OTUS

ОНЛАЙН-ОБРАЗОВАНИЕ

OTUS

Дисковая подсистема: RAID

Курс «Администратор Linux»

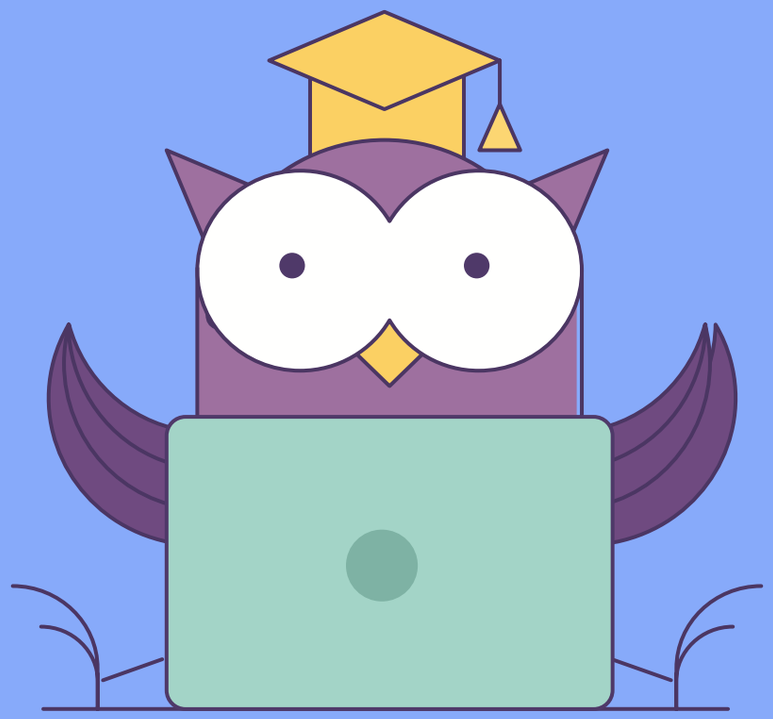
Занятие № 2

Цыкунов Алексей

Молчанов Дмитрий

Ожегов Григорий

Меня хорошо слышно && видно?



Напишите в чат, если есть проблемы!

Ставьте `+` если все хорошо
Ставьте `-` если есть проблемы

Цели вебинара

По итогам занятия вы сможете:

- рассказать какие бывают RAID массивы и чем они отличаются;
- получить информацию о дисковой подсистеме на любом сервере с ОС Linux;
- собрать программный рейд и восстановить его после сбоя.

Маршрут вебинара

Дисковая подсистема

Блочные устройства в UNIX и Linux — это устройства хранения с произвольным доступом, над которыми размещаются файловые системы.

Обеспечивает интерфейс к устройству.

```
[root@mdadm ~]$ ls -la /dev/sd*
brw-rw----. 1 root disk 8,  0 Oct 22 05:32 sda
brw-rw----. 1 root disk 8,  1 Oct 22 05:32 sda1
brw-rw----. 1 root disk 8,  2 Oct 22 05:32 sda2
brw-rw----. 1 root disk 8,  3 Oct 22 05:32 sda3
```

Дисковая подсистема

Блочное устройство обеспечивает обмен блоками данных.

Блок (chunk)— это единица данных фиксированного размера. Размер блока определяется ядром, но чаще всего он совпадает с размером страницы аппаратной архитектуры, и для 32-битной архитектуры x86 составляет 4096 байт.

```
[root@mdadm ~]$ fdisk -l /dev/sda | grep -i sector
Disk /dev/sda: 42.9 GB, 42949672960 bytes, 83886080 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
```

Дисковая подсистема

В настоящее время существует два подхода к организации /dev –

- Статическая организация - специальные файлы для всех возможных устройств вне зависимости от того, загружен драйвер соответствующего устройства или нет

- Динамическая организация - специальные файлы в /dev создаются по мере инициализации устройств и загрузки драйверов, и удаляются при выгрузке соответствующего драйвера или удалении устройства.

Процесс работы со статическим /dev особых проблем не вызывает – системный администратор при необходимости просто создает отсутствующие файлы командой **mknod.**

Дисковая подсистема - DevFS

Ядро монтирует к каталогу /dev специальную файловую систему, называемую devfs:

- Целиком находится в оперативной памяти
- Драйвер devfs динамически создает специальный файл
- Динамически же удаляет его

Дисковая подсистема - Udev

В отличие от devsfd, который требовал поддержки со стороны ядра, udev такой поддержки не требует.

- Постоянно закрепленные за устройствами имена, которые не зависят от того, какое положение они занимают в дереве устройств.
- Уведомление внешних по отношению к ядру программ, если устройство было заменено
- Гибкие правила именования устройств.

Дисковая подсистема - devtmpfs

На текущий момент используется devtmpfs + udev

После монтирования корневой файловой системы, этот экземпляр tmpfs перемонтируется ядром в каталог /dev

devtmpfs отвечает за заполнение каталога /dev

udev за права доступа, необходимые симлинки и запуск скриптов пользователя

Задачи стоящие перед дисковой подсистемой

1. Максимальный объем
2. Максимальная скорость
3. Максимальная надежность
4. Минимальная стоимость

Практически все эти задачи решаются разными уровнями RAID (Redundant Array of Independent/Inexpensive Disks), но, к сожалению, не все сразу.

