нетология

Операционная система Linux: **Дисковые системы**

Артем
Поневин

Артем Поневин

Инженер

Luxoft

Артем Поневин

Предисловие

На этом занятии мы поговорим о:

- хранилищах и протоколах;
- видах и различиях RAID;
- LVM и его настройка в Linux;
- возможностях диагностики ввода вывода в Linux.

По итогу занятия вы получите представление о дисковой подсистеме Linux, научитесь настраивать программный RAID и LVM, получите представление о диагностике ввода вывода в Linux.

План занятия

1. Предисловие
2. Устройства хранения данных. Таблица Разделов.
3. Интерфейсы и протоколы
4. RAID
5. LVM
6. Диагностика ввода-вывода в Linux
7. Итоги
8. Домашнее задание

Устройства хранения данных. Таблица разделов.

Символьные и блочные устройства

Устройства ввода-вывода	
Блочные устройства	Символьные устройства
HDD	Терминал
CD-ROM	Клавиатура
Flash	Принтер

Символьное устройство не обладает возможностью произвольного доступа. Его удобно представить как поток.

Блочное устройство обладает возможностью произвольного доступа. При работе с ним данные буферизируются.

Внутренние и внешние накопители

Накопитель информации – устройство, осуществляющее чтение и/или запись информации. Внутренний накопитель подключается к материнской плате и не имеет собственного корпуса.

Накопитель	**Вид накопителя**	**Носитель**
Оперативная память	внутренний	модуль ОЗУ
HDD	внутренний/внешний	жесткий диск
Flash-карта	внешний	Flash-карта
CD-ROM	внутренний/внешний	CD-диск

Разделы жесткого диска

Раздел — часть долговременной памяти жёсткого диска или флеш-накопителя, выделенная для удобства работы, и состоящая из смежных блоков. На одном устройстве хранения может быть несколько разделов.

Использование разделов дает возможность:

- хранить информацию в разных файловых системах, или в одинаковых файловых системах, но с разным размером кластера;
- на одном жёстком диске можно установить несколько операционных систем;

Master Boot Record

MBR (Master Boot Record, главная загрузочная запись) содержит сведения о структуре жесткого диска и код для запуска операционной системы.

Источник изображения: https://www.starusrecovery.ru

GUID partition table

GUID (Globally Unique Identifier) **Partition Table** (GPT) — стандарт формата размещения таблиц разделов на физическом жестком диске. Является частью расширяемого микропрограммного интерфейса, Extensible Firmware Interface, EFI.

Источник изображения: https://www.starusrecovery.ru

Команды Linux для работы с разделами и местом на дисках

- **lsblk** – утилита выводит список блочных устройств с информацией о них;
- **df** – утилита выводит информацию о файловых системах, их размере, занятом и свободном пространстве и точках монтирования;
- **du** – утилита позволяет получить информацию об использовании дискового пространства заданными файлами и иерархией каталогов;
- **blkid** – утилита отображает информацию об уникальных идентификаторах блочных устройств;
- **fdisk** – утилита выводит информацию об имеющихся разделах и позволяет управлять ими.

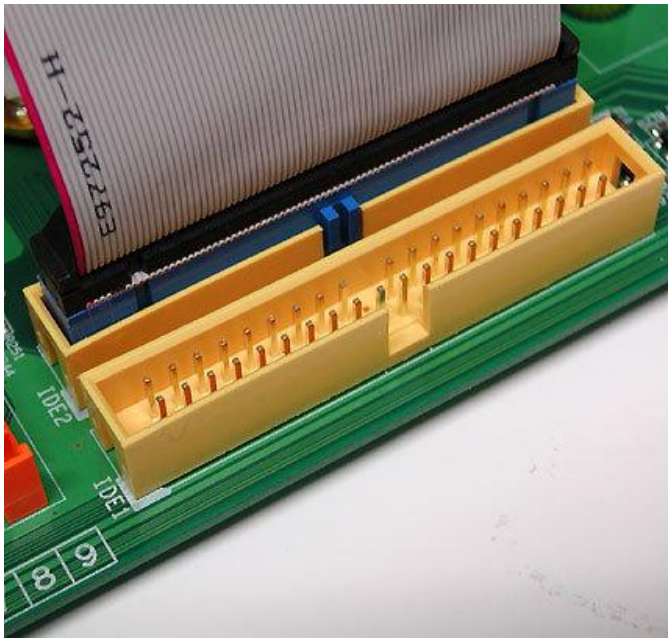
Интерфейсы и протоколы

SCSI

SCSI (Small Computer System Interface) – это интерфейс, предназначенный объединения на одной шине устройств различных классов: жестких дисков, CD-ROM, приводов CD, DVD, стримеров, сканеров, принтеров и т. д.

IDE (PATA)

PATA (Parallel Advanced Technology Attachment) или **IDE** (Integrated Drive Electronics) – параллельный интерфейс подключения накопителей (гибких дисков, жёстких дисков и оптических дисководов) к компьютеру.

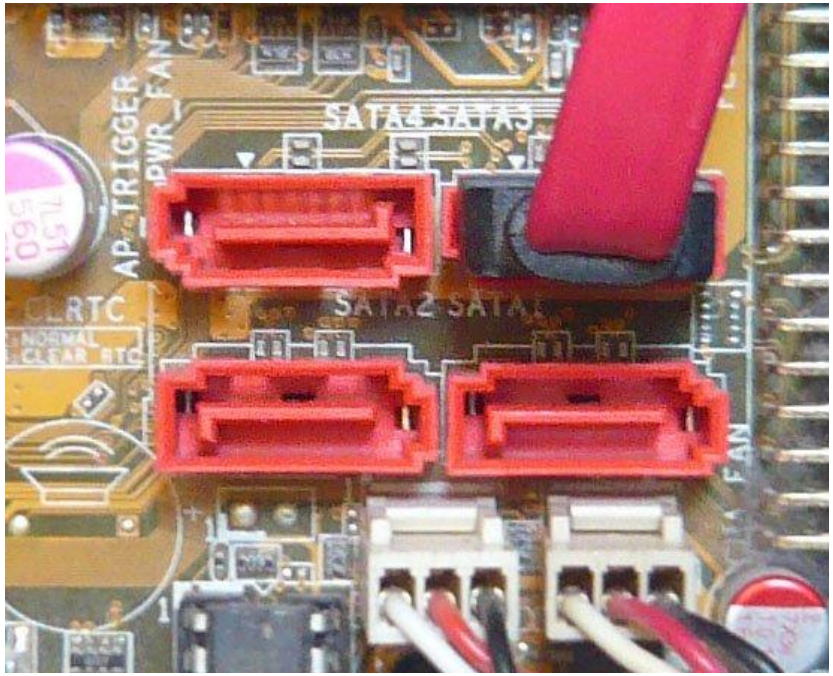


Источник изображения: http://www.xtechx.ru/

SATA

SATA (Serial ATA) – последовательный интерфейс обмена данными с накопителями информации.

SATA является развитием параллельного интерфейса ATA (IDE).



