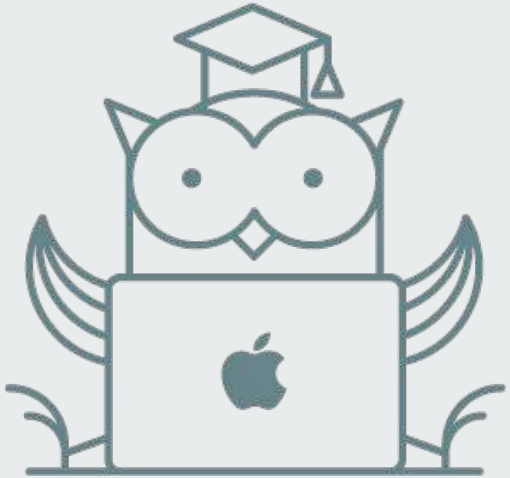
Курс «Администратор Linux»

Дисковая подсистема и RAID

Занятие # 2

Дмитрий Молчанов
Григорий Ожегов



Дисковая подсистема

- Немного истории
- Hardware vs Software RAID
- RAID
  - RAID0
  - RAID1
  - RAID5
  - RAID6
  - RAID10
- Таблицы разделов
- mdraid

Немного истории

- Скорость дисков
- Оптимальность работы: PATA vs SCSI
  - PATA - Parallel ATA.
    - − До 2х устройств на канал.
    - − Неоптимальное чтение-запись.
    - \+ Низкая стоимость (в сравнении со SCSI)
  - SCSI - Small Computers System Interface
    - \+ Не только диски.
    - \+ Более оптимальное чтение
    - \+ RAID-контроллеры
      - Кэш + Батарейка
- SATA и SAS
- Производительность дисков обороты растут, скорость шины растет
- Аппетиты на хранение данных растут
- Рост популярности web-приложений, запрос на дешевые хранилища

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.

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.

Hardware RAID

+ 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)

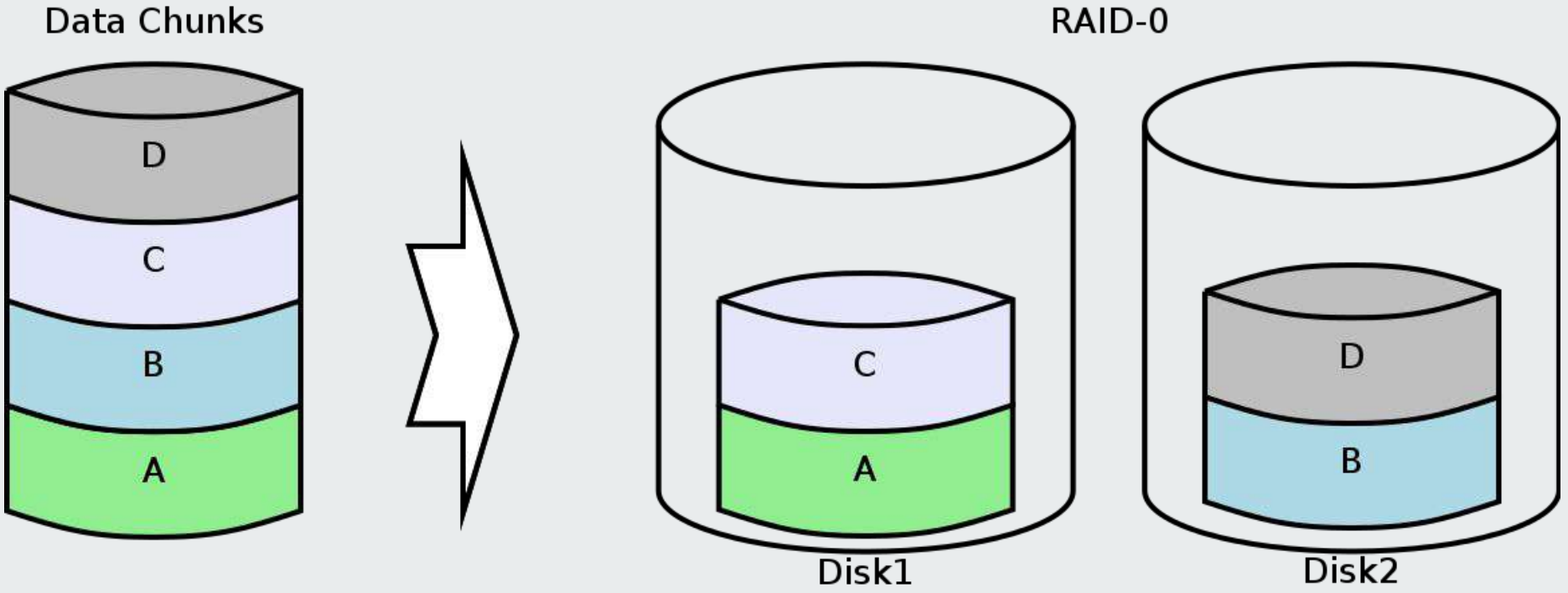
- Высокая стоимость
- Высокая сложность
- Разнообразность интерфейсов управления и драйверов
- Низкая «мобильность»/переносимость
- Привязка к железу
- Бóльший простой по времени при аварии
- Очень дорогой ремонт, необходимость закупать впрок контроллеры, которые потом могут прекратить выпускать