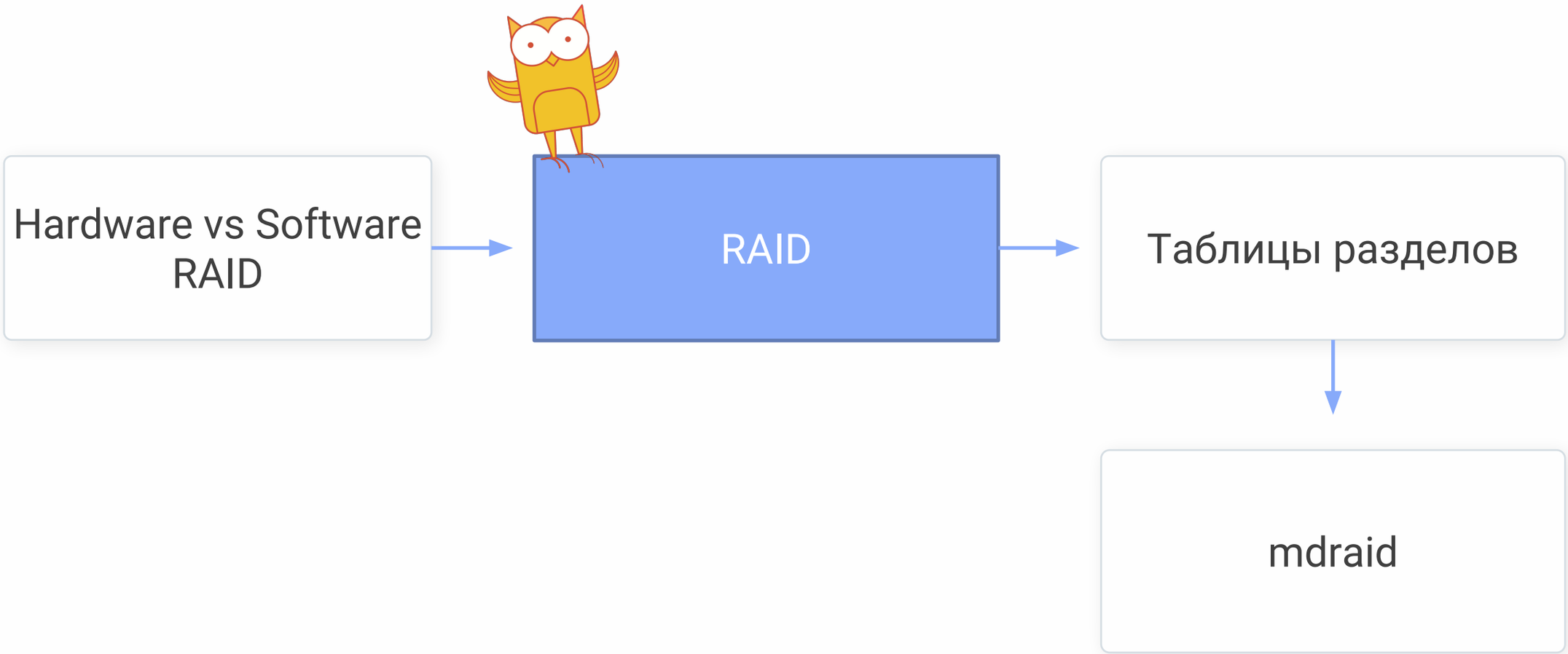
Задачи стоящие перед дисковой подсистемой

- Например, RAID-0 дает максимум объема, при отсутствующей надежности, а RAID-1 — максимальную надежность, при отсутствии выигрыша в пространстве.

- В случае RAID-0 и RAID-1 цифра в названии равна вероятности восстановления в случае отказа диска в массиве.

- Но есть и другие уровни RAID — 5,6,10, все это в том или ином виды — компромиссы.

Маршрут вебинара



Типы RAID

- Аппаратный - отдельный контроллер со своим процессором, cache, портами, батарейкой.

- Программный - все через CPU.

- Интегрированный - встроенный в матплату с отдельным чипом для вычисления XOR.

Hardware RAID

Преимущества

+ Аппаратное решение, не влияющее на производительность основной системы
+ Выделенный CPU
+ Выделенная память для Кэшей
+ Возможность использовать BBU (Battery Backup Unit)
+ Возможность подключения большого количества дисков
+ Прозрачность для загрузчиков (возможность грузиться с любого массива)

Недостатки

- Высокая стоимость
- Высокая сложность
- Разнообразность интерфейсов управления и драйверов
- Низкая «мобильность» /переносимость
- Привязка к железу
- Бóльший простой по времени при аварии
- Очень дорогой ремонт, необходимость закупать впрок контроллеры, которые потом могут прекратить выпускать

Software RAID (mdraid)

