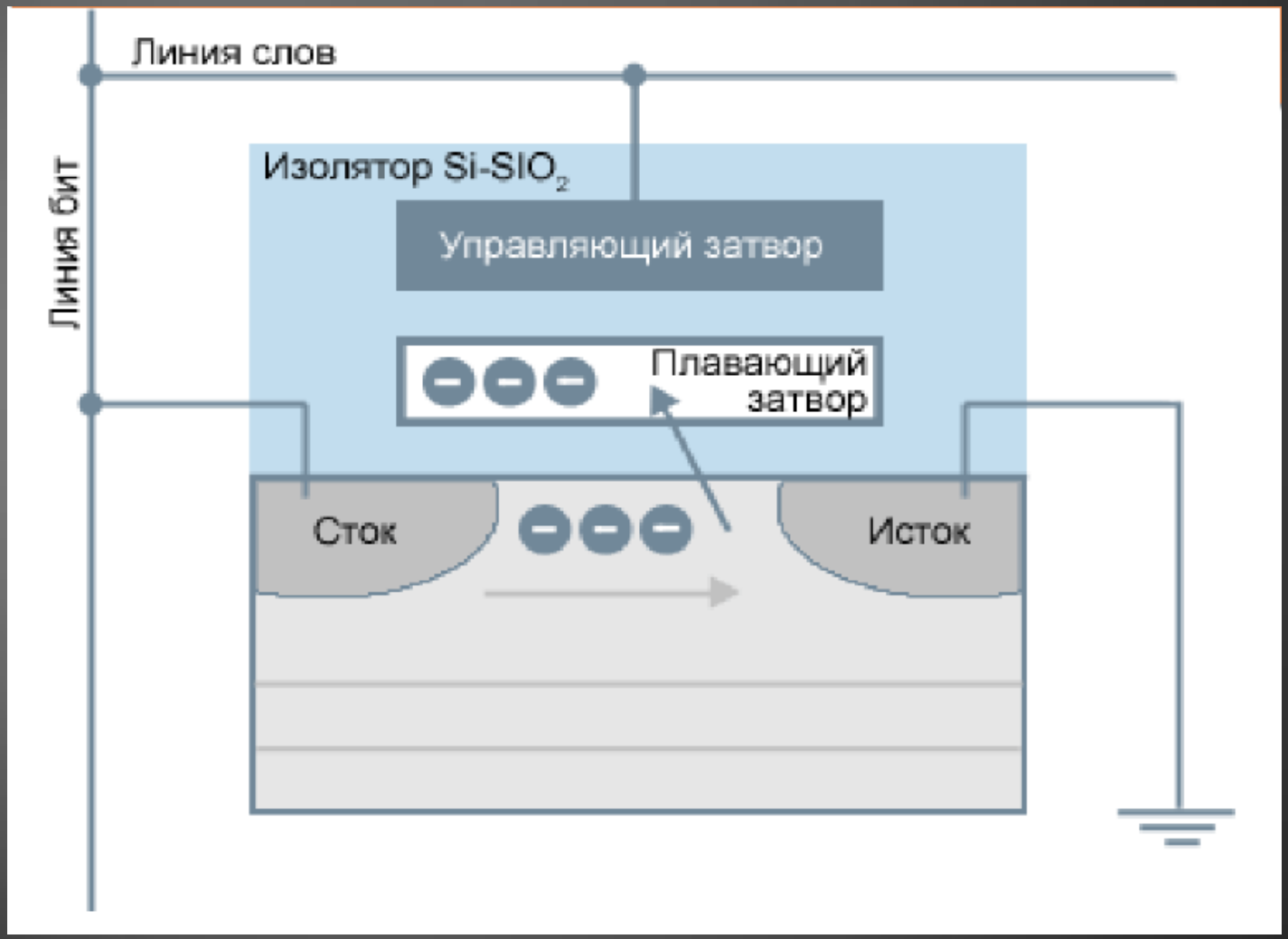
Структура NOR состоит из параллельно включенных элементарных ячеек хранения информации.

Такая организация ячеек обеспечивает произвольный

доступ к данным и побайтную запись информации.



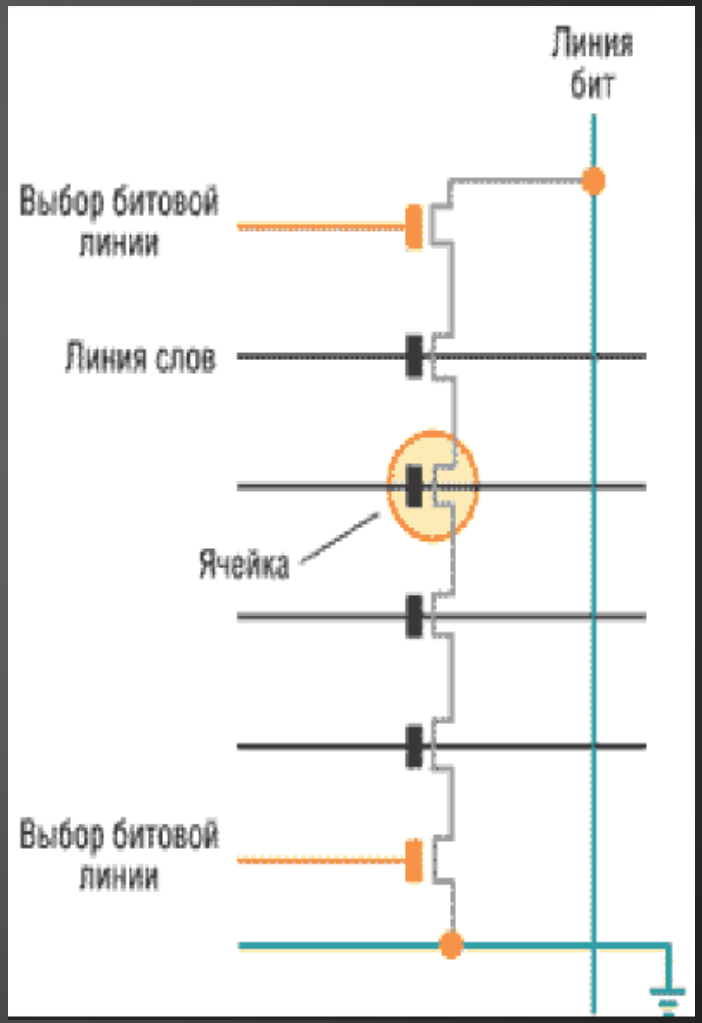
1. Схема ячейки NOR



1. Схема ячейки NAND

В основе структуры NAND лежит принцип последовательного соединения элементарных ячеек, образующих группы (по 16 ячеек в одной группе), которые объединяются в страницы, а страницы - в блоки. При таком построении массива памяти **обращение к отдельным ячейкам невозможно.**

Программирование выполняется одновременно только в пределах одной страницы, а при стирании обращение происходит к блокам или к группам блоков.



1. Различия структуры памяти NOR и NAND

Различия в организации структуры между памятью NOR и NAND находят свое отражение в их характеристиках.

При работе со сравнительно большими массивами данных процессы записи/стирания в памяти NAND выполняются значительно быстрее, чем в памяти NOR. Поскольку 16 прилегающих друг к другу ячеек памяти NAND соединены последовательно, без контактных промежутков, достигается высокая плотность размещения ячеек на кристалле, что позволяет получить большую емкость при одинаковых технологических нормах. Последовательная организация ячеек обеспечивает высокую степень масштабируемости, что делает NAND-флэш лидером в гонке наращивания объемов памяти.

1. Различия структуры памяти NOR и NAND

В структуре флэш-памяти для хранения 1 бита информации задействуется только один элемент (транзистор), в то время как в энергозависимых типах памяти для этого требуется несколько транзисторов и конденсатор. Это позволяет существенно уменьшить размеры выпускаемых микросхем, упростить технологический процесс, а следовательно, снизить себестоимость. Но и 1 бит - далеко не предел.

Еще в 1992 г. команда инженеров корпорации Intel начала разработку устройства флэш-памяти, одна ячейка которого хранила бы более одного бита информации. Еще в сентябре 1997 г. была анонсирована микросхема памяти Intel StrataFlash емкостью 64 Мбит, одна ячейка которой могла хранить 2 бита данных.

1. Применение Flash-памяти

Современные технологии производства флэш-памяти позволяют использовать ее для различных целей. Непосредственно в компьютере эту память применяют для хранения BIOS (базовой системы ввода-вывода), что позволяет, при необходимости, производить обновление последней, прямо на рабочей машине.

Распространение получили, так называемые, USB-Flash накопители, эмулирующие работу внешних винчестеров. Эти устройства подключается, обычно, к шине USB и состоит из собственно флэш-памяти, эмулятора контроллера дисковода и контроллера шины USB. При включении его в систему (допускается "горячее" подключение и отключение) устройство с точки зрения пользователя ведет себя как обычный (съемный) жесткий диск.

Конечно, производительность его меньше, чем у жесткого диска.

