



Ай-ТИ Фреш

ТЕХНИЧЕСКИЙ РАЗБОР

Восстановление RAID5 после двойного отказа: методология ITfresh

Посекторные образы ddrescue, анализ Events counter,
degraded-сборка mdadm и переход на RAID6

Июль 2026

itfresh.ru · ИТ-аутсорсинг для юридических лиц

Суть проблемы

Файловый сервер офиса на RAID5 после отказа второго диска останавливается целиком: массив уходит в inactive, сетевые папки недоступны, а «бэкап» нередко лежит на том же томе. Мы разбираем нашу процедуру: как без риска для оригиналов снять посекторные образы, собрать degraded-массив по метаданным mdadm и перейти на конфигурацию, где двойной отказ уже не фатален.

Почему это важно бизнесу

- Простой всей компании: без файлового сервера договоры, сканы и рабочие документы недоступны каждому сотруднику
- Неверная первая команда админа (--create, ребилд по оригиналам) превращает восстановимый массив в потерю данных навсегда
- Лабораторное восстановление стоит на порядок дороже программного и занимает недели вместо дней
- Диски одной закупки стареют синхронно: второй отказ приходит в окно ребилда, когда избыточности уже нет
- Мониторинг за 5 минут (mdmonitor + MAILADDR) снимает главный риск — незамеченный первый отказ

Ключевые параметры реализации

1.30

GNU ddrescue для посекторных образов: первый проход --no-scrape, добор битых зон -r3 по mapfile
по докам GNU ddrescue

7.5

smartmontools: smartd опрашивает атрибуты 5/197/198 каждые 30 минут, алерт при любом росте
по докам smartmontools

7 с

TLER у WD Red Pro (Error Recovery Control): диск отдаёт ошибку за секунды, а не виснет минутами
WD Red Pro WD4003FFBX

50 / 200

наши пороги Reallocated_Sector_Ct: warning и critical — диск в замену, не дожидаясь отказа
наш стандарт

1 раз/мес

scrub: echo check >
/sys/block/md0/md/sync_action — ловим латентные bad-блоки до ребилда
md(4), kernel.org

55 000 ч

порог плановой замены дисков по Power_On_Hours (~6,3 года) — до лавины отказов одной партии
наш стандарт

Триаж и посекторные образы через ddrescue

Что настраиваем

Файловый сервер на mdadm RAID5 4×WD Red Pro 4 ТБ (WD4003FFBX), массив в inactive: два диска помечены faulty

Как мы это делаем

- 1 Останавливаем ввод-вывод: ничего не пересобираем на живую, снимаем mdadm --detail и /proc/mdstat, фиксируем, какие слоты выпали и когда
- 2 Собираем SMART со всех дисков: smartctl -H и -A, ключевые атрибуты 5 (Reallocated), 197 (Current_Pending), 198 (Offline_Uncorrectable), 9 (Power_On_Hours)
- 3 Подключаем rescue-бокс (объём $\geq$ суммы всех дисков, USB 3.1 Gen 2) и клонируем худший диск первым: ddrescue -d --no-scrape /dev/sdX образ mapfile
- 4 Вторым проходом добираем битые зоны: ddrescue -d -r3 по тому же mapfile; «живые» диски тоже клонируем — под нагрузкой копирования они могут посыпаться

РЕЗУЛЬТАТ

Все дальнейшие операции идут по образам: оригиналы больше не читаются и не деградируют. Mapfile ddrescue позволяет прерывать и возобновлять копирование, а по errsize ещё до сборки массива видно, сколько данных реально не вычиталось.

КЛЮЧЕВОЙ НЮАНС

Порядок проходов критичен: --no-scrape сначала вычитывает всё лёгкое, пока диск жив, и только потом скрейпим битые области. Начинаем с самого повреждённого диска — у него меньше всего времени.



