Ай-ТИ Фреш

ТЕХНИЧЕСКИЙ РАЗБОР

Ceph: распределённое хранилище для production

Проектирование, развёртывание через cephadm и эксплуатация кластера — наша методология

Июль 2026

itfresh.ru · ИТ-аутсорсинг для юридических лиц

О ЧЁМ ЭТОТ РАЗБОР

Суть проблемы

Классический файловый сервер или RAID-массив с NFS — единая точка отказа: смерть контроллера или переполнение тома останавливает 1С, файлы и виртуальные машины на часы, а расширение означает покупку нового сервера и миграцию с простоем. Мы решаем это кластером Ceph: диски 3–5+ нод объединяются в самовосстанавливающийся пул, ёмкость растёт добавлением ноды без остановки.

Почему это важно бизнесу

- Простой хранилища = простой всего офиса: 1С, файловые шары и диски VM обычно живут на одном массиве — час простоя дороже внедрения кластера
- Отказ ноды Ceph клиент не замечает: восстановление автоматическое, без инженера в 3 ночи и без «ждём новый контроллер две недели»
- Масштабирование без миграций: добавили ноду — ёмкость и производительность выросли, данные ребалансируются сами
- Одна платформа вместо трёх: блочные диски VM (RBD), файловая шара (CephFS) и S3-объекты (RGW) на одном и том же железе

ЧТО МЫ ЗАКЛАДЫВАЕМ

Ключевые параметры реализации

19.2 / 20.2

Squid 19.2 — базовая ветка наших прод-кластеров; Tentacle 20.2 обкатываем на стендах перед кли...

релизы Ceph, docs.ceph.com

3 / 2

size/min_size реплицированных пулов: три копии данных, при падении до одной живой реплики запи...

наш стандарт + доки RADOS

85 / 95%

nearfull/full ratio: с 85% планируем расширение, на 95% кластер сам останавливает запись — дов...

mon_osd_*_ratio, Ceph docs

4 ГиБ

osd_memory_target на каждый BlueStore-OSD; RAM ноды считаем как 4 ГиБ × число OSD плюс MON/MGR...

BlueStore, docs.ceph.com

600 с

mon_osd_down_out_interval: через 10 минут упавший OSD помечается out и стартует автоматический...

RADOS config, docs.ceph.com

2×10G

раздельные public и cluster network: репликация и recovery не конкурируют с клиентским трафико...

наш стандарт сети

РЕАЛИЗАЦИЯ 1 ИЗ 3

Развёртывание кластера через cephadm

Что настраиваем

3–5 нод: гиперконвергенция с Proxmox VE или отдельный storage-слой; MON+MGR на первых трёх нодах, OSD — на всех

Как мы это делаем

1. cephadm bootstrap --mon-ip <ip> --cluster-network 10.0.2.0/24 на первой ноде: MON, MGR и dashboard поднимаются контейнерами
2. ceph orch host add + метки _admin/mon/osd; MON строго 3 (кворум), MGR — 2 в active/standby
3. OSD через service spec (ceph orch apply -i osd.yml): HDD → data, NVMe → block.db под WAL/DB, а не --all-available-devices вслепую
4. CRUSH failure domain = host; pg_autoscale_mode on, балансировка — MGR-модуль balancer в режиме upmap
5. Приёмка: ceph -s до HEALTH_OK, ceph osd tree — все OSD up/in и разложены по хостам, тест выключения ноды

РЕЗУЛЬТАТ

Кластер разворачивается за один день по воспроизводимому spec-файлу, все демоны под управлением оркестратора: замена диска или добавление ноды — одна команда ceph orch, без ручной правки конфигов на хостах.

КЛЮЧЕВОЙ НЮАНС

Кворум MON — это нечётное число и минимум 3 физических узла; «Ceph на двух серверах» мы не строим в принципе — при потере одного узла кластер обязан продолжить работу, а не встать вместе с ним.