1. Заключение о Flash-памяти

Существуют вполне реальные планы перехода от динамической регенерируемой памяти к памяти энергонезависимой (NV-RAM) – FeRAM, MRAM, PCM (PCRAM) и т.д.

Память типа NOR Flash, которая ныне применяется во многих мобильных устройствах, на эту роль не подходит ввиду высокой технологической сложности и недостаточной надежности.

2. История ленточных накопителей

Первый ленточный накопитель с лентой шириной 1⁄2” (12,5 мм) был применен еще в 1951 году для подключения к ЭВМ UNIVAC I. На металлическую ленту с никелевым покрытием записывались 8 дорожек (6 информационных) со скоростью 128 Кб/с.

В 50-80-х годах компания IBM стала одним из лидеров в разработке различных ленточных накопителей с использованием гибкой ленты с ферритовым покрытием (оксиды железа). Применялись 1⁄2” ленты на бобинах (без кассет или картриджей) с различным количеством дорожек.

Начиная с середины 70-х ленточные накопители применяются для ЭВМ малого класса, рабочих станций и даже персональных ЭВМ. Появляются форматы ленты меньшей ширины, упакованной кассеты, улучшаются методы записи, разметки, перемотки и т.д.

Расцвет технологии пришелся на середину 90-х.

2. Принцип записи на магнитную ленту

Магнитная лента представляет собой гибкую ленту из того или иного материала, на которую с одной стороны нанесено покрытие из ферромагнетика. Применение вакуумного напыления позволяет надежно фиксировать покрытие (без адгезивного материала), что повышает устойчивость ленты к износу.

Запись данных на ленту выполняется в виде дорожек, разделенных зазорами.

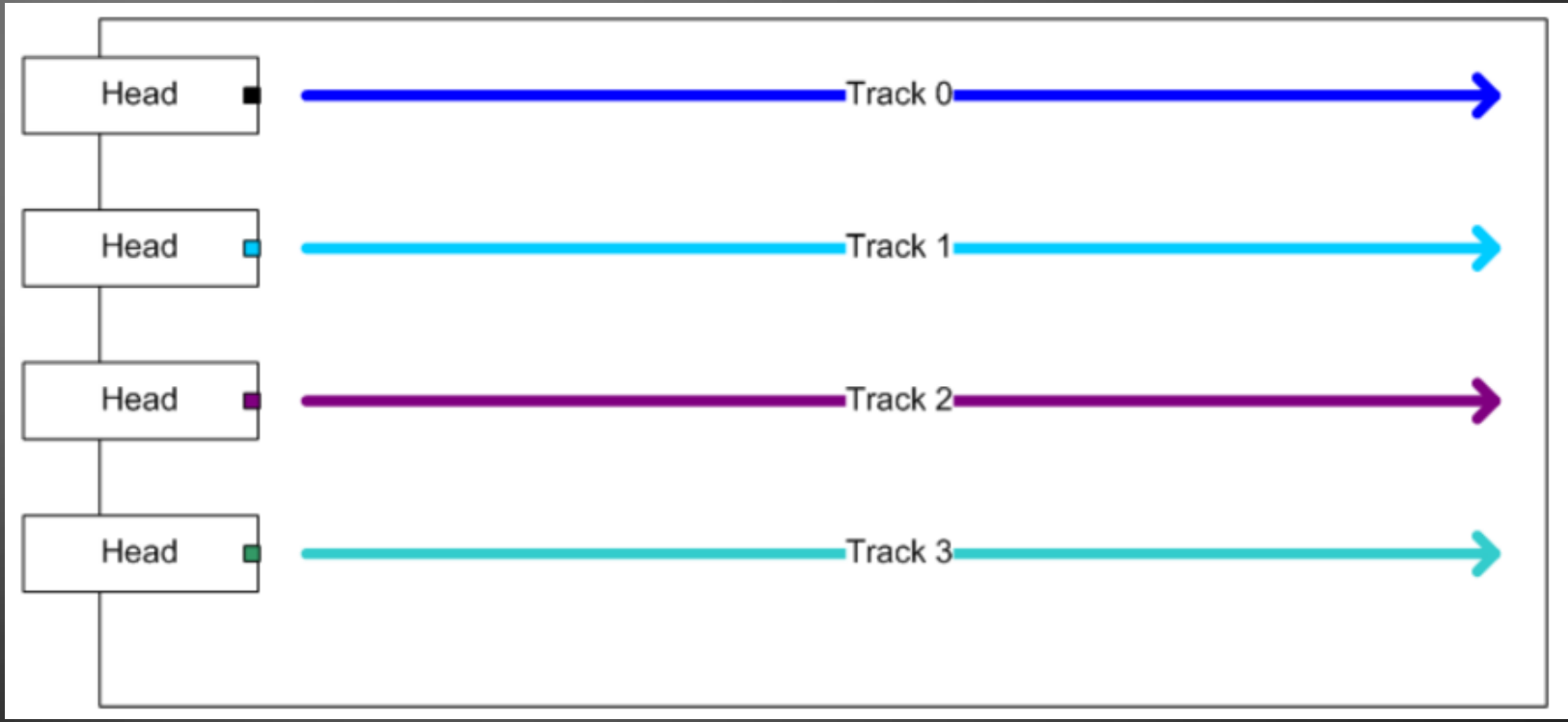
Разработчики из разных компаний не только предлагали свой метод записи, свой **типоразмер и конструкцию** картриджа или кассеты, но и в ходе наращивания емкости носителей меняли (зачастую кардинально) основные **параметры устройств**, сохраняя в основном только типоразмеры (и то не всегда). За совместимостью и выработкой единого стандарта никто не следил (за исключением последних лет).

Устройства можно разделить по двум признакам:

1. Линейный или наклонно-строчный способ записи.
2. Ширина ленты.

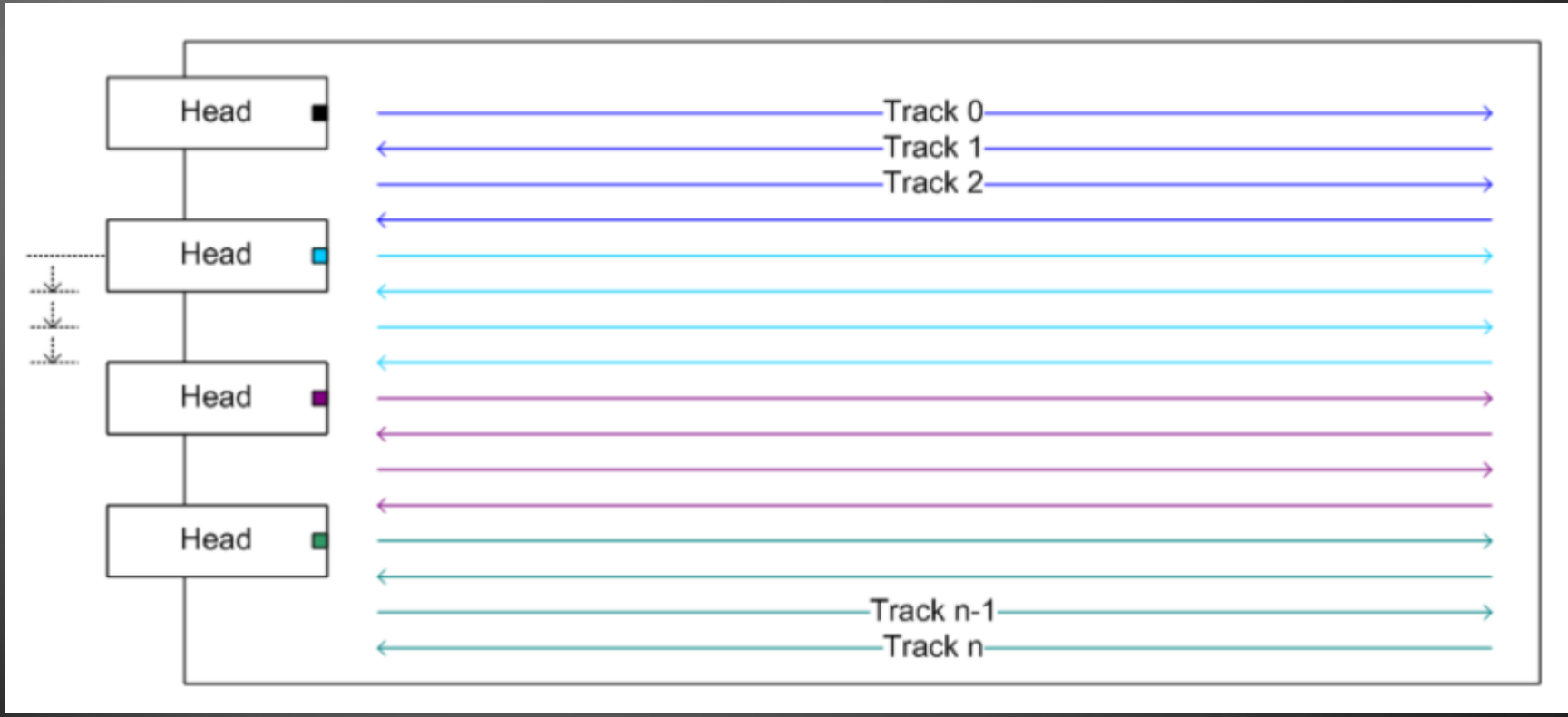
2. Линейная запись

Неподвижная головка формирует дорожку вдоль всей ленты. Как правило, на ленте можно разместить несколько дорожек с зазорами между ними (в один проход несколькими головками или в несколько проходов).



