Создание программных RAID-массивов в Linux

- RAID (англ. redundant array of independent/inexpensive disks — избыточный массив независимых/недорогих жёстких дисков) — массив из нескольких дисков, управляемых контроллером, взаимосвязанных скоростными каналами и воспринимаемых внешней системой как единое целое.
- В зависимости от типа используемого массива может обеспечивать различные степени отказоустойчивости и быстродействия.
- Служит для повышения надёжности хранения данных и/или для повышения скорости чтения/записи информации (RAID 0).

уровни спецификации RAID

- RAID 0 представлен как неотказоустойчивый дисковый массив.
- RAID 1 определён как зеркальный дисковый массив.
- RAID 2 зарезервирован для массивов, которые применяют код Хемминга.
- RAID 3, 4, 5 используют чётность для защиты данных от одиночных неисправностей.
- RAID 6 используют чётность для защиты данных от двойных неисправностей

RAID 0: Дисковый массив без отказоустойчивости (Striped Disk Array)

Диск 1	Диск 2	Диск 3	Диск 4
A	B	C	D
E	F	G	H
I	J	K	L
M	N	O	P
…	…	…	…

RAID 1: Дисковый массив с зеркалированием

Диск 1	Диск 2
A	A
B	B
C	C
D	D
…	…

RAID 2: Отказоустойчивый дисковый массив с использованием кода Хемминга

Диск 1	Диск 2	Диск 3	Диск 4	Диск 5	Диск 6	Диск 7
A	B	C	D	ECC1: ABCD	ECC2: ABCD	ECC3: ABCD
E	F	G	H	ECC1: EFGH	ECC2: EFGH	ECC3: EFGH
I	J	K	L	ECC1: IJKL	ECC2: IJKL	ECC3: IJKL
M	N	O	P	ECC1: MNOP	ECC2: MNOP	ECC3: MNOP
…	…	…	…	…	…	…

RAID 3: Отказоустойчивый массив с параллельной передачей данных и четностью

Диск 1	Диск 2	Диск 3	Диск 4	Диск 5
A.1	A.2	A.3	A.4	ECC: A
B.1	B.2	B.3	B.4	ECC: B
C.1	C.2	C.3	C.4	ECC: C
D.1	D.2	D.3	D.4	ECC: D
...	...	...	...	...

XOR: исключающее ИЛИ

X	Y	X⊕Y
0	0	0
0	1	1
1	0	1
1	1	0

RAID 4: Массив независимых дисков с разделяемым диском четности

Диск 1	Диск 2	Диск 3	Диск 4	Диск 5
A	B	C	D	ECC: ABCD
E	F	G	H	ECC: EFGH
I	J	K	L	ECC: IJKL
M	N	O	P	ECC: MNOP
...	...	...	...	...

RAID 5: Отказоустойчивый массив независимых дисков с распределенной четностью

Диск 1	Диск 2	Диск 3	Диск 4	Диск 5
A	B	C	D	ECC: ABCD
E	F	G	ECC: EFGH	H
I	J	ECC: IJKL	K	L
M	ECC: MNOP	N	O	P
…	…	…	…	…

RAID 6: Отказоустойчивый массив независимых дисков с двумя независимыми распределенными схемами четности

- Обеспечивает надежное хранение данных при выходе из строя до **двух** дисков
- Два основных подхода
  - ECC независимые по данным
  - ECC независимые по алгоритмам
- Несколько различных реализаций:
  - EVENODD
  - X-Code
  - С кодами Рида-Соломона (Reed-Solomon)
  - …

RAID 6: EVENODD

Диск 1	Диск 2	Диск 3	Диск 4	Диск 5	Диск 6
A	B	C	D	P:ABCD	Q:ALOS
E	F	G	H	P:EFGH	Q:BEPS
I	J	K	L	P:IJKL	Q:CFIS
M	N	O	P	P:MNOP	Q:DGJMS

$S=H \oplus K \oplus N$

RAID 6: EVENODD

- Коды четности распределены по дискам
- P – XOR внутри горизонтальных групп
- Q – XOR внутри диагональных групп
- Случайная запись вызывает 6 операций ввода/вывода для 13 блоков и 12 для 3 блоков