Анализ суперблоков и сборка degraded-массива

Что настраиваем

Образы на rescue-хранилище, подключённые через losetup как loop-устройства; mdadm metadata 1.2, ext4 поверх /dev/md127

Как мы это делаем

- 1 losetup на каждый образ, затем mdadm --examine: сверяем Events counter, Array State и Update Time — видно, какой диск выпал первым и насколько отстал
- 2 В сборку берём члены с максимальным и близким Events; диск, отставший на тысячи событий (часы данных), исключаем — он даст битую ФС
- 3 Собираем штатно: mdadm --assemble --force /dev/md127 из трёх элементов, четвёртый слот missing; --create --assume-clean держим как крайнюю меру
- 4 ФС проверяем сначала read-only: e2fsck -nv, после оценки масштаба — e2fsck -yfv; монтируем mount -o ro и валидируем выборку файлов (pdftinfo, контрольные открытия)

РЕЗУЛЬТАТ

Массив поднимается в состоянии clean, degraded без перезаписи parity. Разбор lost+found даёт точную цену сбоя: обычно там lock-файлы Office, а не рабочие документы. Клиент получает конкретный список потерь, а не «что-то восстановили».

КЛЮЧЕВОЙ НЮАНС

Events counter — главный критерий выбора дисков: разрыв в десятки событий допустим, в тысячи — часы устаревших данных. mdadm --create без точного chunk, layout и порядка дисков уничтожает массив.



Пересборка в RAID6 и мониторинг после инцидента

Что настраиваем

Новый массив из 6 дисков (смешиваем WD Red Pro и IronWolf Pro), mdmonitor + Zabbix 7.0 LTS на отдельной ВМ

Как мы это делаем

- 1 Создаём RAID6: `mdadm --create --level=6 --raid-devices=6 --chunk=256 --metadata=1.2`; диски двух производителей, партии разводим по serial range
- 2 `mkfs.ext4 -m 1`, перенос данных `rsync -a` с образов, сверка количества файлов и выборочных контрольных сумм
- 3 Базовый алертинг за 5 минут: MAILADDR в `/etc/mdadm/mdadm.conf`, `systemctl enable --now mdmonitor`, тест `mdadm --monitor --scan --test --oneshot`
- 4 Zabbix: шаблон SMART by Zabbix agent 2 (smartctl LLD), триггеры на рост Reallocated/Pending и state из `/proc/mdstat`, алерты в Telegram
- 5 Cron на ежемесячный scrub (`sync_action check`) и офлайн-бэкап на отдельный носитель — массив перестаёт быть единственной копией

РЕЗУЛЬТАТ

Массив переживает одновременный отказ двух дисков, отказ первого виден в Telegram за минуты, латентные bad-блоки вычищаются scrub'ом до того, как сорвут ребилд, а копия данных живёт вне сервера.

КЛЮЧЕВОЙ НЮАНС

RAID5 на дисках 4 ТБ+ после отказа одного диска работает без избыточности всё окно ребилда (12-36 ч) — именно туда статистически попадает второй отказ. Поэтому наш базовый уровень для 4+ дисков — RAID6.



Подводные камни

✗ **mdadm --create по живому массиву**

Пересчёт parity затирает данные безвозвратно. Мы начинаем с `--assemble --force`, а `--create --assume-clean` применяем только по образам

✗ **Работа с оригиналами дисков**

Каждое чтение умирающего диска добывает поверхность. Все эксперименты — только на ddrescue-образах через `losetup`, оригиналы убираем

✗ **Сборка с отставшим диском**

Диск с меньшим Events counter несёт устаревшие данные — ФС соберётся битой. Перед сборкой сверяем `mdadm --examine` по всем членам

✗ **Desktop-диски в массиве**

Без TLER диск при ошибке чтения уходит в recovery на минуты, и его выкидывает из массива. Только NAS-класс: Red Pro, IronWolf Pro, Toshiba MG

✗ **Все диски одной партии**

Одинаковый ресурс — отказы пачкой в узком окне. Смешиваем производителей и партии, при закупке смотрим serial range

✗ **e2fsck -у сразу после сборки**

Автопочинка по неверно собранному массиву множит повреждения. Сначала `e2fsck -n`, оцениваем масштаб и только потом запускаем `-yfv`

✗ **Бэкап на том же RAID**

RAID защищает от отказа диска, но не от шифровальщика, удаления и двойного отказа. Копия — на отдельное устройство вне сервера (3-2-1)

✗ **Массив без scrub и мониторинга**

Латентные bad-блоки всплывают в ребилд и валят его. Ежемесячный `sync_action check` плюс `smartd` закрывают эту проблему