Software RAID (mdraid)

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

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

RAID-0



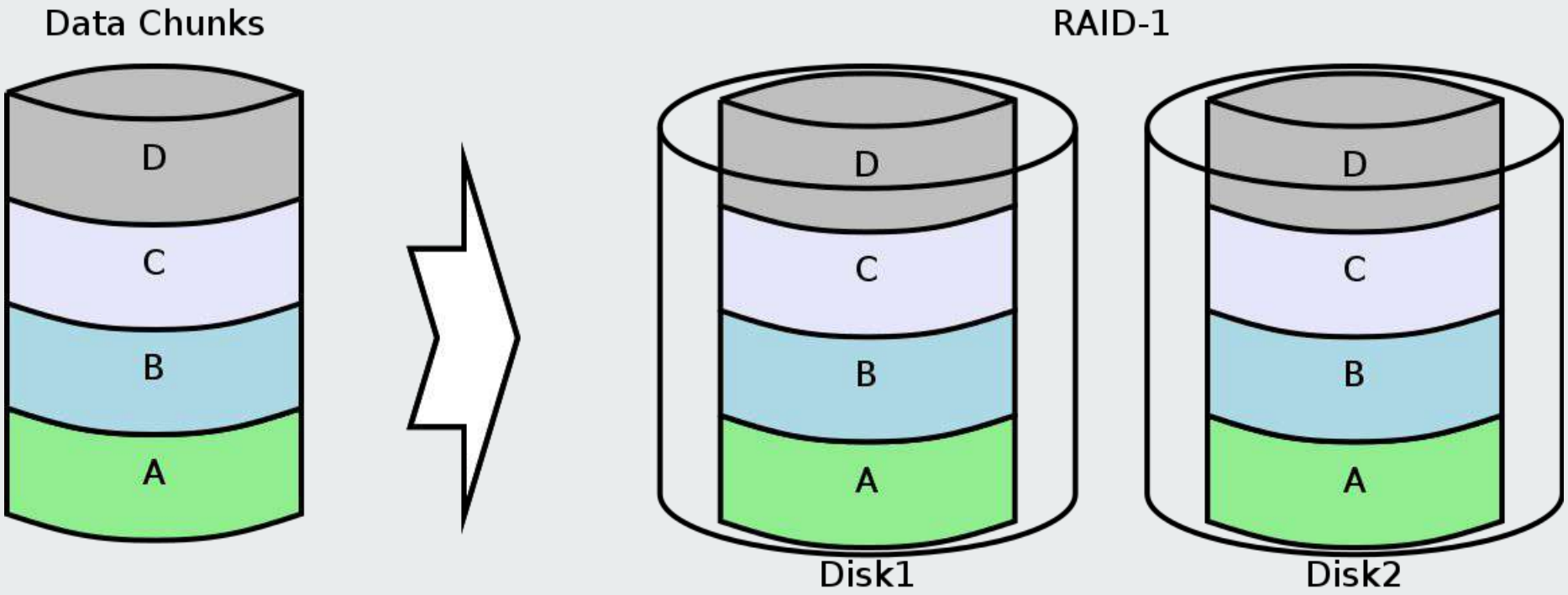
$$\mathcal{V}_t = \mathcal{V}_1 * \mathcal{N}$$