RAID 6: X-Code

Диск 1	Диск 2	Диск 3	Диск 4	Диск 5
A	B	C	D	E
F	G	H	I	J
K	L	M	N	O
P:CIO	P:DJK	P:EFL	P:AGM	P:BHN
Q:DHL	Q:EIM	Q:AJN	Q:BFO	Q:CGK

RAID 6: X-code

- Количество дисков должно быть простым числом
- P – XOR внутри диагональных групп слева направо
- Q – XOR внутри диагональных групп справа налево
- Случайная запись вызывает 6 операций ввода/вывода

RAID 6: Reed-Solomon

Диск 1	Диск 2	Диск 3	Диск 4	Диск 5	Диск 6
A	B	C	D	XOR ABCD	R-S ABCD
E	F	G	XOR EFGH	R-S EFGH	H
I	J	XOR IJKL	R-S IJKL	K	L
M	XOR MNOP	R-S MNOP	N	O	P
XOR QRST	R-S QRST	Q	R	S	T

RAID 6: Reed-Solomon

- XOR внутри горизонтальных групп
- R-S внутри горизонтальных групп
- Случайная запись вызывает 6 операций ввода/вывода
- Может быть расширен для обеспечения надежного хранения данных в случае отказа большего числа дисков (>2)

RAID 1+0: Отказоустойчивый массив с дублированием и параллельной обработкой

Диск 1	Диск 2	Диск 3	Диск 4	Диск 5	Диск 6
A	A	B	B	C	C
D	D	E	E	F	F
G	G	H	H	I	I
…	…	…	…	…	…

RAID 0+1: Отказоустойчивый массив с параллельной обработкой и зеркалированием

Диск 1	Диск 2	Диск 3	Диск 4	Диск 5	Диск 6
A	B	C	A	B	C
D	E	F	D	E	F
G	H	I	G	H	I
…	…	…	…	…	…

RAID 5+0. Отказоустойчивый массив с распределенными блоками четности и повышенной производительностью

Диск 1	Диск 2	Диск 3	Диск 4	Диск 5	Диск 6
A	B	P: AB	C	D	P:CD
E	P:EF	F	G	P:GH	H
P:IJ	I	J	P:KL	K	L
…	…	…	…	…	…

RAID 1E: Отказоустойчивый массив с двунаправленным зеркалированием

Диск 1	Диск 2	Диск 3
A	A	B
B	C	C
D	D	E
E	F	F
...	...	...