Источник изображения: https://remontcompa.ru/

Ревизии SATA

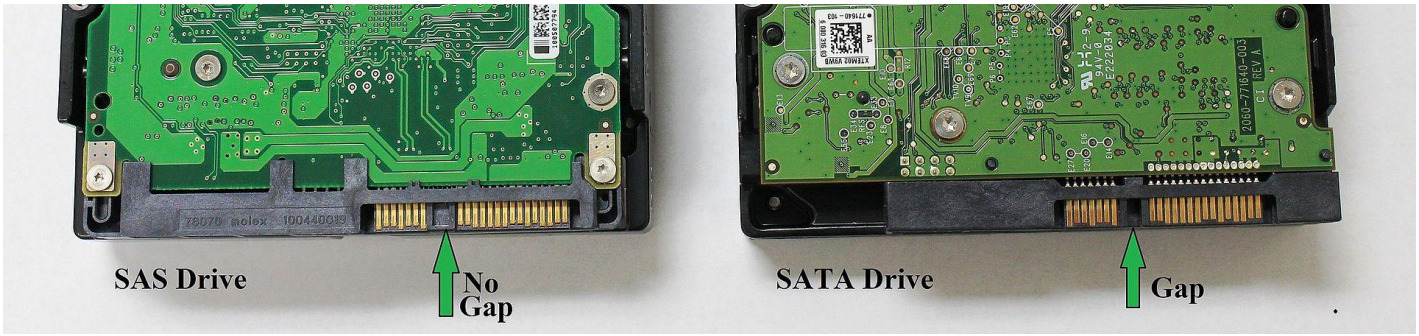
Интерфейс	Пропускная способность		Частота
	биты	байты	
SATA 1.x 1.5Gb/s	1,2 Гбит/с	150 МБ/с	1.5 Ггц
SATA 2.x 3Gb/s	2,4 Гбит/с	300 МБ/с	3.0 Ггц
SATA 3.x 6Gb/s	4,8 Гбит/с	600 МБ/с	6.0Ггц

SAS

Serial Attached SCSI (SAS) – последовательный компьютерный интерфейс, разработанный для подключения различных устройств хранения данных, например, жёстких дисков и ленточных накопителей.

SAS разработан для замены параллельного интерфейса SCSI и основывается во многом на терминологии и наборах команд SCSI.

SAS обратно совместим с интерфейсом SATA: устройства 3 Гбит/с и 6 Гбит/с SATA могут быть подключены к контроллеру SAS, но не наоборот.



Источник изображения: https://ru.wikipedia.org/

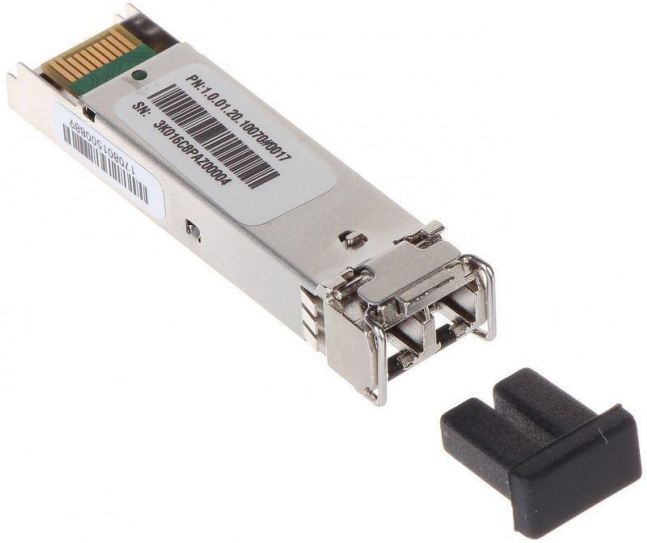
Сравнение интерфейсов SAS и SATA

SAS	SATA
Для серверных систем	Преимущественно для настольных и мобильных систем
Использует набор команд SCSI	Использует набор команд ATA
Минимальная скорость вращения шпинделя HDD 7200 RPM, максимальная – 15000 RPM	Минимум 5400 RPM, максимум 7200 RPM
Два дуплексных порта	Один полудуплексный порт
Поддерживается Multipath I/O	Подключение по типу «точка – точка»
Очередь команд до 256	Очередь команд до 32
Можно использовать кабели до 10 м	Длина кабелей не более 1 м
Пропускная способность шины до 12 Гбит/с (в перспективе – 24 Гбит/с)	Пропускная способность 6 Гбит/с (SATA III)
Стоимость накопителей выше, иногда значительно	Дешевле в пересчете на цену за 1 Гб

Fibre Channel

Fibre channel (волоконный канал) — семейство протоколов для высокоскоростной передачи данных.

Fibre Channel Protocol (FCP) — транспортный протокол (как TCP в IP-сетях), инкапсулирующий протокол SCSI по сетям Fibre Channel. Является основой построения сетей хранения данных.



Источник изображения: https://www.citilink.ru

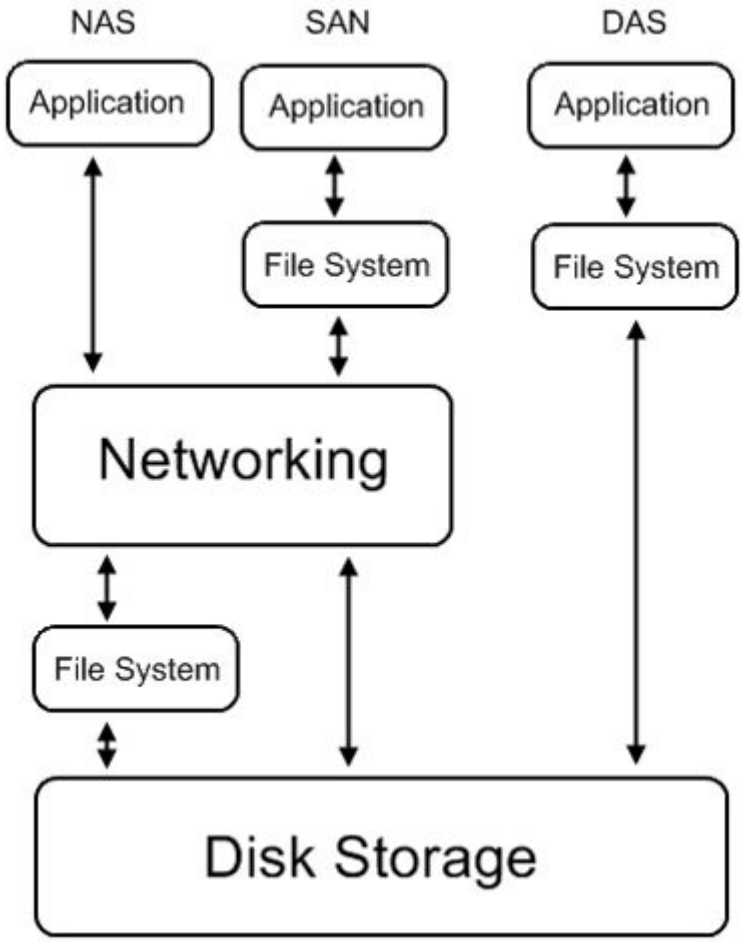
SAN, NAS, DAS

DAS – блочное устройство с диска, который физически, напрямую, подключен к хост-машине.
Необходимо поместить на нее файловую систему, прежде чем ее можно будет использовать.
Технологии для этого включают IDE, SCSI, SATA и т.д.