РЕАЛИЗАЦИЯ 2 ИЗ 3

RBD: блочный слой под диски виртуальных машин

Что настраиваем

Пул vm-disks на SSD-классе устройств, интеграция с Proxmox VE через librbd; архивный пул на HDD отдельно

Как мы это делаем

1. ceph osd pool create vm-disks + application enable rbd + rbd pool init; size=3, min_size=2 фиксируем явно
2. CRUSH device class: правило replicated_ssd — диски VM ложатся только на SSD-OSD, архивы — на HDD-пул с EC 4+2
3. Отдельный ключ клиента: ceph auth get-or-create client.pve с caps profile rbd, а не admin-ключ в Proxmox
4. В Proxmox storage.cfg: тип rbd, monhost всех трёх MON, krbd 0 для librbd; снапшоты rbd snap create перед обновлениями гостей

РЕЗУЛЬТАТ

Диски VM переезжают между нодами Proxmox мгновенно (общее хранилище), снапшот перед обновлением делается секунды и откатывается командой rbd snap rollback; потеря любой ноды не трогает работающие машины.

КЛЮЧЕВОЙ НЮАНС

Erasure coding для дисков VM не используем: мелкая случайная перезапись на EC-пуле кратно медленнее репликации. EC 4+2 (overhead 50% вместо 200%) оставляем холодным данным — архивам, бэкапам, S3.

РЕАЛИЗАЦИЯ 3 ИЗ 3

Мониторинг, scrub и обновления без остановки

Что настраиваем

MGR-модуль prometheus на порту 9283 + Alertmanager; регламент целостности и апгрейдов на весь кластер

Как мы это делаем

1. ceph mgr module enable prometheus; алерты: ceph_health_status != 0, sum(ceph_osd_up) < sum(ceph_osd_in), заполнение пула > 80%
2. Scrub по расписанию: лёгкий ежесуточно, deep-scrub каждые 7 дней (osd_deep_scrub_interval), окно osd_scrub_begin_hour/osd_scrub_end_hour — на ночь
3. Приоритет клиентов при recovery: mClock-планировщик, профиль osd_mclock_profile high_client_ops в рабочие часы
4. Обновления: ceph orch upgrade start --image quay.io/ceph/ceph:v19.2.x — оркестратор перезапускает демоны по одному, кластер живёт
5. RBD-снапшоты выгружаем rbd export-diff на внешнее бэкап-хранилище; /etc/ceph и ключи — в шифрованный оффсайт

РЕЗУЛЬТАТ

Деградацию видим до того, как её заметят пользователи: медленный OSD, отставший deep-scrub или рост latency ловятся алертом. Минорные обновления проходят в рабочее время без окна простоя.

КЛЮЧЕВОЙ НЮАНС

Снапшоты внутри Ceph — не бэкап: они умирают вместе с кластером. Внешняя копия через export-diff или rbd-mirror на другую площадку обязательна, это мы закладываем в проект с первого дня.

ЧТО МЫ ВЫУЧИЛИ НА ПРАКТИКЕ

Подводные камни

- **Кластер из двух нод**
  MON-кворум требует 3 узла: на двух нодах отказ любой из них останавливает кластер целиком. Минимум 3 ноды, MON нечётным числом.
- **min_size=1 «ради аптайма»**
  Запись в единственную копию: сбой этого OSD = безвозвратная потеря свежих данных. Держим min_size=2 — пауза I/O лучше потери.
- **Recovery по общей 1G-сети**
  Ребилд после сбоя забивает канал и «кладёт» VM. Выносим репликацию в отдельную cluster network 10G+ с jumbo frames 9000.
- **HDD-OSD без SSD под WAL/DB**
  Метаданные BlueStore на том же HDD режут IOPS в разы. Выносим block.db на NVMe: 1–2% объёма data-диска для RBD, от 4% для RGW.
- **Заполнение выше 85%**
  После сбоя ноды backfill не помещается в остаток, кластер уходит в full и блокирует запись. Расширяемся с порога nearfull.
- **Failure domain = osd**
  Все три реплики могут лечь на один хост — его отказ теряет данные. Проверяем CRUSH rule: chooseleaf firstn 0 type host.
- **Desktop-SSD без PLP**
  Sync-запись BlueStore без конденсаторов проседает до сотен IOPS. Только enterprise SSD/NVMe с power-loss protection.
- **EC-пул под диски VM**
  Частичная перезапись на erasure-пуле требует read-modify-write и тормозит. EC 4+2 — для архивов и S3, VM — на репликации.