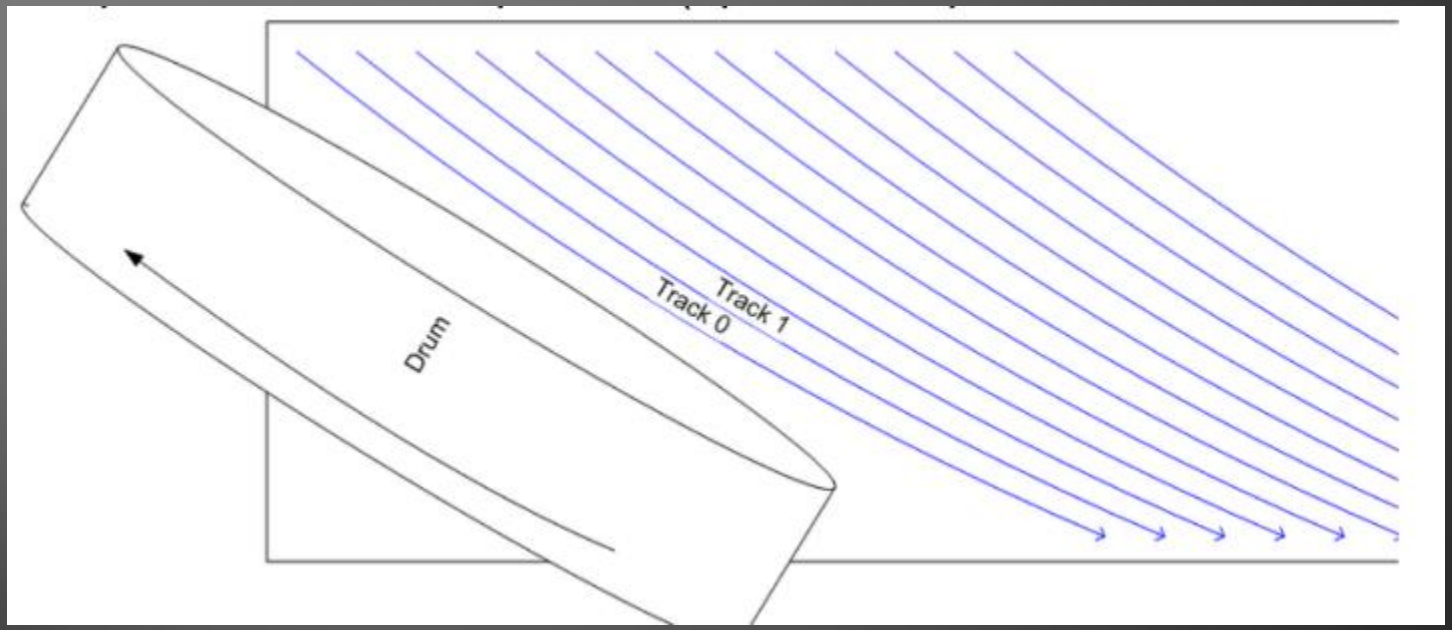
2. Линейная с серпантином

Вариант линейной записи – головка выполняет запись в обратном направлении, но с небольшим смещением. Таким образом можно записать более 100 дорожек на ленту.



2. Наклонно-строчная запись

(Helical) Пишущие головки выполняют запись диагональных штрихов, что позволяет за один проход заполнять максимальную полезную площадь ленты. Для реализации этого метода головка выполнена в виде вращающегося барабана (против направления движения ленты).



2. LTO – универсальный стандарт

Сегодня самым распространенным (и практически единственным) стандартом ленточных накопителей стал стандарт LTO (Linear Tape-Open). Он появился в 2000 году, и изначально разрабатывался компаниями IBM, Hewlett-Packard и Quantum. Позже к этому стандарту присоединились другие производители. Поэтому картриджи и приводы (накопители) LTO разных производителей полностью совместимы.

За прошедшее время сменилось несколько поколений LTO. Сейчас (2022 год) последним является девятое поколение – LTO-9.

Стандарт	LTO-5	LTO-6	LTO-7	LTO-8	LTO-9	LTO-10
Год появления	2010	2012	2015	2017	2021	план 2022
Физическая ёмкость	1.5TB	2.5TB	6TB	12.8TB	18TB	36TB
Максимальная скорость (МБ/с)	140	160	300	900-920	1000	2750

Картриджи девятого поколения имеют 18ТВ физической ёмкости. Информация может записываться на картриджи со сжатием (упаковкой). Считается, что в коэффициент сжатия общем случае составляет 2,5:1, поэтому иногда для картриджей указывается их ёмкость с учетом сжатия, например – 45ТБ для картриджей LTO-9.

Стандарт LTO регламентирует для приводов поддержку чтения на 2 поколения назад и записи на 1 поколение назад. Поэтому привод LTO-9, например, умеет читать картриджи LTO-9, LTO-8 и LTO-7, а писать на картриджи стандартов LTO-9 и LTO-8.

2. Конструктивные особенности

1. Шпиндель («бобина») – лента накручена на втулку, для ее обработки требуется закрепить свободный конец на вращающейся втулке устройства.	**2. Картридж** – пластиковая упаковка с одной втулкой, на которую накручена лента. Вторая втулка находится внутри устройства, привод вращает обе втулки при работе с лентой. Является развитием первого варианта.	**3. Кассеты** – в одном корпусе две втулки с лентой. Обеспечивает лучшую защиту для ленты при транспортировке. Вращение ленты обеспечивается либо за счет непосредственного вращения втулок, либо за счет другого типа привода (например, трения вращающегося ремня).
		

2. Картриджи

В устройствах LTO используются картриджи (кассеты) следующих типов:

- **RW** (англ. ReWritable) — лента для многократной записи.
- **WORM** (англ. Write Once, Read Many) — картриджи со специальной электронной схемой, допускающей только однократную запись и многократное чтение.
- **UCC** (англ. Universal Cleaning Cartridge — чистящие картриджи), совместимые со всеми устройствами, для проведения технического обслуживания привода (чистка головок чтения/записи привода).

Магнитная лента картриджа содержит по своей ширине несколько сотен (и даже тысяч) дорожек. За один проход головки привода захватывают несколько десятков дорожек. Таким образом, чтобы полностью пройти всю ленту (весь картридж), требуется несколько десятков или даже сотен проходов. На торце картриджа может быть наклеена метка (label) - штрих-код, который маркирует картридж и используется в библиотеках для выбора картриджа (о библиотеках см. ниже). Наработка на отказ одного картриджа – около 250 циклов (полных проходов чтения/записи всей ленты). Наработка на отказ чистящих картриджей – 50 циклов.

2. Приводы

Приводы LTO состоят из лентопротяжного механизма и головок чтения/записи. Приводы различаются по:

- **стандарту LTO**. Стандарт LTO требует от приводов читать картриджи на 2 поколения назад и записывать на 1 поколение назад. Поэтому привод LTO-9, например, умеет читать картриджи LTO-9, LTO-8 и LTO-7, и писать на картриджи стандартов LTO-9 и LTO-8.
- **интерфейсу подключения**. Приводы LTO имеют один из двух интерфейсов подключения: SAS 6Gb или FC 8Gb. Более быстрых стандартов не бывает, т,к. привод не может писать быстрее, и использование быстрого интерфейса не требуется.
- **форм-фактору**. Приводы бывают двух форм-факторов: FH и HH. Привод форм-фактора FH (Full-Height) занимает два стандартных отсека 5,25”. Привод HH (Half-Height) занимает только один такой отсек. Раньше приводы FH были более производительными, но сейчас разницы в производительности уже нет.

2. Типы устройств

Все устройства хранения на магнитной ленте можно разделить на несколько типов:

- внутренние;
- внешние одиночные приводы;
- автозагрузчики;
- библиотеки.