SAN – блочное устройство, которое доставляется по сети.
Как и DAS, вы все равно должны разместить на нем файловую систему, прежде чем она сможет использоваться.
Технологии для этого включают FibreChannel, iSCSI, FoE и т. Д.

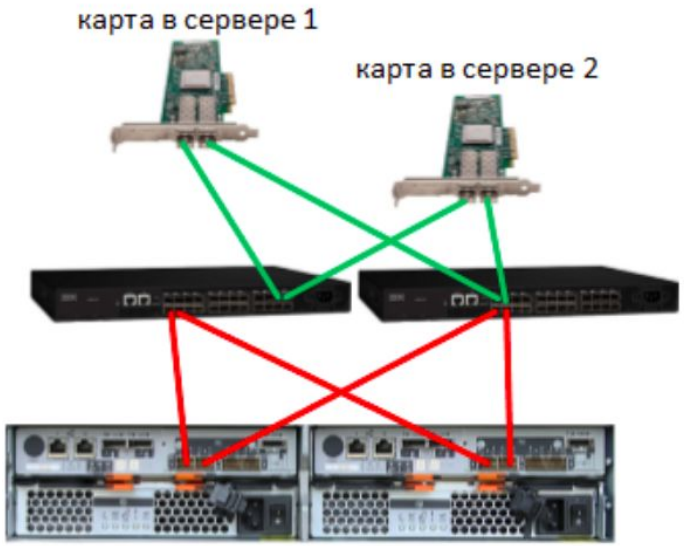
NAS – файловая система, доставляемая по сети.
Технологии для этого включают NFS, CIFS, AFS и т.д.

Сравнение SAN, NAS, DAS



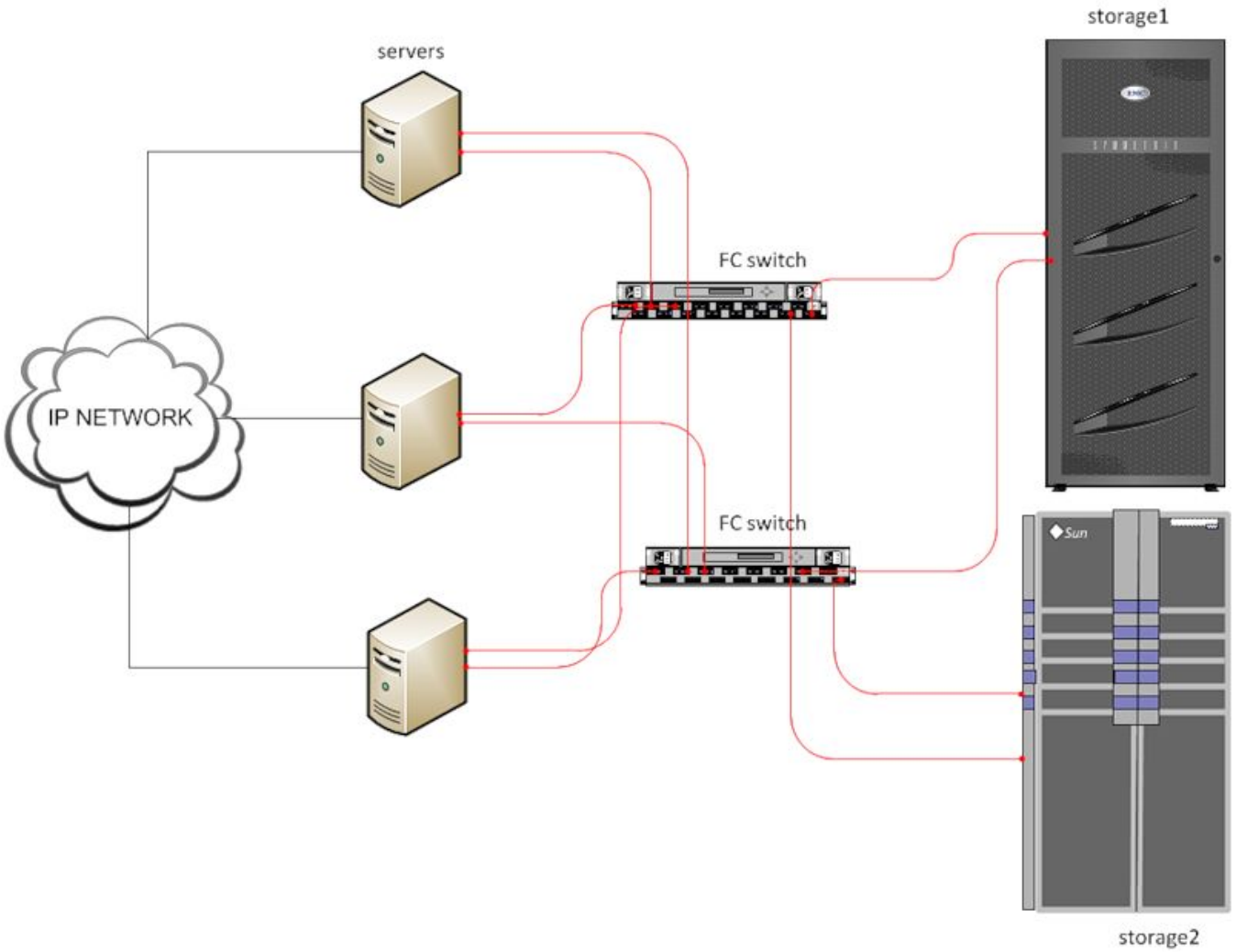
Пример СХД

СХД (сеть хранения данных, SAN, storage area network) – архитектурное решение для подключения внешних устройств хранения данных (дисковые массивы, ленточные библиотеки) к серверам таким образом, чтобы операционная система распознала подключённые ресурсы как локальные.



Источник изображения: http://itsave.ru

Архитектура СХД



Источник изображения: https://habr.com/

Протоколы в СХД

- **FCP**, транспорт SCSI через Fibre Channel;

Наиболее часто используемый на данный момент протокол. Существует в вариантах 1 Gbit/s, 2 Gbit/s, 4 Gbit/s, 8 Gbit/s и 10 Gbit/s.

- **iSCSI**, транспорт SCSI через TCP/IP;
- **FCoE**, транспортировка FCP/SCSI поверх «чистого» Ethernet;
- **FCIP** и **iFCP**, инкапсуляция и передача FCP/SCSI в пакетах IP;
- **HyperSCSI**, транспорт SCSI через Ethernet;
- **FICON,** транспорт через Fibre Channel (используется только мейнфреймами);
- **ATA over Ethernet,** транспорт ATA через Ethernet;
- **SCSI** и/или **TCP/IP,** транспорт через InfiniBand (IB).