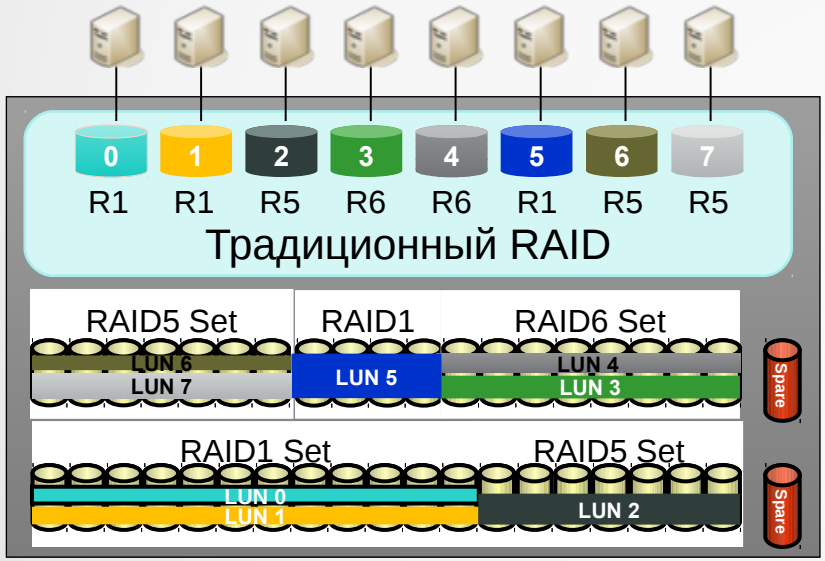
Hot Spare HDDs: Диски «горячей» подмены

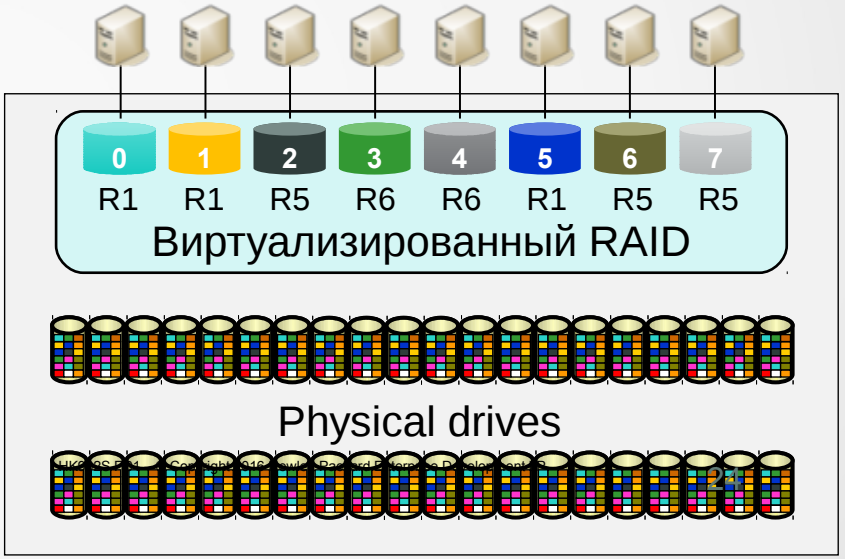
- Предназначены для замены вышедших из строя HDD без участия человека
- В нормальном режиме работы не используются
- Могут быть общими для всех групп в комбинированных уровнях RAID

RAID 5EE: Отказоустойчивый массив независимых дисков с распределенными четностью и диском горячей подмены

Диск 1	Диск 2	Диск 3	Диск 4	Диск 5	Диск 6
A	B	C	D	XOR ABCD	Hot spare
E	F	G	XOR EFGH	Hot spare	H
I	J	XOR IJKL	Hot spare	K	L
M	XOR MNOP	Hot spare	N	O	P
XOR QRST	Hot spare	Q	R	S	T

Виртуализация RAID





Реализации RAID

- Аппаратный RAID-контроллер
  - управляет всем так, что дисковый массив виден как один диск даже на самом низком уровне.
- Программный RAID,
  - использует ПО операционной системы для объединения отдельных устройств в RAID-массив.
- Гибридный RAID («фальшивым» (fake-raid))

- Управление программным RAID-массивом в Linux выполняется с помощью программы **mdadm**
- Режимы работы **mdadm**
  - Assemble (сборка)
    - Собрать компоненты ранее созданного массива в массив.
  - Build (построение)
    - Собрать массив из компонентов, у которых нет суперблоков.
  - Create (создание)
    - Создать новый массив на основе указанных устройств.

- Режимы работы **mdadm** (продолжение)
  - Monitor (наблюдение)
    - Следить на изменением состояния устройств. Для RAID0 этот режим не имеет смысла.
  - Grow (расширение или уменьшение)
    - Расширение или уменьшение массива, включаются или удаляются новые диски.
  - Incremental Assembly (инкрементальная сборка)
    - Добавление диска в массив.
  - Manage (управление)
    - Разнообразные операции по управлению массивом, такие как замена диска и пометка как сбойного.
  - Misc (разное)
    - Действия, которые не относятся ни к одному из перечисленных выше режимов работы.
  - Auto-detect (автооообнаружение)
    - Активация автоматически обнаруживаемых массивов в ядре Linux.