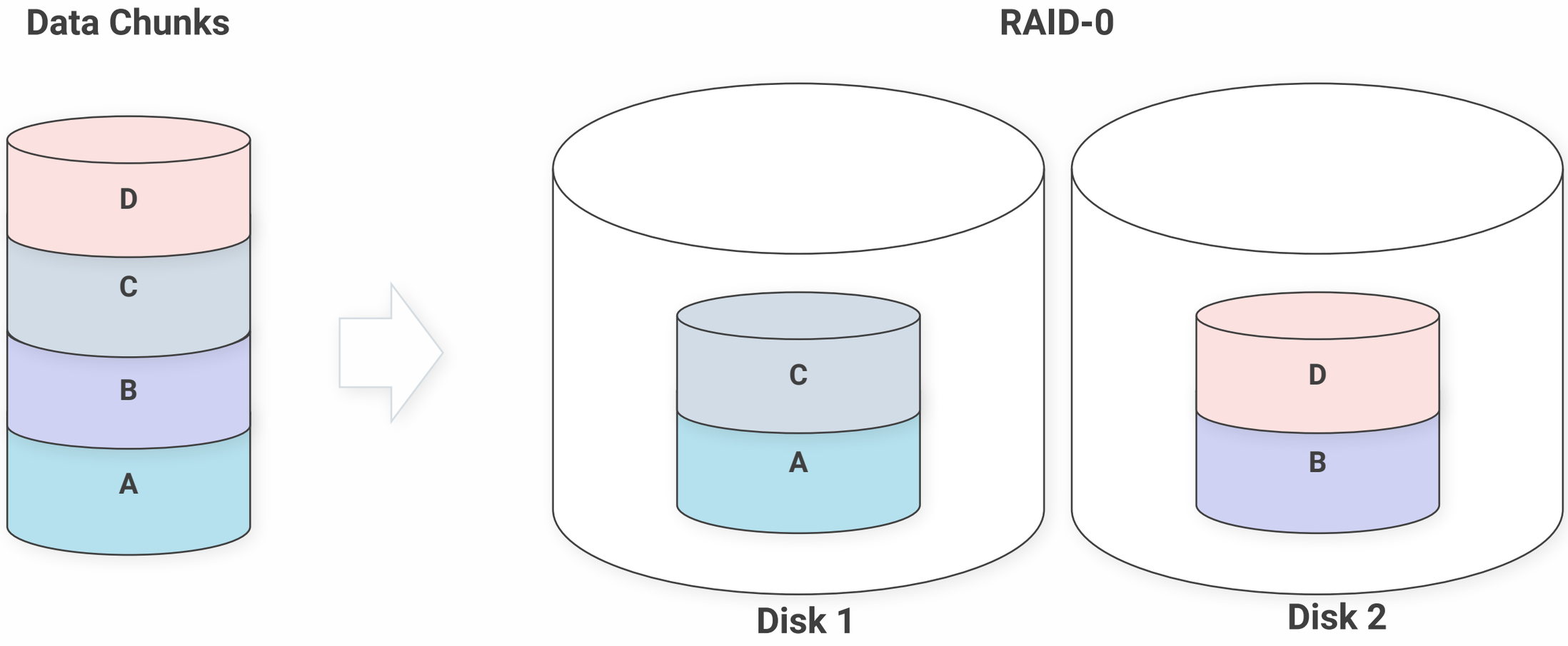
Преимущества

+ Бесплатно
+ Отсутствие привязки к конкретному железу
+ Прозрачность конфигурации
+ Примерно одинаковый интерфейс управления в любом linux.
+ Легкая переносимость между компьютерами.
+ Гибкость конфигурации

Недостатки

- Отсутствие BBU
- Отсутствие выделенного кэша
- Отсутствие службы поддержки :-)

RAID-0



$$V_t = V_1 * N$$

RAID-0

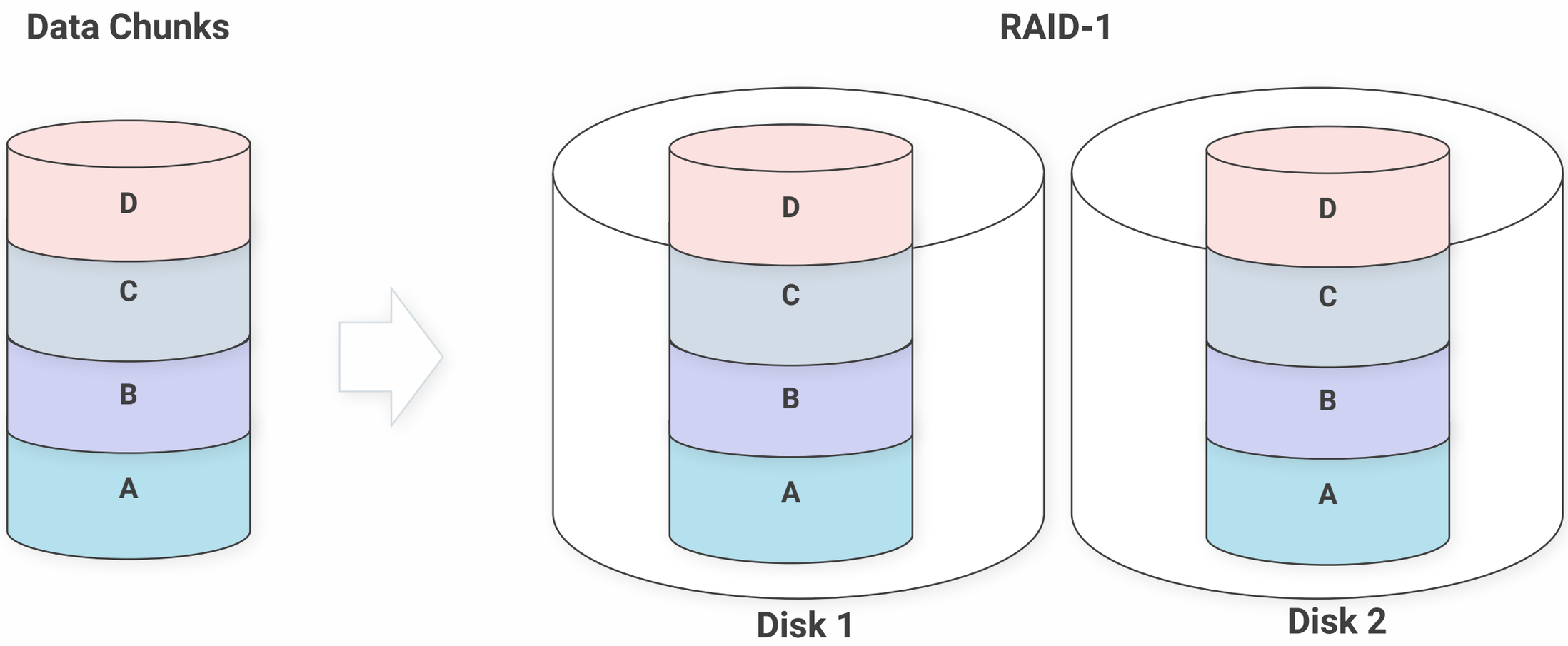
Преимущества

+ Самое быстрое чтение
+ Очень простой
+ Максимальная эффективность использования дискового пространства

Недостатки

- Не «настоящий» RAID, нет отказоустойчивости: отказ одного диска влечет за собой потерю всех данных массива