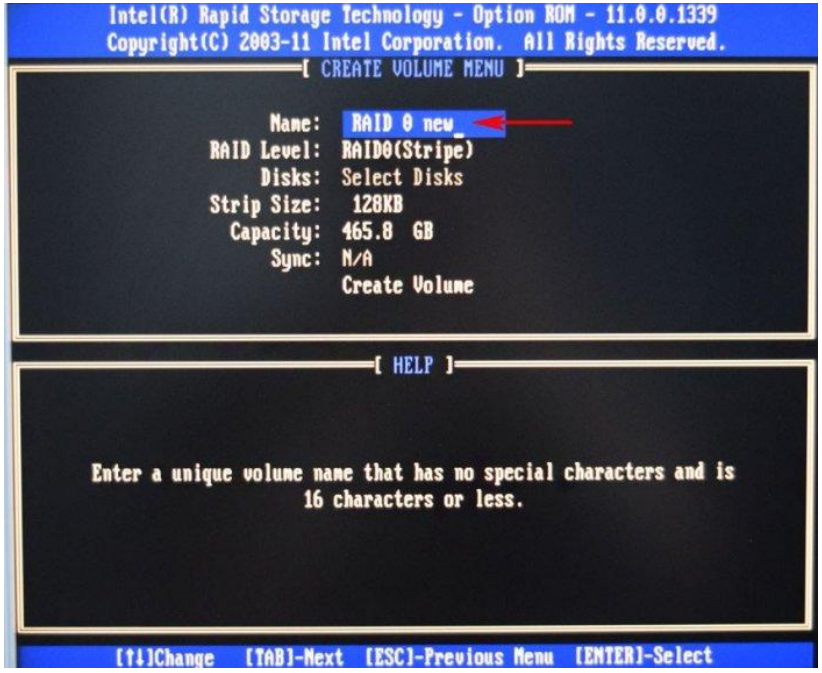
Команды Linux для работы с аппаратным обеспечением

- **lshw** – утилита выводит информацию об имеющемся аппаратном обеспечении;
- **hdparm** – утилита, предназначенная для регулировки и просмотра параметров жёстких дисков с интерфейсом ATA;
- **/dev** – каталог специального назначения, который содержит файлы физических устройств.

RAID

RAID

RAID (Redundant Array of Independent Disks, избыточный массив независимых дисков) — технология виртуализации данных, которая объединяет несколько дисков в логический элемент для повышения производительности и отказоустойчивости.



Виды RAID контроллеров

- **Аппаратный** – отдельное устройство.

Настройка производится через специальное ПО.

- **Полуаппаратный** – встроен в материнскую плату.

Настройка производится через BIOS.

- **Программный** – работает из ОС.

Настраивается утилитами из ОС: mdadm для Linux, составные тома для Windows.

Источник изображения: https://www.nix.ru

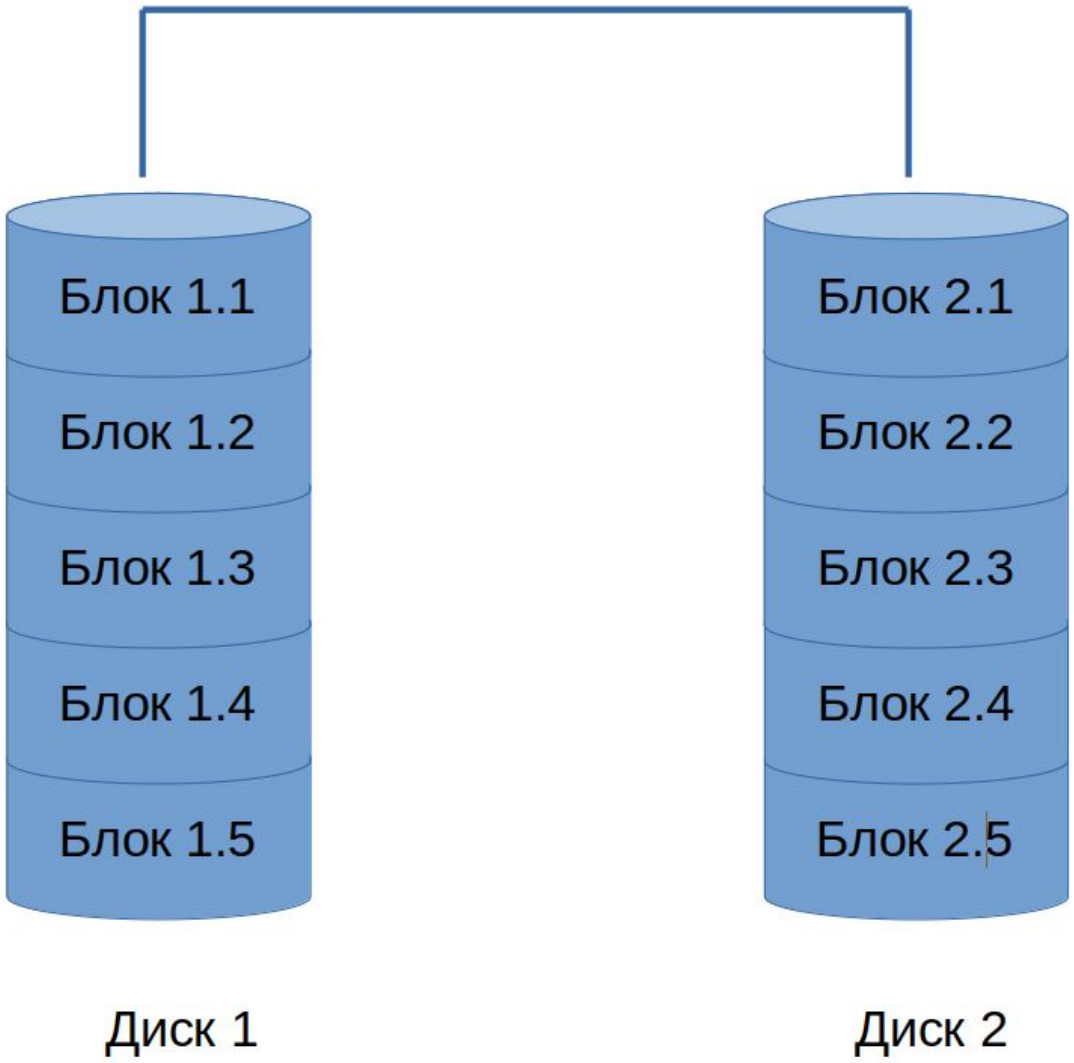
RAID JBOD*

В конфигурации **JBOD** данные на дисках хранятся последовательно. Например, сначала данные записываются на Диск 1. После заполнения Диска 1 данные записываются на Диск 2, затем на Диск 3 и т. д.

Два преимущества данного уровня RAID — это доступность всей общей емкости хранения дисков и простота расширения. Однако если один диск выйдет из строя, все данные будут утеряны.