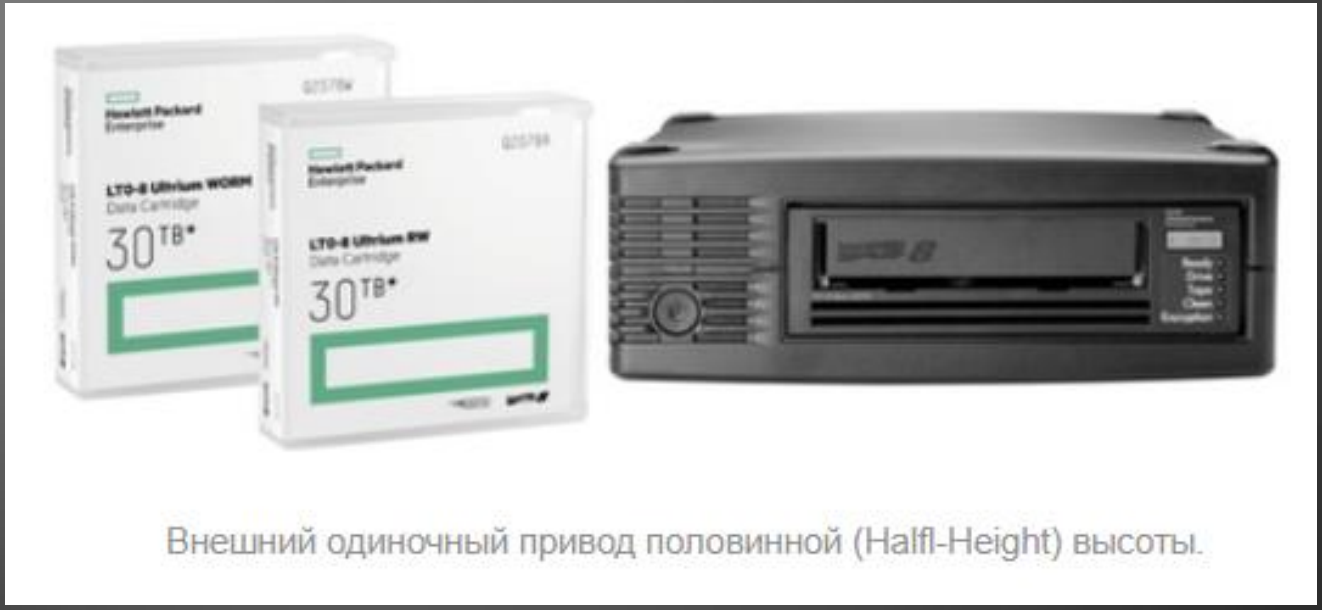
2. Внутренние приводы

Внутренние приводы (Internal tape drive) устанавливаются непосредственно в корпус сервера или компьютера. Последнее время они почти не используются, так как в стоечных серверах перестали делать отсеки 5.25”, а в случае напольных серверов конечные пользователи предпочитают использовать другие носители: внешние диски или USB-накопители.

Внутренний привод полной (Full-Height) высоты.

2. Внешние одиночные приводы

Внешние одиночные приводы (Standalone tape drive) выполнены в виде отдельного выносного корпуса. Они бывают только с подключением SAS, и подключаются с помощью SAS-кабеля к SAS HBA, установленному в сервере. По необходимости картридж заменяется вручную. Устройство не устанавливается в стойку и существует только настольном варианте.



Внешний одиночный привод половинной (Halfl-Height) высоты.

2. Автозагрузчики

Внутренние и внешние одиночные приводы не имеют автоматизации. Каждый раз пользователь должен вставлять и вынимать картриджи самостоятельно.

Автозагрузчик (autoloader) – это устройство с одним приводом и автоматизированной системой хранения картриджей. Система может хранить несколько картриджей в магазинах. Внутри устройства робот-автомат выбирает картриджи из магазинов и вставляет в привод для чтения или записи.

Обычно на картриджи наклеиваются штрих-коды (label) – метки. По этим штрих-кодам робот-автомат различает картриджи (считывает фотоэлементом), а управляющее программное обеспечение ведет по этим штрих-кодам каталог (какая информация хранится на каждом картридже) и отдаёт соответствующие команды роботу.

Автозагрузчики обеспечивают автоматизацию процессов резервного копирования, и поэтому применяются в тех случаях, когда объём информации не очень высок (для больших объёмов используются библиотеки), но требуется частое резервное копирование.

Приводы, которые устанавливаются в автозагрузчики, могут иметь интерфейс SAS либо FC 8Gb.

2. Библиотеки

Если автозагрузчик всегда имеет только один привод чтения/записи, то библиотека рассчитана на использование нескольких приводов. Кроме того, они обычно рассчитаны на большее количество картриджей и имеют возможности расширения путём подключения модулей расширения с картриджами и дополнительными приводами.

Например, головной блок библиотеки начального уровня HPE MSL 3040 (на фото) имеет 40 слотов под картриджи, а с помощью шести модулей расширения количество картриджей можно увеличить до 280 штук. Кроме того, каждый модуль расширения может также содержать приводы, и их количество может быть увеличено до 21 штуки (3 штуки в головном блоке библиотеки и по 3 штуки в каждом из 6 модулей расширения). Приводы, которые устанавливаются в библиотеку, могут иметь интерфейс SAS либо FC 8Gb – причем в одной библиотеке могут использоваться приводы с разными интерфейсами.

2. Распространенные форматы

1. Линейный формат записи:

1. Лента 12.7 мм (1/2”): DLT (Digital Linear Tape), LTO Ultrium (Linear Tape Open), T1000 (StorageTek), Jaguar (IBM).

2. Лента 8 мм (1/3”): Travan (3M).

3. Лента 6.35 мм (1/4”): QIC/SLR (Quarter-Inch Cartridge/Scalable Linear Recording).

2. Наклонно-строчный формат записи:

1. Лента 12.7 мм (1/2”): SAIT (Sony) - Advanced Intelligent Tape.

2. Лента 8 мм (1/3”): AIT (Sony), VXA (Exabyte).

3. Лента 3.8 мм (1/8”): DDS/DAT (Digital Audio Tape).

Наиболее распространенным сегодня форматом (по количеству существующих решений):

- в области «больших» ленточных библиотек является LTO Ultrium (ближайший конкурент – DLT);
- в области локальных и компактных систем резервного копирования – DDS/DAT.

3. Дисковые массивы

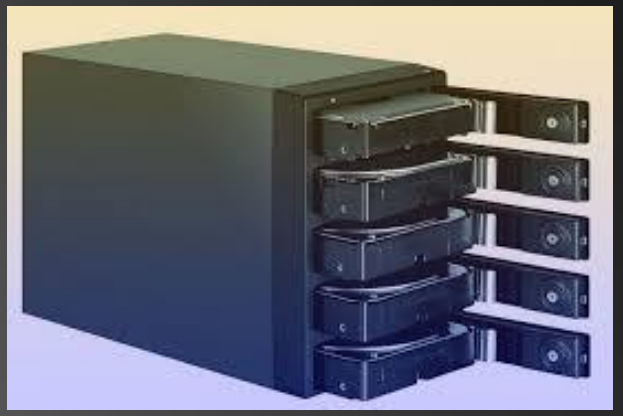
Дисковым массивом (Disc Array) называют набор жестких дисков, подключенных к одному многопортовому контроллеру. В простейшем случае контроллер интерпретирует их как независимые накопители, которые ОС может использовать для размещения логических разделов. Такой массив называется **JBOD** (Just a Bunch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

3. 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.

3. 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 %.

Поговорим далее о:

функциональный сбой и скрытая (или отложенная) ошибка.

3. 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.

3. 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,
- коррозийные 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.

3. 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.

3. 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.

3. 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.

3. 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).

3. 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 A, B, C 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.

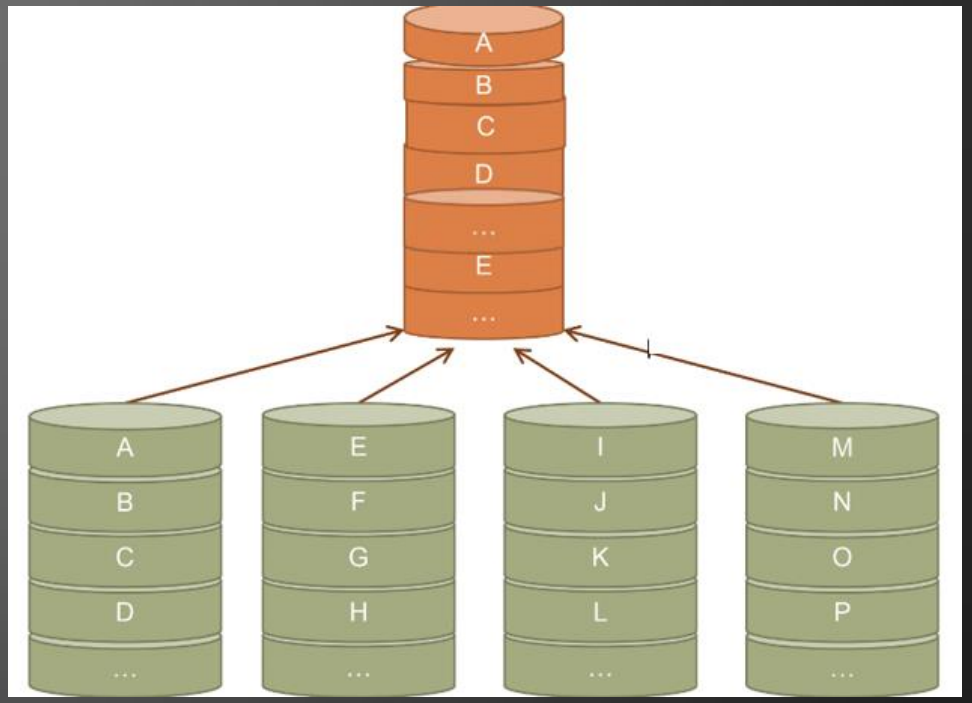
Страйп (stripe, по аналогии bit-byte) – это сумма всех стрипов с каждого из дисков массива. Размер страйпа также важен, т.к. он определяет, запрос какого размера может быть выполнен параллельно всеми дисками.