RAID-0

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

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

RAID-1



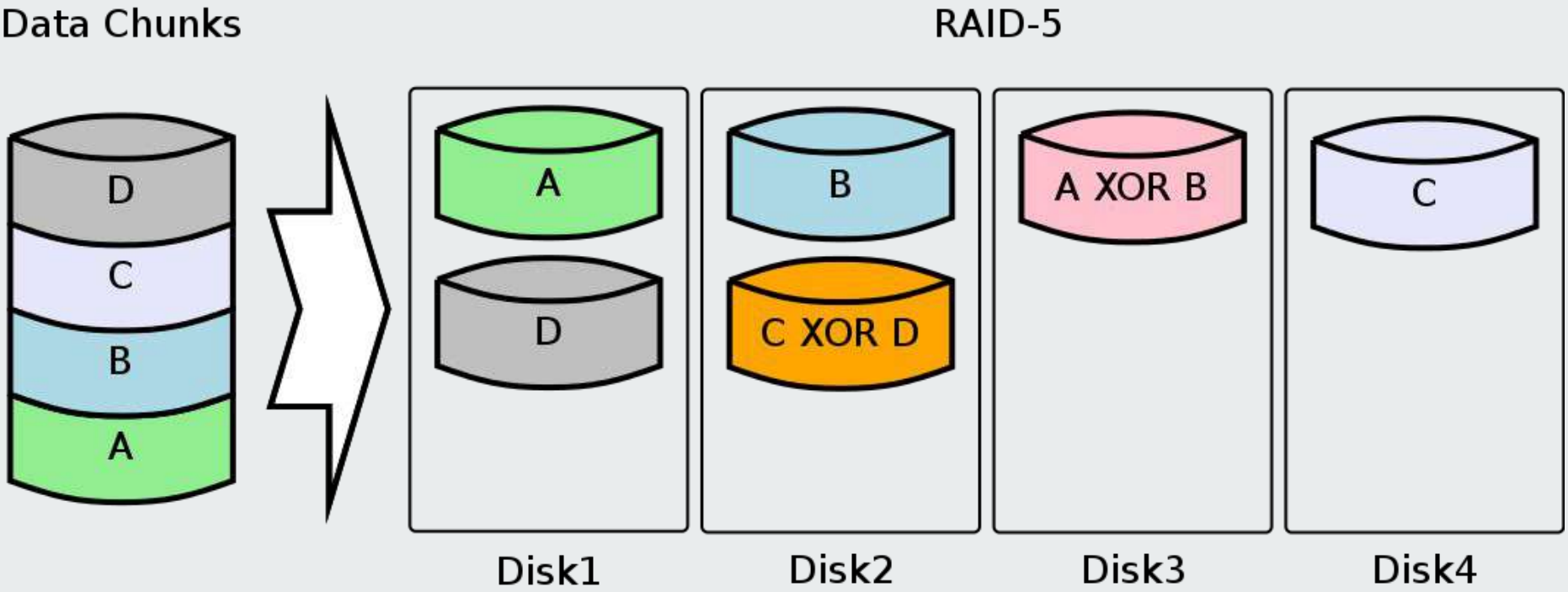
$$V_t = V_1$$

RAID-1

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

– Высокая стоимость на единицу объема: 100% избыточность