*RAID JBOD – Just a Bunch Of Disks.

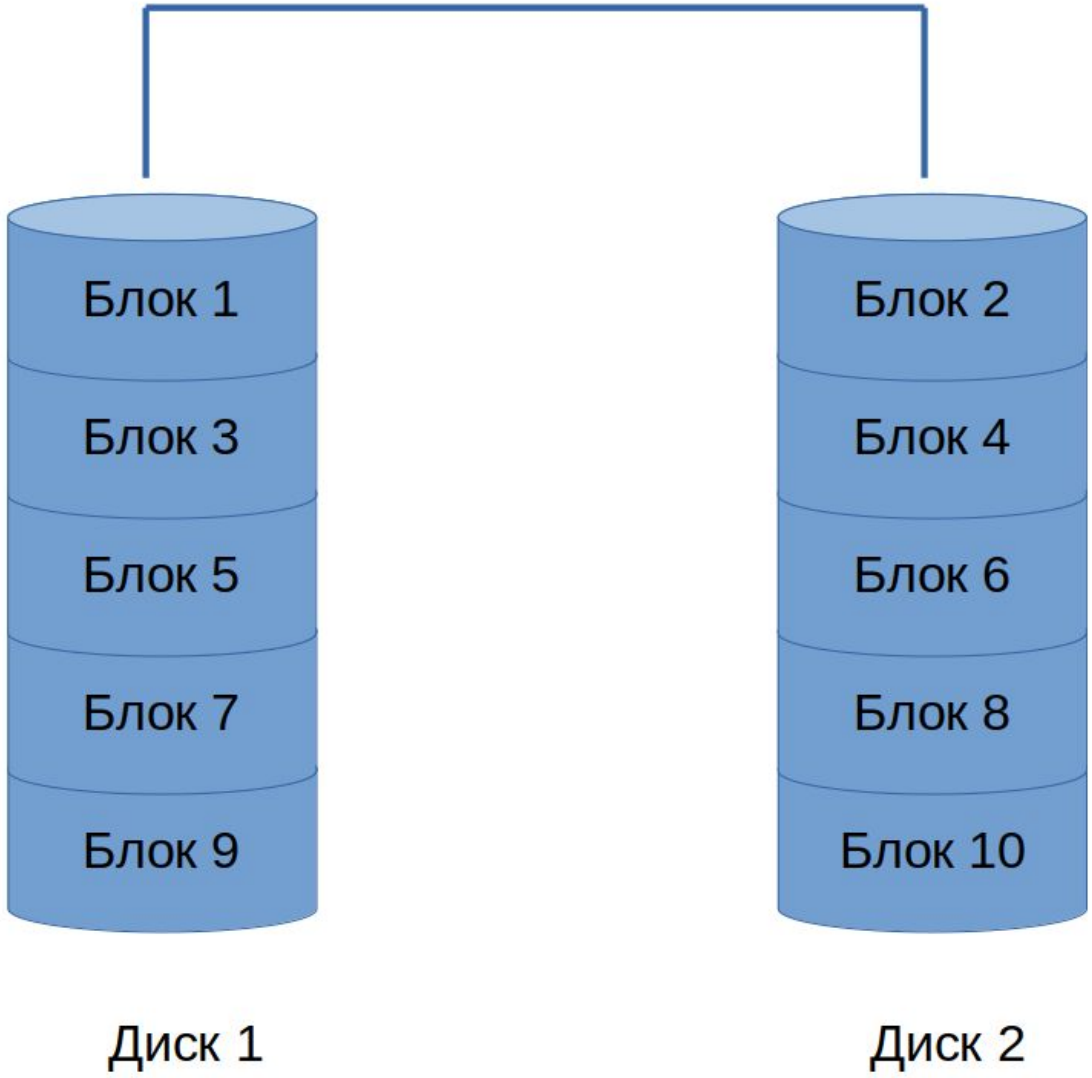
JBOD


Блок 1.1
Блок 1.2
Блок 1.3
Блок 1.4
Блок 1.5
Блок 2.1
Блок 2.2
Блок 2.3
Блок 2.4
Блок 2.5
Диск 1
Диск 2

RAID 0

RAID 0 (striping, чередование) — дисковый массив из двух или более жёстких дисков без резервирования.
Информация разбивается на одинаковые по длине блоки, а затем записывается поочерёдно на каждый диск в структуре.
Основное предназначение такой системы — увеличение производительности, при этом вам будет доступен полный объем всех дисков. Если один диск выходит из строя, все данные становятся недоступными.

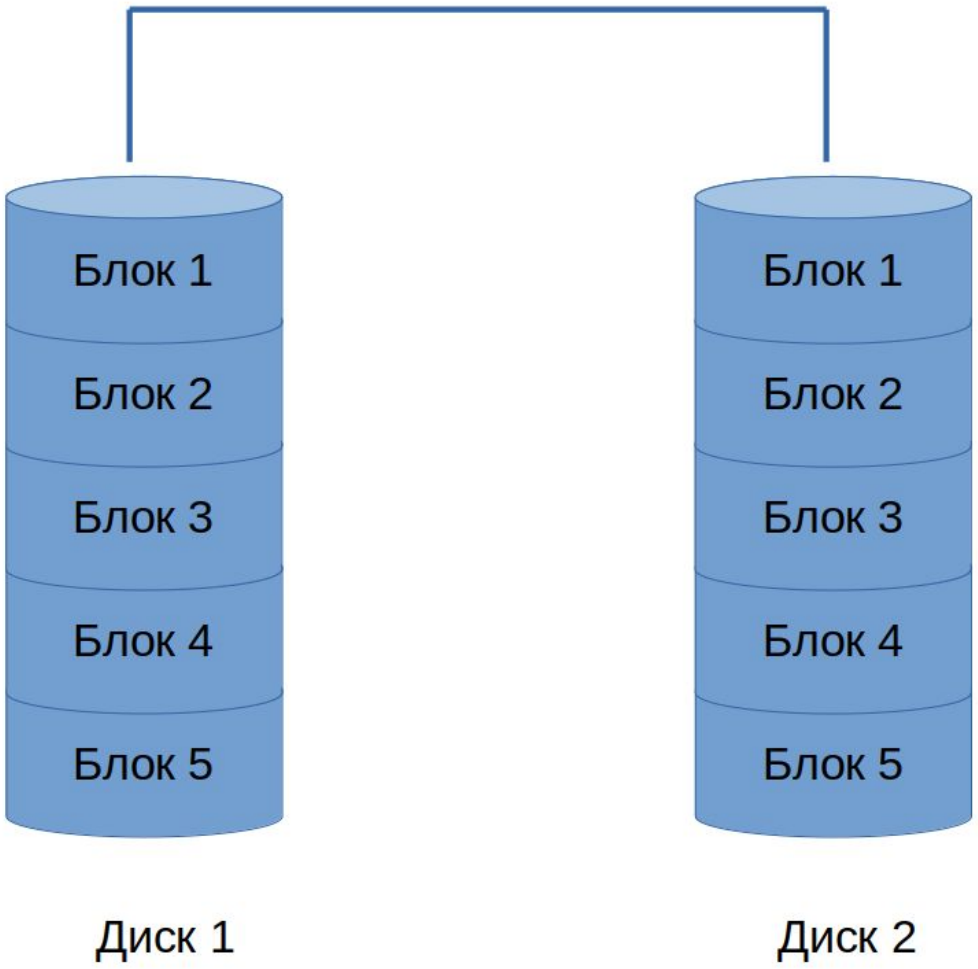
RAID 0


Блок 1
Блок 3
Блок 5
Блок 7
Блок 9
Блок 2
Блок 4
Блок 6
Блок 8
Блок 10
Диск 1
Диск 2