RAID-1



$$V_t = V_1$$

RAID-1

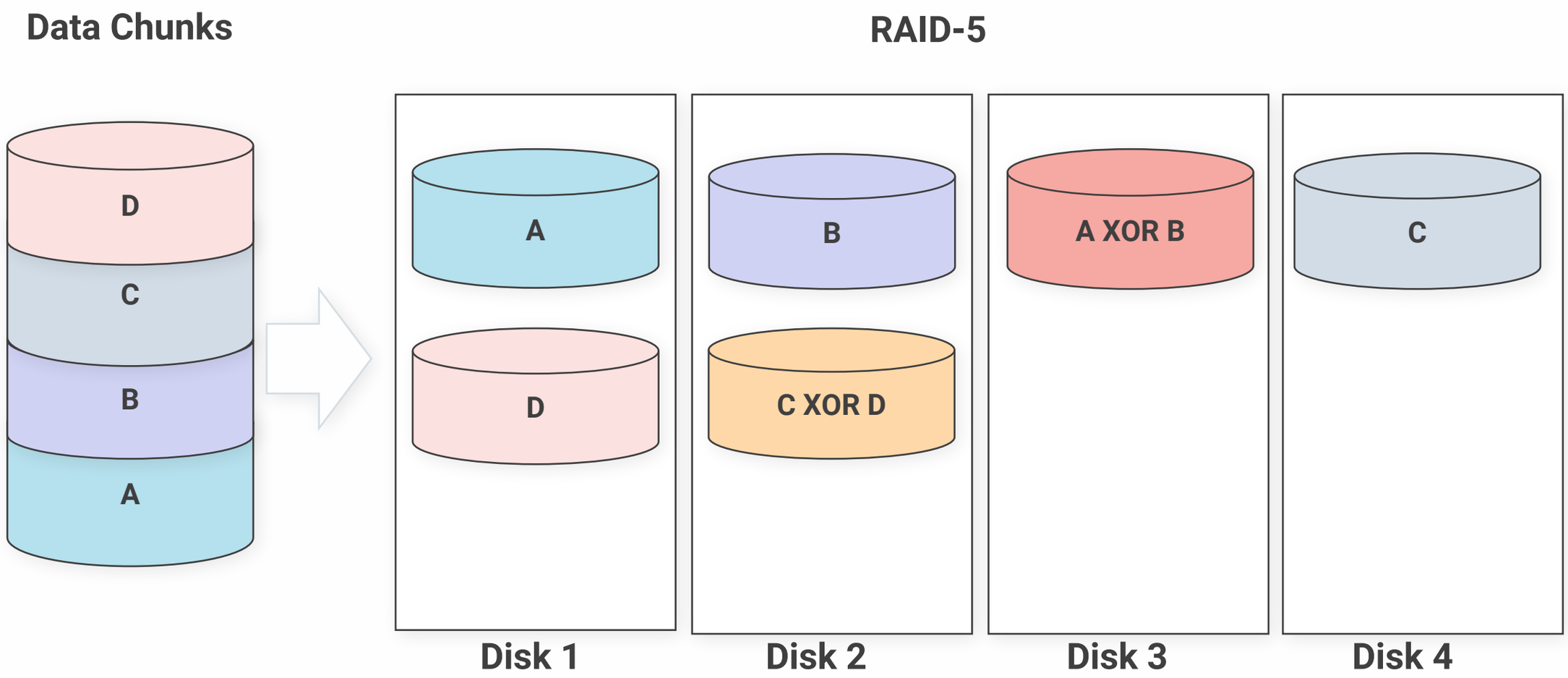
Преимущества

+ Простота реализации
+ Простота восстановления: перекопировать все данные с «выжившего» диска
+ Высокая скорость на чтение

Недостатки

- Высокая стоимость на единицу объема: 100% избыточность

RAID-5



$$V_t = V_1 * (N-1)$$

RAID-5

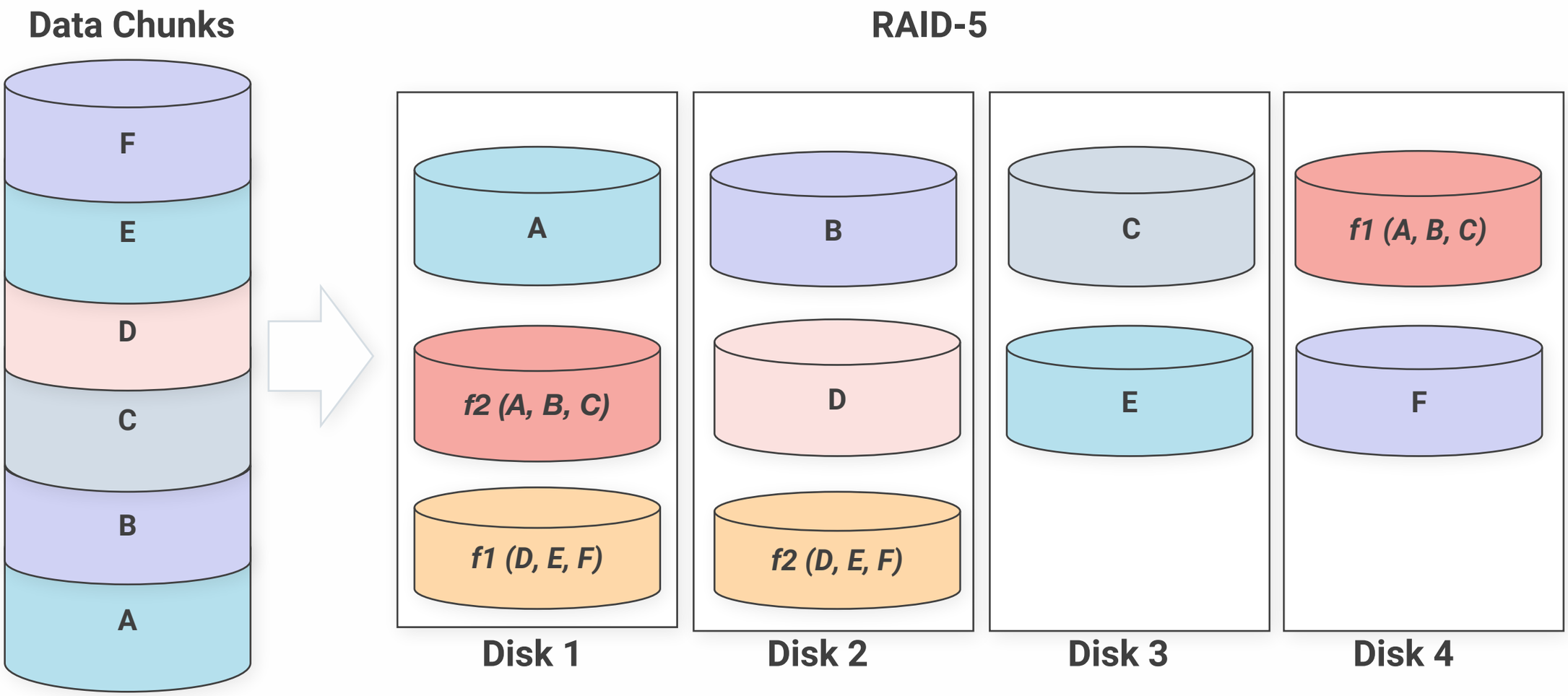
Преимущества

+ Высокая скорость записи данных
+ Достаточно высокая скорость чтения данных
+ Высокая производительность при большой интенсивности запросов чтения/записи данных
+ Малые накладные расходы для реализации избыточности