3. 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.

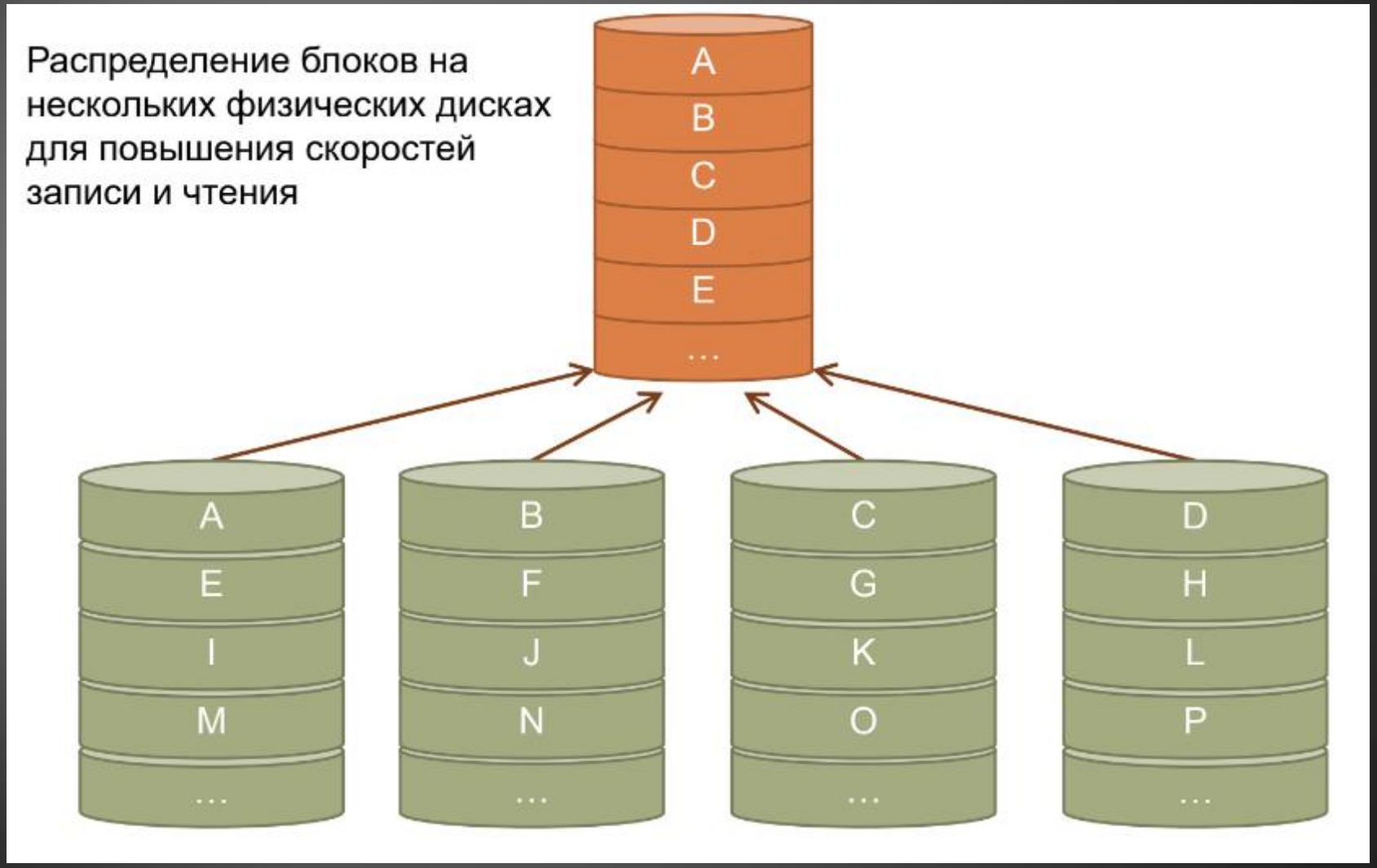


3. Основные понятия

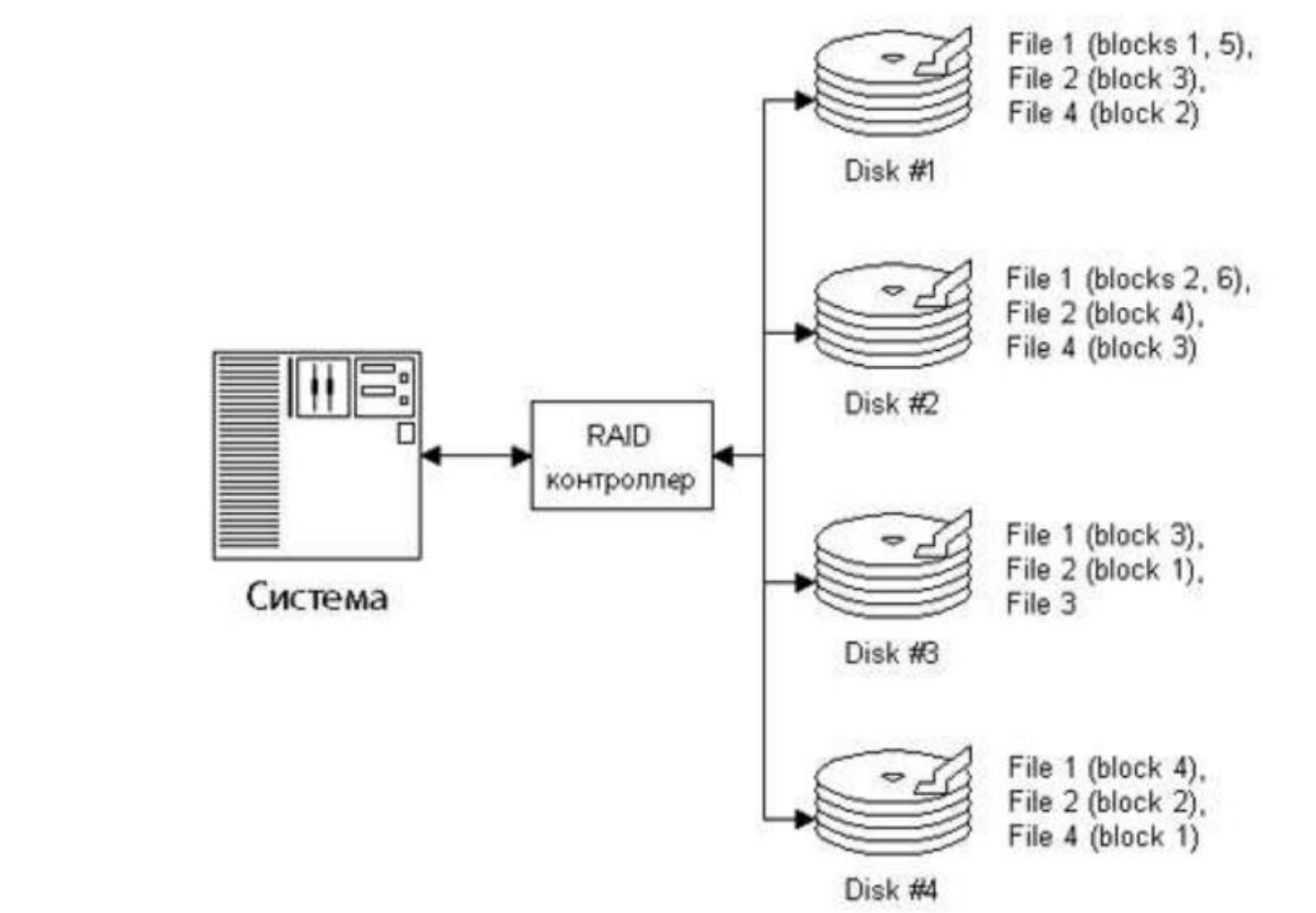
В основе функционирования RAID-массивов лежит несколько базовых терминов, без которых нельзя понять принципы работы этой технологии.

1. **Массив** — объединение нескольких физических или виртуальных накопителей в один большой диск с возможностью единой настройки, форматирования и управления.
2. **Метод зеркалирования** — способ повысить надежность хранения информации через создание копии исходного диска на другом носителе, входящем в массив.
3. **Дуплекс** — один из методов зеркалирования, в котором используется вдвое большее количество накопителей для создания копий.
4. **Чередование** — увеличение производительности диска, благодаря блочной разбивке данных при записи.
5. **Четность** — технология, сочетающая в себе чередование и зеркалирование.