RAID 1

RAID 1 (mirroring, зеркалирование) — массив из двух или более дисков, являющихся полными копиями друг друга.
Обеспечивается бесперебойность работы даже если один из дисков выйдет из строя.
Производительность остается на прежнем уровне, объем равен меньшему из дисков в массиве.

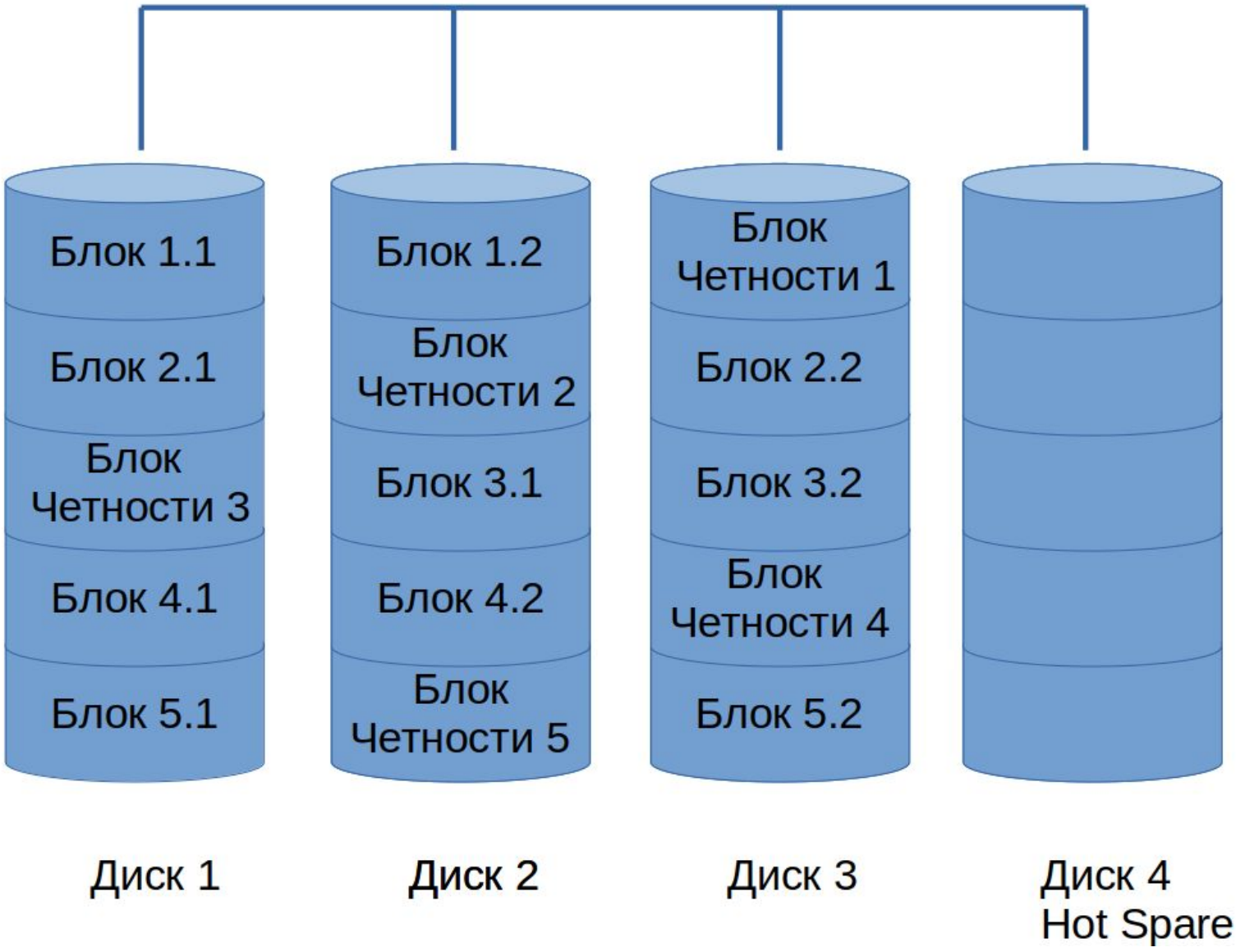
RAID 1


Блок 1
Блок 2
Блок 3
Блок 4
Блок 5
Блок 1
Блок 2
Блок 3
Блок 4
Блок 5
Диск 1
Диск 2

RAID 5

- Данные записываются на все диски в томе и в блок четности для каждого блока данных.
- Если один физический диск выходит из строя, данные из неисправного диска можно восстановить на запасной диск.
- Данные сохраняются при выходе из строя одного диска, но в случае выхода из строя второго диска до того, как данные смогли быть восстановлены на запасной диск, все данные будут утеряны.
- Для создания тома RAID 5 требуется минимум три диска.

RAID 5


Блок 1.1
Блок 2.1
Блок Четности 3
Блок 4.1
Блок 5.1
Блок 1.2
Блок Четности 2
Блок 3.1
Блок 4.2
Блок Четности 5
Блок Четности 1
Блок 2.2
Блок 3.2
Блок Четности 4
Блок 5.2
Диск 1
Диск 2
Диск 3
Диск 4
Hot Spare