Недостатки

- Низкая скорость чтения/записи данных малого объема при единичных запросах
- Достаточно сложная реализация
- Сложное восстановление данных

RAID-6



$$V_t = V_1 * (N-2)$$

RAID-6

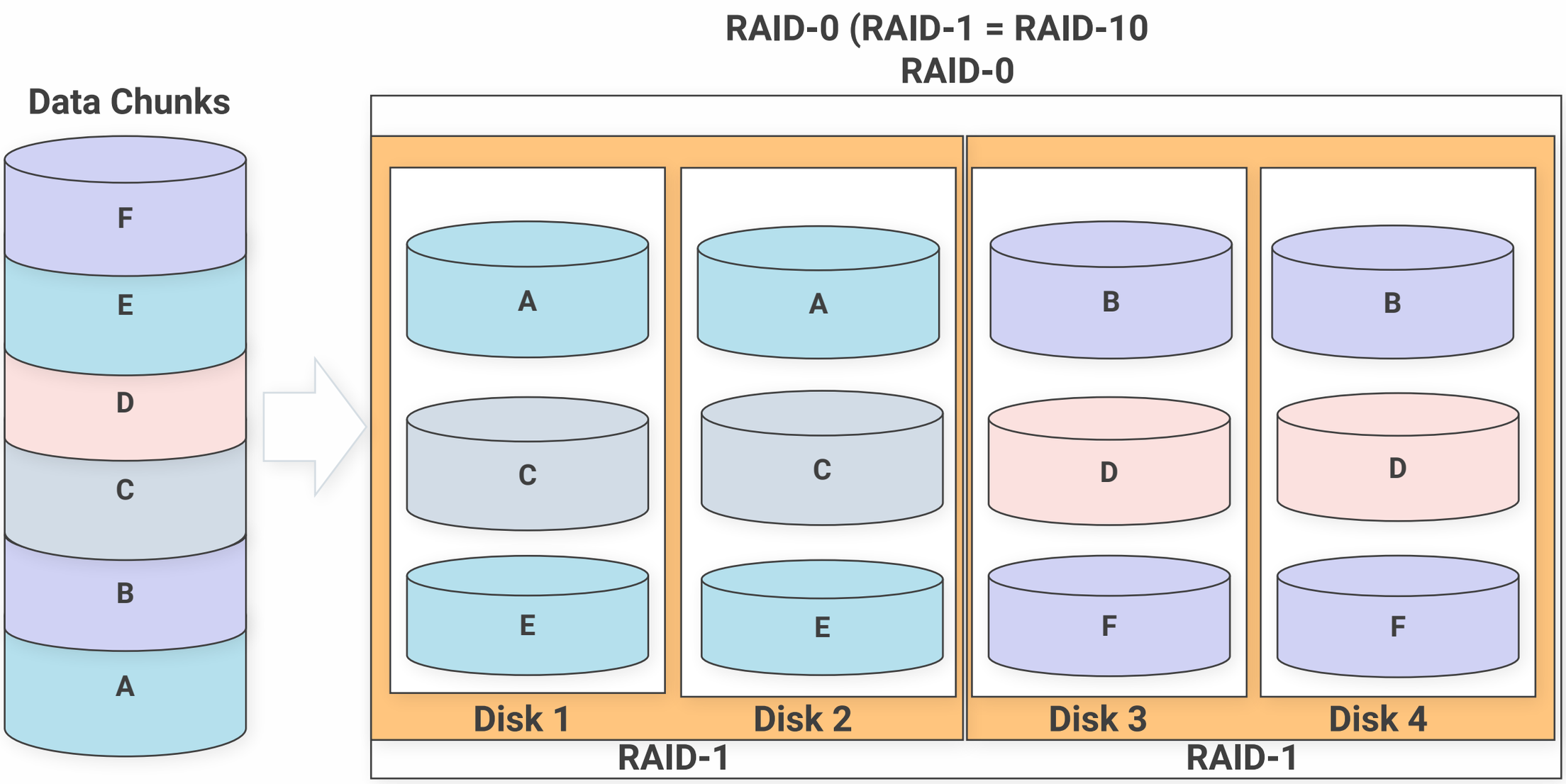
Преимущества

+ Высокая отказоустойчивость
+ Достаточно высокая скорость обработки запросов
+ Относительно малые накладные расходы для реализации избыточности

Недостатки

- Очень сложная реализация
- Сложное восстановление данных
- Очень низкая скорость записи данных

RAID-10



$$V_t = V_1 * N/2$$

RAID-10

Преимущества

+ Самая высокая отказоустойчивость
+ Самая высокая производительность
+ Сочетает в себе преимущества R0 и R1