RAID-5



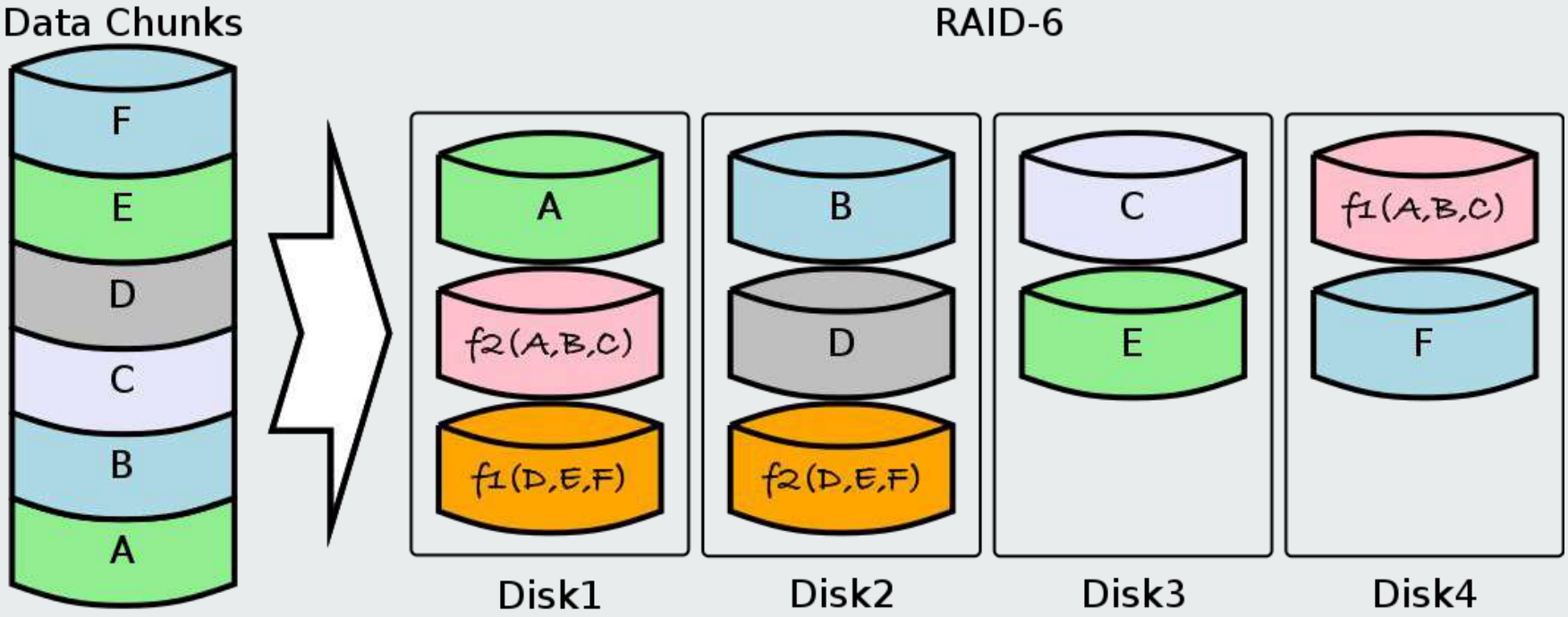
$$V_t = V_1 * (N-1)$$

RAID-5

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

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

RAID-6



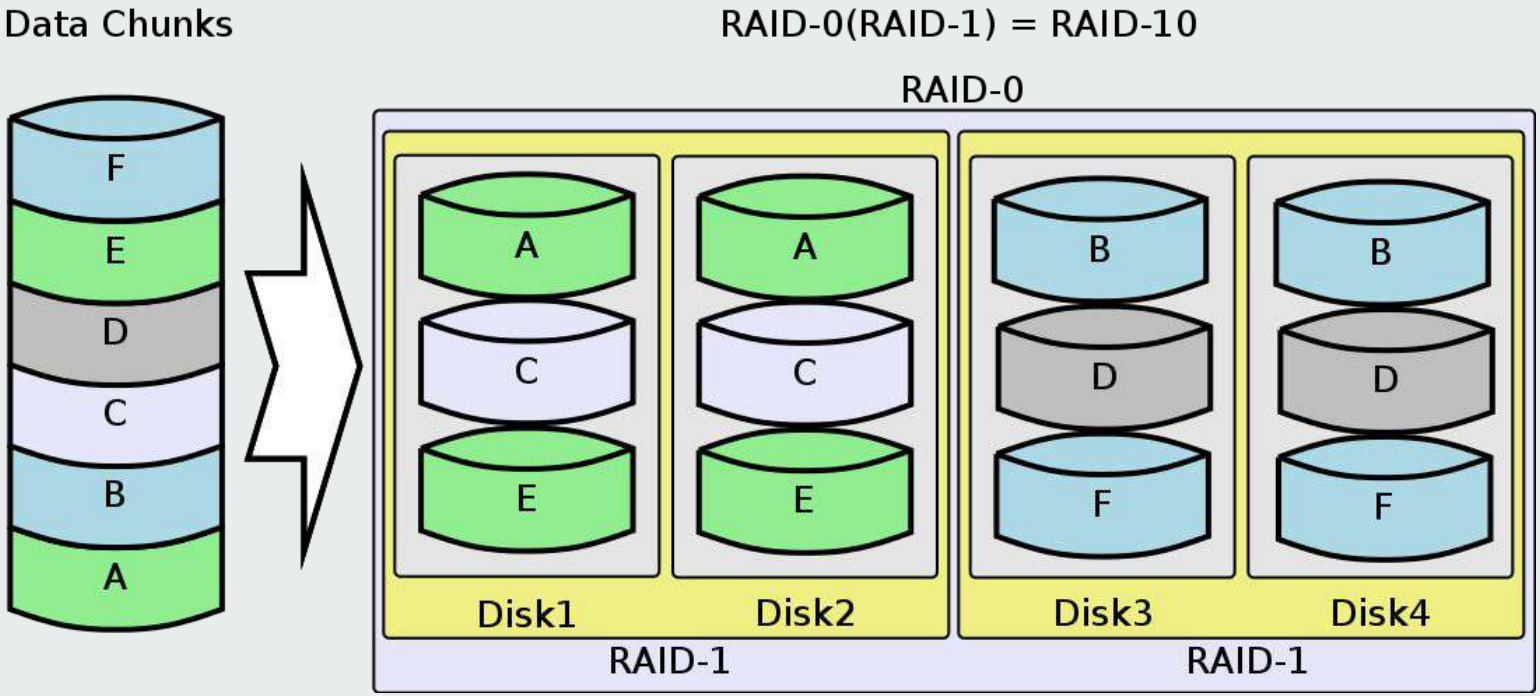
$$V_t = V_1 * (N-2)$$

RAID-6

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