3. RAID 0 (Striping — «чередование»)



3. Пример RAID 0



3. 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.

3. Вероятность выхода из строя RAID 0 - 2 HDD

p **-** вероятность выхода из строя HDD

q=1-p - вероятность работоспособного состояния .

A- событие выхода RAID0 из строя

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

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

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

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

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

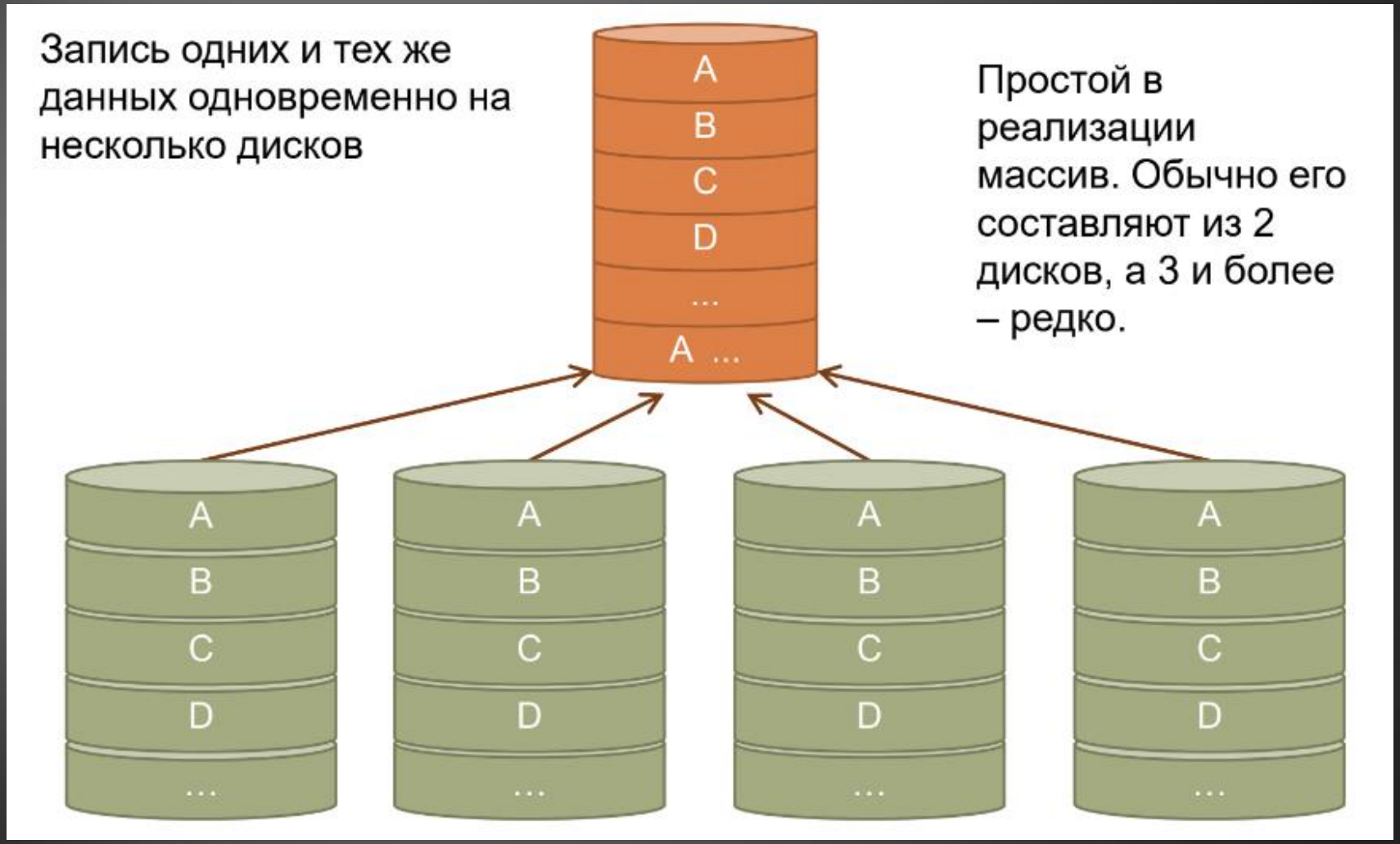
$$P(A) = 1 - P(\overline{A}_1)P(\overline{A}_2) = 1 - q^2$$

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%.

3. RAID 1 (Mirroring - «зеркалирование»)



3. 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.

3. Вероятность выхода из строя RAID 1 – 2 HDD

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

p - вероятность выхода из строя HDD

q=1-p - вероятность работоспособного состояния .

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

A- событие выхода RAID1 из строя

$$\overline{A}_1\overline{A}_2 \qquad A_1\overline{A}_2 \qquad A_1\overline{A}_2 \qquad A_1A_2$$

$$P(A) = P(A_1A_2) = P(A_1)P(A_2) = p^2 \quad (3)$$

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

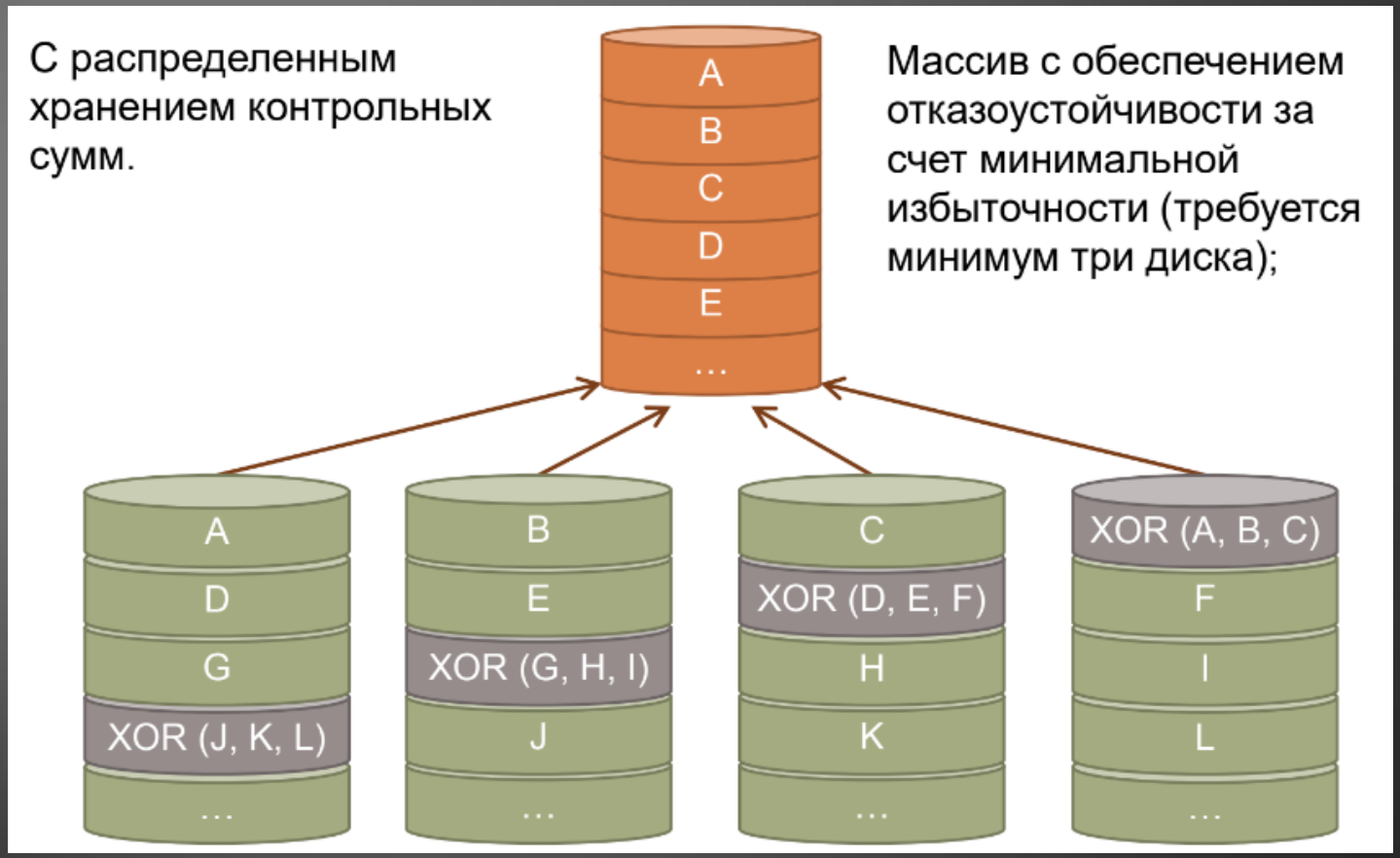
$$p = 0.03$$

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

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

3. RAID 5 (Striping with parity)

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



3. RAID 5

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 = d1 XOR d2* XOR ... XOR *dn–1*.