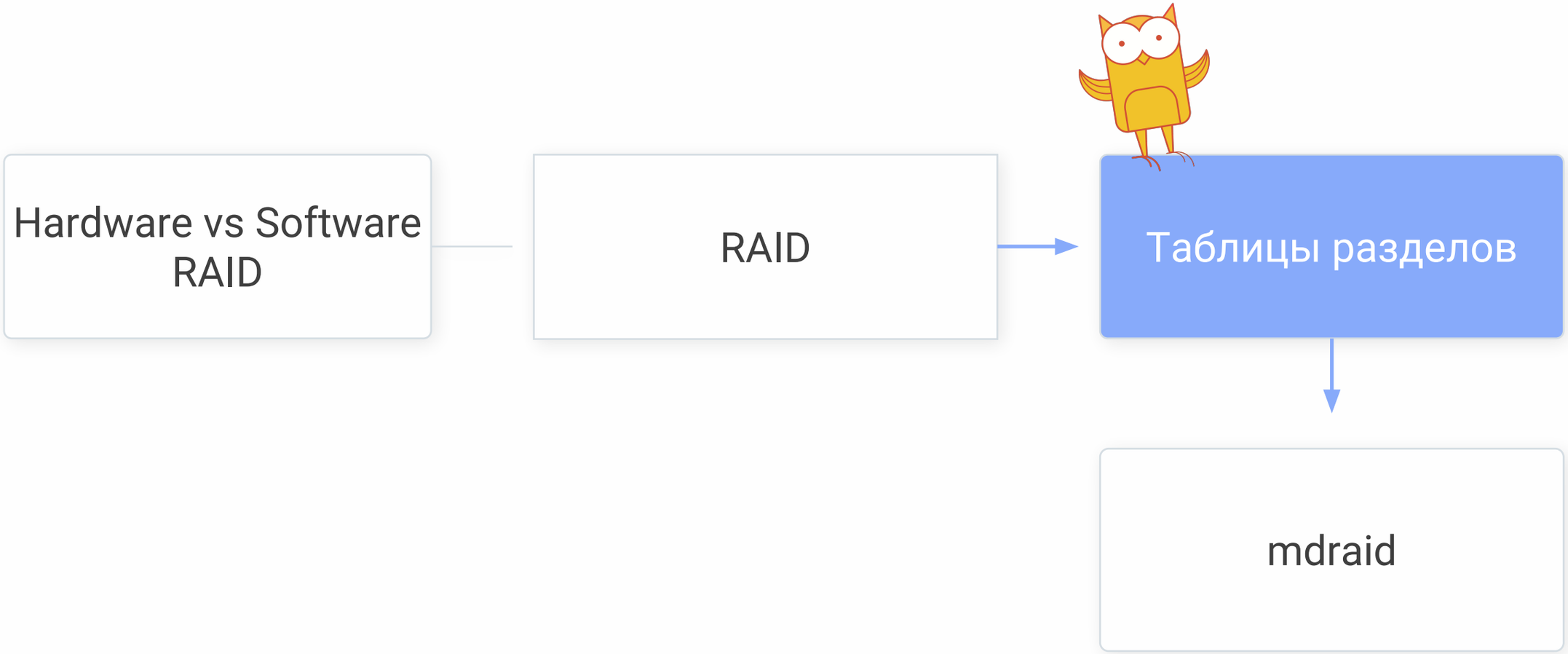
Недостатки

- Двойная стоимость пространства

Экзотические RAID

- 1E (запись по очереди в один из дисков, копия в следующий)
- 5+0 (RAID0 из RAID5)
- 5+1 (RAID1 из RAID5)
- 6+0 (RAID0 из RAID6)
- 6+1 (RAID1 из RAID6)

Маршрут вебинара



Таблицы разделов: MBR

Преимущества

+ Самый популярный и совместимый

Недостатки

- Максимум 4 раздела на диске
- Если первые сектора диска повреждены, диск перестает читаться
- Максимальный раздел — 2.1 Tb

MBR - Master boot record

Таблицы разделов: MBR

- 512 байт
- Смещение Длина Описание
  000h 446 Код загрузчика
  1BEh 64 Таблица разделов
  16 Раздел 1
  1CEh 16 Раздел 2
  1DEh 16 Раздел 3
  1EEh 16 Раздел 4
  1FEh 2 Сигнатура (55h AAh)

Таблицы разделов

16 байтный блок раздела

00h	1	Признак активности раздела
01h	1	Начало раздела - головка
02h	1	Начало раздела - сектор (биты 0-5), дорожка (биты 6,7)
03h	1	Начало раздела - дорожка (старшие биты 8,9 хранятся в байте номера сектора)
04h	1	Код типа раздела
05h	1	Конец раздела - головка
06h	1	Конец раздела - сектор (биты 0-5), дорожка (биты 6,7)
07h	1	Конец раздела - дорожка (старшие биты 8,9 хранятся в байте номера сектора)
08h	4	Смещение первого сектора
0Ch	4	Количество секторов раздела

Таблицы разделов

- Backup MBR
- **# dd if=/dev/sdX of=/tmp/sda.mbr bs=512 count=1**
- Восстановление MBR с сохранением таблицы разделов
- **# dd if=/tmp/sda.mbr of=/dev/sdb bs=446 count=1**
- Утилиты для работы