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

RAID-10



$$\mathcal{V}_t = \mathcal{V}_1 * \mathcal{N}/2$$

RAID-10

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

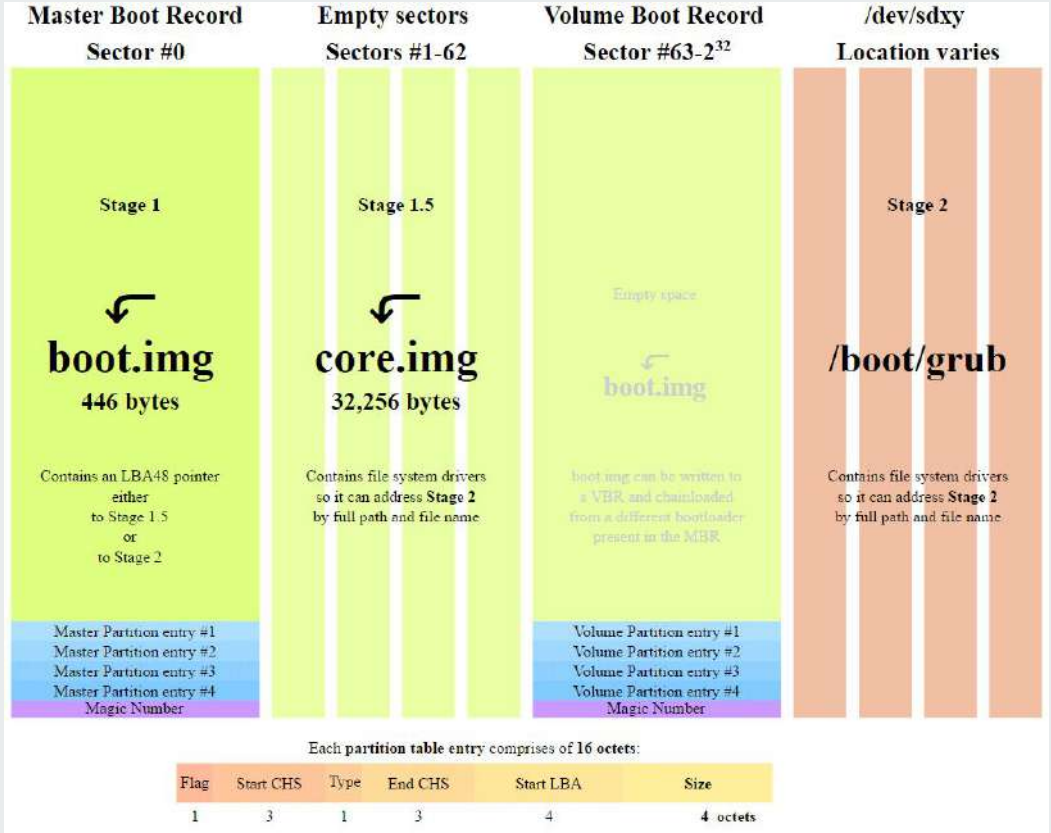
– Двойная стоимость пространства

Экзотические RAID