RAID 10

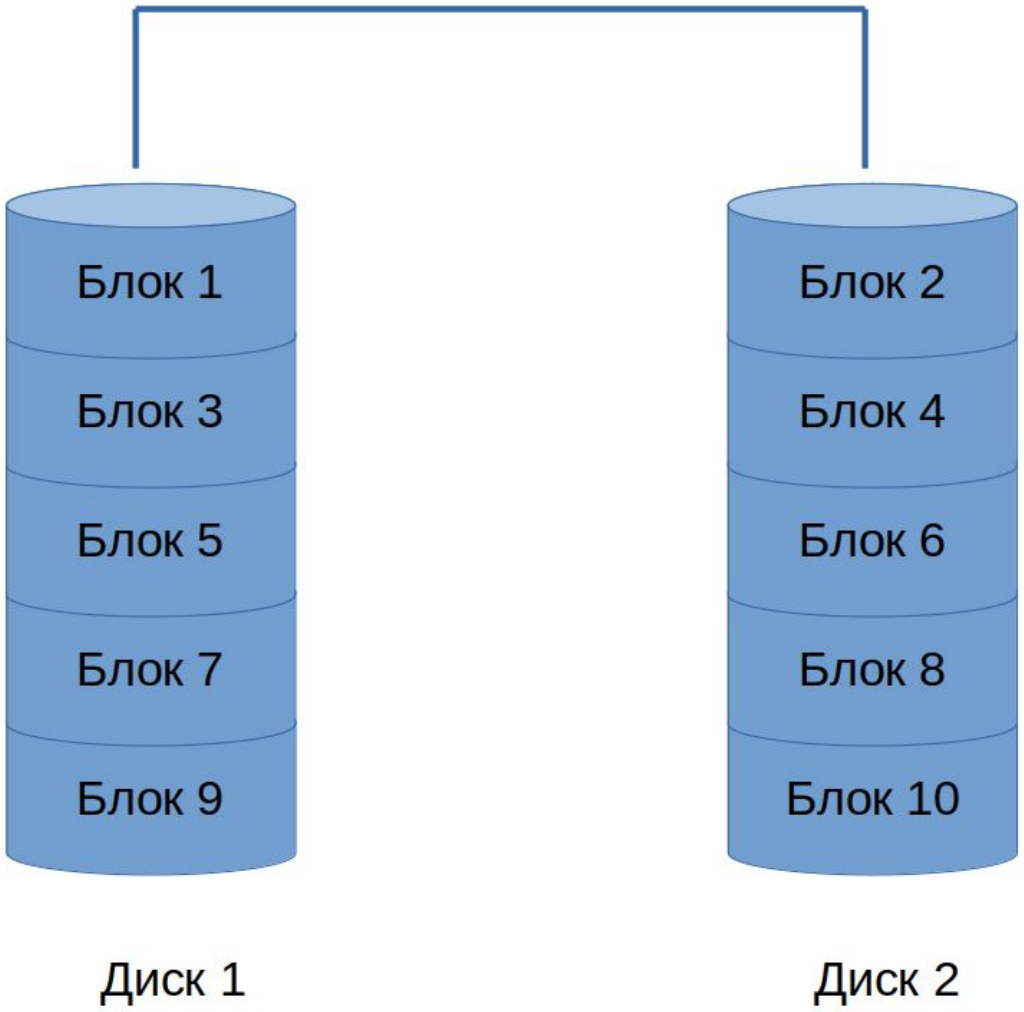
В режиме RAID 10 сочетаются защита режима RAID 1 и производительность режима RAID 0.

➡️ При использовании четырех дисков в режиме RAID 10 создается два сегмента RAID 1, которые объединяются в страйп RAID 0.

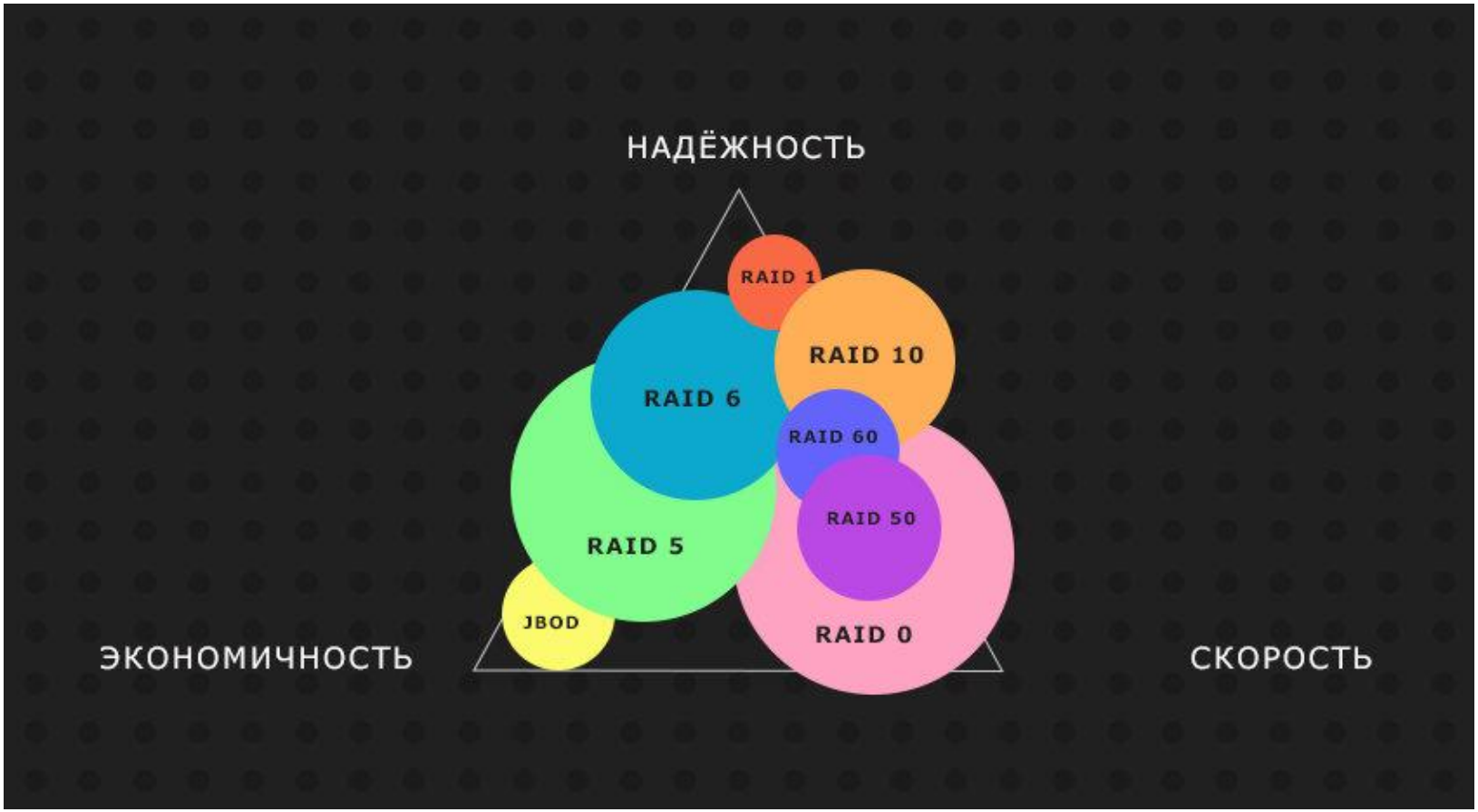
➡️ При использовании восьми дисков в страйпе RAID 0 будет уже четыре сегмента RAID 1.

В режиме RAID 10 могут выйти из строя до 2х дисков в двух сегментах RAID 1.

RAID 0


Блок 1
Блок 3
Блок 5
Блок 7
Блок 9
Блок 2
Блок 4
Блок 6
Блок 8
Блок 10
Диск 1
Диск 2

Выбор RAID



Источник изображения: https://www.nix.ru/

Команды Linux для работы с RAID

- **sudo yum (apt-get) install mdadm** — установка утилиты;
- **sudo mdadm --create /dev/md0 -l 1 -n 2 /dev/sd{b,c}** — создание нового массива;
- **cat /proc/mdstat** — текущее состояние;
- **/etc/mdadm.conf** — файл конфигурации.