  fdisk, sfdisk - создание/удаление партиций

Таблицы разделов: GPT или GUID

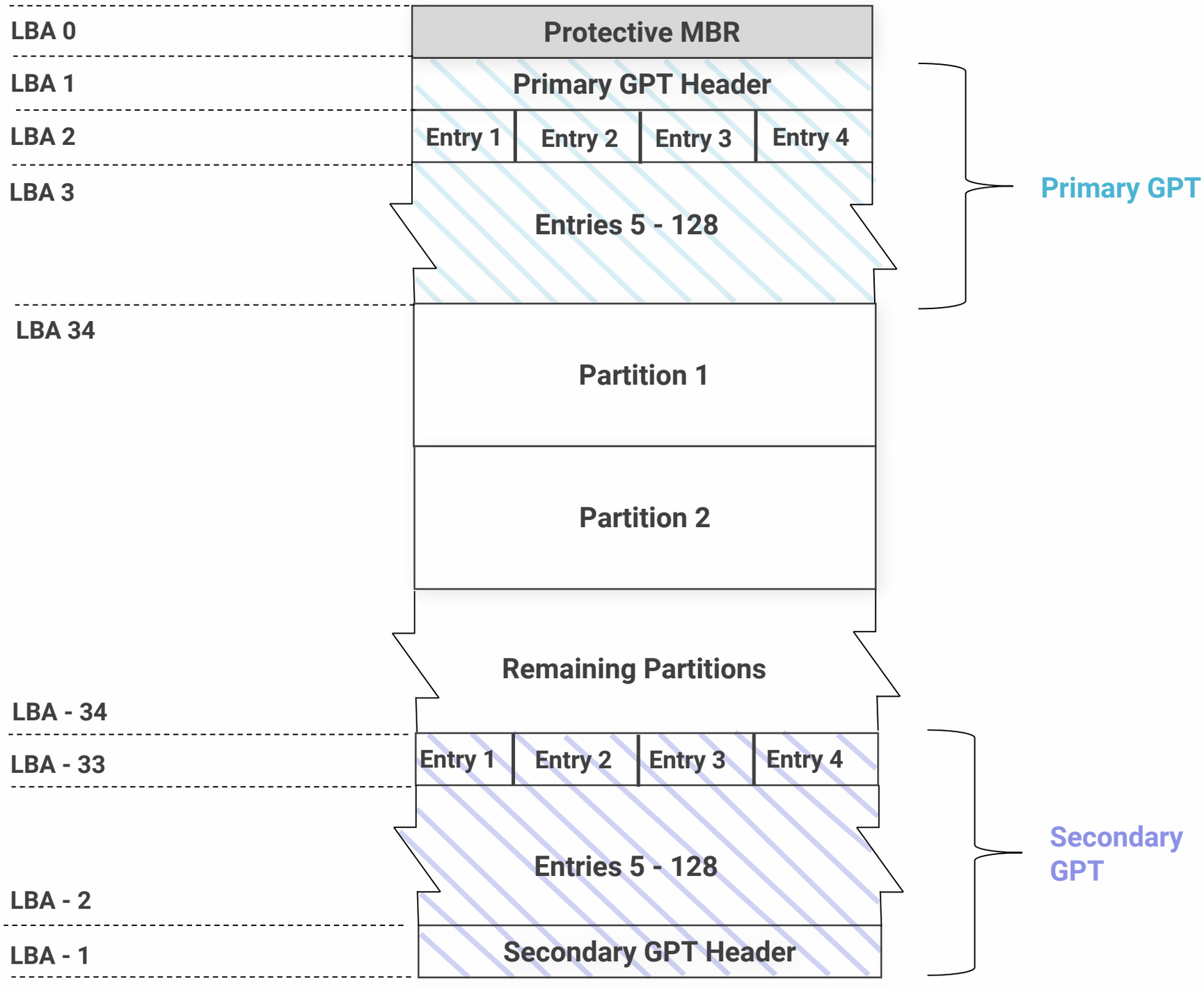
Преимущества

+ Неограниченное кол-во разделов
+ Очень большие ограничения на объем
+ Раздел зарезервирован, хранятся контрольные CRC-суммы, в случае проблем возможно восстановление

Недостатки

- Не поддерживается старыми системами

Таблицы разделов



Таблицы разделов

- Backup GPT
- **# sgdisk --backup=sda.gpt.bkp /dev/sda**
- Восстановление GPT
- **# sgdisk --load-backup=sda.gpt.bkp /dev/sda**
- Утилиты для работы:

  gdisk, sgdisk

Полезные команды

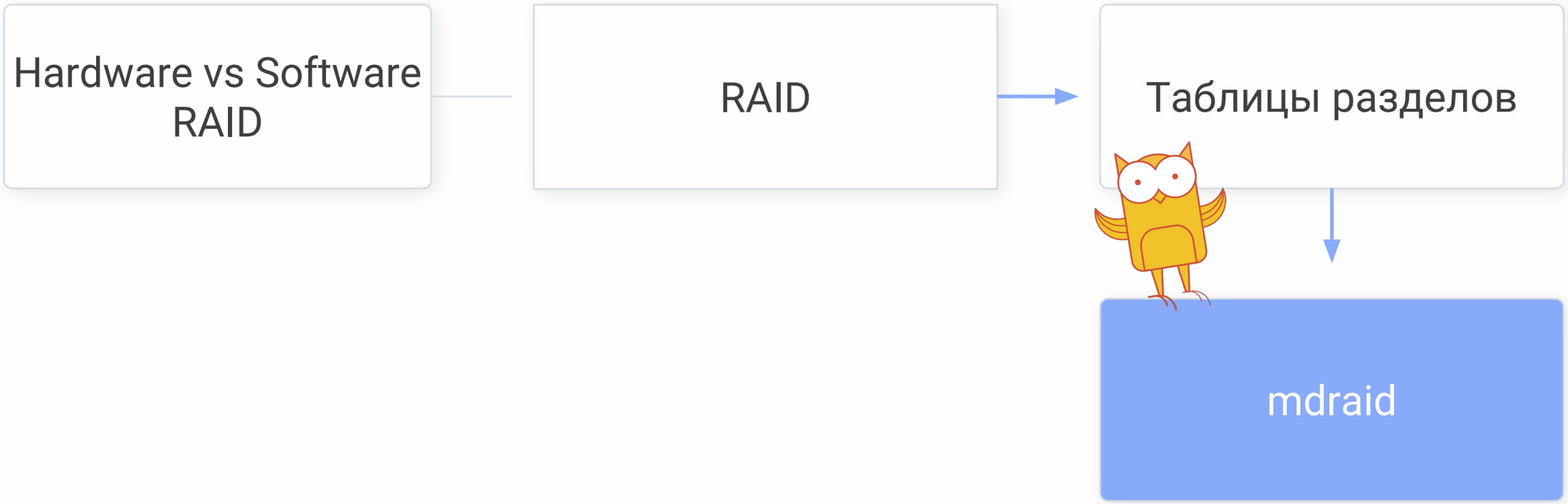
Утилиты для работы с партициями

- gdisk
- parted
- partx
- partprobe

Утилиты для работы с информацией о железе

- smartctl
- hdparm
- dmesg
- dmidecode

Маршрут вебинара



mdraid: элементы

Модули ядра

Утилита управления - mdadm

Утилита мониторинга

mdraid: элементы

Блочные устройства

- разделы
- диски
- тома lvm

Метаданные

- 0.9, 1.0 - конец устройства (необходимо для загрузки в некоторых случаях)
- 1.1 - начало
- 1.2 - 4K от начала устройства