- mdadm [mode] [array] [options]
- Режимы:
  - -A, --assemble — режим сборки
  - -B, --build — режим построения
  - -C, --create — режим создания
  - -F, --follow, --monitor — режим наблюдения
  - -G, --grow — режим расширения
  - -I, --incremental — режим инкрементальной сборки

Пример создания RAID 5

- сначала необходимо установить и правильно настроить необходимое оборудование.
- Разбиение на разделы для RAID отличается от разбиения одного отдельного диска:
  - Вместо типа раздела «Linux» (тип 83) или «подкачки Linux» (тип 82), все разделы, которые станут частью RAID-массива, должны иметь тип **«Linux raid auto» (тип FD)**.
- Понадобятся минимум три раздела
- например /dev/sda1 /dev/sdb1 /dev/sdc1

Пример создания RAID 5

```
mdadm --create --verbose /dev/md0 --level=5 \
--raid-devices=3 /dev/sda1 /dev/sdb1 /dev/sdc1
```

- `--create` создание RAID-массива
- `--level` для того чтобы создать RAID-массив 5 уровня.
- `--raid-devices` устройства, поверх которых будет собираться RAID-массив.

Проверка правильности сборки

- Убедиться, что RAID-массив проинициализирован корректно можно просмотрев файл /proc/mdstat. В этом файле отражается текущее состояние RAID-массива.

```
cat /proc/mdstat

    Personalities : [raid5]

    read_ahead 1024 sectors

    md0 : active raid5 sda1[2] sdb1[1] sdc1[0]

        4120448 blocks level 5, 32k chunk, algorithm 3 [3/3
  ] [UUU]



    unused devices: <none>
```

Или командой

```
mdadm --detail /dev/md0
```

Создание конфигурационного файла mdadm.conf

- Команда mdadm не нуждается в файле конфигурации, но будет использовать его, если он указан.
- Рекомендуется создать файл конфигурации, поскольку он позволяет документировать конфигурацию RAID.
- команда:

```
mdadm --detail --scan —verbose
```

Пример:

```
echo "DEVICE partitions" > /etc/mdadm/mdadm.conf

mdadm --detail --scan --verbose | awk '/ARRAY/ {print}'
  >> /etc/mdadm/mdadm.conf
```

работа с массивом

- Пометка диска как сбойного
  - `mdadm /dev/md0 --fail /dev/sda1`
  - `mdadm /dev/md0 -f /dev/sda1`
- Удаление сбойного диска
  - `mdadm /dev/md0 --remove /dev/sda1`
  - `mdadm /dev/md0 -r /dev/sda1`
- Добавление нового диска
  - `mdadm /dev/md0 --add /dev/sda1`
  - `mdadm /dev/md0 -a /dev/sda1`

работа с массивом

- Сборка существующего массива
  - `mdadm --assemble /dev/md0 /dev/sda1 /dev/sdb1 /dev/sdc1`
  - `mdadm --assemble –scan`
- Мониторинг функционирования массива:
  - `mdadm --monitor --mail=sysadmin --delay=300 /dev/md0`
- Удаление массива
  - `mdadm -S /dev/md0`
  - `mdadm --zero-superblock /dev/sda1`
  - `mdadm --zero-superblock /dev/sdb1`

работа с массивом

- Расширение массива
  - Сначала добавляется диск
    - `mdadm /dev/md0 --add /dev/sdd1`
  - Проверяем, что диск (раздел) добавился
    - `mdadm --detail /dev/hdh2`
    - `cat /proc/mdstat`
  - Если раздел действительно добавился, мы можем расширить массив
    - mdadm -G /dev/md0 --raid-devices=4
  - Убедитесь, что массив расширился
    - `cat /proc/mdstat`
  - обновить конфигурационный файл
    - `mdadm --detail --scan >> /etc/mdadm/mdadm.conf`
    - `vi /etc/mdadm/mdadm.conf`

Создание многоканального устройства с mdadm

- Команда mdadm также может использоваться для работы с оборудованием, поддерживающим любое число путей ввода/ вывода к отдельным дискам SCSI LUN.
- Главным назначением многоканального устройства хранения является обеспечение постоянного доступа к данным в случае сбоя оборудования.
- Команда mdadm включает дополнительный параметр опции level для определения отдельного устройства, которое будет доступным в случае сбоя пути ввода/ вывода.

- Команда создания многоканального устройства аналогична команде создания RAID устройства с единственной разницей — параметр уровня RAID будет замещен параметром ***multipath***.

```
mdadm -C /dev/md0 --level=multipath --raid-
  devices=4 /dev/sda1 /dev/sdb1

 /dev/sdc1 /dev/sdd1

Continue creating array? yes

mdadm: array /dev/md0 started.
```