Страйп *d1* записывается на первый диск, страйп d2 — на второй и так далее вплоть до страйпа dn–1, который записывается на *(n–1)* -диск.

Далее на n-й диск записывается контрольная сумма pn, и процесс циклически повторяется с первого диска, на который записывается страйп *dn*.

Если один из дисков, например второй, вышел из строя, то блок *d2* окажется недоступным при считывании. Однако его значение легко восстановить по контрольной сумме и по значениям остальных блоков с помощью все той же операции *XOR: d2 = d1 XOR p XOR...XORdn–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.

3. 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.

3. Вероятность выхода из строя RAID 5 – 3 HDD

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

p - вероятность выхода из строя HDD

q=1-p - вероятность работоспособного состояния .

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

A- событие выхода RAID5 из строя

$$\overline{A}_1\overline{A}_2\overline{A}_3 \quad A_1\overline{A}_2\overline{A}_3 \quad \overline{A}_1A_2\overline{A}_3 \quad \overline{A}_1\overline{A}_2A_3$$

$$A_1A_2\overline{A}_3 \quad A_1\overline{A}_2A_3 \quad \overline{A}_1A_2A_3 \quad A_1A_2A_3 \quad (3)$$

C_3^2 Два диска из трех

C_3^3 три из трех

$$P(A_1A_2\overline{A}_3) = P(A_1\overline{A}_2A_3) = P(\overline{A}_1A_2A_3) = p^2q$$

$$P(A_1A_2A_3) = p^3$$

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

3. Пример

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

Найдем вероятность разрушения RAID5 на трех HDD

$$p = 0.03$$

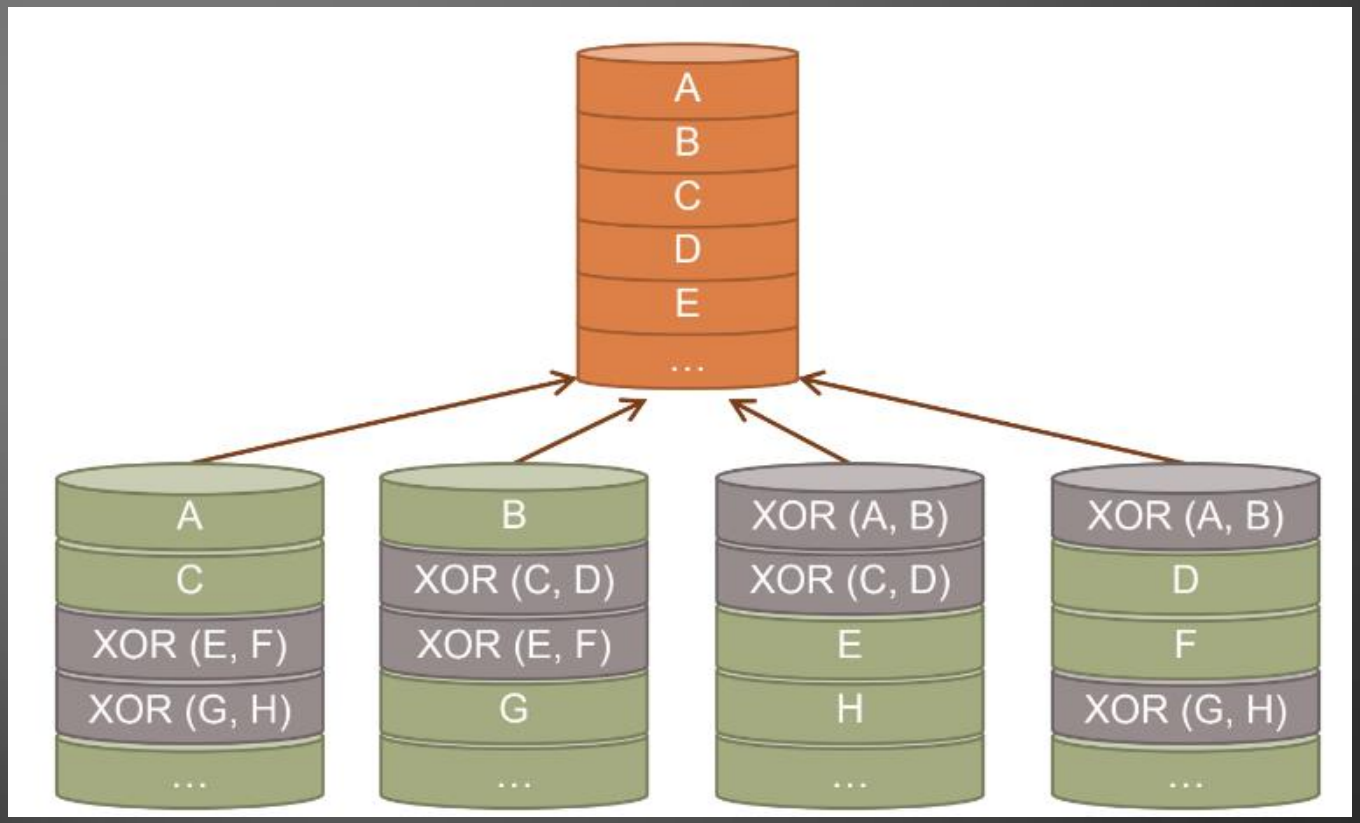
$$q = 1 - p = 0.97$$

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

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

3. 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)



3. Вероятность выхода из строя RAID 6 – *n* HDD

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

p **-** вероятность выхода из строя HDD

q=1-p - вероятность работоспособного состояния .

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

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:

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

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

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

3. Пример

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

Найдем вероятность разрушения RAID6 из четырех HDD

$$N = 4$$

$$p = 0.03$$

$$q = 1 - p = 0.97$$

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

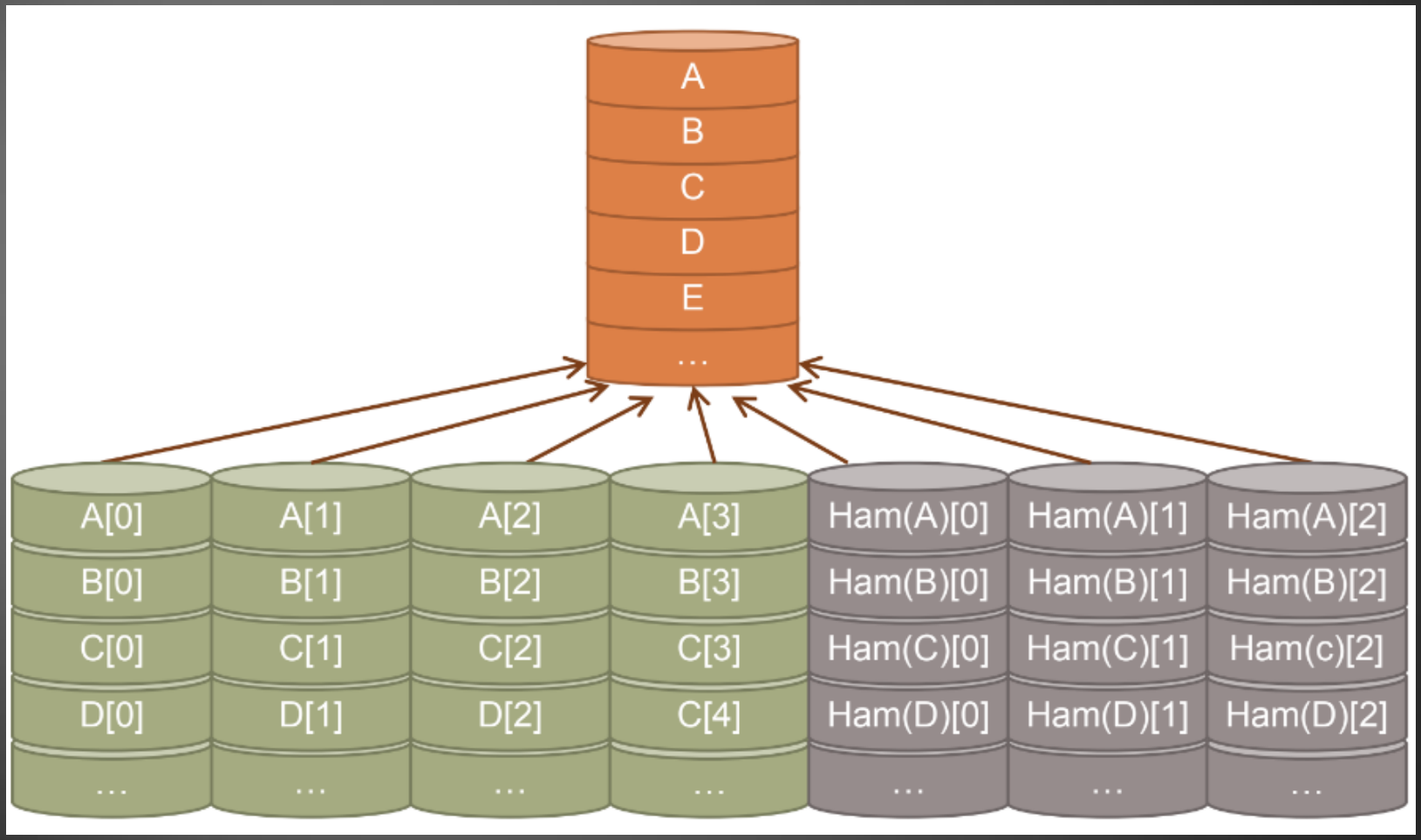
Ответ: Вероятность разрушения RAID6 равняется ,0,01%.