- 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)

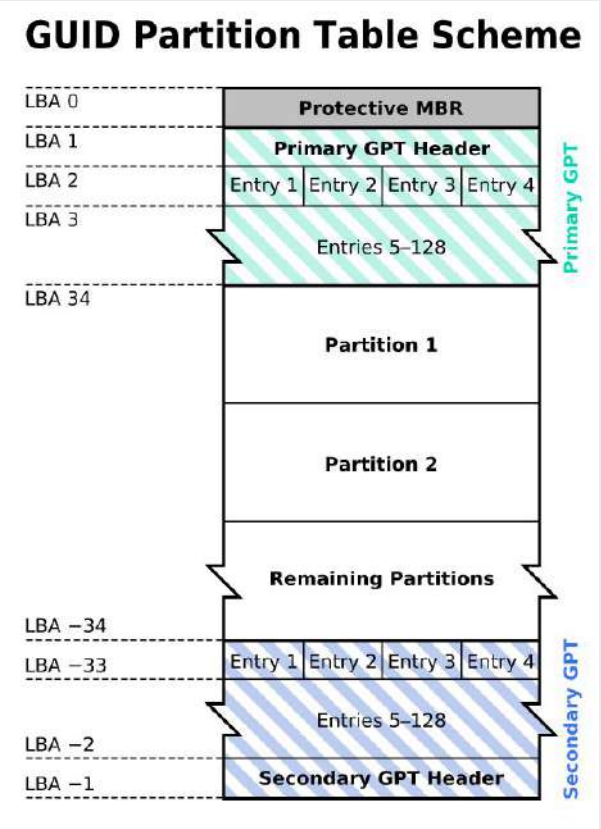
Таблицы разделов

- MBR (Master boot record)
  - \+ Самый популярный и совместимый
  - – Максимум 4 раздела на диске
  - – Если первые сектора диска повреждены, диск перестает читаться
  - – Максимальный раздел — 2.1 Tb