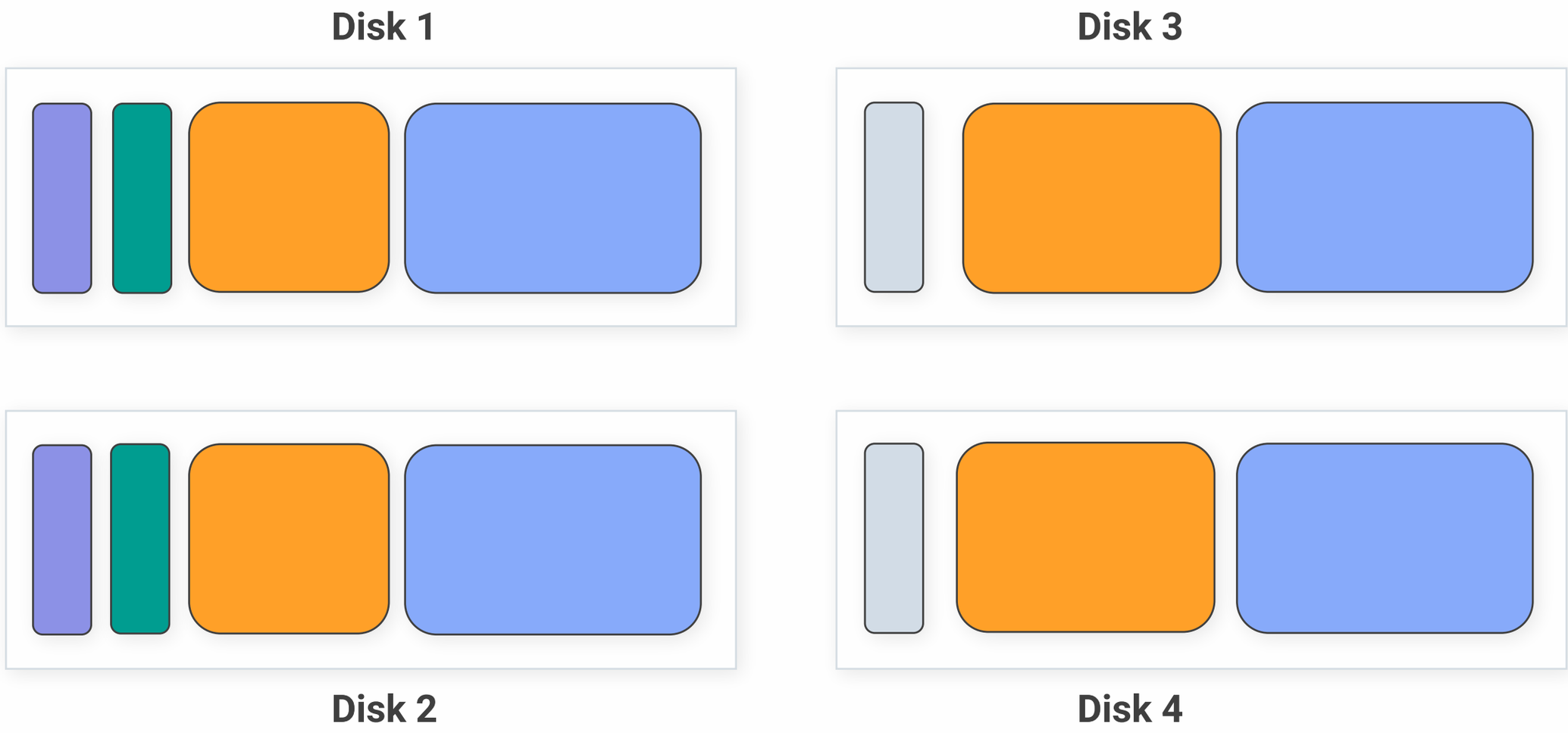
mdraid

Массивы mdraid можно создавать из любых блочных устройств:

- Дисков,
- Разделов,
- Томов lvm.

Наиболее **безопасно** создавать RAID поверх разделов, это может помочь сгладить разный размер дисков, позволит на одном наборе дисков создать разные RAID.

Пример организации дисков



Как использовать mdraid

- Подготовка к созданию, “занулить суперблок”

  mdadm --zero-superblock $dev_list

- Создание массива

  mdadm --create $raiddev -l $level -n $numdev $dev_list

- Остановка массива

  mdadm -S $raiddev

- Информация о массиве

  mdadm --detail $raiddev

- Генерация данных для конфигурационного файла

  mdadm --examine --scan

  mdadm --detail --scan

Как использовать mdraid

- Информация о массиве

  cat /proc/mdstat

- Запуск/остановка проверки

  echo (check|idle) > /sys/block/md${N}/device/action

- Изменение ограничений скорости ребилда

 # grep . /proc/sys/dev/raid/speed_limit_m*

/proc/sys/dev/raid/speed_limit_max:200000

/proc/sys/dev/raid/speed_limit_min:1000

 # echo 10000 > /proc/sys/dev/raid/speed_limit_min

Ваши вопросы?

Домашнее задание: работа с mdadm

Задание:

- добавить в Vagrantfile еще дисков
- сломать/починить raid
- собрать R0/R5/R10 на выбор
- прописать собранный рейд в конф, чтобы рейд собирался при загрузке
- создать GPT раздел и 5 партиций

В качестве проверки принимаются - измененный Vagrantfile, скрипт для создания рейда, конф для автосборки рейда при загрузке.

* доп. задание - Vagrantfile, который сразу собирает систему с подключенным рейдом

Рефлексия

Напишите, пожалуйста, свое впечатление о вебинаре.

- Отметьте 3 пункта, которые вам запомнились с вебинара.
- Что вы будете применять в работе из сегодняшнего вебинара?

Заполните, пожалуйста, опрос в ЛК о занятии

Спасибо
за внимание!
До встречи в Slack и на вебинаре