УРОВНИ РЕАЛИЗАЦИИ

Как правильно

МИНИМУМ

- 3 ноды, 3 MON, репликация size=3/min_size=2, failure domain = host
- Отдельная cluster network от 10G, jumbo frames 9000
- SSD/NVMe под block.db для каждого HDD-OSD
- ceph -s и dashboard в ежедневном регламенте админа

НОРМАЛЬНО

- cephadm + service spec в git: кластер воспроизводим с нуля
- MGR prometheus + Alertmanager: health, OSD down, пулы > 80%
- Ночные окна scrub, deep-scrub всех PG за 7 дней без просрочек
- RBD-снапшоты + export-diff на внешнее бэкап-хранилище

ХОРОШО

- 5+ нод, device class SSD/HDD, EC 4+2 для холодных данных
- rbd-mirror (snapshot-based) на резервную площадку
- Стенд для обкатки релизов + регламент ceph orch upgrade
- Учения: плановое отключение ноды с контролем recovery

ПРОВЕРЬТЕ У СЕБЯ

Чек-лист самопроверки

- [ ] MON ровно 3 (или 5) и все на разных физических нодах?
- [ ] min_size=2 на всех реплицированных пулах — проверено через ceph osd pool ls detail?
- [ ] Recovery идёт по отдельной cluster network и не делит канал с клиентами?
- [ ] WAL/DB всех HDD-OSD вынесены на серверные SSD/NVMe с PLP?
- [ ] Алерт nearfull 85% приходит инженеру, а не только светится в dashboard?
- [ ] Плановое расширение стартует до 75% заполнения raw-ёмкости?
- [ ] Deep-scrub успевает пройти все PG за 7 дней без HEALTH_WARN?
- [ ] Отказ целой ноды протестирован: recovery завершился, пользователи не заметили?
- [ ] Есть копия данных ЗА пределами кластера (export-diff / rbd-mirror)?
- [ ] План обновлений определён: минорные без окна, мажорные через стенд?

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

ЧТО МЫ ДЕЛАЕМ

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 текущего хранилища и расчёт кластера: ноды, диски, сеть, usable-ёмкость под ваш рост на 3 года
- Развёртывание Ceph под ключ: cephadm, пулы RBD/CephFS/RGW, интеграция с Proxmox VE
- Мониторинг 24/7: Prometheus/Alertmanager + Zabbix, реакция на деградацию OSD и рост latency
- Регламентная эксплуатация: scrub, обновления, замена дисков, расширение без простоя
- Миграция данных с NFS/RAID-сервера на Ceph без остановки рабочих сервисов

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

ДОКУМЕНТАЦИЯ И НАШ СТЕК

Техническая база

01 Cephadm — Deploying a New Ceph Cluster (docs.ceph.com — 19.2)

02 RADOS Configuration (mon_osd_*_ratio, mClock, heartbeat) (docs.ceph.com — 19.2)

03 BlueStore Configuration Reference (osd_memory_target, block.db) (docs.ceph.com — 19.2)

04 CRUSH Maps и Erasure Code Profiles (docs.ceph.com — 19.2)

05 MGR Prometheus Module / Ceph Dashboard (docs.ceph.com — 19.2)

06 Release Notes: Squid 19.2.x, Tentacle 20.2.x (ceph.io — 2026)

07 Deploy Hyper-Converged Ceph Cluster (pve.proxmox.com — PVE 8/9)

08 Наш шаблон osd service spec + alert rules (itfresh.ru — 2026)

Основано на официальной документации продуктов и нашей практике внедрения.