Как правильно

МИНИМУМ

- mdmonitor с MAILADDR: письмо при деградации массива (настройка — 5 минут)
- smartd из smartmontools 7.5: контроль атрибутов 5/197/198 на всех дисках
- Бэкап критичных данных на отдельный носитель, не на сам массив

НОРМАЛЬНО

- RAID6 вместо RAID5 для массивов из 4+ дисков объёмом от 4 ТБ
- Ежемесячный scrub (sync_action check) по cron + контроль mismatch_cnt
- Zabbix-шаблон SMART by Zabbix agent 2, алерты в Telegram
- Диски двух производителей/партий в одном массиве

ХОРОШО

- Hot spare в массивах от 6 дисков — ребилд стартует без человека
- Плановая замена дисков после 55 000 ч Power_On_Hours, по одному
- Бэкап 3-2-1 с офлайн-копией и тестовым восстановлением раз в квартал
- Rescue-комплект наготове: бокс под образы объёмом $\geq$ массива, свежий ddrescue



Чек-лист самопроверки

- | | |
|---|--|
| ☐ Знаете ли вы прямо сейчас состояние массива (clean/degraded) без захода на сервер? | ☐ Известен ли возраст дисков (Power_On_Hours) и есть ли план замены после 55 000 ч? |
| ☐ Придёт ли алерт в течение часа, если один диск в RAID выйдет из строя? | ☐ Диски в массиве — из разных партий или от разных производителей? |
| ☐ Есть ли копия данных вне самого массива и вне сервера? | ☐ Переживёт ли ваш массив одновременный отказ двух дисков (RAID6/RAID10)? |
| ☐ Проверялись ли SMART-атрибуты 5/197/198 всех дисков за последний квартал? | ☐ Есть ли под рукой хранилище, куда поместятся посекторные образы всех дисков? |
| ☐ Выполняется ли scrub массива хотя бы раз в месяц? | ☐ Документированы ли chunk size, layout и порядок дисков на случай ручной сборки? |

Если хотя бы на два вопроса ответ «нет» или «не знаю» — тема требует внимания.



Как поможет ITFresh

ITFresh — ИТ-аутсорсинг для юридических лиц до 50 рабочих мест в Москве и области. 15+ лет практики, собственная инфраструктура в дата-центре МТС (8 серверов Dell Xeon Platinum).

- Аудит дискового массива за 2–3 часа: SMART всех дисков, состояние RAID, корректность бэкапов, смета работ
- Аварийное восстановление RAID: образы ddrescue, degraded-сборка, проверка ФС — без риска для оригиналов
- Миграция RAID5 → RAID6 с подбором дисков разных партий и переносом данных без потерь
- Мониторинг под ключ: mdmonitor + Zabbix с алертами в Telegram, ежемесячный scrub по cron
- Бэкап-контур 3-2-1 с офлайн-копией и регулярным тестовым восстановлением

15+

лет в ИТ-поддержке

50

рабочих мест — наш профиль

МТС

дата-центр, Москва



КОНТАКТЫ

Обсудить вашу задачу

Сайт **itfresh.ru**

Телефон **+7 903 729-62-41**

Telegram **@ITfresh_Boss**

Бесплатно посмотрим вашу инфраструктуру по этому чек-листу и скажем, где тонко — без обязательств.



itfresh.ru



Техническая база

- 01** GNU ddrescue Manual: фазы копирования, --no-scrape, mapfile (gnu.org — 1.30)
- 02** Linux RAID Wiki: Assemble Run / RAID Recovery (raid.wiki.kernel.org — 2025)
- 03** smartctl / smartd Reference (smartmontools.org — 7.5)
- 04** WD Red Pro Series Spec Sheet (WD4003FFBX) (westerndigital.com — 2024)
- 05** md(4): sync_action, mismatch_cnt, scrub (man7.org — 2025)
- 06** Zabbix integration: SMART by Zabbix agent 2 (zabbix.com — 7.0)
- 07** Шаблон ITfresh: регламент аварийного снятия образов (rescue-бокс) (itfresh.ru — 2026)

Основано на официальной документации продуктов и нашей практике внедрения.