LVM

LVM

LVM (Logical Volume Manager, Менеджер логических томов) — это дополнительный слой абстракции между "железом" и файловой системой, позволяющий использовать разные области одного жёсткого диска и/или области с разных жёстких дисков как один логический том.

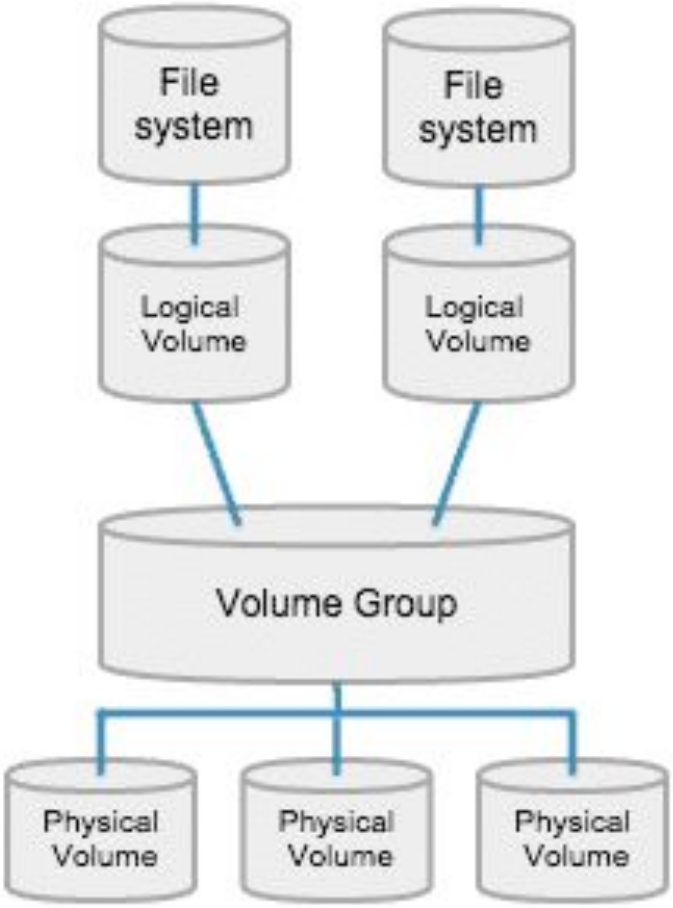
LVM

LVM (Logical Volume Manager, Менеджер логических томов) — это дополнительный слой абстракции между "железом" и файловой системой, позволяющий использовать разные области одного жёсткого диска и/или области с разных жёстких дисков как один логический том.

Уровни абстракции LVM

- **PV** (Physical Volumes, физические тома) — разделы или целые «неразбитые» диски.
- **VG** (Volumes Group, группа томов) — объединенные в группу физические тома (PV), из которых создается единый логический диск, который в дальнейшем можно разбивать по своему усмотрению.
- **LV** (Logical Volumes, логические разделы) — собственно раздел нового «единого диска», полученного из группы томов, который можно форматировать и использовать как обычный раздел, обычного жёсткого диска.

Уровни абстракции LVM



Device mapper

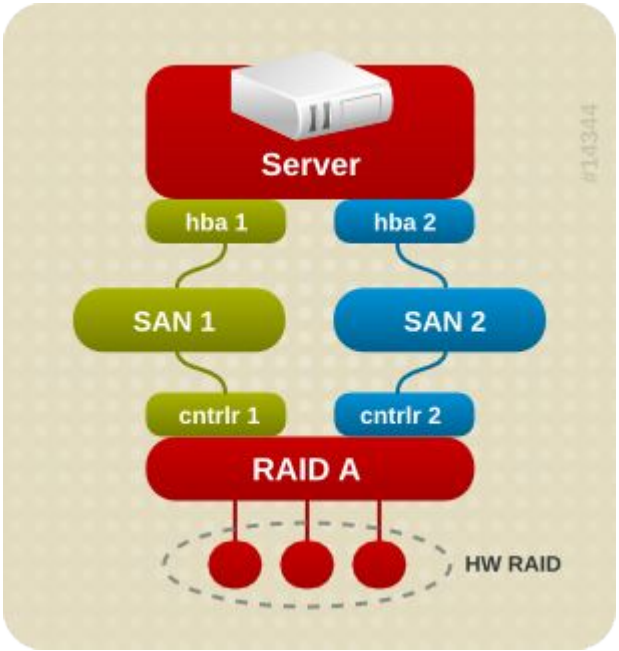
Подсистема ядра Linux, которая позволяет мэппить одно блочное устройство на другое или несколько других.

➡️ Это позволяет реализовать объединение нескольких реальных устройств в одно виртуальное, создавать snapshot'ы, организовывать балансировку между устройствами, делать шифрованные тома и т.д.

➡️ Через device-mapper, в качестве низкоуровнего API, работают LVM, cryptsetup, dm-multipath и другие утилиты.

Multipath

DM-Multipath позволяет объединить несколько маршрутов ввода-вывода между серверами и дисковыми массивами в единое целое. Маршруты в этом случае представляют собой физические SAN-соединения, которые могут включать отдельные кабели, переключатели и контроллеры. В результате агрегации будет создано новое устройство.



Команды Linux для изменения LVM

- **pvcreate** — позволяет создать физический том на жестком диске;
- **vgcreate** — позволяет создать группу томов из физических томов;
- **lvcreate** — позволяет создать логический том в группе томов;

Команды Linux для получения состояния LVM

- **pvdisplay\pvs** — позволяет отобразить информацию о физических томах;
- **vgdisplay\vgs** — позволяет отобразить информацию о группах томов в ОС;
- **lvdisplay\lvs** — позволяет отобразить информацию о логических томах;
- **vmstat** — отображает информацию об использовании CPU, памяти, дисков.

Диагностика ввода-вывода в Linux

Команды Linux для диагностики нагрузки на дисковую подсистему

- **top** — выводит информацию о работающих в системе процессах и информацию о них;
- **iostat** — мониторинг использования дисковых разделов;
- **iotop** — аналогична утилите top, но вместо использования процесами CPU и памяти показывает работу процессов с дисками;
- **vmstat** — утилита отображает информацию об использовании CPU, памяти, дисков.

Итоги

Итоги

Сегодня мы рассмотрели работу с дисками в Linux:

- Виды носителей и протоколы;
- Устройство LVM;
- Виды RAID массивов;
- Утилиты для диагностики ввода-вывода.

Домашнее задание

Давайте посмотрим ваше домашнее задание.

- Вопросы по домашней работе задавайте **в чате** мессенджера Slack.
- Задачи можно сдавать **по частям**.
- Зачёт по домашней работе проставляется после того, как **приняты все задачи**.

нетология

Задавайте вопросы и пишите отзыв о лекции!

Поневин Артем

Имя Фамилия