Таблицы разделов

- GPT
  - \+ Неограниченное кол-во разделов
  - \+ Очень большие ограничения на объем
  - \+ Раздел зарезервирован, хранятся контрольные CRC-суммы, в случае проблем возможно восстановление
  - – Не поддерживается старыми системами



mdraid: компоненты

- Модули ядра
- Утилита управления
- Утилита мониторинга

mdraid: элементы

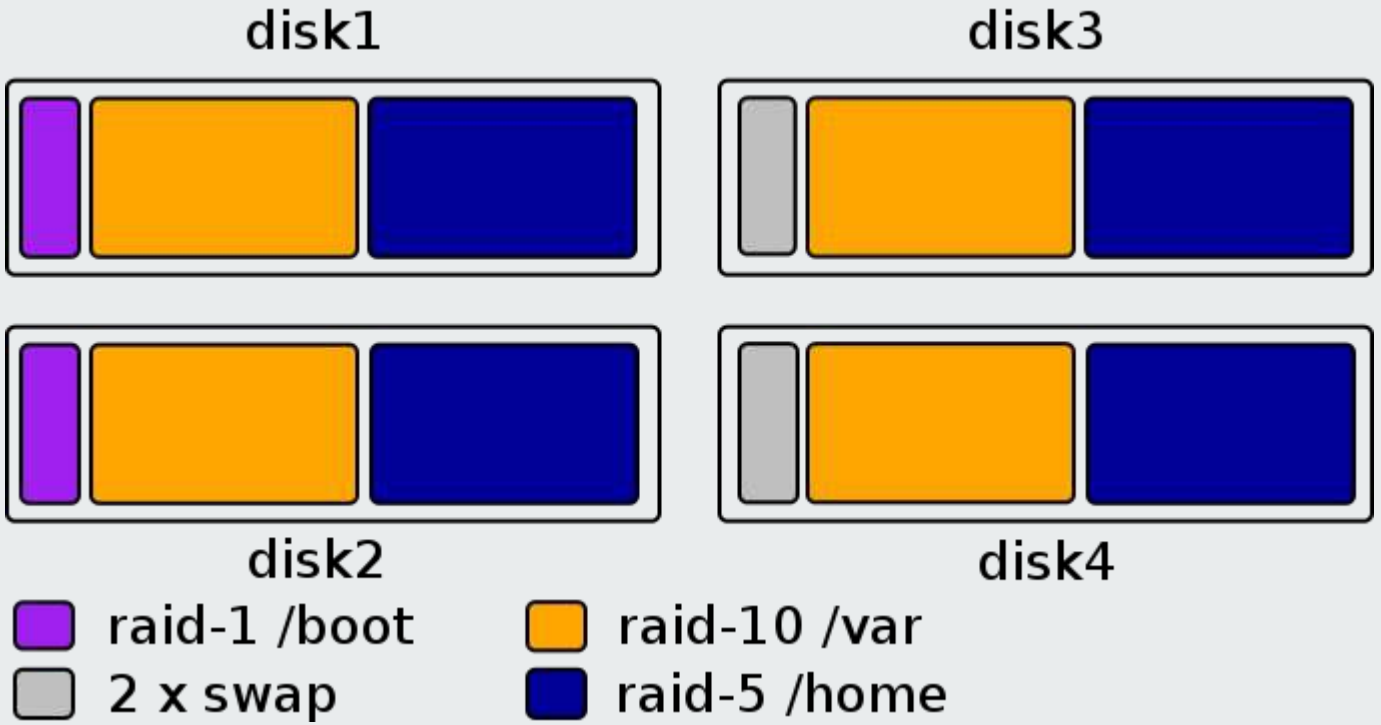
- Блочные устройства
  - разделы
  - диски
  - тома lvm
- Метаданные
  - 0.9,1.0 - конец устройства (необходимо для загрузки в некоторых случаях)
  - 1.1 - начало
  - 1.2 - 4K от начала устройства

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
```

Спасибо за внимание

Дмитрий Молчанов
Григорий Ожегов