3. 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.

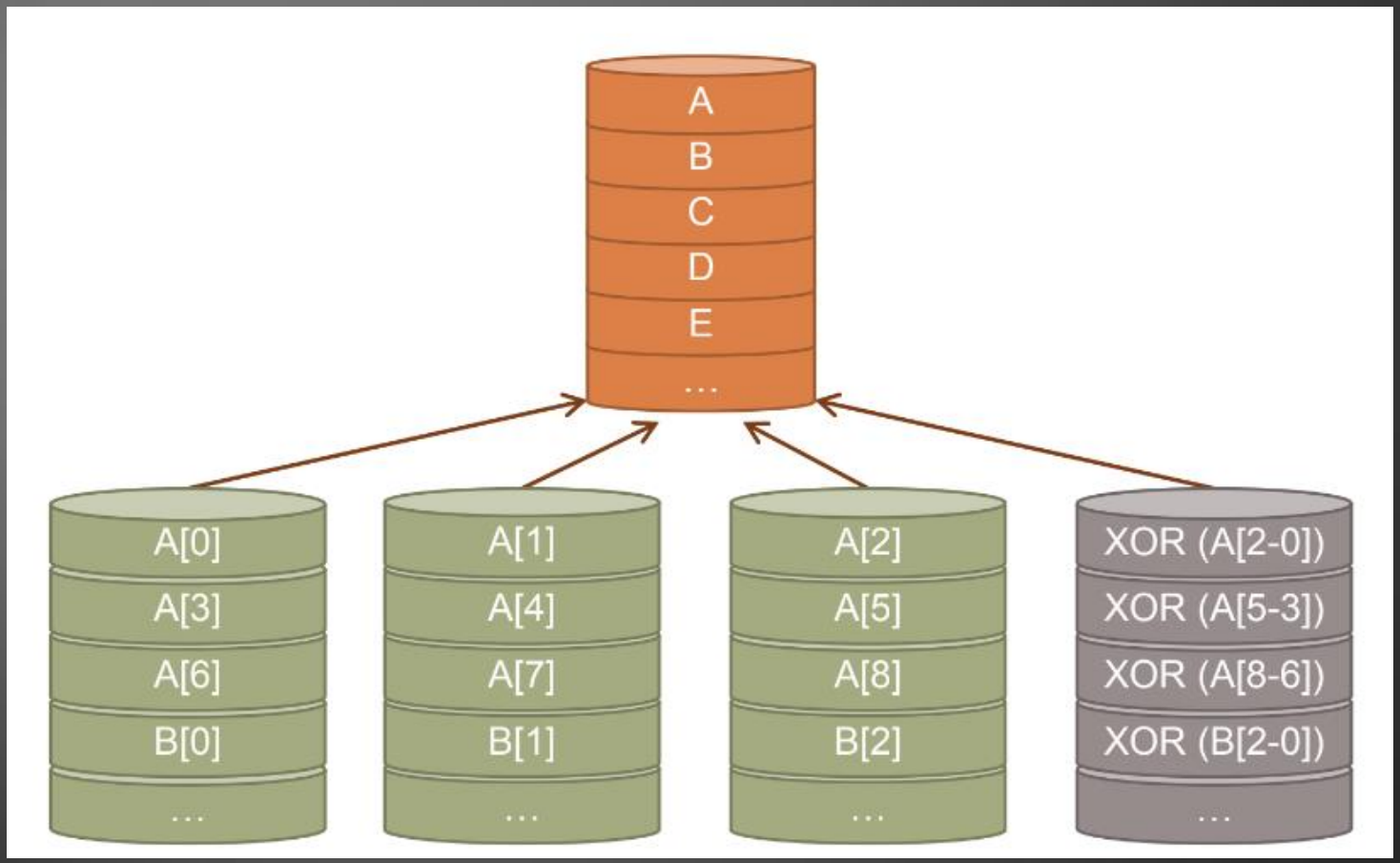
Реализация RAID 2, RAID 3 и RAID 4 практически не встречается в современных контроллерах жестких дисков ввиду высокой технической сложности и отсутствии явных преимуществ перед уровнем RAID 5.

3. RAID 2 (Bit-striping with Hamming code)



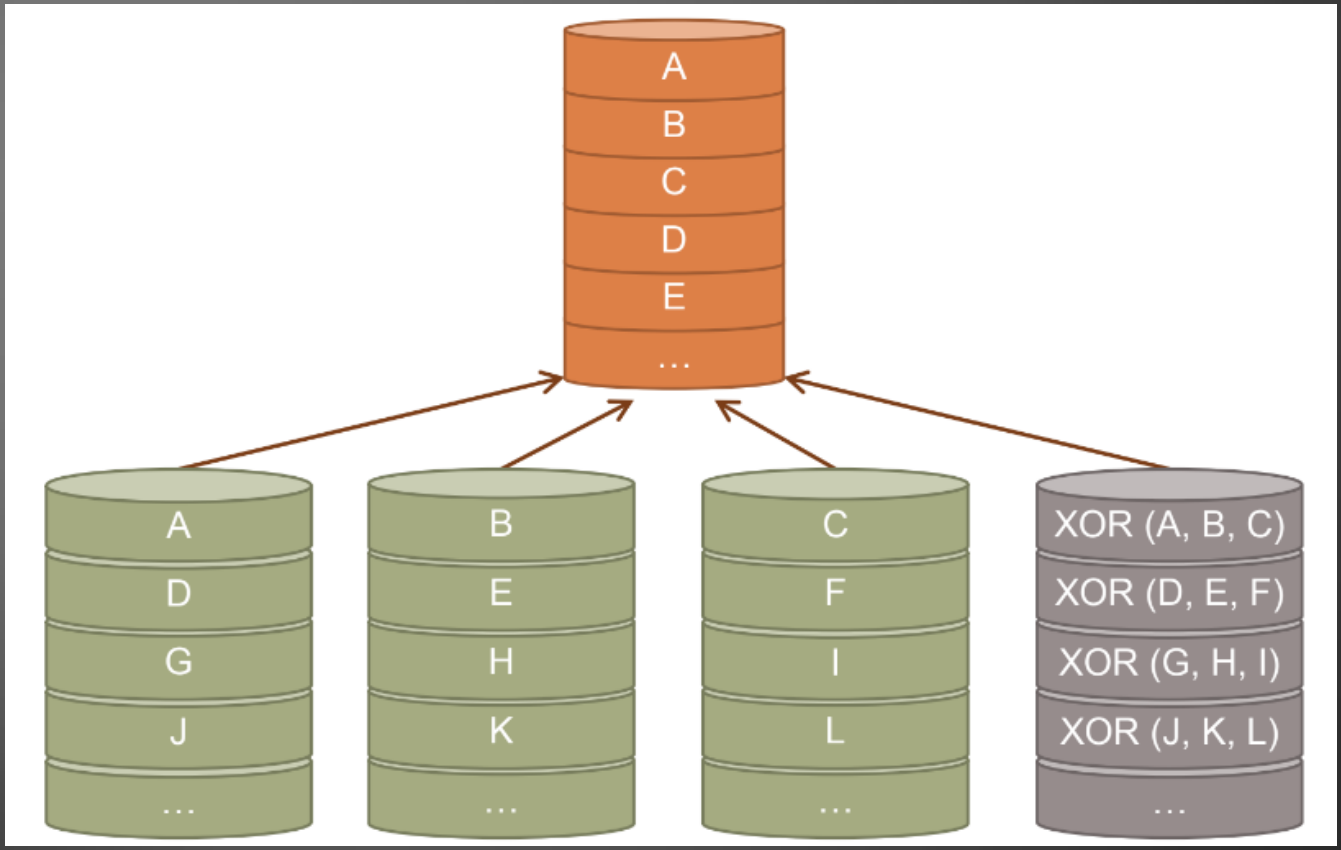
3. RAID 3 (Parallel transfer with parity)

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



3. RAID 4 (Striping with dedicated parity)

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



3. 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.

3. 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

3. Контроллер 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.

3. Контроллер RAID

Классический RAID-контроллер представляет собой либо плату расширения, либо модуль во внешнем исполнении (например, 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.

3. Функции RAID-контроллера

- Обслуживание массивов: создание, поддержка работы, мониторинг, восстановление с применением резервного диска, конвертирование в другой уровень реальном 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.